SETSCI conference proceedings
SETSCI Conference Proceedings
9, 508-511, 2019

4th International Symposium on Innovative Approaches in Engineering and Natural Sciences
November 22-24, 2019, Samsun, Turkey

https://doi.org/10.36287/setsci.4.6.145
2687-5527/ © 2019 The Authors. Published by SETSCI

ISAS WINTER - 2019

Classification of Breast Cancer by Machine Learning Methods

Hakan Kör[1]*+
[1]Department of Computer Technology, Sungurlu Vocational School, University of Hitit, Corum, Turkey
*Corresponding author: hakankor19@gmail.com
Presentation/Paper Type: Oral / Full Paper

Abstract **–** In this study, the classification success of breast cancer was tested on R and python programming language with 569 records and 31 columns. The first stage consists of clearing the data set, removing null values, and checking the data. Then, the correlation between each variable was examined and presented with graph. Principal component analysis was applied to the variables, and the results were given by graphs and then variables were grouped. 70% of the data set was selected as training data and 30% was selected as test data. In this study, different machine learning methods were applied in R and python programs. Svm method has been found to have the highest value with 97,66% accurate classification rate of breast cancer as benign and malignant.

Keywords *– Machine learning methods, Breast cancer and machine learning, Classification with machine learning, Binary classification of breast cancer*

Meme Kanserinin Makine Öğrenmesi Yöntemleri İle İkili Sınıflandırılması

Hakan Kör[1*]

[1]Bilgisayar Programcılığı/ Sungurlu Meslek Yüksekokulu, Hitit Üniversitesi, Çorum, Turkey
*Sorumlu yazar: hakankor19@gmail.com
Sunum / Türü : Sözlü / Tam Metin

Özet **–** Bu çalışmada, 569 kayıt ve 31 sütundan oluşan veri setiyle Python ve R programlama dilleri üzerinde meme kanserinin ikili sınıflama başarısı test edilmiştir. İlk aşama veri setinin temizlenmesi, boş değerlerin kaldırılması ve verilerin denetlenmesinden oluşmaktadır. Daha sonra her bir değişken arasındaki korelasyon incelenmiş ve görsel olarak sunulmuştur. Değişkenler için temel bileşen analizi uygulanarak sonuçlar grafiklerle verilmiş ve değişkeler gruplandırılmıştır. Veri setinin % 70'i eğitim verisi % 30'u ise test verisi olarak seçilmiştir. Çalışmada, R ve python programında farklı makine öğrenmesi yöntemleri uygulandı. Svm metodunun meme kanserini iyi huylu ve kötü huylu olarak % 97,66 doğru sınıflama oranı ile en yüksek değere sahip olduğu tespit edilmiştir.

Anahtar Kelimeler *– Makine öğrenmesi yöntemleri, Meme kanseri ve makine öğrenmesi, Makine öğrenmesi ile ikili sınıflama, Meme kanserinin ikili sınıflandırılması*

I. GİRİŞ

Günümüzde bilgisayarların, depolama, matematiksel hesap yapma hızı gibi özelliklerinin çok gelişmiş olması önceden uygulanması zor işlemlerin dijital ortama taşınmasını kolaylaştırmıştır. Özellikle yapay zeka alanında yürütülen çalışmalar için hızlı işlem yapan bilgisayarların katkısı ciddi derecede artmıştır. Yapay zeka alanında yapılan çalışmalar incelendiğinde tıp alanında yapılan çalışmaların daha fazla arttığı literatürden takip edilmektedir.

Bilgisayar bilimi alanında en çok duyduğumuz kavramlardan biri yapay zeka ve alt dallarıyla ilgili kavramlardır. Bilgisayar grafikleri, veritabanı yönetim sistemleri ve yapay zeka alanlarındaki araştırmalar, daha hızlı ve daha güçlü donanım platformlarının geliştirilmesiyle birlikte, bilgisayar mühendisliği problemlerinin çözümü için kullanımını hızlandırdı ve daha geniş bir alana yayıldı [1]. En basit ifadeyle yapay zeka, insan zekasını örnek alarak topladıkları verilerle yinelemeli olarak kendilerini iyileştirebilen sistemler olarak tanımlanabilir[2]. Yapay zeka, makine öğrenmesi ve derin öğrenme gibi kavramlar birbirleri ile ilişkili kavramlar olup, makine öğrenmesi yapay zekanın bir alt dalı olarak kabul edilmektedir.

Makine öğrenmesi, önceki gözlemlerden yararlanılarak doğru tahminler yapabilmek amacıyla sistematik tekniklerin geliştirilmesidir [3]. Diğer bir ifadeyle makine öğrenmesi, deneyimlerden yola çıkarak öğrenme işlemini otomatik olarak geliştiren ve yürüten bilgisayar sistemlerinin geliştirilmesidir [4]. Bu alanda yapılan araştırmaların artmasıyla beraber yapay sinir ağları, hesaplamalı öğrenme teorisi, istatistik ve örüntü tanıma gibi araştırma alanları arasında ilişki kurulmuş ve bu disiplinler bilimsel çalışmalarını birlikte yürütmüşlerdir. Böylece makine öğrenmesi çalışma alanı genişleyerek yüz

tanıma gibi daha geleneksel problemlerin yanı sıra veritabanlarında bilgi keşfi, doğal dil işleme ve robot kontrolü gibi yeni problemlerin çözmünde kullanılmaya başlanmıştır [5]. Son zamanlarda makine öğrenmesi tekniklerilerinin yaygın olarak uygulandığı alanlardan biride tıp bilimidir. Özellikle mevcut olan hastalık verilerinden geleceğe dönük tahminlerde makine öğrenmesi ciddi başarı oranlarına sahiptir. Bu çalışmada da Wisconsin Üniversitesi tarafından paylaşılan meme kanseri hastalığı verilerine makine öğrenmesi algoritmaları uygulanmıştır. Araştırmalara göre kadınlarda en sık rastlanan ve ölüme neden olan kanser türü meme kanseridir[6]. Son zamanlarda, erken tanı ve tedavi imkanları ile kanserli hastada yaşam süresinin uzamasını sağlamaktadır [7, 8]. Literatür incelediğinde, meme kanserinin teşhisinin doğru tahmininde makine öğrenmesi yönteminin kullanıldığı çok sayıda yayına rastlanmaktadır. Bu çalışmada, makine öğrenmesi algoritmaları kullanarak mevcut meme kanseri verilerinden kanserin iyi huylu mu kötü huylu mu olduğunu sınıflama işlemi yapılmıştır.

Table 1. Meme Kanseri Teşhisinde Makine Öğrenmesi Kullanan Çalışmalar

Yazar	Yöntem	Doğruluk / Oranı
Burcu Bektaş, Sebahattin Babur, 2016	Makine Öğrenmesi Teknikleri Kullanılarak Meme Kanseri Teşhisinin Performans Değerlendirmesi	DVM (Lineer), 67,01
Onur Sevli, 2019	Göğüs Kanseri Teşhisinde Farklı Makine Öğrenmesi Tekniklerinin Performans Karşılaştırması	Lojistik regresyon %98,24
Polat ve Güneş, 2007	Breast cancer diagnosis using least square support vector machine	En küçük kare destek vektör makinesi (LS-SVM), 98
Alharbi ve Tchier, 2017	Using a genetic-fuzzy algorithm as a computer aided diagnosis tool on Saudi Arabian breast cancer	Bulanık-genetik hibrit algoritma, 97
Hidayet Takçı, 2016	Centroid sınıflayıcılar yardımıyla meme Kanseri teşhisi	Euclidian tabanlı centroid, 99,04
Hülya OLMUS, Semra	Bayes Ağlarda Kümeleme Metotunu Kullanarak Meme Kanseri Tanisinin Modellenmesi	Bayes
AKYOL, K.,2018	Meme Kanseri Tanısı İçin Özniteliklerin Öneminin Değerlendirilmesi Üzerine Bir	Özyinelemeli Özellik Elemesi
Papageorgi ou vd.,2015	A risk management model for familial breast cancer: A new application using Fuzzy Cognitive	Fuzzy Cognite Map (FCM), 95

Tablo 1 incelendiğinde, meme kanserinin teşhisine yönelik araştırmalarda destek vektör makinesi, lineer algoritmalar, bulanık mantık, bayes, öz yinelemeli özellik elemesi ve lojistik regresyon gibi algoritmaların yer aldığı görülmektedir [9, 10, 11, 12, 13, 14, 15, 16].

II. MATERYAL VE YÖNTEM

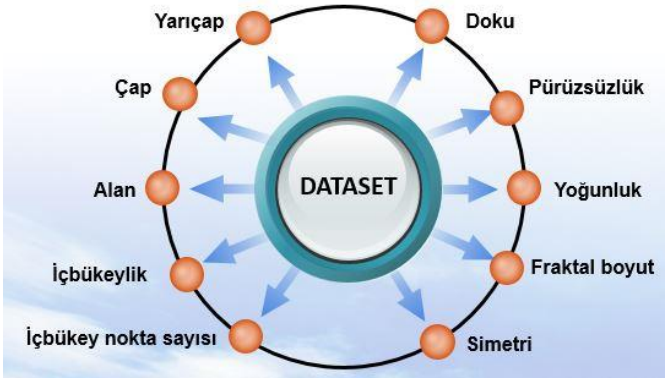
Çalışmanın bu bölümünde kulllanlan veri setine, verinin işlenmesi için kullanılan program ve gerekli kütüphanelere, veri işleme adımlarına yer verilmiştir.

A. Veri Seti ve Düzenlenmesi

Çalışmanın verileri, 1992 yılında Wisconsin Üniversitesi tarafından paylaşıma açılan ve birçok araştırmacı tarafından üzerinden çalışılan bir veri setidir. Veri seti 569 kayıt ve 31 özellikten oluşmaktadır[17].

B. Veri İşleme Adımları

Araştırma verisinde toplam 32 değişken yer almaktadır. Bu değişkenlerden ilk 2 sütun id ve teşhis özellilerini içermektedir.



Şekil 1. Meme Kanseri Ölçümü Yapılan Veriler

Şekil 1'de ölçülen 10 farklı meme kanseri verisi yer almaktadır. Geriye kalan 20 değişken ölçülen bu değerlerden elde edilmiştir. Veri işleme adımları şöyle devam etmektedir; veri temizleme işlemi, teşhis tanımlama(etiketleme), kötü huylu [Malignant], iyi huylu [Benig], algoritmaların çalıştırılması, sonuçların kıyaslanması ve en iyi sonucun seçilmesi. Veri setinin % 63 ü iyi huylu, %37 kötü huylu tümörden oluştuğu bilinmektedir.

C. Kütüphane ve Veri Setinin Eklenmesi

Araştırmada Python ve R olmak üzere veri işlemede en çok tercih edilen iki dil kullanılmıştır. Aşağıda yer alan tablo 2'de veri işleme için gerekli kütüphanelere ve veri setinin programa dahil edildiği kod parçalarına yer verilmiştir.

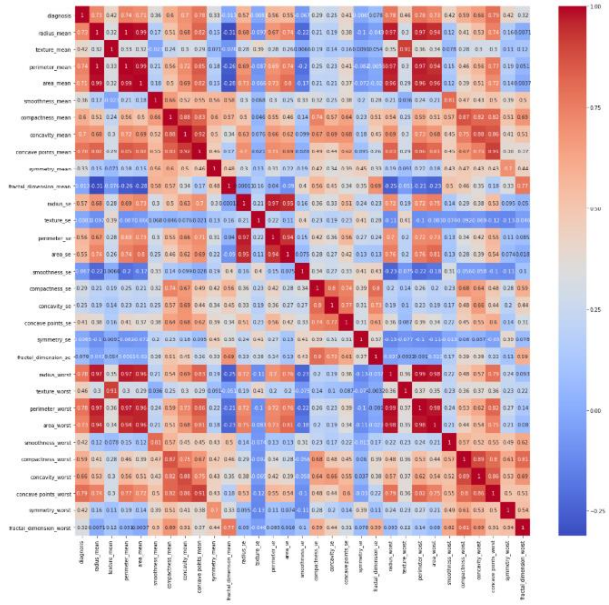
Tablo 2. Python ile Veri İşlemenin İlk Adımı

Kod
import pandas as pd import numpy as np import seaborn as sns import matplotlib.pyplot as plt %matplotlib inline import warnings warnings.filterwarnings('ignore') data = pd.read_csv('../input/breast-cancer-wisconsin-data/data.csv')

	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean
0	1	17.99	10.38	122.80	1001.0	0.11840	0.27760
1	1	20.57	17.77	132.90	1326.0	0.08474	0.07864
2	1	19.69	21.25	130.00	1203.0	0.10960	0.15990
3	1	11.42	20.38	77.58	386.1	0.14250	0.28390
4	1	20.29	14.34	135.10	1297.0	0.10030	0.13280

Şekil 2. Veri Seti Görünümü

Şekil 2' de verilerin ön işleme basamağı sonucunda elde edilen ilk görünümü yer almaktadır.



Şekil 3. Değişkenlerin Birbirleri Arasındaki İlişki

Şekil 3'te ise değişkenleri birbirleri arasındaki ilişkilerine yer verilmiştir. Veri setinin % 70'i eğitim, % 30'u ise test verisi olarak kullanılmıştır.

III. BULGULAR

Çalışma veri setinin 2 farklı dil ve 11 farklı algortima ile işlenmesi sonucunda tablo 3'te yer alan bulgular elde edilmiştir.

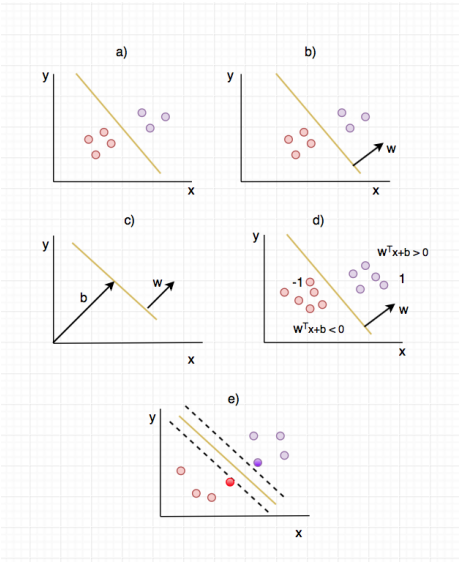
Tablo 3. Makine Öğrenmesi Algortimalarının Doğru Sınıflama Değerleri

Model	**Başarı Oranı**
Logistic Regression	97.36
K Nearest Neighbors Classifier	94.73
Support Vector Classifier (SVC)	96.49
Decision Tree Classifier	94.73
Random Forest Classifier	95.61
Svm	**97,66**
C5	94,15
Rpart	95,32
OneR	87,72
NaviBayes	92,4
K-Means	83,04

Tablo 3'te yer alan algoritmalardan ilk 5 tanesi python geriye kalan 6 tanesi R veri işleme programları ile işleme tabi tutulmuştur. Tablodan en başarılı sınıflama başarısına Support Vector Machine (svm) algoritmasının sahip olduğu görülmektedir. Türkçe destek vektör makineleri olarak ifade edilen bu algoritma 1963 yılında Vladimir Vapnik ve Alexey Chervonenkis tarafından temelleri atılan "Destek Vektör Makineleri (DVM)" istatiksel öğrenme teorisine dayanan bir gözetimli öğrenme algoritmasıdır. Bu algoritma temelleri 60'lı yıllara dayansa da 1995 yılında Vladir Vapnik, Berhard Boser ve Isabelle Guyon tarafından geliştirilmiştir. Destek Vektör Makineleri esasında iki sınıfa ait verileri birbirinden en uygun şekilde ayırmak amacıyla kullanılır. Bilişim alanında, yüz tanıma sistemleri ve ses analizi gibi işlemlerde sıklıkla tercih edilen algoritmalardan biridir[18].

Destek Vektör Makinelerinin işlevi şöyle ifade edilebilir;

$$\mathcal{D} = \{(\mathbf{x}_i, c_i)|\mathbf{x}_i \in \mathbb{R}^p, c_i \in \{-1, 1\}\}_{i=1}^{n}$$



Şekil 4. Destek Vektör Makinelerinin İşlevi

Şekil 4'te destek vektör makinelerinin grafik gösterimi yer almaktadır. Destek vektörlerinin üzerinde kesikli çizgilerle gösterilmiş düzlemlere sınır düzlemleri denir. Her iki düzleme de eşit uzaklıkta bulunan ve sınır düzlemlerinin tam ortasından geçen düzlem ise hiper düzlem olarak ifade edilir. Şekilde (-1, +1) sınıf etiketlerini, w ağırlık vektörünü ve b ise eğilim değerini göstermektedir [19, 20].

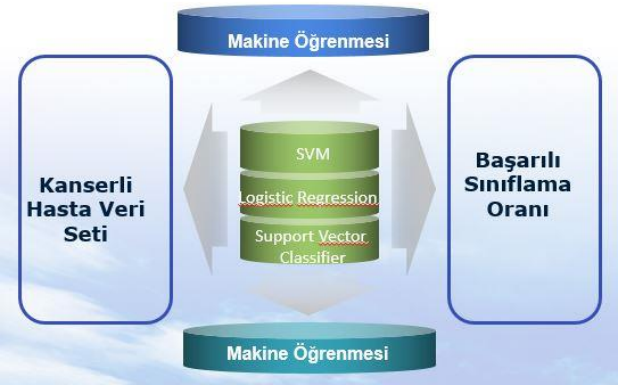
IV. TARTIŞMA

Bu çalışma sonucunda da görüldüğü gibi bilgisayar bilimlerinde yaşanan gelişmelerden tıp bilimide kendine düşün payı fazlasıyla almaktadır. Tahmin ve sınıflama algoritmalarının önceleri düşük seviyelerde olan yetenekleri bilgisayarların yazılım ve donanımsal olarak üstün yeteneklere sahip olmasıyla yüksek seviyelerde sınıflama ve tahmin başarısına ulaşmıştır. Bu tür çalışmalarda araştırmacılar için en çok sıkıntı çekilen konu kuşkusuz veri elde etme sürecidir. Bu süreçte hastalıklar konusunda kapsamlı bilgiye sahip olması nedeniyle tıp doktorlarının veri işleme basamaklarına ciddi katkıları oldu ifade edilbilir.

Diğer yandan elde edilen verilerin temizlenmesi ve işlemeye hazır hale getirilmesi çalışmaya harcanan zamanın neredeyse yüzde 90'ını almaktadır. Bu tür çok zaman ve emek isteyen çalışma sonuçlarının uygulanabilir platform ve proje haline getirillmesi bilime ve ekonomiye önemli derecede katkı sağlayacaktır.

V. SONUÇ

Bu çalışmada, meme kanserinde en sık kulanılan veri setlerinden biri incelenmiş ve destek vektör makineleri algoritmasının diğer algoritmalardan daha başarılı olduğu tespit edilmiştir. Araştırmanın kavramsal modeli şekil 5'te gösterildiği gibi hastalık verileri giriş değelerini ifade ederken işin merkezinde verileri işleyen makine öğrenmesi algoritmaları görev yapmaktadır. Çıktı olarak ise sınıflama başarısı yer almaktadır.



Şekil 5. Çalışmanın Kavramal Modeli

Çok fazla sayıda güncel meme kanseri hastalık verilerini içeren ve çok merkezli bir yapıdan elde edilecek veri setiyle farklı sonuçların alınabileceği düşünülmektedir.

REFERENCES

[1] Krishnamoorthy, C. S., & Rajeev, S. (2018). Artificial intelligence and expert systems for engineers. CRC press.

[2] Oracle, 2019, https://www.oracle.com/tr/artificial-intelligence/ Erişim Tarihi: 1 Kasım 2019.

[3] Schapire, R. E. , 2003. The boosting approach to machine learning: An overview. In Nonlinear estimation and classification (pp. 149-171). Springer, New York.

[4] Ayodele, T. O. , 2010. Machine learning overview, Intech Open Access Publisher.

[5] Dietterich, T. G. , 1997. Machine-learning research, AI magazine, 18(4), 97.

[6] Siegel RL, Miller KD, Jemal A. Cancer statistics, 2015. CA Cancer J Clin. 2015;65(1):5-29.

[7] Mystakidou K, Tsilika E, Parpa E, Galanos A, Vlahos B. Caregivers of advanced cancer patients:feelings of hopelessness and depression. Cancer Nurs 2007; 30:412–8.

[8] 10. Kitrungroter L, Cohen MZ. Quality of life of family caregivers of patients with cancer: aliterature review. Oncol Nurs Forum 2006; 33:625–32.

[9] Bektaş, B., & Babur, S. (2016, October). Machine learning based performance development for diagnosis of breast cancer. In 2016 Medical Technologies National Congress (TIPTEKNO) (pp. 1-4). IEEE.

[10] Sevli, O. Göğüs Kanseri Teşhisinde Farklı Makine Öğrenmesi Tekniklerinin Performans Karşılaştırması. Avrupa Bilim ve Teknoloji Dergisi, (16), 176-185.

[11] Polat, K., & Güneş, S. (2007). Breast cancer diagnosis using least square support vector machine. Digital signal processing, 17(4), 694-701.

[12] Alharbi, A., & Tchier, F. (2017). Using a genetic-fuzzy algorithm as a computer aided diagnosis tool on Saudi Arabian breast cancer database. Mathematical biosciences, 286, 39-48.

[13] Takcı, H. (2016). Centroid Sınıflayıcılar Yardımıyla Meme Kanseri Teşhisi. Gazi Üniversitesi Mühendislik-Mimarlık Fakültesi Dergisi, 31(2).

[14] Olmus, H., & Erbas, S. (2012). Bayes Ağlarda Kümeleme Metotunu Kullanarak Meme Kanseri Tanisinin Modellenmesi. Turkiye Klinikleri Journal of Biostatistics, 4(1).

[15] AKYOL, K. (2018). Meme Kanseri Tanısı İçin Özniteliklerin Öneminin Değerlendirilmesi Üzerine Bir Çalışma. Academic Platform Journal of Engineering and Science, 6(2), 109-115.

[16] Papageorgiou, E. I., Subramanian, J., Karmegam, A., & Papandrianos, N. (2015). 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.

[17] Wolberg, W. H., Street, W. N., & Mangasarian, O. L. (1992). Breast cancer Wisconsin (diagnostic) data set. UCI Machine Learning Repository [http://archive. ics. uci. edu/ml/].

[18] DATA -Veri Madenciliği Veri Analizi (Haldun Akpınar), Papatya Bilim, 2014.

[19] Osuna, E.E., Freund, R., Girosi, F. (1997). "Support Vector Machines: Training and Applications, A.I. Massachusetts Institute of Technology and Artificial Intelligence Laboratory, Massachusetts".

[20] Kavzoğlu, T., ve Çölkesen, İ.(2010). "Destek Vektör Makineleri ile Uydu Görüntülerinin Sınıflandırılmasında Kernel Fonksiyonlarının Etkilerinin İncelenmesi"