Avrupa Bilim ve Teknoloji Dergisi
Sayı 16, S. 176-185, Ağustos 2019
© Telif hakkı EJOSAT'a aittir
Araştırma Makalesi



www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
No. 16, pp. 176-185, August 2019
Copyright © 2019 EJOSAT
Research Article

Göğüs Kanseri Teşhisinde Farklı Makine Öğrenmesi Tekniklerinin Performans Karşılaştırması

Onur Sevli[1*]

[1] Burdur Mehmet Akif Ersoy Üniversitesi, Mühendilsik Mimarlık Fakültesi, Bilgisayar Mühendisliği Bölümü, Burdur, Türkiye (ORCID: 0000-0002-8933-8395)

(İlk Geliş Tarihi 14 Nisan 2019 ve Kabul Tarihi 22 Mayıs 2019)

(**DOI:** 10.31590/ejosat.553549)

ATIF/REFERENCE: Sevli, O. (2019). Göğüs Kanseri Teşhisinde Farklı Makine Öğrenmesi Tekniklerinin Performans Karşılaştırması. *Avrupa Bilim ve Teknoloji Dergisi*, (16), 176-185.

Öz

Bilgisayarlar insanlara nazaran daha hızlı işlem yapabilmektedir ancak karar verme yetenekleri kısıtlıdır. Günümüz bilgisayarlarının daha iyi analizler yapıp tahminlerde bulunabilmeleri için çeşitli makine öğrenmesi teknikleri geliştirilmektedir. Bu teknikler bilgisayarların karar verme güçlerini arttırmakta ve farklı sahalarda uzmanlara destek sistemlerin geliştirilmesine olanak sağlamaktadır. Makine öğrenmesi tekniklerinin, başarılı sınıflama ve tanılama yetenekleri ile hastalık teşhisinde medikal uzmanlara yardımcı olarak kullanımları hızla artmaktadır. Kanser teşhisinde de kullanımı hızla artan makine öğrenmesi ile başarılı çalışmalar yapılabilmektedir. Göğüs kanseri dünya genelinde en yaygın görülen ikinci kanser türü olup kadınlar arasında kanser kaynaklı en yüksek oranda ölüme sebep olan hastalıktır. Diğer tüm kanser türlerinde olduğu gibi göğüs kanserinin de erken teşhisi ölüm oranını azaltmada kritik bir öneme sahiptir. Göğüs kanseri tanısı, test sonuçların yorumlanarak teşhis edilmesi uzman insan bilgisine ihtiyaç duymaktadır ancak gelişen makine öğrenmesi teknikleri ile göğüs kanseri teşhisinde başarılı çalışmalar yürütülmektedir. Makine öğrenmesi bilgisayarların mevcut verilerden öğrenerek karmaşık ve büyük veri setleri içerisindeki desenleri hızlı bir şekilde tespit etmesini sağlayan bir yapay zekâ dalıdır. Bu yeteneğinden dolayı makine öğrenmesi kanser tanı ve teşhisinde özellikle göğüs kanseri konusunda da yaygın kullanım alanı bulmaktadır. Bu çalışmada her biri 30 adet özellik içeren ve 569 örnekten oluşan Wisconsin Üniversitesi göğüs kanseri veri seti, beş farklı makine öğrenmesi tekniği ile sınıflandırılmıştır. Veriler rastgele olarak eğitim ve test setlerine ayrılmıştır. Destek vektör makinesi, Naïve Bayes, rastgele orman, K en yakın komşu ve lojistik regresyon metotları ile gerçekleştirilen eğitim sürecinin ardından confusion matrisleri ve roc eğrileri oluşturulmuştur. Her bir tekniğin başarısı karşılaştırılmıştır. Bu karşılaştırmanın sonucunda lojistik regresyonun %98.24 doğruluk ile en başarılı yöntem olduğu ortaya konmuştur.

Anahtar Kelimeler: Göğüs kanseri, Makine öğrenmesi, Yapay zekâ

Performance Comparison of Different Machine Learning Techniques in Diagnosis of Breast Cancer

Abstract

Computers are able to process faster than people, but their ability to make decisions is limited. Various machine learning techniques are being developed for today's computers to make better analyzes and predictions. These techniques increase the decision-making power of computers and enable the development of support systems for experts in different fields. Machine learning techniques are being used rapidly to assist medical specialists in diagnosing diseases with their successful classification and diagnostic capabilities. Successful work can be done with machine learning, which is rapidly increasing in the use of cancer diagnosis. Breast cancer is the second most common type of cancer in the world and is the most common cancer related cause of death among women. As with all other types of cancer, early diagnosis of breast cancer is critical in reducing the mortality rate. Diagnosis of breast cancer, diagnosis and interpretation of test results require specialized human knowledge, but successful studies are being carried out in the diagnosis of breast cancer by developing machine learning techniques. Machine learning is an artificial intelligence branch that allows computers to quickly identify

[*] Sorumlu Yazar: Burdur Mehmet Akif Ersoy Üniversitesi, Mühendislik Mimarlık Fakültesi, Bilgisayar Mühendisliği Bölümü, Burdur, Türkiye, ORCID: 0000-0002-8933-8395, onursevli@mehmetakif.edu.tr

patterns within complex and large data sets by learning from existing data. Due to this ability, machine learning is widely used in diagnosis of cancer, especially in breast cancer. In this study, the University of Wisconsin breast cancer data set, which consists of 569 samples, each with 30 features, was classified by five different machine learning techniques.Data was randomly splitted as training and test set. After the training process of Support Vector Machine, Naïve Bayes, Random Forest, K-Nearest Neighbour and Logistic Regression methods, confusion matrices and roc curves were created and the success of each method has been compared. As a result of this comparison, it has been shown that Logistic Regression is the most successful model with 98.24% accuracy.

Keywords: Breast cancer, Machine learning, Artificial Intelligence

1. Giriş

Günümüz bilgisayarları insana nazaran daha hızlı işlem yapabilmektedir ancak karar verme yetenekleri insana göre daha düşüktür. Bu nedenle bilgisayarların daha iyi analizler yapıp kararlar verebilmelerini sağlayan farklı makine öğrenmesi teknikleri geliştirilmiş ve geliştirilmektedir. Kümeleme, sınıflama yöntemleri, karar ağaçları, yapay sinir ağları gibi pek çok teknik ile veriden anlam çıkarımı ve tahminleme yapılabilmektedir [1]. Makine öğrenmesi tekniklerinin, başarılı sınıflama ve tanılama yetenekleri ile hastalık teşhisinde medikal uzmanlara yardımcı olarak kullanımları da hızla artmaktadır.

Göğüs kanseri özellikle 40 – 49 yaş arası kadınlarda sıklıkla görülen ve dünya genelinde kadınlar arasında kanser kaynaklı en yüksek oranda ölüme sebep olan hastalıktır [2]. 2018 yılında tüm dünyada tespit edilen 18.1 milyon kanser vakası içerisinde %11.6 oranla akciğer kanserinden sonra ikinci sırada yer almaktadır [3]. Göğüs dokusunda özellikle süt kanalları ve bezlerinde küçük tümör ya da kitleler şeklinde görülür. Kitlenin pürüzsüz olması ve sınırlarının belli olması iyi huylu olduğuna, sınırlarının düzensiz olması ve pürüzlü yapıda olması kötü huylu yani kanser riski taşımasına işarettir [4]. Diğer tüm kanser türlerinde olduğu gibi göğüs kanserinin de erken teşhisi ölüm oranını azaltmada kritik bir öneme sahiptir [5]. Göğüs kanseri tanısı, test sonuçların yorumlanarak teşhis edilmesi uzman insan bilgisine ihtiyaç duymaktadır. Gelişen makine öğrenmesi teknikleri ile göğüs kanseri teşhisinde başarılı çalışmalar yürütülmektedir.

Makine öğrenmesi içerisinde çeşitli istatistiki, olasılıksal teknikleri ve optimizasyon tekniklerini barındıran; bilgisayarların mevcut verilerden öğrenerek karmaşık ve büyük veri setleri içerisindeki desenleri hızlı bir şekilde tespit etmesini sağlayan bir yapay zeka dalıdır. Bu yeteneğinden dolayı makine öğrenmesi kanser tanı ve teşhisinde de yaygın kullanım alanı bulmaktadır [6].

Göğüs kanserine dair her yıl yaklaşık olarak 1.38 milyon yeni vaka görülmektedir [7]. Makine öğrenmesi tekniklerinin klinik alanda özellikle kanser teşhisine yönelik kullanımı giderek artarken [8] göğüs kanserinin tespitini kolaylaştırmaya yönelik olarak yapılan çeşitli çalışmalar bulunmaktadır. Göğüs kanseri teşhisinde farklı makine öğrenmesi algoritmaları kullanılmakla birlikte yaygın olarak kümeleme [9 – 11], yapay sinir ağları [12 – 14], destek vektör makineleri [15 – 17], bulanık ve yapay bulanık mantık [18 – 19] ve hibrit teknikler kullanılmaktadır.

Ravdin ve Clark düşük ve yüksek risk taşıyan göğüs kanseri hastalarının tespitine yönelik bir yapay sinir ağı modeli ortaya koymuşlardır [20]. Mangasarian vd. ise kötü huylu tümörler için tekrarlamayan vakaları, tekrarlayan vakalar için ise tekrarlama zamanlarını tahmin etmeye yönelik doğrusal programlama tabanlı bir sistem geliştirmişlerdir [21]. Ravi ve Zimmermann tümör veri seti üzerinde geliştirdikleri üç fazlı bulanık veri işleme modelinde önce özellik uzayında boyut indirgeme yoluna gitmiş, ardından bulanık kuralları otomatik olarak oluşturup daha az kuralla daha yüksek bir sınıflama gücü elde etmişlerdir [22]. Delen vd. geniş bir göğüs kanseri veri seti üzerinde iki popüler veri madenciliği algoritması olan yapay sinir ağları ve karar ağaçlarını kullanarak tahmin modelleri geliştirmişlerdir. Karar ağaçları ile %93.6, yapay sinir ağı modeli ile %91.2 doğruluk elde etmişlerdir [23]. Polat ve Güneş, En küçük kare destek vektör makinesi (LS-SVM) sınıflama algoritması kullanarak göğüs kanseri verileri üzerinde %98 oranında başarı elde etmişlerdir [24]. Khan vd. geliştirdikleri bulanık karar ağaçları ile göğüs kanseri verilerini sınıflamış ve bağımsız sınıflayıcılara göre daha başarılı olduklarını ortaya koymuşlardır [25].

Chauhan vd. yapay sinir ağlarında parametre ayarlamaları için diferansiyel evrim modeli kullanarak gerçekleştirdikleri sistemi, göğüs kanseri veri seti dâhil üç farklı veri seti ile test ederek, geleneksel yapay sinir ağı modelinden daha başarılı olduğunu ortaya koymuşlardır [26]. Karabatak ve İnce, ilişki kuralları ile yapay sinir ağlarını birleştirerek ürettikleri hibrit model ile göğüs kanseri verilerini sınıflamışlar ve modellerinin %95.6 oranda doğru sınıflama yaptığını ortaya koymuşlardır [27]. Powel vd. Kaliforniya'da yaşayan ve içerisinde yüksek oranda göğüs kanseri olan, hiç doğum yapmamış ve geç doğum yapmış kadınlara ait veriler içeren veri seti ile Breast Cancer Risk Assessment Tool (BCRAT) [28], International Breast Intervention Study (IBIS) [29] ve BRCAPRO [30] göğüs kanseri risk değerlendirme modellerini 5 yıl boyunca yaptıkları uygulamalarla karşılaştırmışlardır ve performanlarını test etmişlerdir.[31].

Papageorgiou vd. çalışmalarında Fuzzy Cognite Map (FCM) kullanarak geliştirdikleri bir sağlık asistanı ile 40 adet hastanın verilerini işleyerek %95 oranında doğruluk elde etmişlerdir [4]. Kolay ve Erdoğmuş çalışmalarında göğüs kanseri veri setini herhangi bir ön işleme yapmadan, Matlab ve Weka programları üzerinde K-means yöntemi ile sınıflandırmış ve çeşitli parametre değişimleri ile %45 ile %79 arasında değişen başarılar elde etmişlerdir [1]. Alharbi ve Tchier, yaptıkları çalışmada bulanık mantık ve evrimsel genetik algoritma tabanlı göğüs kanserinin erken teşhisine yardımcı olan bir sistem geliştirerek Suudi Arabistan göğüs kanseri teşhis veri tabanı üzerinde uygulamışlardır. Bulanık-genetik hibrit algoritma ile %97 doğruluk ve %91 güvenilirlik elde etmişlerdir [32]. Akyol, 2018

yılında yaptığı çalışmada Özyinelemeli Özelik Elemesi yöntemi ile meme kanseri veri seti üzerinde öznitelik tespiti yapıp rastgele orman yöntemini uygulayarak sınıflama yapmış ve %98 başarı elde etmiştir [33].

Bu çalışmada her biri 30 adet özellik içeren 569 göğüs kanseri veri örneği, 5 ayrı makine öğrenmesi tekniği ile sınıflandırılarak, modellerin başarıları karşılaştırılmıştır. Sonraki bölümlerde materyal ve metot, bulgular ve sonuçlar yer almaktadır.

2. Materyal ve Metot

Kanser, hücrelerin kontrolsüz olarak bölünmesi ile ortaya çıkar ve tümör olarak anılan kitleler oluşturur. Tümörler iyi huylu (benign) ve kötü huylu (malignant) olabilir. Kötü huylu tümörler hızla büyüyerek etrafındaki dokuları işgal eder ve zarar görmelerine neden olur. Göğüs dokusundaki anormallikler, göğüs şekli ve deri rengindeki değişimler göğüs kanseri habercisi olabilir. Tüm kanser türlerinde olduğu gibi göğüs kanserinde de erken teşhis hayati önem taşımaktadır.

Bu çalışmada Wisconsin Üniversitesi hastanesinde Dr. William H. Wolberg tarafından toplanan ve araştırmalar için paylaşılan, göğüs kanseri bulgularını içeren 569 örnekten oluşan veri seti kullanılmıştır. Veriler %80 eğitim seti, %20 test seti olacak şekilde rastgele bölünerek beş farklı makine öğrenmesi modeli ile sınıflanarak test edilmiştir. Destesk Vektör Makinesi (Support Vector Machine – SVM), Naive Bayes (NB), Rastgele Orman (Random Forest – RF), K En yakın Komşu (K Nearest Neighbor – KNN) ve Lojistik Regresyon (Logistic Regression – LR) sınıflayıcılarının test başarıları karşılaştırılmıştır. Modellerin oluşturulması ve test işlemleri Python programlama dili ile gerçekleştirilmiştir.

2.1. Veri Seti

Kullanılan veri seti 569 adet örnek içermektedir. Her bir veri örneği için 30 tanımlayıcı özellik ve bir adet teşhis sınıfı olmak üzere toplam 31 özellik bulunmaktadır. 30 adet tanımlayıcı özellik, göğüste görülen kitle resimlerinin sayısallaştırılması ile elde edilen dokunun çapı, şekli, pürüzsüzlüğü, yüzey alanı gibi verilerden oluşmaktadır. 30 adet özelliğin 10 tanesi tümör hücresinin çekirdeği üzerinden direkt olarak ölçümlenmiş, 20 tanesi ise bunlara bağlı olarak hesaplanmış sayısal değerlerdir. Direkt ölçümlenen 10 adet özellik ise şunlardır:

1) Yarıçap

2) Doku

3) Çap

4) Alan

5) Pürüzsüzlük

6) Yoğunluk

7) İçbükeylik

8) İçbükey nokta sayısı

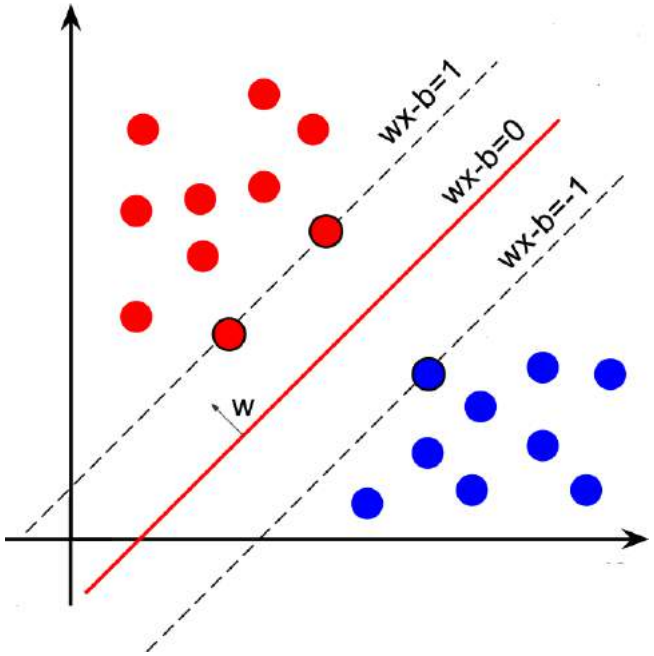
9) Simetri

10) Fraktal boyut

Diğer özellikler ise bu özelliklerden türetilen ortalama, standart hata, en kötü ve en büyük değerlerden oluşmaktadır. Bu değerlere bağlı olarak, tümörün iyi huylu veya kötü huylu olduğunu belirten B (benign) ve M (malignant) etiketi ile ifade edilen teşhis sınıfı yer almaktadır. 569 verinin sınıf dağılımı ise 357 iyi huylu, 212 kötü huylu şeklindedir.

2.2. Destek Vektör Makinesi

Destek vektör makinesi, 1990'lı yıllarda Vapnik ve ekibi tarafından geliştirilen, iki temel sınıfa ait olan verileri birbirinden ayırmak için kullanılan, istatistiksel öğrenme teorisine dayalı, sınıflandırma ve regresyon işlemleri için kullanılabilen gözetimli makine öğrenmesi algoritmasıdır [34]. Eğitim verilerinin yer aldığı düzlemde iki sınıfın üyelerinden en uzak olacak şekilde bir karar sınırının çizilmesini sağlar (Şekil 1).



Şekil 1. Destek Vektör Makinesi

Verinin her bir noktası Eş. 1'de verilen şekilde tanımlanır.

$$\{ (x_i,y_i) \mid x_i \in R^d, y_i \in \{-1,1\} \}_{i=1}^{n} \qquad (1)$$

Formülde x bir girdiyi, y ise -1 ve 1 ile temsil edilen bir sınıfı belirtir. Düzlemde her bir nokta wx-b şeklinde ifade edilir. Burada w düzleme dik olan normal vektörü, b ise kayma miktarıdır. Destek vektör makinesi, karesel optimizasyon yöntemi ile ayırma sınırının bulunmasını sağlar [15].

2.3. Naive Bayes

Naive Bayes sınıflayıcı, İngiliz matematikçi Thomas Bayes'in Eş. 2'de gösterilen teoremine dayanır [35].

$$P(G|X) = (P(X|G)\ P(G))/(P(X)\ ) \qquad (2)$$

Formülde P(G|X), G olayının verilen X olayına göre olma olasılığıdır. P(X|G) ise X olayının G olayı gerçekleştiğinde olma olasılığıdır. P(G) ve P(X) ise G ve X olaylarının önsel olasılıklarıdır.

2.4. Rastgele Orman

Rastgele orman, 2001 yılında Leo Breiman tarafından ortaya atılan bir yaklaşımdır [36]. Birden çok karar ağacının birleşiminden oluşan bir modeldir. Veriler N adet karar ağacı üzerinde işlendikten sonra elde edilen tahminlerin ortalaması alınarak doğru bir tahmin üretilmeye çalışılır. Rastgele orman geleneksel karar ağaçlarında en çok karşılaşılan problemlerden biri olan aşırı uydurma (overfitting) sorununu hem veri seti, hem öznitelikleri çok sayıda parçaya bölüp birden çok ağaç üzerinde işleyerek çözer.

2.5. K En Yakın Komşu

Sınıfı belirlenmek istenen bir noktanın, daha önceden sınıflanmış olan noktalardan, belirlenen K sayısınca en yakın noktaya göre sınıfının tespit edilmesini sağlayan bir modeldir. En yakın noktalar hesaplanırken genelde öklit uzaklığına bakılır. İdeal K değerinin seçimi üzerinde çalışılan veriye bağlı olarak değişiklik gösterir. Büyük K değerleri sınıflamadaki gürültü etkisini azaltırken, sınıflar arasındaki sınırların ayrımını azaltır.

2.6. Lojistik Regresyon

Lojistik regresyon bağımlı değişkenin süreksiz olduğu ikili sınıflama (0 ve 1) durumunda kullanılan bir modeldir. Makine öğrenmesi alanı dışında, diğer uygulamalı bilimlerde, gerçek dünya problemlerinde yaygın olarak kullanılmaktadır [37]. Lojistik regresyon ikili (binary) bir bağımlı değişken ile bir dizi bağımsız değişken arasındaki ilişkiyi açıklamaya yönelik tahminleyici bir analizdir. Bir a+bx denklemi için bir olayın gerçekleşme olasılığı Eş. 3'teki gibidir.

$$p = \frac{e^{a+bx}}{1+e^{a+bx}} \quad (3)$$

Olayın gerçekleşmeme olasılığı da ise 1-p olmak üzere logit fonksiyonu Eş. 4'te verilmiştir:

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) \quad (4)$$

Logistik regresyon logit dönüşümünü tahminlemek için bir formülün katsayılarını üretir.

2.7. Performans Değerlendirme

Toplam 569 adet örnekten oluşan veri setinin %80'i eğitim, %20'si test için kullanılmış; bu ayrımda örneklerin seçimi ise rastgele yapılmıştır. Eğitim sürecinden sonra test verileri ile sınıflama başarısı kontrol edilmiştir. Sistemin ürettiği sınıflar ile test sınıfları karşılaştırıldığında ne kadarlık kısmın doğru tahmin edildiği sistemin genel sınıflandırma doğruluğunu gösterir. Bu durum Eş. 5'te gösterilmiştir. Detaya inildiğinde sınıflanan veri kümesinde dört muhtemel sonuç vardır: Aslı pozitif olan örnek pozitif olarak doğru sınıflandırıldığında doğru pozitif (true positive – TP), aslı pozitif olan örnek negatif olarak yanlış sınıflandırıldığında yanlış negatif (false negative – FN), aslı negatif olan örnek negatif olarak doğru sınıflandırıldığında doğru negatif (true negative – TN), aslı negatif olan örnek pozitif olarak yanlış sınıflandırıldığında yanlış pozitif (false positive – FP) olarak nitelendirilir. Tüm bu olası durumların gösterildiği matrise confusion matrix adı verilir (Tablo 1).

Tablo 1. Confusion matrix

Tahmin edilen değerler	**Gerçek değerler**	
	Positive	*Negative*
Positive	True Positive (TP)	False Positive (FP)
Negative	False Negative (FN)	True Negative (TN)

$$\text{Doğruluk} = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

Doğru pozitif değerlerin, doğru pozitif ve yanlış negatiflerin toplamına oranı duyarlılık (sensitivity) değerini verir (Eş. 6). Duyarlılık sınıflandırmanın doğru pozitifleri tespit etme kabiliyetini gösterir.

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

Doğru negatif değerlerin, yanlış pozitif ve doğru negatiflerin toplamına oranı belirleyicilik (sprecifity) değerini verir (Eş. 7). Bir sınıflandırmanın gerçek negatif oranını yani negatif olan bir sonuca negatif teşhis koyma oranını gösterir.

$$\text{Belirleyicilik} = \frac{TN}{FP+TN} \quad (7)$$

Doğru pozitif değerlerin, doğru pozitif ve yanlış pozitiflerin toplamına oranı kesinlik değerini verir (Eş. 8). Kesinlik sınıflandırmanın yanlış pozitifleri eleme kabiliyetini gösterir.

$$\text{Kesinlik} = \frac{TP}{TP+FP} \quad (8)$$

Sınıflandırıcı performansını test etmek için ROC (reveiver operating characteristic – alıcı işlem karakteristiği) eğrilerinden yararlanılır. ROC eğirişi duyarlılığın kesinliğe oranı ile ortaya çıkmaktadır. Sınıflandırma yöntemleri duyarlılık ile kesinlik arasında dengeyi sağlamaya çalışır. ROC eğrileri duyarlılık ve kesinlik arasındaki dengeyi değerlendirmek için kullanılır. ROC eğrileri ayırt ediciliği göstermekle birlikte, farklı testlerin performans karşılaştırmasında eğri altında kalan alana (AUC – area ender curve) ihtiyaç duyulur. ROC eğrisi altında kalan alanın değeri ROC puanını verir ve bu değerin 1'e yaklaşması pozitiflerin negatiflerden başarılı bir şekilde ayrıldığı anlamına gelir.

3. Bulgular

Göğüs kanseri bulgularını içeren veri seti, toplam 569 adet örneğin %20'sini oluşturan 114 örnek ile Destek vektör makinesi (SVM), Naive bayes (NB), Rastgele orman (RF), K-en yakın komşu (KNN) ve Lojistik regresyon (LR) modelleri ile ayrı ayrı sınıflandırılmıştır. Her bir model için confusion matrix oluşturulmuş, doğruluk, duyarlılık, belirleyicilik, kesinlik değerleri belirlenmiş ve Tablo 2'de gösterilmiştir.

Tablo 2. Modellerin sınıflandırma başarıları (Classification achievements of models)

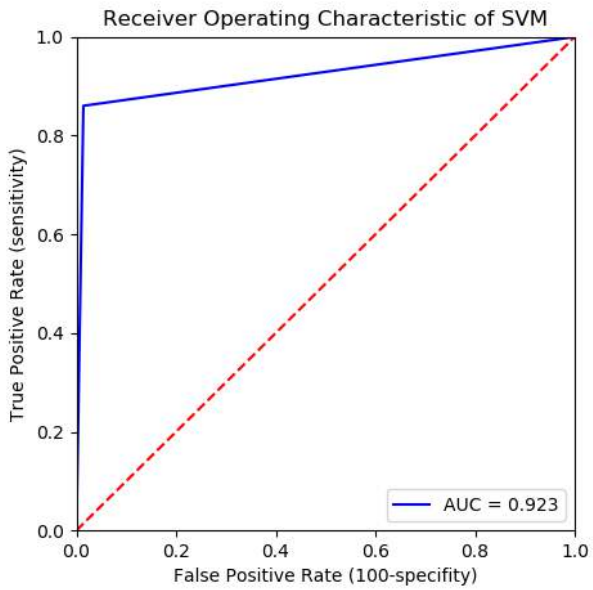
Model	TP	FN	TN	FP	Doğruluk	Duyarlılık	Belirleyicilik	Kesinlik
SVM	37	6	70	1	%93,86	%86,05	%98,59	%97,37
NB	37	6	71	0	%94,74	%86,05	%100	%100
RF	41	2	69	2	%96,49	%95,35	%97,18	%95,35
KNN	35	8	69	2	%91,23	%81,39	%97,18	%94,59
LR	41	2	71	0	%98,24	%95,35	%100	%100

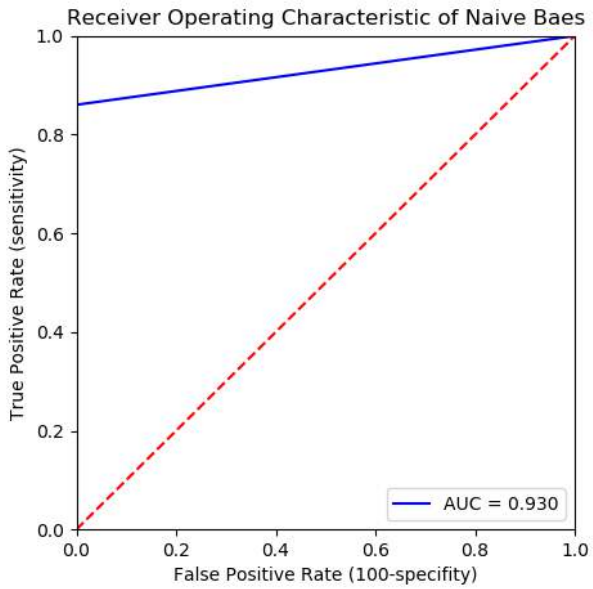
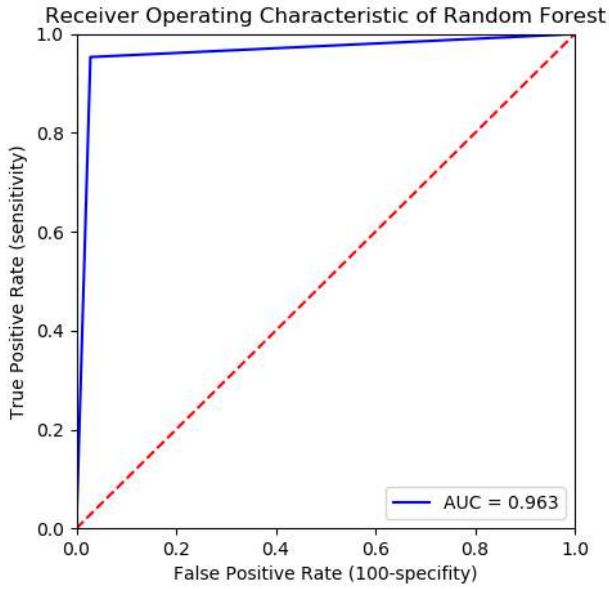
Tablo 2'de modellerin duyarlılıkları karşılaştırıldığında Rastgele orman ve Lojistik regresyon modellerinin doğru pozitifleri tespit etme oranlarının eşit ve diğer modellerden daha yüksek olduğu görülmektedir. Destek vektör makinesi ve Naive bayes eşit oranda ve ikinci sırada, K en yakın komşu modeli ise en düşük orana sahiptir. Bunun yanında doğru negatifleri ayırt etme başarısında tüm modeller genel anlamda yüksek belirleyiciliğe sahip olup Naive bayes ve Lojistik regresyon modelleri %100 belirleyici orana sahiptir. Yanlış pozitifleri eleme oranında da yine Naive bayes ve Lojistik regresyon modelleri %100 kesinlik sağlamaktadır.

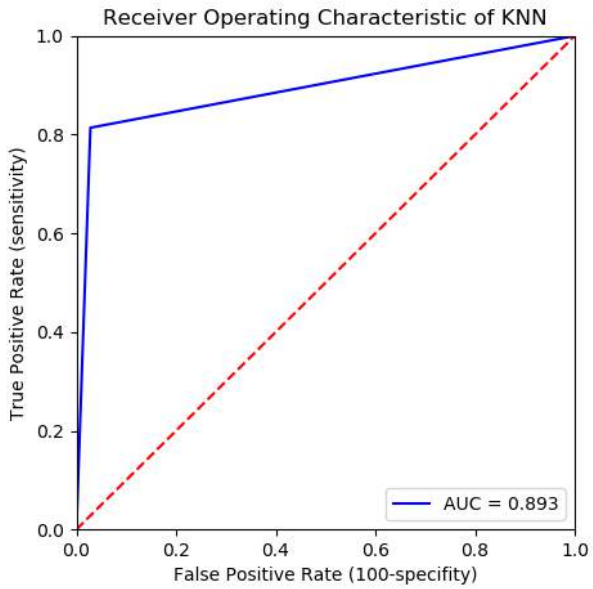
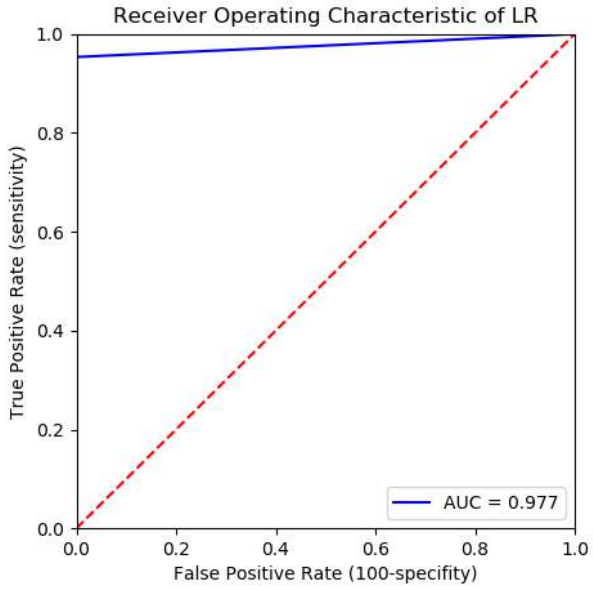
Tüm başarı kriterlerine bağlı olarak en yüksek doğruluğu sağlayan modelin Lojistik regresyon olduğu görülmektedir. Bunun ardından Rastgele orman, Naive bayes ve Destek vektör makinesi gelmektedir. En düşük doğruluğu veren ise K en yakın komşu modelidir.

Modellerin ROC eğrileri ve hesaplanan eğri altındaki alan (AUC) değerleri Tablo 3'te verilmiştir.

Tablo 3. Model ROC eğrileri (Model ROC curves)

Model	ROC eğrisi	AUC
SVM		0,923

NB		0,930
RF		0,963

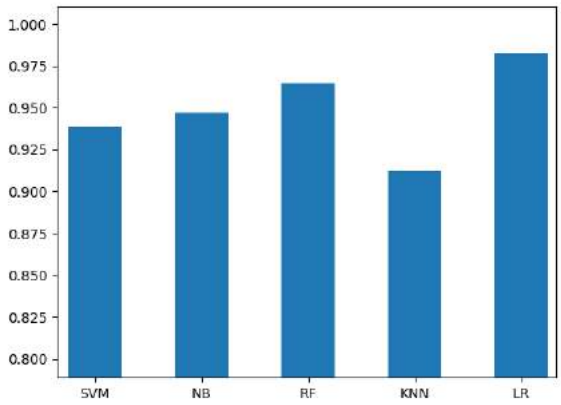
KNN		0,893
LR		0,977

ROC eğrisinin sol üst köşeye yaklaşması, doğru pozitif oranının yüksek ve eğri altında kalan alanın fazla olduğunu gösterir. Buradan yola çıkarak pozitiflerin negatiflerden başarılı bir şekilde ayrılıp ayrılmadığı görülebilir. Tablo 3 incelendiğinde de Lojistik Regresyon (LR) modelinin 0.977 AUC değeri ile diğer modellere göre daha başarılı bir ayrım yaptığı görülmektedir.

4. Sonuç

Makine öğrenmesi pek çok farklı alanda olduğu gibi tıp alanında da yaygın bir şekilde kullanılmakta, hastalıkların teşhisinde destekleyici bir sistem rolü üstlenmektedir. Özellikle kanser teşhisi konusunda artan bir kullanıma sahiptir. Göğüs kanseri, tüm kanser türleri içerisinde en yaygın görülen ikinci kanser türü olup, doğru ve erken teşhis edilmediği takdirde ölümcül olabilmektedir. Bu nedenle göğüs kanseri teşhisinin doğru ve yüksek başarımlı olarak gerçekleştirilmesi büyük önem taşımaktadır.

Bu çalışmada Wisconsin Üniversitesi hastanesinde toplanan göğüs kanseri bulgularını içeren 569 adet örnekten oluşan veri seti, beş farklı makine öğrenmesi modeli ile sınıflandırılarak, modellerin başarıları karşılaştırılmıştır (Şekil 2).



Şekil 2. Modellerin başarı karşılaştırması

Destek Vektör Makinesi (SVM), Naive Bayes (NB), Rastgele Orman (RF), K En yakın Komşu (KNN) ve Lojistik Regresyon (LR) modelleri kullanılarak yapılan sınıflandırmada veri setinin %20'lik kısmı (114 adet örnek) test için kullanılmış geriye kalan %80'lik kısım ile model eğitilmiştir. Test işlemi sonucunda en başarılı modelin %98,24 doğruluk oranıyla Lojistik regresyon olduğu ortaya konmuştur.

Makine öğrenmesi teknikleri ile bilgisayarlar insan uzmanların verdikleri bilgileri öğrenerek otonom kararlar üretme kabiliyetine kavuşmaktadır. Bu yolla uzmanlara destek birer sistem rolü üstlenmekte, öğrenme verileri artıp çeşitlendikçe daha başarılır sonuçlar üretebilmektedirler.

Kaynakça

[1] Kolay, N., Erdoğmuş, P., The classification of breast cancer with Machine Learning Techniques. In Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT), 1-4, 2016.
[2] Jemal, A., Siegel, R., Xu, J., Ward, E., Cancer statistics 2010, CA: a cancer journal for clinicians, 60(5), 277-300, 2010.
[3] Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A., Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA: a cancer journal for clinicians, 68(6), 394-424, 2018.
[4] Papageorgiou, E. I., Jayashree Subramanian, Karmegam, A., & Papandrianos, N., A risk management model for familial breast cancer: A new application using Fuzzy Cognitive Map method, Computer Methods and Programs in Biomedicine, 122(2), 123–135, 2015.
[5] Tapak, L., Shirmohammadi-Khorram, N., Amini, P., Alafchi, B., Hamidi, O., & Poorolajal, J., Prediction of survival and metastasis in breast cancer patients using machine learning classifiers, Clinical Epidemiology and Global Health, 2018.
[6] Cruz, J. A., Wishart, D. S., Applications of machine learning in cancer prediction and prognosis, Cancer informatics, 2, 2006.
[7] Jemal, A., Bray, F., Center, M. M., Ferlay, J., Ward, E., Forman, D., Global cancer statistics, CA: a cancer journal for clinicians, 61(2), 69-90, 2011.
[8] Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., Fotiadis, D. I., Machine learning applications in cancer prognosis and prediction, Computational and structural biotechnology journal, 13, 8-17, 2015.
[9] Thyagarajan, R., Murugavalli, S., Segmentation of Digital Breast Tomograms using clustering techniques, In India Conference (INDICON), 2012 Annual IEEE, 1090-1094, 2012.
[10] Heriana, O., Soesanti, I., Tumor size classification of breast thermal image using fuzzy C-Means algorithm, In Radar, Antenna, Microwave, Electronics, and Telecommunications (ICRAMET), 2015 International Conference on, 98-103, 2015.
[11] Belciug, S., Salem, A. B., Gorunescu, F., Gorunescu, M., Clustering-based approach for detecting breast cancer recurrence, In Intelligent Systems Design and Applications (ISDA), 2010 10th International Conference on, 533-538, 2010.
[12] Abbosh, Y. M., Yahya, A. F., Abbosh, A., Neural networks for the detection and localization of breast cancer, In Communications and Information Technology (ICCIT), 2011 International Conference on, 156-159, 2011.
[13] Isa, N. A. M., Hamid, N. H. A., Sakim, H. A. M., Mashor, M. Y., Zamli, K. Z., Intelligent classification system for cancer data based on artificial neural network, In Cybernetics and Intelligent Systems, 2004 IEEE Conference on, 196-201, 2004.
[14] Pawar, P. S., Patil, D. R., Breast cancer detection using neural network models, In Communication Systems and Network Technologies (CSNT), 2013 International Conference on, 568-572, 2013.
[15] Shen, L., Chen, H., Yu, Z., Kang, W., Zhang, B., Li, H., Liu, D., Evolving support vector machines using fruit fly optimization for medical data classification, Knowledge-Based Systems, 96, 61-75, 2016.

[16] Banu, G. S., Fareeth, A., Hundewale, N., Prediction of breast cancer in mammagram image using support vector machine and fuzzy C-means, In Biomedical and Health Informatics (BHI), 2012 IEEE-EMBS International Conference on, 573-576, 2012.
[17] Majid, A., Ali, S., Iqbal, M., Kausar, N., Prediction of human breast and colon cancers from imbalanced data using nearest neighbor and support vector machines, Computer methods and programs in biomedicine, 113(3), 792-808, 2014.
[18] Ribeiro, A. C., Silva, D. P., Araujo, E., Fuzzy breast cancer risk assessment, In Fuzzy Systems (FUZZ-IEEE), 2014 IEEE International Conference on, 1083-1087, 2014.
[19] Keleş, A., Keleş, A., Yavuz, U., Expert system based on neuro-fuzzy rules for diagnosis breast cancer, Expert systems with applications, 38(5), 5719-5726, 2011.
[20] Ravdin, P. M., Clark, G. M., A practical application of neural network analysis for predicting outcome of individual breast cancer patients, Breast cancer research and treatment, 22(3), 285-293, 1992.
[21] Mangasarian, O. L., Street, W. N., Wolberg, W. H., Breast cancer diagnosis and prognosis via linear programming, Operations Research, 43(4), 570-577, 1995.
[22] Ravi, V., Zimmermann, H. J., Fuzzy rule based classification with FeatureSelector and modified threshold accepting, European Journal of Operational Research, 123(1), 16-28, 2000.
[23] Delen, D., Walker, G., Kadam, A., Predicting breast cancer survivability: a comparison of three data mining methods. Artificial intelligence in medicine, 34(2), 113-127, 2005.
[24] Polat, K., Güneş, S., Breast cancer diagnosis using least square support vector machine, Digital signal processing, 17(4), 694-701, 2007.
[25] Khan, M. U., Choi, J. P., Shin, H., Kim, M., Predicting breast cancer survivability using fuzzy decision trees for personalized healthcare, In Engineering in Medicine and Biology Society 30th Annual International Conference of the IEEE , 5148-5151, 2008.
[26] Chauhan, N., Ravi, V., Chandra, D. K., Differential evolution trained wavelet neural networks: Application to bankruptcy prediction in banks, Expert Systems with Applications, 36(4), 7659-7665, 2009.
[27] Karabatak, M., Ince, M. C., An expert system for detection of breast cancer based on association rules and neural network, Expert systems with Applications, 36(2), 3465-3469, 2009.
[28] Costantino, J. P., Gail, M. H., Pee, D., Anderson, S., Redmond, C. K., Benichou, J., Wieand, H. S., Validation studies for models projecting the risk of invasive and total breast cancer incidence, Journal of the National Cancer Institute, 91(18), 1541-1548, 1999.
[29] Tyrer, J., Duffy, S. W., Cuzick, J., A breast cancer prediction model incorporating familial and personal risk factors, Statistics in medicine, 23(7), 1111-1130, 2004.
[30] Parmigiani, G., Berry, D. A., Aguilar, O., Determining carrier probabilities for breast cancer–susceptibility genes BRCA1 and BRCA2, The American Journal of Human Genetics, 62(1), 145-158, 1998.
[31] Powell, M., Jamshidian, F., Cheyne, K., Nititham, J., Prebil, L. A., Ereman, R., Assessing breast cancer risk models in Marin County, a population with high rates of delayed childbirth, Clinical breast cancer, 14(3), 212-220, 2014.
[32] Alharbi, A., Tchier, F., Using a genetic-fuzzy algorithm as a computer aided diagnosis tool on Saudi Arabian breast cancer database, Mathematical biosciences, 286, 39-48, 2017.
[33] AKYOL, K., Meme Kanseri Tanısı İçin Özniteliklerin Öneminin Değerlendirilmesi Üzerine Bir Çalışma, Academic Platform Journal of Engineering and Science, 6(2), 109–115, 2018.
[34] Cortes, C., Vapnik, V., Support-vector networks, Machine learning, 20(3), 273-297, 1995.
[35] Wood, A., Shpilrain, V., Najarian, K., Kahrobaei, D., Private naive bayes classification of personal biomedical data: Application in cancer data analysis. Computers in biology and medicine, 105, 144-150, 2019.
[36] Breiman, L., Random forests. Machine learning, 45(1), 5-32, 2001.
[37] Yang, Y., Loog, M., A benchmark and comparison of active learning for logistic regression, Pattern Recognition, 83, 401-415, 2018.