Araştırma Makalesi
Zeki Sistemler Teori ve Uygulamaları Dergisi 5(1) (2022) 35-41
DOI: 10.38016/jista.966517

Göğüs Kanseri Verileri Üzerinde Makine Öğrenmesi Yöntemlerinin Uygulanması

Ebru Aydındağ Bayrak[1]*, Pınar Kırcı[2], Tolga Ensari[3], Engin Seven[4], Mustafa Dağtekin[5]

[1] İstanbul Üniversitesi-Cerrahpaşa, Mühendislik Bilimleri Bölümü, İstanbul, Türkiye

[2] Bursa Uludağ Üniversitesi, Bilgisayar Mühendisliği Bölümü, Bursa, Türkiye

[3] Arkansas Tech University, Department of Computer and Information Science, Russellville, USA

[4,5] İstanbul Üniversitesi-Cerrahpaşa, Bilgisayar Mühendisliği Bölümü, İstanbul, Türkiye

ebruaydindag@gmail.com, pinarkirci@uludag.edu.tr, tensari@atu.edu, engin.seven@ogr.iu.edu.tr, dagtekin@iuc.edu.tr

Öz

Kanser dünya genelinde pek çok insanın ölümüne sebep olan en önemli hastalıklardan biridir. Özellikle göğüs kanseri kadınlar arasında en çok rastlanan hastalıkların başında yer almaktadır. Bu sebeple kanser hastalığının teşhisi ile alakalı herhangi bir gelişme insanların sağlıklı bir yaşam sürmesi açısından oldukça önemlidir. Günümüzde makine öğrenmesi yöntemlerinin kullanılması, kanser hastalığının erken teşhisi ve tahmini için yapılan çalışmalara büyük katkılar sağlamaktadır. Bu çalışmada da k-En Yakın Komşu, Destek Vektör Makinaları, Naive Bayes, Karar ağaçları ve Yapay Sinir Ağları gibi beş farklı makine öğrenmesi yöntemleri Kaggle platformunda yer alan iki farklı göğüs kanseri veri kümesi üzerinde uygulanmıştır. Elde edilen sonuçlar doğruluk değerleri ve karmaşıklık matrisi değerleri ile verilerek karşılaştırılmıştır. Birinci göğüs kanseri veri kümesi içinde %98,2456 doğruluk oranıyla ve ikinci göğüs kanseri veri kümesinde %93,8596 doğruluk oranıyla Yapay Sinir Ağları (YSA) yönteminde en yüksek doğruluk değerleri elde edilmiştir.

Anahtar kelimeler: Makine öğrenmesi, göğüs kanseri, sınıflandırma, erken teşhis.

Diagnosing Breast Cancer Using Machine Learning Methods

Abstract

Cancer is one of the most important diseases that cause the death of many people around the world. Especially, breast cancer is one of the most common diseases among women. For this reason, any development related to the diagnosis of cancer is critical for people to live healthy lives. Today, the use of machine learning methods makes great contributions to studies for the early diagnosis and prediction of cancer disease. In this study, five different machine learning methods such as k-Nearest Neighbor, Support Vector Machines, Naive Bayes, Decision Trees, and Artificial Neural Networks were applied on two other breast cancer datasets on the Kaggle platform. The obtained results were compared by giving accuracy values and confusion matrix values. The highest accuracy values were obtained in Artificial Neural Networks (ANN) method with an accuracy rate of 98.2456% in the first breast cancer dataset and 93.8596% in the second breast cancer dataset.

Keywords: Machine learning, breast cancer, classification, early diagnosis.

1. Giriş (Introduction)

Kanser dünya genelinde en çok insan ölümüne sebep olan hastalık türleri arasında ikinci sırada yer almaktadır ve 2018 yılında dünya genelinde yaklaşık olarak 9,6 milyon insan kanser hastalığından dolayı hayatını kaybetmiştir. Yapılan araştırmalarda dünyadaki her 6 ölümden 1 tanesinin kanser yüzünden gerçekleştiği görülmektedir. Özellikle, az ve orta gelişmiş ülkelerde meydana gelen ölümlerin %70'i de yine kanser hastalığından kaynaklanmaktadır (Cancer, 2021).

Kadınlar arasındaki en yaygın kanser türleri ise göğüs, akciğer ve kalın bağırsak kanserleri olup, bu üç kanser türü kadınlarda meydana gelen tüm kanser vakalarının yarısını oluşturmaktadır. Ayrıca meme kanseri kadınlarda teşhis edilen yeni kanser vakalarının %30'luk gibi büyük bir bölümüne karşılık gelmektedir (Siegel vd., 2018).

Bu derece yaygın olan bir hastalığın erkenden teşhis edilmesiyle alakalı yapılacak çalışmaların oldukça önemli olduğu açıkça görülmektedir. Özellikle kanser hastalığı üzerinde gerçekleştirilen makine öğrenmesi uygulamaları incelendiğinde, makine öğrenmesi

* Sorumlu Yazar.
E-posta adresi: ebruaydindag@gmail.com

Alındı : 08 Temmuz 2021
Revizyon : 04 Ekim 2021
Kabul : 30 Ekim 2021

tekniklerinin kanser hastalığının erken teşhis edilebilmesi ve öngörülebilmesi açısından oldukça kullanışlı olduğu açıktır. Makine öğrenmesi yöntemleri var olan verilerin analiz edilmesini ve veri kümesinde var olan ilişkileri ve önemli bilgilerin karakteristik özelliklerinin elde edilmesini sağlar. Ayrıca, verinin iyi şekilde tanımlanabilmesi sağlayan bir hesaplamalı bir model üretir. (Maity ve Das, 2017).

Poyraz (2012) çalışmasında Wisconsin göğüs kanseri veri seti üzerinde veri madenciliği metotlarını uygulayarak, sonuçları başarım ölçütlerine göre karşılaştırmıştır. J48 algoritması Karar ağacı algoritması, Naive Bayes, Lojistik Regresyon ve örnek tabanlı sınıflandırma algoritması olarak K-Star metotları WEKA çalışma ortamında kullanılmıştır. Çalışmanın sonucunda doğruluk değerleri açısından Lojistik Regresyon algoritmasının diğer algoritmalara oranla daha iyi sonuç verdiği görülmüştür (Poyraz, 2012).

Ahmad vd., (2013) çeşitli makine öğrenmesi tekniklerini göğüs kanseri hastalığının iki yıl içinde yeniden ortaya çıkabilme durumunun tahmin edilmesiyle ilgili kullanmışlardır. Çalışmalarında Tahran Ulusal Kanser Enstitüsü veri tabanında yer alan ve 1997-2008 yılları arasını kapsayan ICBC (Iranian Center for Breast Cancer) göğüs kanseri veri setini incelemişlerdir. Veri setinde eksik olan değerler Beklenti Maksimizasyonu algoritması kullanılarak düzenlenmiştir. Makine öğrenmesi yöntemlerinden Karar Ağacı (C4.5), Destek Vektör Makinaları (DVM) ve Yapay Sinir Ağları (YSA) uygulanmıştır. Göğüs kanserinin yeniden tekrarlanmasının tahmini üzerine yapılan bu çalışmada en yüksek doğruluk ve en az hata oranı DVM yönteminde elde edilmiştir.

İşeri (2014) mamogram görüntülerine makine öğrenmesi yöntemlerini uygulayarak göğüs kanserinin teşhis edilebilmesi üzerinde çalışmıştır. Çalışma mamogram görüntülerindeki mikro kireçlenme bölgelerinin tespiti ve bu bölgelerin kötü ya da iyi huylu olma durumlarına göre sınıflandırılması olacak şekilde iki aşamada gerçekleştirilmiştir. MATLAB ortamında göğüs kanseri tespit sistemi isimli (BCDS:Breast Cancer Detection System) bir yazılım geliştirilmiştir. Dört adet özellik çıkarım yöntemi ile Çok Katmanlı Yapay Sinir Ağı ve Destek Vektör Makinaları sınıflandırıcı olarak kullanılarak göğüs kanseri bulgularının tespiti amaçlanmıştır.

Şık (2014) çalışmasında kanser hastalığının erken teşhis edilebilmesinde veri madenciliği uygulamalarının etkisini araştırmıştır. Wisconsin Göğüs Kanseri veri setine WEKA ortamında Bayes Ağı, Naive Bayes, Çok Katmanlı Algılayıcı, Basit Lojistik, Olasılıksal Gradyan İniş, Sıralı Minimal Optimizasyon, IB1, K-Yıldız, PART, Lojistik Model Ağaçları ve Rassal Ormanlar gibi çeşitli sınıflandırma yöntemlerini uygulamıştır. Sınıflandırma sonuçlarını karşılaştırılırken Kappa istatistiği, doğruluk, kesinlik, duyarlılık, F-ölçütü ve ROC alanı gibi parametreler göz önüne alınmıştır. 0,94 Kappa istatistiği, %97,40 doğruluk, 0.97 kesinlik, 0,99 duyarlılık 0,98 F-ölçütü ve 1,00 ROC alanı sonuçlarına göre Basit Lojistik sınıflandırma yöntemi en iyi sonucu vermiştir.

Asri vd., (2016) UCI Makine Öğrenmesi Veri Havuzunda yer alan Wisconsin meme kanseri veri setine Destek Vektör Makinaları, Karar Ağacı (C4.5), Naive Bayes ve k-En Yakın Komşu makine öğrenmesi algoritmalarını WEKA ortamında uygulamışlardır. Yapmış oldukları çalışmada sınıflandırma modellerini değerlendirirken doğruluk, hassasiyet, duyarlılık ve özgüllük parametreleri kullanılmıştır. Uygulamanın sonucunda Destek Vektör Makinaları %97,13 gibi yüksek doğruluk oranı ve %0.02 hata oranıyla en iyi sonucu vermiştir.

Bazazeh ve Shubair (2016) göğüs kanserinin erken teşhis edilmesiyle ilgili yapmış oldukları bu çalışmada Destek Vektör Makinaları, Rastgele Orman ve Bayes Ağı yöntemlerini Wisconsin göğüs kanseri veri setine uygulamışlardır. WEKA yazılımının kullanıldığı bu çalışmada duyarlılık ve hassasiyet değerlerine göre Bayes Ağı en iyi performansı göstermiştir. ROC eğrisi parametresi dikkate alındığında ise Rastgele Orman yöntemi en iyi sonucu vermiştir. Doğruluk, özgünlük ve hassasiyet cinsinden ise en iyi performansı Destek Vektör Makinaları göstermiştir.

Anwer (2017) tez çalışmasında Python ortamında Wisconsin göğüs kanseri veri seti üzerinde çeşitli derin öğrenme algoritmalarını uygulayarak performans sonuçları üzerinden karşılaştırma yapmıştır. Tam bağlantılı sinir ağları, konvolüsyon sinir ağları, basit tekrarlayan sinir ağları, uzun kısa dönem yapay sinir ağları ve kapılı yinelenen birim sinir ağları gibi çeşitli derin öğrenme yöntemleri kullanılmıştır. Ayrıca çalışmada Naive Bayes, k-En Yakın Komşu, Lojistik Regresyon ve Karar Ağacı gibi klasik makine öğrenmesi yöntemleri de uygulanmıştır. Çalışmanın sonucunda derin öğrenme yöntemlerinin klasik makine öğrenmesi yöntemlerine göre daha üstün çalıştığı sonucu elde edilmiştir. Konvolüsyon sinir ağı yönteminde %99,30 ile en yüksek doğruluk değeri elde edilmiştir.

Maity ve Das (2017) Image J programını kullanarak göğüs kanseri hücre görüntülerinden özellik çıkarımı gerçekleştirmiş ve Yapay Sinir Ağı (YSA) algoritmasını uygulamışlardır. Çalışmanın sonucunda %90 doğruluk oranı ile göğüs kanseri hücre görüntüleri doğru şekilde sınıflandırılabilmiştir.

Turgut (2017) yapmış olduğu tez çalışmasında Python ortamında çeşitli makine öğrenmesi yöntemlerini iki farklı mikro dizi göğüs kanseri veri setlerine uygulamıştır. Çalışmada öznitelik seçimleri yapılarak makine öğrenmesi metotlarıyla yüksek doğrulukta tahmin yapılabilmesi amaçlanmıştır. Çalışmada DVM, YSA, k-EYK, Karar Ağaçları, Rastgele Orman, Lojistik Regresyon, Adaboost ve Gradyan Boosting Makina algoritmaları kullanılmıştır. İki göğüs kanseri veri kümesinde de öznitelik yöntemlerin uygulandıktan sonra en yüksek doğruluk DVM yönteminde, en düşük doğruluk Karar Ağaçları yönteminde elde edilmiştir. Ayrıca iki veri kümesinde de kullanılan aynı öznitelik yöntemleri çalışmada

uygulanan tüm makine öğrenmesi algoritmalarında birbirine yakın sonuçlar vermiştir.

Sherafatiyan (2018) çalışmasında göğüs kanseri hastalarının miRNA ekspresyon veri setlerini kullanıp minimal biyo-belirteçleri belirlemek için ağaç tabanlı sınıflandırma modellerinden yararlanmıştır. Önerilen biyo-belirteçlere ek olarak göğüs kanseri sınıflandırmasındaki en önemli mikro RNA'larda açıklanmıştır.

Turgut vd. (2018) çeşitli makine öğrenmesi yöntemlerini iki farklı mikro dizi göğüs kanseri veri seti üzerinde uygulayarak veri sınıflandırması yapmışlardır. Rastgele lojistik regresyon ve yinelemeli öznitelik eleme özellik seçim yöntemleri kullanılarak yüksek doğrulukta kanser teşhisinin yapılması amaçlanmıştır. İki farklı özellik seçim yöntemi uygulandıktan sonra iki mikro dizi göğüs kanseri veri seti içinde destek vektör makinaları en iyi performansı göstermiştir.

Aydındağ Bayrak vd., (2019) Wisconsin (orijinal) göğüs kanseri veri seti üzerinde yapmış olduğu çalışmada, WEKA ortamında Destek Vektör Makinaları ve Yapay Sinir Ağı makine öğrenmesi yöntemlerini uygulamışlardır. Algoritmaların sonuçları doğruluk, hassasiyet, duyarlılık ve ROC alanı gibi performans metriklerine göre karşılaştırıldığında Destek Vektör Makinaları (SMO algoritması) en iyi performansı göstermiştir.

Dhahri vd. (2019) makine öğrenmesi algoritmalarına dayanarak otomatik olarak göğüs kanserinin teşhis edilebilmesi üzerine çalışmışlardır. Özellik tabanlı ön işleme yöntemleri ve sınıflandırma algoritmalarının birleştirilerek kullanılmasının göğüs kanserinin teşhisi üzerinde daha iyi sonuç verebileceğini açıklamışlardır.

Ganggayah vd. (2019) makine öğrenmesi modellerini kullanarak göğüs kanseri hastalığından hayatta kalabilmek için önemli olan prognostik faktörleri belirlemeye çalışmışlardır. Destek Vektör Makinası, Rassal Ağaç, Yapay Sinir Ağları, Ekstrem Boost, Lojistik Regresyon ve Karar Ağacı gibi pek çok makine öğrenmesi algoritması çalışmada kullanılmıştır. Rastgele Orman yöntemi diğer yöntemlere oranla biraz daha yüksek performans sergilemiştir. Buna rağmen çalışmada kullanılan tüm algoritmalar birbirine yakın doğruluk değerleri vermiştir.

Tapak vd. (2019) göğüs kanseri hastaları üzerinde yapmış oldukları çalışmada makine öğrenmesi yöntemlerini kullanarak hayatta kalma ve metastaz tahmininde bulunmuşlardır. Naive Bayes, Rassal Orman, AdaBoost, Destek Vektör Makinası En Küçük Kareler Destek Vektör Makinası, Adabag, Lojistik Regresyon ve Lineer Diskriminant Analizi yöntemlerini uygulamışlardır.

Tseng vd. (2019) makine öğrenmesi teknolojilerini kullanarak göğüs kanseri metastazını belirlemek üzerine çalışma yapmışlardır. Rastgele Orman temelli makine öğrenmesi modelinin göğüs kanseri metastazını en az üç ay önceden tahmin etmek için en uygun yöntem olduğunu belirlemişlerdir.

Magna vd. (2020) hastaların tıbbi geçmiş bilgilerinden faydalanarak göğüs kanserinin sınıflandırmasında makine öğrenmesi, derin öğrenme ve kelime yerleştirme uygulamalarının kullanılması üzerinde çalışma yapmışlardır. Hekimin karar vermesini destekleyen bir öneri sistemi ortaya koymaya çalışmışlardır.

Reddy vd. (2020) destek değerine sahip derin sinir ağı (DNNS) yöntemini göğüs kanseri teşhisi için kullanmıştır. Deneysel sonuçlara göre, önerilen DNNS'nin mevcut yöntemlerden daha iyi sonuçlar verdiği kanıtlanmıştır.

Saxena ve Gyanchandani (2020) histopatolojiyi kullanarak bilgisayar destekli göğüs kanseri teşhisi yapabilmek için makine öğrenmesi yöntemlerini incelemişlerdir. Pek çok farklı yaklaşımı inceledikten sonra göğüs kanseri üzerine yapılan makine öğrenmesi çalışmalarının genellikle derin öğrenme konusunda yoğunlaştığı görülmüştür.

Bu çalışmada da popüler makine öğrenmesi yöntemlerinden olan k-En Yakın Komşu, Destek Vektör Makinaları, Navie Bayes, Karar Ağacı ve Yapay Sinir Ağları Kaggle platformunda yer alan iki farklı göğüs kanseri veri kümesine uygulanmıştır. Uygulamada elde edilen sonuçlar doğruluk performans metriği ve karmaşıklık matrisine göre karşılaştırılmıştır. Çalışmanın devamında, ikinci bölümde kullanılan veri kümelerinin ve makine öğrenmesi yöntemleri kısaca açıklanmaktadır. Ardından uygulamanın deneysel sonuçları, tartışma ve sonuç bölümleri ile çalışma sona ermektedir.

2. Materyal ve Yöntem **(Material and Method)**

2.1. Veri seti (Data set)

Bu çalışmada göğüs kanseri sınıflandırması için Kaggle platformunda yer alan iki farklı göğüs kanseri veri kümesi kullanılmıştır. Birinci veri kümesinde 30 adet özellik içermekte olup toplamda 569 kayıt yer almaktadır. İkinci veri kümesinde ise 569 kayıt yer alıp 5 adet özellik içermektedir. Bu niteliklerden bazıları ortalama yarıçap, ortalama doku, ortalama çevre, ortalama alan ve ortalama yumuşaklıktır. Veri seti Wisconsin-Madison Üniversitesi Hastanesinden elde edilmiştir (Kaggle, 2020).

2.2. Kullanılan makine öğrenmesi yöntemleri (Utilized machine learning methods)

Yapılan çalışmada iki farklı göğüs kanseri veri kümesi için kategorik verilerin nümerik verilere dönüştürülmesi, gereksiz verilerin atılması, tekrarlanan verilerin kaldırılması, normalleştirme gibi veri ön işleme basamakları gerçekleştirilmiştir. Veri ön işleme basamağından sonra elde edilen verilere k-En Yakın Komşu, Destek Vektör Makinaları, Karar Ağacı, Naive Bayes ve Yapay Sinir Ağları gibi popüler makine öğrenmesi yöntemleri Jüpiter notebook ortamında

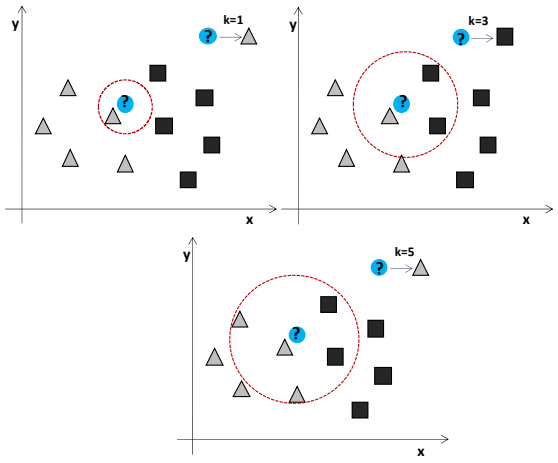
uygulanmıştır. Aşağıda çalışmada kullanılan makine öğrenmesi sınıflandırma yöntemleri kısaca açıklanmaktadır.

2.2.1 k-En Yakın Komşu (k-Nearest Neighbor)

k-En Yakın Komşu algoritması basit ve etkili, aynı zamanda da güçlü bir sınıflandırma yöntemidir. Verilerin sınıflandırılmasında, verilerin birbirleri arasındaki mesafe ölçümü kavramı kullanılmaktadır. Bu yöntem bir denetimli öğrenme yöntemidir, bu nedenle tüm veriler etiketlidir ve her bir veri parçasının hangi sınıfa girmesi gerektiği bilinmektedir. Etiketsiz yani yeni bir veri parçası bize verildiğinde ise, sınıflandırılması için yöntemde uygulanan adımlar aşağıdaki gibi özetlenmektedir (Harrington, 2012):

1. k parametresi belirlenir. k, yeni verilere en yakın olan komşuların sayısıdır.
2. Yeni veri (test) ile mevcut veri (eğitim) arasındaki mesafeler hesaplanır.
3. En yakın mesafe değerleri seçilir. (En yakın komşu bulunur.)
4. Hangi sınıfta en fazla sayıda benzer veri bulursa yeni veriler o sınıfa düşer.

Çalışmada k-en yakın komşu algoritması uygulanırken, en yakın komşu parametresi k=3 olarak belirlenmiştir.

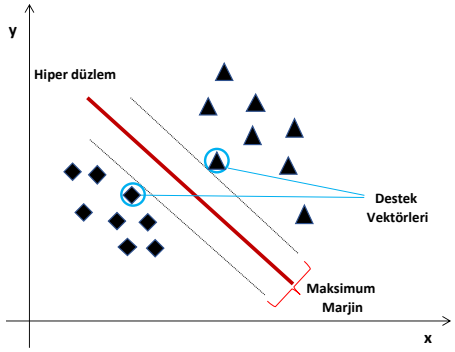


Şekil 1: k-en yakın komşu algoritmasının mimarisi (Ruan vd., 2017'den uyarlanmıştır) (Architecture of the k-nearest neighbor algorithm)

2.2.2 Destek Vektör Makinaları (Support Vector Machines)

Destek Vektör Makinaları (DVM) algoritması ilk kez Vladimir Vapnik (1995) tarafından geliştirilmiştir. DVM en temel haliyle destek vektörlerini kullanarak veri sınıflarını birbirinden ayırmak için en uygun hiper düzlemi bulmaya çalışan, sınıflandırma ve regresyon için kullanılabilen bir makine öğrenme yöntemi olarak açıklanmaktadır. Bu yöntemde öncelikle iki veri sınıfını ayırmak için en uygun hiper düzlemin bulunması amaçlanır ve ardından veri sınıflarının aralarındaki marjin maksimize edildiğinde iki sınıf birbirinden ayrılmaktadır. Eğer sınıflar basit bir hiper düzlemle birbirinden ayrılmazsa, veriler daha yüksek boyutlu yeni bir alana aktarılır ve verileri ayırabilmek için hiper düzlemin bulunması amaçlanır (Burakgazi, 2017).

Çalışmada DVM yöntemi uygulanırken C düzenleme parametresi 1 olarak belirlenmiştir. Ayrıca çekirdek olarak radyal tabanlı kernel fonksiyonu ve gamma parametresinin değeri ise otomatik olarak tercih edilmiştir.



Şekil 2: Destek Vektör Makinaları algoritmasının mimarisi (Alpaydın (2014)'ten uyarlanmıştır) (Architecture of the Support Vector Machines algorithm)

2.2.3. Sade Bayes (Naive Bayes)

Naive Bayes Sınıflandırıcı, Bayes teoremine dayanan en popüler sınıflandırma yöntemlerinden bir tanesidir. Çok basit bir yöntemdir, öyle ki sadece az miktarda eğitim verisiyle bile verilen örnekler sınıflandırabilmektedir. Mevcut ve geçmiş frekans olaylarını hesaplamak için kullanılan Naive Bayes algoritması aşağıdaki gibi açıklanabilir (Umadevi ve Marseline, 2017):

$$P(A|B) = (P(B|A) * P(A))/P(B) \quad (1)$$

P(A): A'nın önsel olasılığı, yalnızca A'nın oluşumlarını saymaktadır.
P(A|B): B verildiğinde A'nın koşullu olasılığıdır. Ayrıca sonsal olasılık olarak adlandırılıp, A'nın B'den türetildiği anlamına gelmektedir.
P(B|A): A verildiğinde B'nin koşullu olasılığıdır.
P(B): B'nin önsel olasılığıdır.

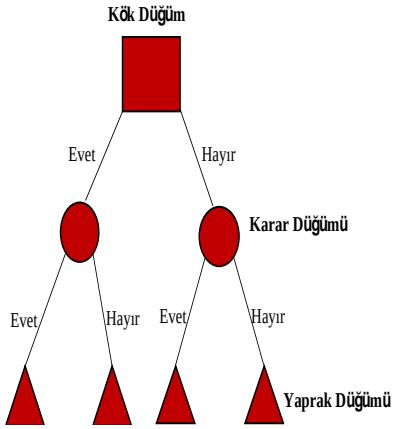
2.2.4. Karar Ağaçları (Decision Trees)

Karar ağaçları algoritmasının temeli böl ve yönet stratejisine dayanmaktadır. Karar düğümleri ve yapraklardan oluşan hiyerarşik bir yapıya sahiptir (Umadevi ve Marseline, 2017).

Karar ağacı yönteminde öncelikle verileri bölmek için hangi özelliğin kullanılacağına karar verilmelidir.

Bunun için her özellik ve ölçüm denenmeli ve ardından elde edilen en iyi sonuçlara göre veri kümelerini alt kümelere bölebiliriz. Yöntemin uygulama adımları aşağıdaki gibi özetlenebilir (Harrington, 2012):

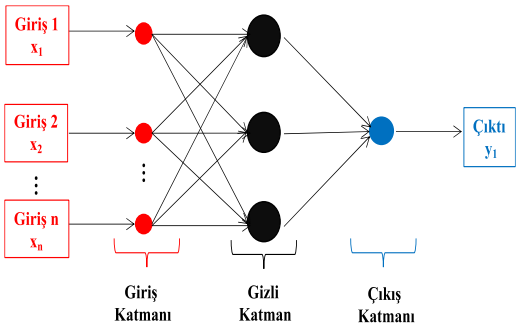
1. Öncelikle tüm veri seti kullanılır.
2. Veri kümesi, bir özelliğin değerine göre iki alt gruba ayrılır. (Bölünen en iyi özellik)
3. Özelliğin tümü aynı sınıfta olana kadar her alt kümeye aynı prosedür uygulanır. Aksi halde bölme işlemine devam edilir.



Şekil 3: Karar Ağacı algoritmasının yapısı gösterilmektedir (Alpaydın (2014)'ten uyarlanmıştır) (The structure of the Decision Tree algorithm)

2.2.5. Yapay Sinir Ağları (Artificial Neural Networks)

Yapay Sinir Ağlarının (YSA) mühendislik çalışmalarındaki amacı sadece insan beynin modellenmesi değildir. Amaç yapay sinir ağlarını kullanarak daha iyi bilgisayarlar yapmak ve insanlara fayda sağlamaktır. İnsan beyni görüntü, konuşma, öğrenme ve tanıma gibi yetenekleri açısından bir mühendislik ürününden fazlasıdır ve bu yeteneklerin yapay zekâ ağları aracılığıyla bilgisayarlara uygulanması oldukça önemlidir (Umadevi ve Marseline, 2017).



Şekil 4: Yapay Sinir Ağlarının (YSA) temel yapısı (Atalay ve Çelik (2017)'den uyarlanmıştır) (Basic structure of Artificial Neural Networks (ANN))

Çalışmada YSA uygulanırken Keras'taki Sequential model tipi kullanılmıştır. Giriş ve çıkış katmanları dahil olmak üzere toplamda 5 katmanlı bir YSA modeli inşa edilmiştir. 3 gizli katmanda nöron sayısı sırasıyla 32, 16 ve 8 olarak belirlenmiş ve Doğrultulmuş Doğrusal Birim (ReLu-Rectified Linear Unit) aktivasyon fonksiyonu kullanılmıştır. Döngü sayısı (epoch) 50, küme büyüklüğü (batch size) ise 10 olarak belirlenmiştir. Ayrıca diğer katmanlarda Doğrultulmuş Doğrusal Birim (ReLu-Rectified Linear Unit) ile Sigmoid aktivasyon fonksiyonu da kullanılmıştır.

YSA modeli derlenirken optimizer için adam algoritması, loss fonksiyonu için binary_crossentropy algoritması ve metrik parametresi için doğruluk değeri kullanılmıştır.

3. Bulgular ve Tartışma (Results and Discussion)

Yapılan çalışmada iki farklı göğüs kanseri veri kümesi için yukarıda bahsedilen beş farklı makine öğrenmesi yöntemleri kullanılmıştır. Veri kümeleri için bazı veri ön işleme basamakları gerçekleştirilmiştir. Veri ön işlemenin ardından yapılan uygulamalarda, birinci veri kümesi için en yüksek doğruluk oranı ile Yapay Sinir Ağları (YSA) yönteminde elde edilmiştir. Tablo 1'de uygulanan beş makine öğrenmesi yönteminin doğruluk değerleri karşılaştırılmıştır. Sonuçlara bakıldığında, YSA yönteminde doğruluk değeri %98,2456 olarak hesaplanmıştır. YSA'dan sonra en yüksek doğruluk değeri ise Destek Vektör Makinalarında (DVM) elde edilmiştir. Bu sonuçları sırasıyla k-En Yakın Komşu, Naive Bayes ve Karar Ağacı algoritmaları takip etmektedir.

Tablo 1: Birinci veri kümesi için makine öğrenmesi yöntemlerinin doğruluk değerlerinin gösterilmesi (The accuracy values of machine learning methods for the first dataset)

Uygulanan Makine Öğrenmesi Yöntemleri	Doğruluk Oranı (%)
k-En Yakın Komşu	94,7368
Destek Vektör Makinaları	97,3684
Naive Bayes	92,1053
Karar Ağacı	86,8421
Yapay Sinir Ağları	**98,2456**

Tablo 2'de ise YSA yöntemine ait karmaşıklık matrisi (confusion matrix) gösterilmiştir. YSA yöntemi için yapılan testlerde ise performans metriklerinden f1 skoru 0,961538 olarak bulunmuştur. Ayrıca duyarlılık ve hassasiyet değerleri 0,961538 olarak bulunmuştur.

Tablo 2: Birinci göğüs kanseri veri kümesine uygulanan YSA yönteminin karmaşıklık matrisi (The complexity matrix of the ANN method applied to the first breast cancer dataset).

		TAHMİN		METRİKLER
		Pozitif	Negatif	
GERÇEK	Pozitif	TP=25	FP=1	Hassasiyet= $\frac{TP}{TP+FP}$=0.961538
	Negatif	FN=1	TN=87	Duyarlılık= $\frac{TP}{TP+FN}$=0.961538
Doğruluk $= \frac{TP+TN}{TP+TN+FP+FN}$=0.982456				F1 Skoru $= 2 * \frac{Hassayiyet * Duyarlılık}{Hassayiyet + Duyarlılık}$=0.961538

İkinci göğüs kanseri veri kümesi için en yüksek doğruluk oranı yine Yapay Sinir Ağları (YSA) yöntemi ile elde edilmiştir. Tablo 3'te de uygulanan beş farklı sınıflandırma yöntemlerinin doğruluk değerleri karşılaştırılmıştır. %93,8596 doğruluk değeriyle YSA yönteminde en yüksek doğruluk değeri hesaplanmıştır. YSA'dan sonra en yüksek doğruluk değeri birinci veri kümesinde olduğu gibi Destek Vektör Makinalarında elde edilmiştir. k-En Yakın Komşu, Naive Bayes ve Karar Ağacı yöntemleri sırasıyla doğruluk değerleri sıralamasında yer almaktadır.

Tablo 3*:* İkinci veri kümesi için makine öğrenmesi yöntemlerinin doğruluk değerlerinin gösterilmesi (Demonstrating the accuracy values of machine learning methods for the second dataset)

Uygulanan Makine Öğrenmesi Yöntemleri	Doğruluk Oranı (%)
k-En Yakın Komşu	90,3509
Destek Vektör Makinaları	92,1053
Naive Bayes	89,4737
Karar Ağacı	85,0877
Yapay Sinir Ağları	**93,8596**

İkinci göğüs kanseri veri kümesi için Yapay Sinir Ağları (YSA) ile yapılan testlerde f1 skoru 0,959537 olarak bulunmuştur. Ayrıca duyarlılık değeri 0,943181 ve hassasiyet değerleri 0,976470 olarak hesaplanmıştır. Tablo 4'te YSA yöntemine ait karmaşıklık (confusion) matrisi gösterilmiştir.

Tablo 4: İkinci göğüs kanseri veri kümesine uygulanan YSA yönteminin karmaşıklık matrisi (The complexity matrix of the ANN method applied to the second breast cancer dataset).

		TAHMİN		METRİKLER
		Pozitif	Negatif	
GERÇEK	Pozitif	TP=83	FP=2	Hassasiyet= $\frac{TP}{TP+FP}$=0.976470
	Negatif	FN=5	TN=24	Duyarlılık= $\frac{TP}{TP+FN}$=0.943181
Doğruluk= $\frac{TP+TN}{TP+TN+FP+FN}$=0.938596				F1 Skoru $= 2 * \frac{Hassayiyet * Duyarlılık}{Hassayiyet + Duyarlılık}$=0.959537

Uygulanan makine öğrenmesi yöntemlerinin performans sonuçları doğruluk değerlerine göre karşılaştırılmıştır. Bu sonuçlara göre Yapay Sinir Ağları yöntemi çalışmada kullanılan diğer makine öğrenmesi yöntemlerine göre bu problemin sınıflandırılmasında daha iyi performans sergilemiştir. İki farklı göğüs kanseri veri kümesi içinde yapılan çalışmada YSA yönteminde en yüksek doğruluk değerleri elde edilmiştir.

Kanser hastalığının erken teşhisi ve tanısı ile ilgili bilgisayar destekli çalışmaların yüksek doğruluk oranıyla gerçekleşmesi hastalığın tedavisi açısından önemli bir adımdır. Kanser hastalığından kaynaklı kayıplar düşünüldüğünde erken teşhis ile ilgili herhangi bir gelişme oldukça büyük önem taşımaktadır.

Kanser hastalığının erken teşhis edilebilmesine katkı sağlamak adına yapılan bu çalışmada Kaggle platformunda yer alan iki farklı göğüs kanseri veri kümesi kullanılarak, farklı makine öğrenmesi yöntemleriyle sınıflandırma işlemi gerçekleştirilmiştir. k-En Yakın Komşu, Destek Vektör Makinaları, Navie Bayes, Karar Ağacı ve Yapay Sinir Ağları iki farklı göğüs kanseri veri kümesine uygulanarak elde edilen sonuçlar doğruluk performans metriği ve karmaşıklık matrisine göre karşılaştırılmıştır. Uygulanan tüm makine öğrenmesi yöntemlerinin doğruluk değerleri genel olarak yüksek hesaplanmıştır. İki farklı göğüs kanseri veri kümesi içinde Yapay Sinir Ağları (YSA) yönteminde diğer sınıflandırma algoritmalarına kıyasla daha yüksek doğruluk değeri elde edilmiştir. Birinci göğüs kanseri veri kümesi için YSA yönteminde %98,2456, ikinci göğüs kanseri veri kümesi içinde %93,8596 doğruluk değerleri hesaplanmıştır.

Kaynaklar (References)

Ahmad, L. G., Eshlaghy, A. T., Poorebrahimi, A., Ebrahimi, M., & Razavi, A. R. (2013). Using three

machine learning techniques for predicting breast cancer recurrence. *J Health Med Inform*, *4*(124), 3.

Alpaydın, E. (2013). *Yapay öğrenme*, 2. Baskı, Boğaziçi Üniversitesi Yayınevi, ISBN-13: 978-6-054-23849-1.

Alpaydın, E. (2014). Introduction to Machine Learning. MIT Press.

Anwer, A. M. O., (2017). *Derin Öğrenme Yöntemleri ile Göğüs Kanseri Teşhisi.* Yüksek Lisans Tezi, Türk Hava Kurumu Üniversitesi, Fen Bilimleri Enstitüsü.

Asri, H., Mousannif, H., Al Moatassime, H., & Noel, T. (2016). Using machine learning algorithms for breast cancer risk prediction and diagnosis. *Procedia Computer Science*, *83*, 1064-1069.

Atalay, M., & Çelik, Ö. G. E. (2017). Artificial Intelligence and Machine Learning Applications in Big Data Analysis. Mehmet Akif Ersoy University Journal of Social Sciences Institute, 9(22), 155–172.

Bayrak, E. A., Kırcı, P., & Ensari, T. (2019, Nisan). Comparison of machine learning methods for breast cancer diagnosis. In *2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT)* (pp. 1-3). IEEE.

Bazazeh, D., & Shubair, R. (2016, December). Comparative study of machine learning algorithms for breast cancer detection and diagnosis. *Electronic Devices, Systems and Applications (ICEDSA), 2016 5th International Conference on* (pp. 1-4). IEEE.

Burakgazi, Y., 2017, *Identification of Breast Cancer Sub-Types by Using Machine Learning Techniques*, M.Sc Thesis, Dokuz Eylül University, Graduate School of Natural and Applied Sciences.

Cancer, 2021, https://www.who.int/en/news-room/fact-sheets/detail/cancer, 01.04.2021.

Dhahri, H., Al Maghayreh, E., Mahmood, A., Elkilani, W., & Faisal Nagi, M. (2019). Automated Breast Cancer Diagnosis Based on Machine Learning Algorithms. *Journal of Healthcare Engineering*, *2019*.

Ganggayah, M. D., Taib, N. A., Har, Y. C., Lio, P., & Dhillon, S. K. (2019). Predicting factors for survival of breast cancer patients using machine learning techniques. *BMC medical informatics and decision making*, *19*(1), 48.

Harrington, P. (2012). *Machine learning in Action*, Vol. 5, Greenwich, CT: Manning.

İşeri, İ. (2014). *Mamogram Görüntülerinden Makine Öğrenmesi Yöntemleri ile Meme Kanseri Teşhisi,* Doktora Tezi, Sakarya Üniversitesi, Fen Bilimleri Enstitüsü.

Kaggle, 2020, https://www.kaggle.com/yuqing01/breast-cancer, https://www.kaggle.com/merishnasuwal/breast-cancer-prediction-dataset, 01.04.2021.

Magna, A. A. R., Allende-Cid, H., Taramasco, C., Becerra, C., & Figueroa, R. L. (2020). Application of Machine Learning and Word Embeddings in the Classification of Cancer Diagnosis Using Patient Anamnesis. *IEEE Access*, *8*, 106198-106213.

Maity, N. G., & Das, S. (2017). Machine learning for improved diagnosis and prognosis in healthcare. In *2017 IEEE Aerospace Conference*, pp. 1-9.

Poyraz, O. (2012). *Tıp'da Veri Madenciliği Uygulamaları: Meme Kanseri Veri Seti Analizi. Yüksek* Lisans Tezi, Trakya Üniversitesi, Fen Bilimleri, Enstitüsü.

Reddy, A., Soni, B., & Reddy, S. (2020). Breast cancer detection by leveraging Machine Learning. *ICT Express*.

Ruan, Y., Xue, X., Liu, H., Tan, J., & Li, X. (2017). Quantum algorithm for k-nearest neighbors classification based on the metric of hamming distance. International Journal of Theoretical Physics, 56(11), 3496–3507. Doi:10.100710773-017-3514-4.

Saxena, S., & Gyanchandani, M. (2020). Machine Learning Methods for Computer-Aided Breast Cancer Diagnosis Using Histopathology: A Narrative Review. *Journal of Medical Imaging and Radiation Sciences*, *51*(1), 182-193.

Sherafatian, M. (2018). Tree-based machine learning algorithms identified minimal set of miRNA biomarkers for breast cancer diagnosis and molecular subtyping. *Gene*, *677*, 111-118.

Siegel, R. L., Miller, K. D., & Jemal, A. (2018). Cancer statistics, *Ca-a Cancer Journal for Clinicians*, *68* (1), pp. 7-30.

Şık, M. Ş., 2014, *Veri Madenciliği ve Kanser Erken Teşhisinde Kullanımı,* Yüksek Lisans Tezi, İnönü Üniversitesi, Sosyal Bilimler Enstitüsü.

Tapak, L., Shirmohammadi-Khorram, N., Amini, P., Alafchi, B., Hamidi, O., & Poorolajal, J. (2019). Prediction of survival and metastasis in breast cancer patients using machine learning classifiers. *Clinical Epidemiology and Global Health*, *7*(3), 293-299.

Tseng, Y. J., Huang, C. E., Wen, C. N., Lai, P. Y., Wu, M. H., Sun, Y. C., & Lu, J. J. (2019). Predicting breast cancer metastasis by using serum biomarkers and clinicopathological data with machine learning technologies. *International journal of medical informatics*, *128*, 79-86.

Turgut, S. (2017). *Makine Öğrenmesi Yöntemleri Kullanarak Kanser Teşhisi,* Yüksek Lisans Tezi, Erciyes Üniversitesi, Fen Bilimleri Enstitüsü.

Turgut, S., Dağtekin, M. and Ensari, T. (2018). "Microarray breast cancer data classification using machine learning methods," *2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT)*, Istanbul, pp. 1-3, doi: 10.1109/EBBT.2018.8391468.

Umadevi, S., & Marseline, K. J. (2017, July). A survey on data mining classification algorithms. In *2017 International Conference on Signal Processing and Communication (ICSPC)* (pp. 264-268). IEEE