**Makine Öğrenmesi ile Bireylerde**

**Var Olan Kalp Hastalığının Teşhisi**

**Nadire BAY**

Fırat Üniversitesi, Mühendislik Fakültesi, Yazılım Mühendisliği, Elazığ, Türkiye

nadirebay8@gmail.com

***ÖZET****---* Başlıca görevi vücuda kan pompalamak olan kalp, metabolizma faaliyetleri sonrası oluşan artık ürünlerin vücuttan uzaklaştırılması, vücut ısısının düzenlenmesi, asit-baz dengesinin korunması, hormonlar ve enzimlerin vücudun gerekli bölgelerine taşınması gibi görevleri yapar. Hareketsiz yaşam tarzı, aşırı yağlı yiyecek tüketimi, alkol, sigara kullanımı gibi etmenler kalbin çalışmasını olumsuz etkileyerek kalp hastalıklarının oluşmasına neden olabilmektedir. İnsan vücudu için birinci derecede hayati öneme sahip kalpte oluşan hastalıkların erken bir aşamada tespit edilmesi oldukça önemlidir. Bu çalışmada ise kullanılan veri seti ile birlikte hedef alanındaki hastalarda kalp hastalığına rastlanıp rastlanılmadığını tespit etmektir. Bu tespit için elde edilen veri seti ile birlikte bazı değerler üzerinde durulmuş ve bu değerler ile doğruluk duyarlılık ve özgüllük oranlarına ulaşılmıştır. Yapılan çalışmalara göre %99,40 duyarlılık, %100 özgüllük ve %99,71 oranında da doğruluk değerleri tespit edilmiştir. Elde edilen sonuçlara bakıldığında da ileriki zamanlarda yapılacak çalışmalar adına oldukça cezbedici bir durumdur.

**Anahtar Kelimeler**: K-En Yakın Komşu, Sınıflandırma, Kalp Hastalığı, Makine Öğrenmesi

***SUMMARY--***The heart, whose main task is to pump blood to the body, performs tasks such as removing waste products from the body after metabolic activities, regulating body temperature, maintaining acid-base balance, and transporting hormones and enzymes to the necessary parts of the body. Factors such as sedentary lifestyle, excessive fatty food consumption, alcohol, and smoking can negatively affect the work of the heart and cause heart diseases. It is very important to detect diseases that occur in the heart, which are of primary importance for the human body, at an early stage. In this study, with the data set used, it is to determine whether heart disease is encountered in the patients in the target area. With the data set obtained for this determination, some values ​​were emphasized and accuracy, sensitivity and specificity ratios were reached with these values. According to the studies, 99.40% sensitivity, 100% specificity and 99.71% accuracy values ​​were determined. Considering the results obtained, it is a very attractive situation for future studies.

**Keywords:** K-Nearest Neighbor, Classification, Heart Disease, Machine Learning

**1.GİRİŞ**

Makine öğrenmesi bir bilgisayarın doğrudan yönergeler olmadan öğrenmesine yardımcı olmak için matematiksel modelleri kullanma işlemidir. Bu yapay zekanın bir alt kümesi olarak kabul edilir. Makine öğrenmesi verilerdeki kalıpları belirlemek için algoritmaları kullanılır. Tahmin yapabilen bir veri modeli oluşturmak için de bu kalıplar kullanılır. Tıpkı insanların daha fazla alıştırma yaptıkça gelişmesi gibi, veri ve deneyim miktarı artıkça makine öğrenmesinin sonuçları da daha doğru hale gelir.[1] Kalp hastalığı, kalbi, damarlarını, kaslarını, kas kasılmasından sorumlu iç elektrik yollarını içeren bir dizi durumu ifade eder. ABD'deki her dört ölümden biri, kalp hastalığının bir sonucu olarak ortaya çıkar. Kalp hastalığı dünyadaki çoğu ülkede hem erkekler hem de kadınlar arasında yaygındır. Bu nedenle, insanlar kalp hastalığı risk faktörlerini göz önünde bulundurmalıdır. Genetik bir rol oynasa da, bazı yaşam tarzı faktörleri, kalp hastalığını önemli derecede etkilemektedir. Kalp hastalığı için bilinen risk faktörleri; yaş, cinsiyet, aile öyküsü, sigara içmek, bazı kemoterapi ilaçları ve kanser için radyasyon tedavisi, kötü beslenme, yüksek tansiyon, yüksek kan kolesterol seviyeleri, diyabet, obezite, fiziksel hareketlilik, stres ve kötü hijyendir [2]. Teknoloji ilerledikçe sağlık alanında yapılan çalışmalarda hız göstermektedir. Tıbbi görüntülemelerde, hastalık teşhislerinde ve bu teşhisler sonucu yapılmakta olan tedavilerde bu gelişmelerle birlikte aynı yönde ilerlemektedir. Günümüzde hastalıkların teşhisinde ise biyomedikal alanda birden çok farklı sınıflandırma yönteminin kullanılması yaygın bir hale gelmiştir. Bunun sebebi ise bilgisayar tabanlı sınıflandırma yöntemleri ile doğruya yakın sonuç elde edilmesi gösterilmektedir.

1259

* 1. **Literatür Çalışması**

Literatürde bu alanda yani kalp hastalığının sınıflandırılması alanında yapılan bazı çalışmalara bakıldığında birçok çalışmalara rastlanılmıştır. Bunlardan bazılarına örnek verecek olursak Amin ve arkadaşları, yaş, diyabet, hipertansiyon, yüksek kolesterol, alkol alımı, obezite veya fiziksel inaktivite gibi başlıca risk faktörlerini temel alan kalp hastalıklarını öngörmektedir. Teknikleri en başarılı iki veri madenciliği aracını, sinir ağları ve genetik algoritmaları içeriyordu. Bu nedenle kalp hastalıklarının öngörülmesi ve tanısı için Naïve Bayes, Yapay Sinir Ağları, Karar Ağacı,Bagging Trees, Çekirdek Yoğunluğu ve DVM uygulamışlardır. Öğrenme, geri yayılma ile karşılaştırıldığında hızlı, daha istikrarlı ve doğru olarak bulunmuştur. Sistem kalp hastalığı riskini tahmin etmek üzere Matlab'da uygulanmıştır. 50 hastanın risk faktörlerine bakılarak hibrid model eğitim doğruluğu ve test doğruluğu değerlerine ulaşılmıştır. Daha sonra Amin ve arkadaşları, kalp hastalıklarının öngörülmesi için hibrit bulanık ve k-en yakın komşu yaklaşımını kullanarak sistemi geliştirdiler. Böylelikle bu hibrit sisteminin avantajlarından biri hastaların kalp hastalıkları ve yan etkilerinden önceki tıbbi muayeneler için maliyet ve zamanlarını azaltmalarına ve kendilerini kontrol etmelerine yardımcı olmaktır. [3] Yapılan bu çalışmalara genel anlamda bakıldığında son aşama olarak kullanılan algoritmalar arasında karışıklık matrisine göre kıyaslama sonucu %99 oranla en yüksek doğruluğu J48’in kaydettiği öğrenildi.

**2.Malzemeler ve Yöntemler**

Bu çalışmada Kaggle’dan almış olduğumuz “Heart Disease Dataset” isimli veri setini kullanarak test sonuçları elde edildi.Kalp hastası olan ve olmayan hastaların tespiti adına kullanılan data setine göre bazı veri kümeleri ele alınmıştır.

**2.1 Veri Seti**

Yapılan bu çalışmada kalp hastalığı tespiti yapılmış ve elde edilen sonuçlar karşılaştırılmıştır. Alınan veri setindeki toplam 13 özellik için katılımcılar üzerinde çeşitli gözlemler yapılmıştır. Bu özellikler ise aşağıdaki tabloda verilmiştir.

**Tablo 1. Kalp Hastalığı Veri Seti Özellikleri**

|  |
| --- |
| **Özellik No Özellik Bilgisi** |

1. Yaş
2. Seks
3. Göğüs Ağrısı Tipi
4. Dinlenme Kan Basıncı
5. Mg/dl cinsinden serum kolestrol
6. Açlık kan şekeri > 120 mg / dl
7. İstiharat elektrokardiyografik sonuçları (0,1,2 değerleri)
8. Ulaşılan Maksimum Kalp Atış Hızı
9. Egzersize Bağlı Anjina
10. Dinlenmeye Göre Egzersizin neden olduğu ST depresyonu
11. ST segmentinin eğimi
12. Floroskopi ile bulunan ana damarların sayısı
13. Tal:0 = normal; 1= sabit kusur; 2=tersine çevrilebilir kusur

|  |
| --- |
|  |

**2.2 Kolektif Sınıflandırıcılar (ENSEMBLE CLASSIFIERS)**

Her bir kolektif sınıflandırıcı tekniği tekil bir sınıflandırıcı kullanmak zorundadır. Bilindiği üzere tekil sınıflandırıcı olarak yaygın bir kullanım alanına sahip karar ağaçları, destek vektör makineleri, Naive Bayes metodu, lineer ayırıcılar ve yapay sinir ağları gibi algoritmalar mevcuttur. Hızlı eğitilebilmeleri, çabuk sınıflandırma yapabilmeleri, doğruluk oranlarının yüksek olması ve şeffaf bir yapıda (white box) olmaları nedeniyle karar ağaçları kolektif yöntemlerde tekil öğrenici olarak kullanılırlar [4]. Verilen data set üzerinde de bu sınıflandırıcılar kullanılarak belli sonuçlara ulaşılmıştır.

***2.2.1 Bagging Algoritması***

Bagging (Bootstrapping Aggregation) metodu,var olan bir eğitim setinden yeni eğitim setleri türeterek temel öğreniciyi yeniden eğiten bir yöntemdir. Bagging’de amaç yeni veri setleri türeterek farklılıkları oluşturmak ve bu sayede toplam sınıflandırma başarısını artırmaktır. Eğitim setinden yaklaşık olarak %63,2 kadar orijinal örnek rastgele alınır ve alınan örneklerden bazıları çoğaltılarak eğitim seti %100’e tamamlanır. Bu yöntemle birbirinden farklı bir miktar eğitim seti elde edilir. Her eğitim seti aynı temel öğreniciye uygulanır ve alınan kararlar ağırlıklı oylama yöntemiyle birleştirilir. Eğitim setinden %63,2 kadar örnek seçilmesindeki neden şu formülle açıklanır:

(1 − 1 𝑛 ) 𝑛 ≈ 𝑒 −1 = 0,368 (1)

Burada n işlem sayısını veya eleman sayısını gösterir. n değeri sonsuza giderken doğal logaritma tabanı olan e sayısının tersi elde edilmiş olur.(1− 0,368 = 0.632) N sonsuza giderken Bu yöntemle orijinal eğitim kümesinden yaklaşık olarak %36’sı büyük ihtimalle hiçbir zaman seçilmemiş olacaktır. Bagging yönteminde seçilmeyenlerin seçilmesini sağlamak için eğitim setinden %63,2 kadar örnek rastgele seçilir ve yeni bir eğitim seti oluşturulur.[5]

***2.2.2 KNN(En Yakın Komşuluk) Algoritması***

K en yakın komşu algoritması, sınıflandırma problemlerinin çözümünde kullanılan örnek tabanlı algoritmalar sınıfındadır. Öğrenme işlemini veri içerisinde tutulan eğitim seti ile gerçekleştirir. Eğitim işlemi en yakın varsayılan k adet veriyi, belirli uzaklık ölçütü çerçevesinde benzerliklerinin hesaplanması ile yapmaktadır [6].

**3.Deneysel Çalışmalar**

Çalışma kapsamında kalp hastalığı tespiti için MATLAB da belli sınıflandırıcılar kullanılarak sonuçlarda karşılaştırmalar yapılmıştır. Kullanılan veri setine bakıldığında 499 kalp hastası ve 526 kişi ise kalp hastası olmayan kişilerin verileri bu çalışmada yer almıştır. Hastalığın tanısında ise veri seti özellikleri baz alınmıştır. Bu veri seti incelemesi sonucu birçok değer elde edilmiştir. Elde edilen sonuçlara bakıldığında;

**Kalp Hastası Olan** **Kalp Hastası Olmayan**

|  |
| --- |
|  |

**Kalp Hastası** **Olan**  499 0

|  |
| --- |
|  |

**Kalp Hastası Olmayan** 3 523

|  |
| --- |
|  |

Yukarıda ki değerler sonucu da elde edilen başarım değerlendirme sonuçları ise %99,4 oranında duyarlılık, %100 oranında özgüllük ve %99,71 oranında doğruluktur. Bu oranların genel anlamda bu kadar yüksek çıkması yapılan veri seti üzerindeki sonuçların yüksek oranda doğruluğunun ispatıdır. Bu oran Bagged Tress sınıflandırıcı altında incelenmiştir. Ulaşılan sonuçlara Komşuluk Matrisi üzerinde incelendiğinde;

**Kalp Hastası Olan** **Kalp Hastası Olmayan**

|  |
| --- |
|  |

**Kalp Hastası** **Olan**  499 0

|  |
| --- |
|  |

**Kalp Hastası Olmayan** 5 521

|  |
| --- |
|  |

Sonuçlarına ulaşılır. Bagging Ağaçları sınıflandırıcılarında ki başlıca avantajlardan biri nitel özelliklerin kolayca ele alınmasıdır. Bagging Ağaçları sınıflandırıcısıyla elde edilen komşuluk matrisindeki verilerle elde edilen sonuçlar %99.01 oranında duyarlılık,%100 oranında özgüllük ve %99.51 oranında da doğruluk değerlerine ulaşılmıştır. Sonuçlara bakıldığında bagged trees sınıflandırıcısı ile KNN sınıflandırıcısı oranları birbirleriyle çokta farklı değildir.

Veri setini Weighted KNN sınıflandırıcısı üzerinde incelediğimizde;

**Kalp Hastası Olan** **Kalp Hastası Olmayan**

|  |
| --- |
|  |

**Kalp Hastası** **Olan**  499 0

|  |
| --- |
|  |

**Kalp Hastası Olmayan** 6 520

komşuluk matrisine ulaşılmıştır. Matris sonucu da %98,8 oranında duyarlılık, %100 oranında özgüllük ve %99,4 oranında da doğruluk oranı elde edilmiştir.

**4.Sonuçlar**

Kalp hastalığının teşhisinde kullanılan yöntemlerden elde edilen verilerde bulunan anlamlı ilişkileri çıkarabilmek hastalığın teşhis ve tedavi süreçlerini olumlu yönde etkilemektedir. Bu çalışmada da görüldüğü üzere kalp hastalığı veri setinin birçok sınıflandırıcı ve algoritmalar yardımıyla yüksek bir başarı oranıyla yapılabileceğinin göstergesi niteliğindedir. Elde edilen sonuçlardaki doğruluk oranlarındaki yüzdelik dilimlere bakıldığında hastalığın teşhis süreçlerinin kısalmasına yardımcı olabileceğini göstermektedir. Çalışmada kullanılan yöntemleri de ele alırsak oldukça iyi bir performans sergilenmiştir.

KAYNAKLAR

[1] <https://azure.microsoft.com/tr-tr/overview/what-is-machine-learning-platform/>

[2] Staff, M.C., Heart disease. Mayo Clinic

[3] Amin, S.U., K. Agarwal, and R. Beg. Genetic neural network based data mining in prediction of heart disease using risk factors. in Information & Communication Technologies (ICT), 2013 IEEE Conference on. 2013. IEEE

[4] Zhou, Zhi-Hua. “Ensemble methods: foundations and algorithms”, CRC Press, New York, ABD, 2012

[5] Breiman L., “Bagging predictors.”, Machine Learning, 24:123–140, 1999. [11] Freund, Y., “

[6] Dudani, S. A. (1976). The distance-weighted k-nearest-neighbor rule. IEEE Transactions on Systems, Man, and Cybernetics, (4), 325-327.