Mühendislik Bilimleri ve Tasarım Dergisi
8(5), 241-254, 2020
e-ISSN: 1308-6693
Araştırma Makalesi

Journal of Engineering Sciences and Design
DOI: 10.21923/jesd.824703
Research Article

Özel Sayı: Uluslararası Mühendislikte Yapay Zeka ve Uygulamalı Matematik Konferansı (UMYMK 2020)
Special Issue: International Conference on Artificial Intelligence and Applied Mathematics in Engineering (ICAIAME 2020)

YAPAY ZEKÂ YÖNTEMLERİ KULLANILARAK KALP HASTALIĞININ TESPİTİ

Özge Ekrem[1*], Osamah Khaled Musleh Salman[1], Bekir Aksoy[1], Seyit Ahmet İnan[1]

[1] Isparta Uygulamalı Bilimler University, Technology Faculty, Department of Mechatronic Engineering, Isparta, Türkiye

Anahtar Kelimeler

Yapay Zekâ, Kalp Hastalığı, Sınıflandırma

Öz

Günümüzde teknolojinin hızla gelişmesi ile birlikte yapay zekâ teknikleri de yaygın bir şekilde kullanılmaktadır. Yapay zekâ yöntemleri mühendislik uygulamaları, eğitim, savunma sanayi gibi birçok alanda sıklıkla kullanılmaktadır. Yapay zekânın önemli kullanım alanlarından birisi de sağlık sektörüdür. Sağlık sektörü alanında gerçekleştirilen bu çalışmada açık erişimli bir internet sitesinden (kaggle) elde edilen veri seti kullanılmıştır. Veri seti üzerinde yapay zekâ yöntemleri kullanılarak kalp hastalığının tespiti gerçekleştirilmiştir. Çalışma kapsamında, Random Forest yöntemi ve Parçacık Sürü Optimizasyonu kullanılarak veri setinde yer alan 303 bireyin kalp hastası olup, olmadığına dair sınıflandırma işlemi gerçekleştirilmiştir. Parçacık Sürü Optimizasyonu yöntemi kullanılarak özellik seçimi yapılmış olup rastgele orman yapay zekâ algoritması ile veri seti eğitilmiştir. Rastgele Orman sınıflandırma modeli; doğruluk, özgüllük, duyarlılık, kesinlik, F-ölçüsü, ROC eğrisi ölçütlerinden oluşan performans değerlendirme kriterlerine göre başarı oranı incelenmiştir. Değerlendirme sonucunda Rastgele Orman sınıflandırmanın %86.88 doğruluk, %85.71 özgüllük, %87.87 duyarlılık, %87.87 kesinlik ve %87.87 F-ölçüsü değeri ile başarılı tahmin gerçekleştirdiği belirlenmiştir.

DETERMINATION OF HEART DISEASE USING ARTIFICIAL INTELLIGENCE METHODS

Keywords

Artificial Intelligence, Heart Disease, Classification

Abstract

Today, with the rapid development of technology, artificial intelligence techniques are also widely used. Artificial intelligence methods are frequently used in many fields such as engineering applications, education and defense industry. One of the important uses of artificial intelligence is the health sector. The data set obtained from an open access website (kaggle) was used in this study conducted in the field of health sector. Heart disease was detected on the data set by using artificial intelligence methods. Within the scope of the study, the classification process was carried out to determine whether 303 individuals in the data set have heart disease or not using the Random Forest method and Particle Swarm Optimization. Feature selection was made using the Particle Swarm Optimization method, and the data set was trained with a random forest artificial intelligence algorithm. Random Forest model; The success rate was examined according to the performance evaluation criteria consisting of accuracy, specificity, sensitivity, precision, F-measure, and ROC curve. As a result of the evaluation, it was determined that Random Forest classification made successful prediction with 86.88% accuracy, 85.71% specificity, 87.87% sensitivity, 87.87% precision and 87.87% F-measure value.

UEAlıntı / Cite

Ekrem, Ö., Salman, O. K. M., Aksoy, B., İnan A. S., (2020). Determination Of Heart Disease Using Artificial Intelligence Methods, Journal of Engineering Sciences and Design, 8(5), 241-254.

Yazar Kimliği / Author ID (ORCID Number)
Ö. Ekrem, 0000-0001-9142-405X
O. K. M. Salman, 0000-0001-6526-4793
B. Aksoy, 0000-0001-8052-9411
A. S. İnan, 0000-0002-9489-7714

Makale Süreci / Article Process	
Başvuru Tarihi / Submission Date	11.11.2020
Revizyon Tarihi / Revision Date	15.12.2020
Kabul Tarihi / Accepted Date	16.12.2020
Yayım Tarihi / Published Date	29.12.2020

[*] İlgili yazar / Corresponding author: yl1930654005@stud.sdu.edu.tr

1. Giriş (Introduction)

Kardiyovasküler hastalıklar olarak adlandırılan kalp ve kan damarı hastalıkları, damarların iç duvarında çeşitli maddelerin birikimine bağlı olarak ortaya çıkan ve plak oluşumuna sebep olan bir problem zinciri ile meydana gelmektedir. Ateroskleroz olarak adlandırılan bu problem zinciri, kalp arterlerinin daralmasına sebebiyet vererek kanın damar içerisinde akışını zorlaştırmaktadır(Opeyemi ve Justice, 2012). Kalp ve damar hastalıklarının oluşmasında başlıca sebepler fiziksel inaktivite, sağlıksız beslenme, hipertansiyon, sigara ve alkol tüketimidir(Bulut,2016). Bu şekilde meydana gelen kalp hastalıkları, dünya çapında başlıca ölüm nedenleri arasında yer almaktadır ve Kalp ve Damar Hastalıklarına (KDH) bağlı ölümlerin 2012 yılındaki 17,5 milyon seviyesinden 2030 yılında 22,2 milyona çıkacağı tahmin edilmektedir. Bununla birlikte yüksek gelir düzeyine sahip ülkelerde KDH'den meydana gelen ölüm sayısı azalırken, birçok düşük ve orta gelirli ülkelerde daha fazla ölüm meydana gelmektedir(Üner vd., 2019). Kalp hastalıklarından dolayı her yıl artan erkek ve kadın hasta sayısının önüne geçilmesi ve ölüm riskinin azaltılmasında hastalığın erken tespit edilmesi ve uygun tedavinin uygulanması önemli rol oynamaktadır(Priyanka ve RaviKuvar, 2017). Günümüzde kalp hastalıklarının tespitinde yapay zekâ uygulamaları ile farklı sınıflandırma yöntemlerinin kullanımı artış göstermektedir(Boyraz vd., 2014). Kullanılan sınıflandırma yöntemleri ile doğru sonuçların elde edilmesi, hastalığın tespiti için gereken sürenin minimum seviyeye düşürülmesi ve insan kaynaklı hataların önüne geçilmesi sağlanmaktadır. Literatürde bu alanda yapılan birçok çalışma yer almaktadır.

2. Kaynak Araştırması (Literature Survey)

Sengur (2007), bu çalışmada temel bileşen analizi (ing. Principal Component Analysis-PCA),yapay bağışıklık sistemi (ing. Artificial Immune System-AIS) ve bulanık k-en yakın komşu (ing. K-Nearest Neighborhood) algoritması kullanımını araştırarak Doppler kalp seslerinden normal ve anormal kalp kapakçıklarını belirlemiştir. Sunulan kalp kapak hastalığı tespit sisteminde ilk olarak görüntü ön işleme uygulanıp gerekli filtreleme, normalleştirme ve beyaz giderme işlemleri bu aşamada gerçekleştirilmiştir. İkinci aşamada özellik çıkarma ile dalgacık paket ayrışımı kullanılmıştır. En son adım olarak, dalgacık entropisi özellikler olarak kabul edilmiştir. Sistemin karmaşıklığını azaltmak için özellik azaltma işleminde PCA kullanılmıştır. Sınıflandırma aşamasında AIS ve bulanık k-NN kullanılmıştır. Önerilen çalışmanın performans değerlendirmesi için 215 örnek içeren bir veri seti kullanılarak karşılaştırmalı bir çalışma yapılmıştır. Yapılan çalışmada % 95.9 duyarlılık ve % 96 özgüllük oranı elde edilmiştir.

Palaniappan ve Awang (2008), yapmış oldukları çalışmada veri madenciliği yöntemlerini (Karar Ağaçları, Naive Bayes ve Sinir Ağı) kullanarak akıllı kalp hastalığı tahmin sistemi (ing. Intelligent Heart Disease Prediction System-IHDPS) prototipi geliştirmişlerdir. Kalp hastası olan kişilerin yaş, cinsiyet, kan basıncı ve kan şekeri gibi özellikleri dikkate alınarak kalp hastalığı olma olasılığı tahmin edilebilmiştir. Elde edilen sonuçlara göre karar ağaçları, sonucu okuma ve yorumlama açısından kolaylık sağlamıştır. Aynı zamanda hasta özelliklerine erişimi sadece karar ağaçları detaylandırmıştır. Naïve Bayes yöntemi, tüm önemli tıbbi tahminleri tanımlayabildiği için, karar ağaçları yöntemine nazaran daha iyi sonuç vermiştir.

Das ve Sengur (2009), tarafından yapılan çalışmada kalp hastalıklarının tespitinde İstatiksel Analiz Sistemi (ing. Statistical Analysis System-SAS) yazılımının 9.1.3 sürümü kullanılmıştır. Bir sinir ağları topluluğu yöntemine dayanan sistemde bulunan değerler birden fazla önceki modelle ilişkilendirilerek yeni modeller oluşturulmuştur. Kalp hastalıklarını tamamen otomatik bir şekilde tahmin etmek için üç ayrı sinir ağı modeli kullanılmıştır. Veri tabanında elde edilen verilerde %89.01 sınıflandırma doğruluğu, %95.91 özgüllük ve %80.95 duyarlılık değerleri elde edilmiştir.

Shao vd. (2014),çalışmalarında kalp hastalığını etkili bir şekilde sınıflandıran bir hibrit akıllı modelleme şeması önermiştir. Modelin gelişimi gösterebilmek için gerçek hastaların oluşturmuş olduğu bir veri seti kullanılmıştır. Önerilen hibrit modelleme, lojistik regresyon (LR), çok değişkenli uyarlanabilir regresyon splineları (ing. Multivariable Adaptive Regression Splines-MARS), yapay sinir ağı (ing. Artificial Neural Network -ANN) ve kaba küme (ing. Rough Cluster-RS) tekniklerinden oluşmuştur. Çalışmada ilk olarak, açıklayıcı değişken kümesini azaltmak için LR, MARS ve RS tekniklerinin kullanılmıştır. Sonraki aşamada ANN yöntemi kullanımı için geri kalan değişkenler girdi olarak atanmıştır. Hibrit modelleme sonuçları ile hibrit şemaların kalp hastalığının sınıflandırılmasında tek aşamalı ANN yönteminden daha etkili olduğu gözlemlenmiştir.

Yan vd. (2006), yaptıkları çalışmada kalp hastalıklarının teşhisini desteklemek için çoklu algılama tabanlı bir karar destek sistemi (ing. Medical Decision Support Systems-MDSS) geliştirmişlerdir. Sistemde, bir hastaya ait eksik veriler ikame ortalama yöntemi kullanılarak değerlendirilmiştir ve sistemi eğitmek için geliştirilmiş bir geri yayılma algoritması kullanılmıştır. Sistemi eğitmek ve test etmek için kalp hastalığı olan 5 hastadan toplam 352

tıbbi kayıt kullanılmıştır. Önerilen karar destek sisteminin tespit doğruluğu %90'dan büyük ve nispeten küçük aralıklar %5'den az olarak elde edilmiştir.

Singh vd. (2016), yapmış oldukları çalışmada California Irvine Üniversitesi (ing. University of California Irvine-UCI) makine öğrenimi veri tabanından Cleveland kalp hastalıkları veri kümesini kullanarak çeşitli veri madenciliği yöntemlerini ele almışlardır. Çalışmada kalp hastalığının sebebi olan farklı özellikler (yaş, cinsiyet, göğüs ağrısı tipi, kan basıncı, kan şekeri vb.) kullanılmıştır. Aprior, FP-Growth, Naive bayes, ZeroR, OneR, J48 ve k-NN gibi farklı veri madenciliği yöntemleri kullanılarak kalp hastalıkları tahmin edilmeye çalışılmıştır. K-NN yönteminin diğerlerinden daha iyi sonuçlar verdiği gözlemlenmiştir.

Priyanka ve RaviKumar (2017), çalışmalarında kalp hastalığının tahmin doğruluğunun yüksek olması için veri madenciliği yöntemlerinden Naïve Bayes & Karar Ağacı tekniklerini karşılaştırmışlardır. Tahmin sistemi için kalp hastalığına ait 13 özellik ele alınmıştır. Çalışma sonucunda karar ağacının kalp hastalığı tahmininde daha kesin sonuçlar verdiği gözlemlenmiştir. Gerçekleştirilen üç test çalışmasında birinci adımda karar ağacı yöntemi ile bayes algoritmaları sırasıyla %98.03 ve %82.35 gibi doğruluk değerleri, ikinci adımda %98.21 ve %82.14 gibi farklı doğruluk değerleri ve üçüncü adımda %90 ve %70 gibi test sonuçları gözlemlenmiştir. Tablo 1'de çalışma ile ilgili gerçekleştirilen akademik literatür araştırması verilmiştir.

Tablo 1. Akademik literatürde Yer Alan Çalışmalar (Studies in the literature)

Çalışmayı gerçekleştirenler	Yöntem	Veri Sayısı	Doğruluk
Sengur (2007)	Temel bileşen analizi(PCA), Yapay bağışıklık sistemi (AIS), Bulanık k-en yakın komşu algoritması(KNN)	215	% 95.9 duyarlılık % 96 özgüllük
Palaniappan ve Awang (2008)	Karar Ağaçları, Naive Bayes, Sinir Ağı	909	Naïve Bayes modeli (%86.12) Sinir Ağı (% 85.68) Karar Ağaçları (% 80.4)
Das ve Sengur (2009)	İstatiksel Analiz Sistemi (SAS) 9.1.3 sürümü	303	%89.01 sınıflandırma doğruluğu, %95.91 özgüllük, %80.95 duyarlılık
Shao vd. (2014)	Hibrit akıllı modelleme (LR, MARS, RS, YSA)	---	MARS-LR (%83.93) RS-LR (83.93)

Yan vd. (2006)	Çoklu algılama tabanlı bir karar destek sistemi(MDSS)	5 hasta (352 kayıt)	>%90
Singh vd. (2016)	FP-Growth, Naive bayes, ZeroR, OneR, J48 ve k-NN	313	ZeroR (%67.2) OneR (%97.31) J48 (%97.85) Naive bayes (%97.58)
Priyanka ve RaviKumar (2017)	Naïve Bayes, Karar Ağacı	13 Özellik	Karar ağacı (%98.03,%98.21,%90) Bayes (%82.35,%82.14,%70)

Çalışmada açık kaynak erişimli internet sitesinden (kaggle.com) elde edilen kalp hastalığı veri kümesinin içerdiği yaş, cinsiyet, göğüs ağrısı tipi, kan basıncı, kolesterol, açlık kan şekeri, elektrokardiyografik sonuçlar, kalp atış hızı vb. gibi veriler işlenerek yapay zekâ yöntemlerinden Rastgele Orman (ing. Random Forest -RF) yöntemi ile %86.88 doğruluk oranında sınıflandırma işlemi gerçekleştirilmiştir.

3. Materyal ve Yöntem (Material and Method)

Yapılan bu çalışmada RF yöntemi ile kalp hastalığı tespiti yapılmıştır. Kalp hastası olan ve olmayan hastaların tespiti için kullanılan veriler açık kaynak erişimli internet sitesindeki (kaggle.com) 'Heart Disease UCI' veri kümesi üzerinden elde edilmiştir.

3.1. Materyal(Material)

3.1.1 RF Yöntemi ile Sınıflandırma(Classification by RF Method)

RF algoritması Leo Breiman (Breiman, 2001) tarafından oluşturulan topluluk (ing. ensemble) sınıflandırma yöntemleri arasında yer almaktadır(Mursalin vd., 2017). Topluluk sınıflandırma teknikleri, sadece bir tane sınıflandırıcı üretmek yerine birden fazla sınıflandırıcı üreten ve üretilen sınıflandırıcıların tahminleri sonucu elde edilen oylar ile yeni veriyi sınıflandıran öğrenme algoritmalardır(Akar ve Güngör, 2012). RF sınıflandırma yönteminde de diğer topluluk öğrenme yöntemlerinde olduğu gibi zayıf öğrenicilerin (tek bir karar ağacı, tek algılayıcı, vb.) performans değerleri bir oylama şemasıyla artırılmaktadır(Ahmad vd., 2017). RF algoritması ile sınıflandırma yöntemi karar ağacı (Özekes, 2003) modeline dayanmaktadır. Karar ağacı modelinde önyükleme örnekleme yaklaşımı ile orijinal veri setinden farklı k eğitim veri alt kümeleri oluşturulmaktadır. Eğitim verisinin 1/3 'ü genelleştirilmiş hatalar (ing. Out Of Bag-OOB) için 2/3 ise karar ağaçlarının oluşturulması için kullanılmaktadır. OOB, oluşturulan ormandaki ağaç sayısı arttıkça girdi değişkeni (özellik) önemini ölçmek ve bir sınıflandırma hata oranı elde etmek kullanılmaktadır(Barrett ve Kulkarni, 2015; Horning, 2010). Eğitilmiş olan k ağaçları, Denklem 1'de tanımlanan bir RF modelinde toplanmaktadır(Chen vd., 2016).

$$H(X, \theta j) = \sum^{k} hi(x, \theta j), (\mathrm{j} = 1, 2, \dots, \mathrm{m}) \qquad (1)$$

Denklemde $H(X, \theta j)$ bir meta karar ağacı sınıflandırıcısıdır. x, Eğitim veri kümesinin girdi özelliği vektörünü temsil etmektedir ve θj, ağacın büyüme sürecini belirleyen bağımsız ve aynı şekilde dağıtılmış rastgele bir vektördür.

Karar ağaçlarında bölünmeler düğümlerde gerçekleşmektedir. Gerekli tüm işlemlerin düğümlerde yapılmasının ardından elde edilen sonuç sınıfları yaprak düğümlerde tutulmaktadır. Düğümlerin bölünme değerlerini belirlemek için entropi, bilgi kazanımı ve Gini indeksi yöntemleri kullanılmaktadır(Yılmaz, 2018). RF algoritması, Denklem 2'de tanımlanan Gini indeksini kullanarak mevcut ağaç dalları içerisinde en iyi dalı belirlemektedir(Kılıçarslan, 2019).

$$\sum \sum j \neq i \, (f \, (Ci, T)/|T|)(f(Cj, T)/|T|) \qquad (2)$$

T,belirli bir eğitim seti olmak üzere, rastgele seçilen bir örneğin Ci sınıfına ait olduğunu belirtmektedir. $f(Ci, T)/|T|$, seçilen vakanın Ci sınıfına ait olma olasılığıdır(Pal,2005).
Bağımsız olarak rastgele oluşturulan k karar ağaçlarının bir araya gelmesi sonucu RF algoritması oluşmaktadır. Test veri kümesine ait örneklerden her biri ormanı oluşturan karar ağaçları tarafından tahmin edilmektedir. Bu ağaçların oylamaları sonucunda da örneklerin sınıflandırma işlemi gerçekleştirilmektedir(Chen vd. 2016).Geleneksel karar ağaçları algoritmasında karşılaşılan en büyük problemler arasında aşırı öğrenme (ing. overfitting) yer almaktadır. RF algoritması, bu problemi ortadan kaldırmak için hem veri setini hem de öznitelikleri çok sayıda parçaya bölerek birden fazla ağaçta işlem gerçekleştirmektedir(Sevli,2019). RF, büyük veri tabanlarında özelliklerin eksiksiz işlenmesinde ve önemli özelliklerin ön plana çıkarılmasında oldukça verimli çalışmaktadır. Sınıflandırma işleminde düşük hatalar üreterek doğruluk oranını artırmaktadır(Gulia vd, 2014).

3.1.2 Parçacık Sürü Optimizasyonu(Particle Swarm Optimization)

Parçacık Sürü Optimizasyonu (PSO), 1995 yılında Kenedy ve Eberhart tarafından, balıklar ve böceklerden esinlenerek geliştirilmiş popülasyon tabanlı stokastik bir optimizasyon yöntemidir. PSO algoritması, sürü halinde hareket eden hayvanların sosyal davranışlarını taklit etmektedir(Wang vd., 2017). Temel prensip olarak kuş ve balık sürülerinin yiyecek ararken yapmış oldukları hareketlerden esinlenilen bu algoritmada, popülasyonu oluşturan her bir parçacık düzensiz hareket ediyor gibi görünmekte fakat kümelenme eğilimindedir(Aydın ve Aşıcı, 2020). Popülasyonda yer alan her bir parçacık, aday bir çözüm olup hız ve konum bilgisi taşımaktadır. Parçacık, kendi pozisyonunu bir önceki tecrübesi ile karşılaştırmakta ve en iyi pozisyona doğru ayarlamaktadır. PSO, temel olarak sürüde bulunan bireylerin pozisyonunun, sürünün en iyi pozisyona sahip olan bireyine yaklaştırılmasına dayanmaktadır(Özsağlam ve Çunkaş, 2008). Böylece kendi ve diğer üyelerin öğrenme deneyimlerine göre arama modelini değiştirmeye devam etmektedir (Wang vd., 2017). Popülasyon tabanlı algoritmalardan biri olarak, PSO popülasyonundaki her parçacık, optimizasyon problemine çözüm sunan konum bilgisine sahip pozisyon vektörü ve yön bilgisine sahip hız vektörü olmak üzere iki vektör ile temsil edilmektedir. d boyutlu bir veri setinde, i parçacığının sahip olduğu konum vektörü $x_i = (x_{i1}, x_{i2}, \ldots x_{id})$, hız vektörü $v_i = (v_{i1}, v_{i2}, \ldots v_{id})$ şeklindedir ve bu bilgiler her iterasyonda güncellenmektedir(Wang vd., 2018; Çınaroğlu ve Bulut, 2018).

PSO algoritmasında, d boyutlu bir sette, i. parçacığının sahip olduğu hız ve konum vektörlerinin her iterasyonda güncellemesine ait matematiksel ifade Denklem 3'de verilmiştir.

$$v_{id}(t+1) = w.v_{id}(t) + c_1.r_1\left(P_{id}(t) - X_{id}(t)\right) + c_2.r_2\left(P_{gd}(t) - X_{id}(t)\right)$$
$$x_{id}(t+1) = x_{id}(t) + v_{id}(t+1) \qquad (3)$$

Denklemde, v_{id}, parçacığın hızını, x_{id} ise parçacığın o anki konumunu ifade etmektedir. w, atalet faktörünü göstermekte ve r_1 ile r_2, popülasyonun çeşitliliğini korumak için kullanılan [0, 1] aralığında değişen iki rastgele sayıyı temsil etmektedir. c1 ve c2 ise öğrenme katsayılarını vermektedir(Wang vd., 2018;Aydın vd., 2018).

3.1.3 Performans Değerlendirme Kriterleri(Performance Evaluation Criteria)

Sınıflandırma sonucunun ve algoritma performansının belirlenmesinde çeşitli yaklaşımlar kullanılmaktadır(Masetic ve Subasi, 2016). RF sınıflandırıcı algoritmasının kalitesinin belirlenmesi için; duyarlılık, özgüllük, kesinlik, doğruluk ve F-ölçüsü kriterleri ele alınmıştır. Bu kriterler; doğru pozitif (ing. True Positive-TP), yanlış pozitif (ing. False Positive-FP), doğru negatif (ing. True Negative-TN) ve yanlış negatif (ing. False Negative-FN) tahmin değerleri kullanılarak değerlendirilmektedir(Ozcift ve Gulten, 2011).

- TP, kalp hastası olarak etiketlenen örneğin gerçekten hasta olması
- FP, kalp hastası olmayan örneğin hasta olarak etiketlenmesi

- TN, kalp hastası olmayan örneğin gerçekten hasta olmaması
- FN, kalp hastası olan örneğin kalp hastası değil olarak etiketlenmesi

anlamlarına gelmektedir.

Sınıflandırma işleminde gerçekleşen ve tahmin edilmesi istenen değerlerin gösterilmesi için Tablo 2'de gösterildiği gibi Karmaşıklık Matris (ing. Confusion Matrix) tablosu kullanılmaktadır. Karmaşıklık Matris (ing. Confusion Matrix), bir sınıflandırma sistemi tarafından yapılan gerçek ve tahmin edilen sınıflandırmalardan oluşmaktadır. Sınıflandırmanın performansı matristeki veriler kullanılarak değerlendirilmektedir(Mursalin vd., 2017).

Tablo 2. Karmaşık Matris(ing. Confusion Matrix)

	Gerçek Pozitif (1)	Gerçek Negatif (0)
Tahmin Pozitif (1)	TP	FP
Tahmin Negatif (0)	FN	TN

3.1.3.1 Duyarlılık(Sensitivity)

Duyarlılık kriteri, gerçekten kalp hastalığı olan örneklerin kaç tanesinin pozitif olarak tahmin edildiğini göstermektedir. Denklem 3'de duyarlılık tahmininin matematiksel ifadesi verilmiştir.

$$\text{Duyarlılık} = TP \,/\, (TP + FN) \tag{4}$$

3.1.3.2 Özgüllük(Specificity)

Sınıflandırıcının hasta olmayan örnekleri tahmin etmedeki etkinliği özgüllük kriteri ile saptanmaktadır. Denklem 4'de özgüllük tahmininin matematiksel ifadesi verilmiştir.

$$\text{Özgüllük} = TN \,/\, (TN + FP) \tag{5}$$

3.1.3.3 Kesinlik(Precision)

Kesinlik kriteri ile kalp hastası olarak tahmin edilen örneklerden gerçekten kaç tanesinin pozitif olduğu saptanmaktadır. Denklem 5'de kesinlik tahmininin matematiksel ifadesi verilmiştir.

$$\text{Kesinlik} = TP \,/\, (TP + FP) \tag{6}$$

3.1.3.4 Doğruluk(Accuracy)

Doğruluk bir modelin başarısını vermektedir. Doğru tahminlerin tüm verilere oranlanması ile bulunmaktadır. Denklem 6 ve 7'de doğruluk ve hata oranı tahminlerinin matematiksel ifadesi verilmiştir.

$$\text{Doğruluk} = (TP + TN) \,/\, (TP + TN + FP + FN) \tag{7}$$

$$\text{Hata Oranı (ERR)} = 1 - ACC \tag{8}$$

3.1.3.5 F-ölçüsü(F-Measure)

F-ölçüsü (ing. F-measure), kesinlik ve duyarlılık değerlendirme ölçülerinin harmonik ortalamasını vermektedir. Denklem 8'de F-ölçüsü tahmininin matematiksel ifadesi verilmiştir.

$$\text{F-Ölçüsü} = 2 * (precision * recall) \,/\, (precision + recall) \tag{9}$$

F1 ölçüsü kriterinin kullanımı ile eşit dağılmayan veri kümelerinde hatalı model seçiminin önüne geçilmektedir.

3.1.3.6 Alıcı İşlem Karakteristikleri Eğrisi (Receiver Operating Characteristic – ROC Curve)

1967 yılında Lusted tarafından önerilen ROC eğrisi,1969 yılında medikal görüntüleme alanında kullanılmaya başlanmıştır ve kullanımın yaygınlaşmasıyla birlikte tıp alanında hastalık teşhislerinde ROC eğrisi analizinin

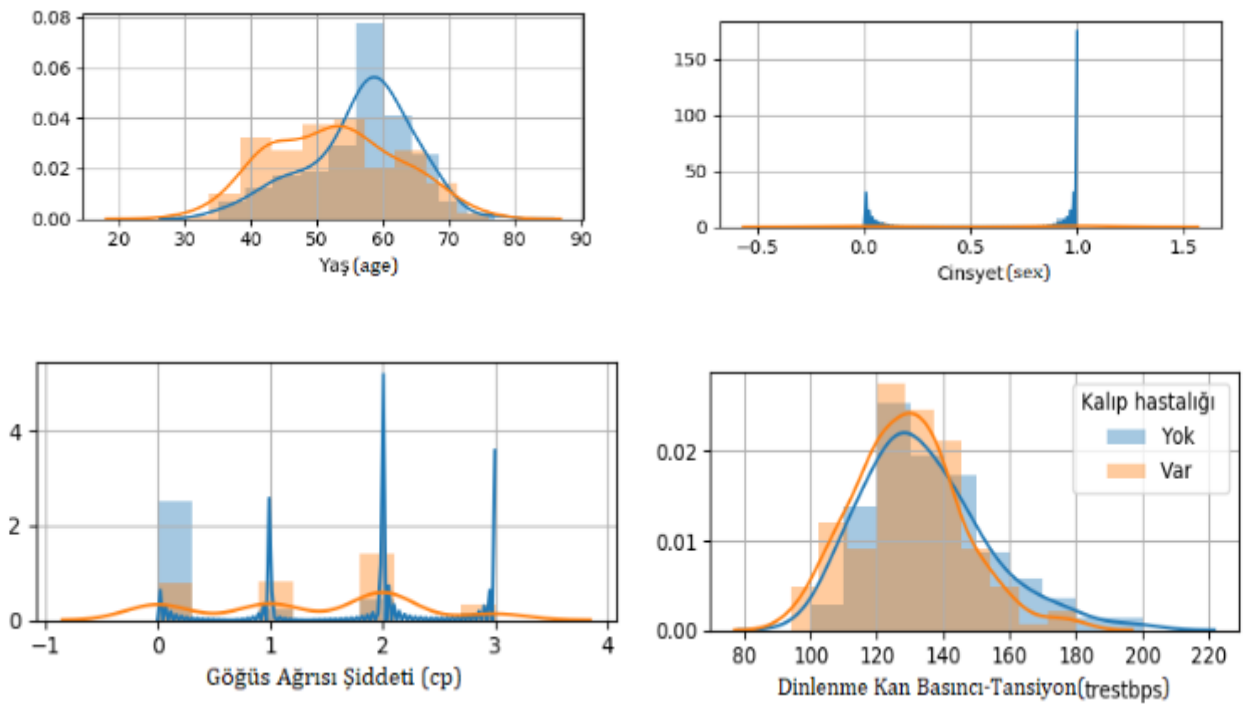
kullanımı artış göstermektedir(Kartal, 2015). ROC eğrisi y eksenindeki TP (hassasiyet) ve x eksenindeki FP (özgüllük) oranının çizilmesi ile elde edilmektedir(Köksel, 2011). Kısacası, TP değerinin FP değerine oranı ROC eğri grafiğini vermektedir. Grafik üzerinde ROC eğrisinin altında kalan alan (AUC) değeri 0'dan 1'e doğru yaklaşması pozitif değerlerin negatif değerlerden başarılı bir şekilde ayrıldığı göstermektedir ve tanı değeri yükselmektedir(Chaovalitwongse vd., 2007).

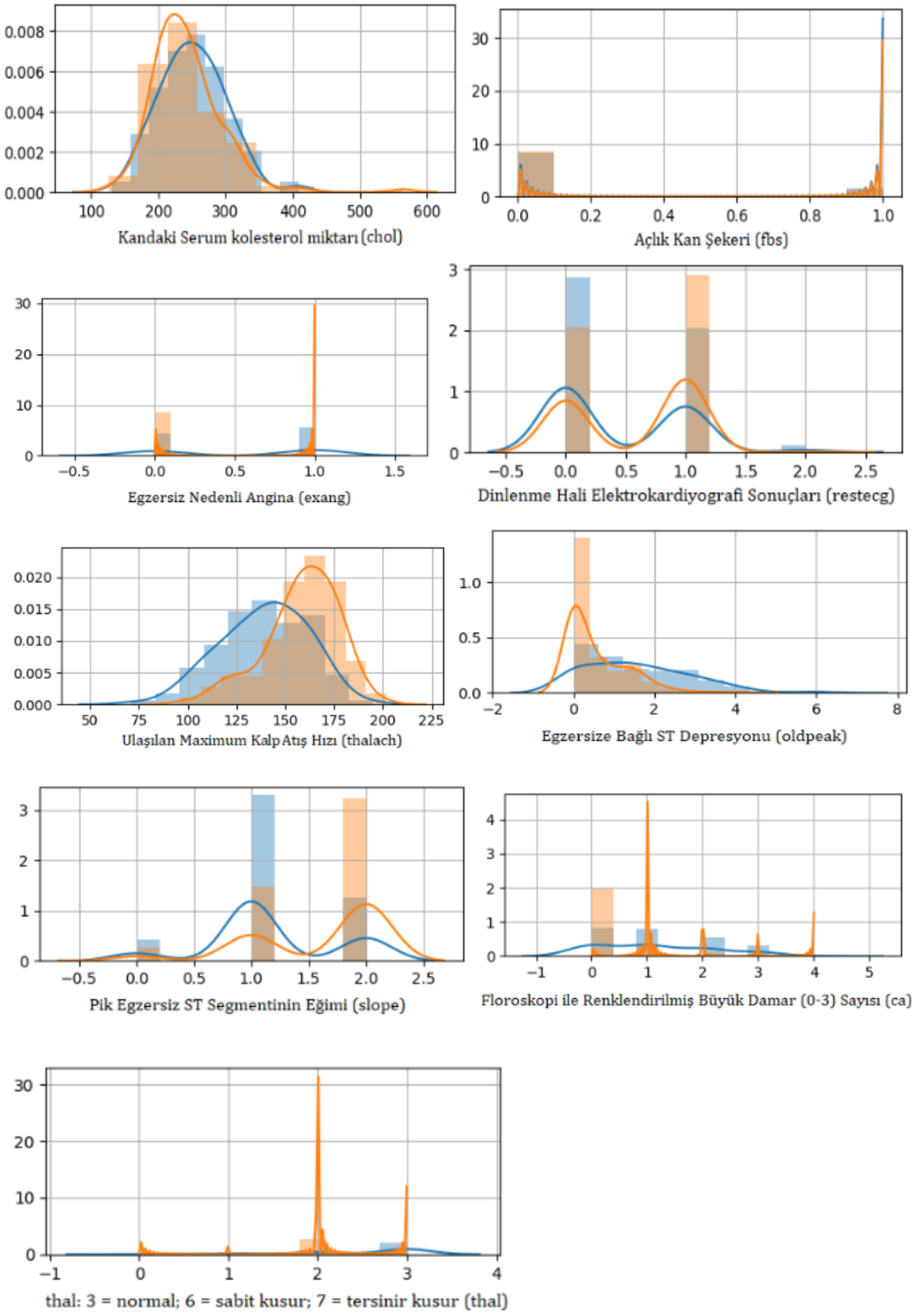
3.2 Yöntem(Method)

Kalp hastalığının tespiti için yapılan çalışmada açık kaynak erişimli internet sitesi olan Kaggle veritabanının 'Heart Disease UCI' veri seti kullanılmıştır. Kullanılan veri seti 167 kalp hastası olmayan ve 136 kalp hastası olan toplam 303 veriden oluşmaktadır. Alınan veri setindeki 13 özellik için her bir kişiye tıbbi tahliller kullanılarak bazı ölçümler yapılmıştır. Tablo 3' de ölçümü gerçekleştirilen özellikler, şekil 1'de ise çıkış parametrelerine göre her bir giriş verisinin dağılımına ait grafikler verilmiştir.

Tablo 3. Kalp Hastalığı Veri Seti Özellikleri(ing. Heart Disease Data Set Features)

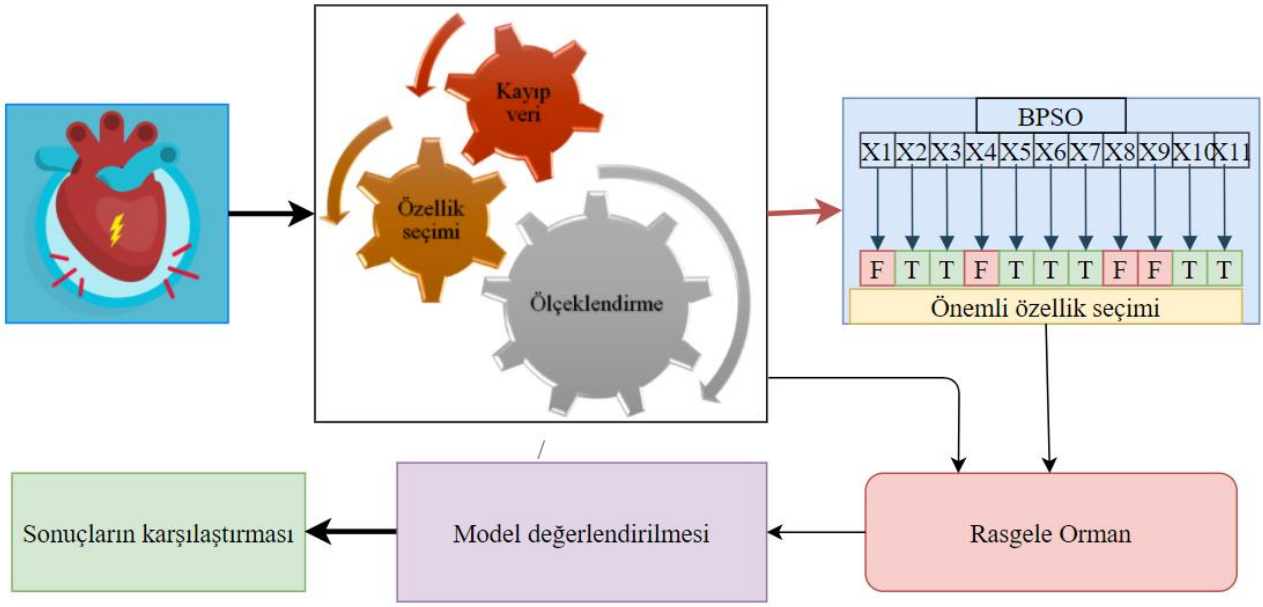
Özellik No	Özellik Bilgisi
1	Hastanın Yaşı (İng. age)
2	Hastanın Cinsiyeti (İng. sex)
3	Göğüs Ağrısı Şiddeti (1-4) (İng. cp)
4	Dinlenme Kan Basıncı (Tansiyon) (İng. trestbps)
5	Kandaki Serum kolesterol miktarı (mg/dl) (İng. chol)
6	Açlık Kan Şekeri > 120 (mg/dl) (İng. fbs)
7	Dinlenme Hali Elektrokardiyografi Sonuçları (0-1-2)(İng. restecg)
8	Ulaşılan Maximum Kan Atış Hızı (İng. thalach)
9	Egzersiz Nedenli Angina (İng. exang)
10	Egzersize Bağlı ST Depresyonu (İng. oldpeak)
11	Pik Egzersiz ST Segmentinin Eğimi (İng. slope)
12	Floroskopi ile Renklendirilmiş Büyük Damar (0-3) Sayısı (İng. ca)
13	thal: 3 = normal; 6 = sabit kusur; 7 = tersinir kusur (İng. thal)





Şekil1.Çıkış Parametresine Göre Giriş Verilerinin Dağılım Grafikleri(ing. Distribution Plots of Input Data According to Output Parameters)

Şekil 2'de ise çalışmanın iş akış diyagramı verilmiştir. İlk aşamada veri setinden alınan tüm veriler arasında kayıp veri, özellik seçimi ve ölçeklendirme işlemleri gerçekleştirilmiştir. İkinci aşamada PSO algoritması ile önemli özelliklerin seçimi sağlanmıştır. Son aşamada ise Python programlama dilinde RF algoritmasını oluşturmak için bir yazılım gerçekleştirilmiştir. Gerçekleştirilen yazılım ile eğitim veri seti RF sınıflandırıcısı için giriş verilerini oluştururken, test veri seti ise modeli test etmek ve değerlendirmek için kullanılmıştır. Elde edilen model doğruluk, özgüllük, duyarlılık, kesinlik, F-ölçüsü ve ROC eğrisi performans değerlendirme ölçütlerine göre değerlendirilmiş ve sonuçlar karşılaştırılmıştır.



Şekil 2. RF Sınıflandırma İş Akış Diyagramı(ing. RF Classification Workflow Diagram)

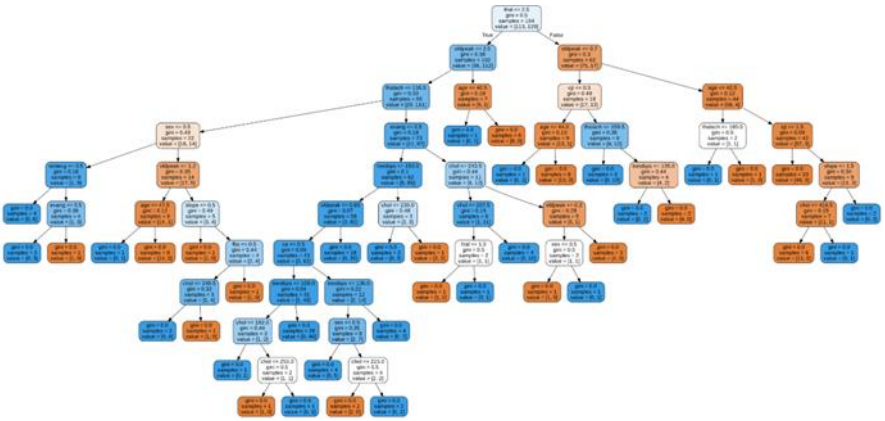
3. Araştırma Bulguları (Research Findings)

Yapılan çalışmada Python programlama dili kullanılarak RF makine öğrenme algoritması ile kalp hastalığı bilgilerini içeren veri setinin 136 adet kalp hastası ve 167 kalp hastası olmayan toplam 303 örnek üzerinde kalp hastalığı teşhisi gerçekleştirilmiştir. Tablo 4'de veri seti içerisinde yer alan ilk 5 veri ve özellikleri verilmiştir.

Tablo 4. İlk 5 Örnek ve Özellikleri(ing. Top 5 Examples And Their Features)

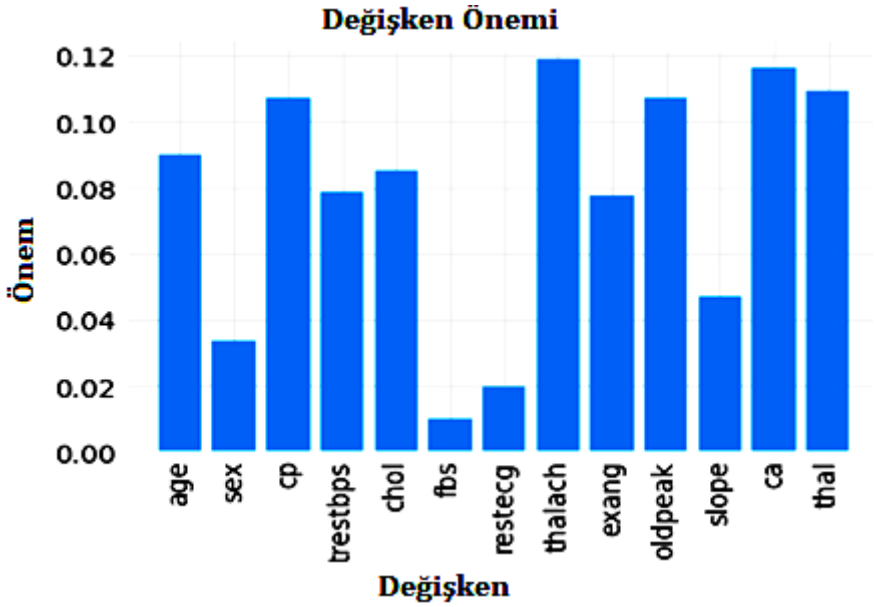
	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

Veri kümesi sınıflaması için oluşturulan karar ağacı Şekil 3'de gösterildiği gibi graphviz ile görselleştirilmiştir. Bu çalışmada bölünme değeri olarak Gini yöntemi kullanılmıştır. Görselleştirme işlemi yapılırken ana düğüm olarak en yüksek gini değeri = 0.5 olan yani en ayırt edici özelliğe sahip kalp atış hızı (ing. thalach) özelliği seçilmiştir.



Şekil 3. Graphviz Modeli(ing. Graphviz Model)

13 özellik üzerinde önem ölçümü gerçekleştirilmiş ve önemli özelliklerin ön plana çıkarılması sağlanmıştır. Özellik önem grafiği Şekil 4'de gösterilmiştir.



Şekil 4. Özellik Önem Grafiği(ing. Feature Significance Chart)

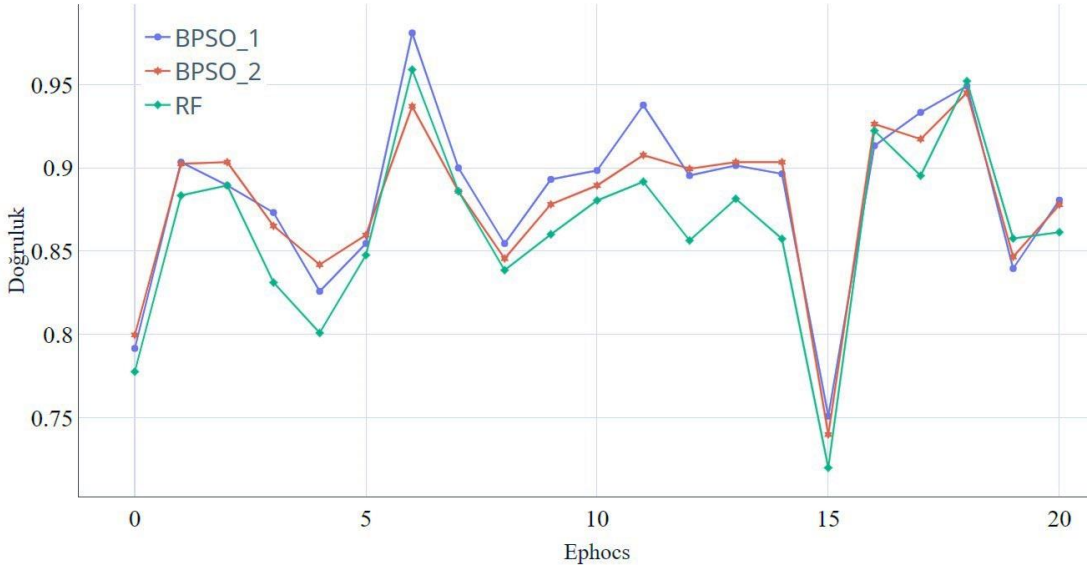
Şekilde görüldüğü gibi thalach değişkeni yani elde edilen maksimum kalp atış hızı değeri kalp hastalığının teşhisinde dikkate alınan en önemli özelliktir.

Çalışmanın bir sonraki aşamasında PSO yöntemi ile veri setine iki farklı popülasyon değeri uygulanmıştır. Tablo 5'de kullanılan yöntem ve öznitelik tablosu verilmiştir. Tablo 5'deki yöntem ve öznitelik tablosu kullanılarak birinci popülasyon değeriyle veri setinden 6 öznitelik ve ikinci popülasyon değeriyle 9 öznitelik seçilerek, Rastgele Orman (ing. Random Forest) yöntemi ile eğitim gerçekleştirilmiştir. Parçacık sürü optimizasyonu kullanılarak özellik seçimi optimize edilmiş ve sisteminin doğruluk üzerindeki etkisi incelenerek Şekil 5'de sonuçlar gösterilmiştir.

Tablo 5. Kullanılan Yöntem ve Öznitelik Tablosu (ing. Method and Attribute Table Used)

Yöntem	Kullanılan Öznitelikler
BPSO_1	[1, 2, 3, 7, 8, 9]

BPSO_2	[0, 1, 2, 3, 6, 7, 8, 9, 10]
RF	[0, 1, 2, 3,4,5, 6, 7, 8, 9, 10,11,12,13]



Şekil 5. Parçacık Sürü Optimizasyon ve Random Forest Doğruluk Grafiği (ing. Particle Swarm Optimization and Random Forest Accuracy Plot)

Çalışmada ayrıca karmaşık matrisi (ing. Confusion Matrix) ile 61 test verisi üzerinde hastalık teşhisi için tahmin sınıflandırma işlemi gerçekleştirilmiştir. Şekil 6'da tasarlanan karmaşıklık Matris (ing. Confusion Matrix) verilmiştir.

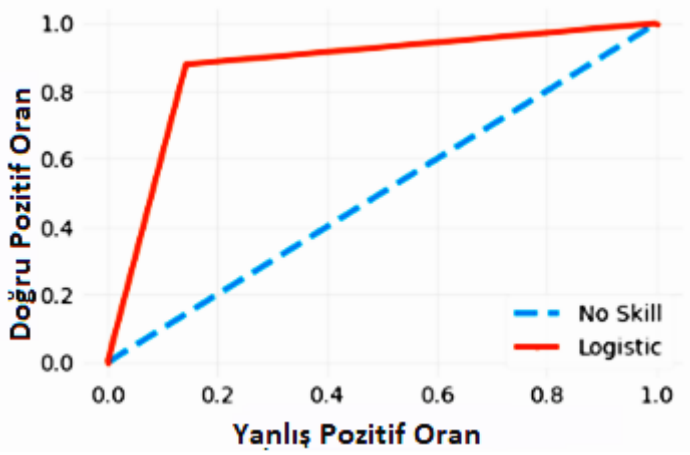
Şekil 6. Karmaşıklık Matris(ing. Confusion Matrix)

Bu çalışma ile bir veri seti içerisinde yer alan 303 örneğin RF algoritması kullanılarak kalp hastası olup olmadığına dair bir sınıflandırma işlemi gerçekleştirilmiştir. Sınıflandırma işlemi farklı performans değerlendirme kriterlerine göre değerlendirilerek sınıflandırıcının başarı oranı incelenmiş ve sonuçlar Tablo 6'da verilmiştir.

Tablo 6. Performans Değerlendirme Kriterleri(ing. Performance Evaluation Criteria)

Doğruluk (ACC)	Duyarlılık (Sensitivity)	Özgüllük (Specificity)	Kesinlik (Precision)	F – ölçüsü (F-masure)	ROC (AUC)
% 86.88	% 87.87	% 85.71	% 87.87	% 87.87	%86.8

Tabloda da görüldüğü gibi AUC değeri %86.8 olarak elde edilmiştir ve bu değerin 1 değerine çok yakın olduğu gözlemlenmiştir. Sonucu grafik üzerinde incelemek için şekil 7'de çalışmaya ait ROC eğrisi verilmiştir. Eğri incelendiğinde kalp hastalıkları tespitinde ideal bir ROC eğrisine yakın sonuç verdiği görülmektedir.



Şekil 7. ROC Eğrisi(ing. ROC Curve)

4.Sonuç(Result)

Yapay zekâ yöntemleri birçok farklı alanda olduğu gibi tıp alanında da önemli bir yere sahiptir ve çeşitli hastalıkların teşhis edilmesinde büyük bir rol oynamaktadır. İnsan hayatını riske atan hastalıkların erken ve doğru bir şekilde teşhis edilmesi bu hastalıklardan kaynaklanan ölüm sayısının önüne geçilmesini sağlayacaktır.
Bu çalışmada RF sınıflandırma yöntemi ile kalp hastalığı teşhisi gerçekleştirilmiş ve sınıflandırma başarısı performans açısından değerlendirilmiştir. Elde edilen sonuçlar aşağıda maddeler halinde verilmiştir.

- İlk olarak RF sınıflandırma modeli için Karmaşıklık Matris (ing. Confusion Matrix) değerlendirme kriterine göre modelin değerlendirilmesi gerçekleştirilmiştir. Modelin değerlendirilmesi ile 61 test verisi içerisinde kalp hastası olan 24 örneği ve kalp hastası olmayan 29 örneği doğru tahmin ettiği gözlemlenmiştir.

- İkinci olarak RF sınıflandırma modelinin; doğruluk, özgüllük, duyarlılık, kesinlik ve F-ölçüsü ölçütlerinden oluşan performans değerlendirme kriterlerine göre başarı oranı incelenmiştir. Değerlendirme sonucunda RF sınıflandırmanın %86.88 doğruluk, %85.71 özgüllük, %87.87 duyarlılık, %87.87 kesinlik ve %87.87 F-ölçüsü değeri ile başarılı tahmin gerçekleştirdiği belirlenmiştir.

- Son olarak sınıflandırma modelinin performansı ROC eğrisi kriteri ile değerlendirilmiştir. Yapılan değerlendirme ile %86.8 AUC değeri ile modelin kalp hastalığı teşhisini %86.88 doğruluk oranında doğru tespit ettiği gözlemlenmiştir.

İleride yapılacak çalışmalarda farklı yapay zekâ yöntemleri kullanılarak doğruluk oranının artırılmasının mümkün olacağı düşünülmektedir.

Teşekkür (Acknowledgement)

Çalışmada 'Heart Disease UCI' veri setini açık kaynak erişimli internet sitesine (kaggle.com) aktaran kişi/kişilere teşekkürlerimizi sunarız.

Çıkar Çatışması (Conflict of Interest)

Yazarlar tarafından herhangi bir çıkar çatışması beyan edilmemiştir. No conflict of interest was declared by the authors.

Kaynaklar (References)

Ahmad, M. W., Mourshed, M., & Rezgui, Y. 2017. Trees vs Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption. Energy and Buildings, 147, 77-89

Akar, Ö., & Güngör, O. 2012. Rastgele orman algoritması kullanılarak çok bantlı görüntülerin sınıflandırılması. Jeodezi ve Jeoinformasyon Dergisi, ss, 139-146.

Aydın, İ., & Aşıcı, B. (2020). İnsan Hareketlerinin Tanınması için Parçacık Sürü Optimizasyonu Tabanlı Topluluk Sınıflandırıcı Yöntemi. Fırat Üniversitesi Mühendislik Bilimleri Dergisi, 32(2), 381-390.

Aydın, İ., Salur, M. U., & Başkaya, F. (2018). Duygu Analizi için Çoklu Populasyon Tabanlı Parçacık Sürü Optimizasyonu. Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi, 11(1), 52-64.

Barrett Lowe and Kulkarni A. D. (2015). Multispectral Image Analysis Using Random Forest, International Journal on Soft Computing, vol. 6, no. 2, pp 1-14

Boyraz, Ö. F., Seymen, V., Bozkurt, M. R., & Çetin, Ö. 2014. Makine Öğrenmesi Algoritmaları Kullanılarak Kalp Hastalığı Tespiti. Icemst 2014, 1260.

Breiman, L. Rastgele Ormanlar. 2001. Machine Learning 45, 5-32.

Bulut, F. 2016, May. Heart attack risk detection using Bagging classifier. In 2016 24th Signal Processing and Communication Application Conference (SIU) (pp. 2013-2016). IEEE.

Chaovalitwongse, W. A., Fan, Y. J., & Sachdeo, R. C. 2007. On the time series $ k $-nearest neighbor classification of abnormal brain activity. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 37(6), 1005-1016.

Chen, J., Li, K., Tang, Z., Bilal, K., Yu, S., Weng, C., & Li, K. 2016. A parallel random forest algorithm for big data in a spark cloud computing environment. IEEE Transactions on Parallel and Distributed Systems, 28(4), 919-933.

Çınaroğlu, S., & Bulut, H. (2018). K-ortalamalar ve parçacık sürü optimizasyonu tabanlı kümeleme algoritmaları için yeni ilklendirme yaklaşımları. Journal of the Faculty of Engineering & Architecture of Gazi University, 33(2).

Das, R., Turkoglu, I., & Sengur, A. 2009. Effective diagnosis of heart disease through neural networks ensembles. Expert systems with applications, 36(4), 7675-7680.

Gulia, A., Vohra, R., & Rani, P. 2014. Liver patient classification using intelligent techniques. International Journal of Computer Science and Information Technologies, 5(4), 5110-5115.

Horning, N. 2010, December. Random Forests: An algorithm for image classification and generation of continuous fields data sets. In Proceedings of the International Conference on Geoinformatics for Spatial Infrastructure Development in Earth and Allied Sciences, Osaka, Japan (Vol. 911).

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

Kılıçarslan, S., & Çelik, M. 2019. Rotasyon Orman Sınıflandırma Algoritması Kullanarak Kronik Böbrek Rahatsızlığının Tahmini. *Dumlupınar Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, (043), 21-34.

Köksal, B. 2011. Regresyon analizinde ROC eğrisi kestirimi ile model seçimi. Yüksek Lisans Tezi. Marmara Üniversitesi, Sosyal Bilimler Enstitüsü. İstanbul.

Masetic, Z., & Subasi, A. 2016. Congestive heart failure detection using random forest classifier. Computer methods and programs in biomedicine, 130, 54-64.

Mursalin, M., Zhang, Y., Chen, Y., & Chawla, N. V. 2017. Automated epileptic seizure detection using improved correlation-based feature selection with random forest classifier. Neurocomputing, 241, 204-214.

Opeyemi, O., & Justice, E. O. 2012. Development of neuro-fuzzy system for early prediction of heart attack. Information Technology and Computer Science, 9(9), 22-28.

Ozcift, A., & Gulten, A. 2011. Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithms. Computer methods and programs in biomedicine, 104(3), 443-451.

Özekes, S. 2003. Veri madenciliği modelleri ve uygulama alanları. İstanbul Ticaret Üniversitesi Dergisi. ss, 65-82.

Pal, M. 2005. Random forest classifier for remote sensing classification. International Journal of Remote Sensing, 26(1), 217-222.

Özsağlam, M. Y., & Çunkaş, M. (2008). Optimizasyon problemlerinin çözümü için parçaçık sürü optimizasyonu algoritması. *Politeknik Dergisi*, *11*(4), 299-305.

Palaniappan, S., & Awang, R. 2008, March. Intelligent heart disease prediction system using data mining techniques. In 2008 IEEE/ACS international conference on computer systems and applications (pp. 108-115). IEEE.

Priyanka, N., & RaviKumar, P. 2017, April. Usage of data mining techniques in predicting the heart diseases—Naïve Bayes & decision tree. In 2017 International Conference on Circuit, Power and Computing Technologies (ICCPCT) (pp. 1-7). IEEE.

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

Sengur, A. 2008. An expert system based on principal component analysis, artificial immune system and fuzzy k-NN for diagnosis of valvular heart diseases. Computers in Biology and Medicine, 38(3), 329-338.

Shao, Y. E., Hou, C. D., & Chiu, C. C. 2014. Hybrid intelligent modeling schemes for heart disease classification. Applied Soft Computing, 14, 47-52.

Singh, J., Kamra, A., & Singh, H. 2016, October. Prediction of heart diseases using associative classification. In 2016 5th International Conference on Wireless Networks and Embedded Systems (WECON) (pp. 1-7). IEEE.

Üner, S., Balcılar, M., & Ergüder, T. 2018. Türkiye hanehalkı sağlık araştırması: bulaşıcı olmayan hastalıkların risk faktörleri prevalansı 2017 (STEPS). Dünya Sağlık Örgütü Türkiye Ofisi, Ankara.

Yan, H., Jiang, Y., Zheng, J., Peng, C., & Li, Q. 2006. A multilayer perceptron-based medical decision support system for heart disease diagnosis. Expert Systems with Applications, 30(2), 272-281.

Yılmaz, M. 2018. Tarımsal Yaz Ürünlerin Sentinel-2 Uydu Görüntülerinden Rastgele Orman Algoritması İle Nesne-Tabanlı Sınıflandırılması. Yüksek Lisans Tezi. Hacettepe Üniversitesi, Fen Bilimleri Enstitüsü. Ankara.

Wang, D., Tan, D., & Liu, L. (2018). Particle swarm optimization algorithm: an overview. Soft Computing, 22(2), 387-408.

Wang, F., Zhang, H., Li, K., Lin, Z., Yang, J., & Shen, X. L. (2018). A hybrid particle swarm optimization algorithm using adaptive learning strategy. Information Sciences, 436, 162-177.