Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi
Dokuz Eylul University Faculty of Engineering Journal of Science and Engineering

Basılı/Printed ISSN: 1302-9304. Elektronik/Online ISSN: 2547-958X

Kalp hastalık risk tahmini için Python aracılığıyla sınıflandırıcı algoritmalarının performans değerlendirmesi

Performance evaluation of classifier algorithms with Python for heart disease risk prediction

Serdar Gündoğdu [1*]

[1] Dokuz Eylül Üniversitesi Bergama Meslek Yüksekokulu Bilgisayar Teknolojileri Bölümü, İzmir, TÜRKİYE (*)
*Sorumlu Yazar / Corresponding Author **: serdar.gundogdu@deu.edu.tr (*)

Geliş Tarihi / Received: 13.09.2020 *Araştırma Makalesi/Research Article*
Kabul Tarihi / Accepted: 20.03.2021 *DOI:10.21205/deufmd.2021236926*
Atıf şekli/ How to cite: GÜNDOĞDU S.(2021). Kalp hastalık risk tahmini için Python aracılığıyla sınıflandırıcı algoritmalarının performans değerlendirmesi. DEÜFMD 23(69), 1005-1013.

Öz

Kalp hastalıkları, günümüzün en büyük sağlık problemlerinden birisidir. Hastalık için erken teşhis, erken ölümlerin önüne geçilebilir. Bu amaçla Kaggle veri tabanından elde dilen veri setinde bulunan 13 bağımsız değişken kullanılarak kalp hastalığı olma olasılığı az (KHOA) ve fazla (KHOF) olan kişiler ayırt edilmeye çalışılmıştır. Çalışmada destek vektör makinaları (DVM), k-en yakın komşu (k-NN), karar ağaçları (KA), lineer diskriminant analiz (LDA), Gausian Naive Bayes (GNB), Gradient Boosting (GB) ve Random Forest (RF) olmak üzere 7 sınıflandırma algoritması kullanılmıştır. Random forest, özgüllük (%100), Matthews korelasyon katsayısı (0.90), Fowlkes-Mallows indeks (0.82), F1 skoru (%89.7) ve doğruluk (%90.2) değerlerine göre çalışmanın en iyi tahminini yapan algoritması olmuştur. Açlık kan şekeri, KHOA ve KHOF grupları arasında istatiksel olarak anlamlı fark saptanamamış ve özellikler arasında en az önemli olduğu bulunmuştur. Bu özellik çıkarılarak yapılan sınıflandırma işlemlerinde önemli bir performans değişikliği görülmemiştir. Sadece işlem zamanları, az da olsa kısalmıştır. Bu çalışma, erken teşhislere destek olacağından dolayı kalp hastalığının tahmininde fayda sağlayacaktır.

Anahtar Kelimeler: *Kalp hastalığı, tahmin, sınıflandırıcılar, Python*

Abstract

Heart diseases are one of the biggest health problems of today. Early diagnosis for the disease can prevent early deaths. For this purpose, by using 13 independent variables in the data set obtained from the Kaggle database, people with low probability of heart disease and people with excess were tried to be distinguished. Seven classification algorithms were used in the study, namely support vector machines (SVM), k-NN, decision trees, linear discriminant analysis (LDA), Gausian Naive Bayes (GNB), Gradient Boosting (GB) and Random Forest (RF). Random forest was the algorithm that made the best estimation of the study according to the values of specificity (100%), Matthews correlation coefficient (0.90), Fowlkes-Mallows index (0.82), F1 score (89.7%) and accuracy (90.2%). There was no statistically significant difference between the groups in fasting blood glucose and it was found to be the least important among the features. No significant performance change was observed in the classification processes made by removing this feature. Only the processing times are slightly shorter. This study will help predict heart disease as it will support early diagnosis.

Keywords: *Heart disease, prediction, classifiers, Python*

1. Giriş

Kardiyovasküler hastalıklar, her yıl yaklaşık (dünyadaki tüm ölümlerin yaklaşık % 31'i) 17,9 milyon ölüme neden olan küresel olarak bir numaralı ölüm nedenidir [1]. Kardiyovasküler hastalıklar, koroner kalp hastalığı, serebrovasküler hastalık, romatizmal kalp hastalıkları gibi kalbin ve damarların tüm hastalıklarını kapsar. Beş hastalık ölümünden dördü kalp krizi ve felçten kaynaklanmaktadır ve bu ölümlerin üçte biri 70 yaşın altındaki kişilerde erken gerçekleşmektedir [1]. Bu kadar önemli bir hastalık türünün erken tespiti ve tedavisi zorunlu hale gelmiştir. Yüksek kardiyovasküler hastalık riski altında olanları tespit ederek uygun tedavi görmelerini sağlamak erken ölümleri önleyebilir.

Makine öğrenmesi diğer medikal uygulamalarda olduğu gibi kalp hastalığı araştırmalarında da önemli bir rol oynamıştır. Mevcut kliniksel verilerde sağlıklı ve kalp hastalıklı bireyler arasındaki farkı bulmak, sınıflandırma çalışmasında güçlü bir yaklaşım olmuştur. Kardiyovasküler hastalıkların sınıflandırılabilmesi, hasta bireylerin tedavisi için kritik bir temel oluşturur. İstatistik ve makine öğrenimi, klinik verilere dayanan kalp hastalığının durumunu tahmin etmek için uygulanan yaklaşımlardandır [2],[3].

Kardiyovasküler hastalıklarla ilgili yapılan araştırmalarda farklı makine öğrenme yöntemlerine başvurulmuştur [4]. Bunlardan birkaç tanesi; karar ağaçları ve destek vektör makinesinin hastalık tahmin etmesi [5], bilgisayarlı kardiyovasküler hastalık tanısı ve kategorizasyon [6], hastalığın tanımlanması için bulanık destek vektör kümelenmesi [7], çok katmanlı algılama topluluğuna dayanan bir komite makinesi [8], hastalıkların sınıflandırılması için önerilen veri füzyon yaklaşımı [9], genetik destek vektör makineleri tabanlı akıllı bir sistem [10], destek vektör makine sınıflandırmasına dayalı bir otomatik algılama sisteminin kullanılması [11] şeklinde sıralanabilir.

UCI ve Kaggle'dan elde edilen ve bu çalışmada kullanılan veritabanı, önceki araştırmacıların da dikkatini çekmiş ve bugüne kadar birçok yayına destek olmuştur. Chen vd. [12], kalp hastalığı tahmini için kullandığı veri setine yapay sinir algoritması uygulayarak %80 civarında bir başarı elde etmişlerdir. Patel vd. [13] çalışmalarında, %56,76 ile en yüksek sınıflandırma doğruluk oranına üç algoritma arasından J48 ile ulaşmışlardır. Başka bir çalışmada kalp hastalığını tahmin etmek için naive Bayes, karar listesi ve k-NN algoritmalı MongoDB adlı veritabanı yazılımı kullanılmıştır [14]. Algoritmaların buldukları doğruluk değerleri sırasıyla naive Bayes, karar listesi ve k-NN için %52.33, %52 ve %45.67 olmuştur. Enriko vd. [15] tarafından aynı veri seti üzerinden k-NN algoritma uygulanan bir tahmin yapmışlardır. 8 özellik kullanarak yaptıkları çalışmada %81.85 doğruluk değerini yakalamışlardır. Shao vd. [16] yaptığı çalışmada kalp hastalığını sınıflandırmak için lojistik regresyon (LR), çok değişkenli adaptif regresyon eğrileri (MARS), kaba küme teorisi (RS) ve yapay sinir ağı (ANN) ve hibrit LR–ANN, MARS–ANN ve RS–ANN modellerini kullanmışlardır. Her ikisi de hibrit model olan RS–LR ve MARS–LR'nin kalp hastalığı sınıflandırması için en iyi modeller olduğu sonucuna varmışlardır. Taşçı ve Şamlı [17] ise WEKA programı üzerinden çalıştırdıkları algoritmalar içinde performans kriter sonuçlarının ortalaması alındığında k-NN'nin en doğru ve en iyi sonucu veren yöntem olduğunu göstermişlerdir.

Tunç vd. [18] derin öğrenme yöntemi uygulayarak kalp krizi geçirme durumunun belirlenmesi için oluşturdukları modelden elde ettikleri doğruluk, duyarlılık ve özgüllük değerleri sırasıyla %81.4, %80.4 ve %82.3 olmuştur. Akalın vd. [19] çalışmalarında, en başarılı sınıflandırma doğruluk oranına %82.50 ile Random Forest algoritması ile gözlemlemişlerdir. Khan ve Mondal [20], kalp hastalığını tahmin etmek için lojistik regresyon, SVM ve Naif Bayes'in bir kombinasyonu olarak çoğunluk oylama algoritmasını önermişlerdir. Bu algoritma % 88,89 değeri ile en iyi doğruluk oranını yakalamıştır. Tougui vd. [21] yaptıkları araştırmada kalp hastalığı sınıflandırması için altı veri madenciliği aracı ve bu araçların her biri kullanılarak veri setine altı makine öğrenimi tekniği uygulamışlardır. Buldukları 85.86% (doğruluk), 83.94 (duyarlılık) ve % 87.50% (özgüllük) değerleri ile Matlab'ın en iyi performans gösteren araç ve Matlab'ın yapay sinir ağı modelinin en iyi performans gösteren teknik olduğunu belirtmişlerdir.

Bu çalışmanın amacı, Phyton yazılımı üzerinde yedi sınıflandırma algoritmasını çalıştırmak ve performanslarını karşılaştırarak daha etkili ve doğru bir kalp hastalığı tahmin sistemi sunmaktır. Ayrıca sınıflandırma girişine uygulanan ve hastalık tahmin başarısını doğrudan etkileyen bağımsız değişkenlerin önem sırasını belirlemektir.

2. Materyal ve Metodoloji

Bu çalışmada Kaggle veri tabanında [22] bulunan kalp hastalığı veri seti kullanılmıştır. Veri setini oluşturan toplam 303 kayıt, 165'i kalp hastalığına yakalanma olasılığı fazla (KHOF) ve 138'i de kalp hastalığına yakalanma olasılığı az olan (KHOA) kişilerin verilerinden oluşmaktadır. Kullanılan 13 özellik yaş, cinsiyet, göğüs ağrısı türü, dinlenme hali kan basıncı, serum kolesterolü, açlık kan şekeri, dinlenme hali elektrokardiyografi (EKG) sonuçları, maksimum kalp hızı, egzersize bağlı anjin, dinlenmeye göre egzersizle indüklenen ST depresyonu, pik egzersiz ST segmentinin eğimi, floroskopi ile renklendirilmiş büyük damar sayısı ve defekt tipidir.

Tablo 1. Kalp hastalığı geçirme olasılığını değerlendirmede kullanılan özellikler.

Özellikler	Değerler	Değişken	
Yaş (yıl)	yaş(nümerik)	Bağımsız	f0
Cinsiyet (0 = kadın; 1 = erkek)	cinsiyet (0,1)	Bağımsız	f1
Göğüs ağrısı türü (0:tipik anjin; 1:atipik anjin; 2:anjinal olmayan; 3:asemptomatik)	cp(0,1,2,3)	Bağımsız	f2
Dinlenme hali kan basıncı (mm Hg)	trestbps(nümerik)	Bağımsız	f3
Serum kolestrolü (mg/dl)	chol(nümerik)	Bağımsız	f4
Açlık kan şekeri > 120 mg/dl (0 = yanlış; 1 = doğru)	fbs (0,1)	Bağımsız	f5
Dinlenme hali EKG sonuçları (0: normal; 1: ST-T anormalliği; 2: Estes kriterlerine göre olası veya kesin sol ventrikül hipertrofisi)	restecg (0,1,2)	Bağımsız	f6
Maksimum kalp atış hızı	thalach(nümerik)	Bağımsız	f7
Egzersize bağlı anjin (0 = hayır; 1 = evet)	exang(0,1)	Bağımsız	f8
Oldpeak = Dinlenmeye göre egzersizle indüklenen ST depresyonu	oldpeak(nümerik)	Bağımsız	f9
Pik egzersiz ST segmentin eğimi (0: yukarı doğru; 1: düz; 2: aşağı doğru)	slope(0,1,2)	Bağımsız	f10
Floroskopi ile renklendirilen büyük damar sayısı	ca(0,1,2,3,4)	Bağımsız	f11
Thal: (1 = normal; 2 = sabit kusur; 3 = tersinir kusur)	thal(1,2,3)	Bağımsız	f12
Hedef: 0 = daha az kalp hastalık olasılığı 1 = daha fazla kalp hastalık olasılığı	(0,1)	Bağımlı	

Kullanılan Cleveland veri kümesinin 14 özelliğinin almış olduğu değerler ve veri tipleri ile ilgili özet bilgiler Tablo 1'de sunulmuştur.

Özelliklerle ilgili tanımlayıcı istatistiklerin belirlenmesi için IBM SPSS Statics 24 programı kullanılırken; sınıflandırıcı yöntemlerini çalıştırmak için de Phyton 3.7 aracından yararlanılmıştır.

2.1. Yöntemler

2.1.1. Destek vektör makineleri

Destek vektör makineleri (DVM), birçok sınıflandırma görevi için uygun yapıda olup doğrusal olmayan bir sınıflandırma yöntemidir. Bir DVM'nin ana avantajı, yüksek boyutlu bir özellik alanında iyi genelleşen bir karar kuralı oluşturma yeteneğidir [23]. Bu, eğitim aşamasında daha önce görülmemiş yeni örnekler için bile yüksek doğruluk sağlar. Bir doğrulama süreci algoritmanın genellemesinin kapsamını ölçmeye yardımcı olabilir [24].

2.1.2. K-en yakın komşu (k-NN)

En basit sınıflandırıcılardan birisi olup farklı uygulamalarda karşımıza çıkmaktadır. Sınıflandırıcının ana fikri, bilinmeyen bir örneğe en yakın k etiketli örnekleri seçmek ve en çok komşu olan sınıf etiketini atamaktır [25].

2.1.3. Karar ağaçları

Bir karar ağacı, ortak değişkenlerle bir sonucu tahmin etmek için kullanılan istatistiksel bir modeldir. Model, verilerin ayrık alt kümelerini, yani verilerin ikili bölümleri dizisi yoluyla hiyerarşik olarak tanımlanan nüfus alt gruplarını tanımlayan bir tahmin kuralı ima eder. Hiyerarşik ikili bölümler kümesi bir ağaç olarak

temsil edilebilir. Her alt kümedeki tahmini sonuç, alt kümedeki bireylerin sonuçlarının ortalaması alınarak belirlenir. Amaç, tahmin edilen ve gerçek değerler arasındaki tutarsızlığı ölçen bir kayıp işlevini en aza indiren bir tahmin kuralı oluşturmaktır. Karar ağaçları, bireysel özelliklerin kombinasyonları ile tanımlanan homojen alt grupları tanımlamak için yararlı bir araçtır [26].

2.1.4. Lineer diskriminant analiz

Lineer diskriminant analizi (LDA), örüntü tanıma ve veri analizinde yaygın olarak kullanılan sınıflar arasında ortak nüfus kovaryans matrisi varsayımı üzerine inşa edilmiş popüler bir sınıflandırıcıdır [27].

LDA, veri sınıflandırması için önemli geleneksel bir modeldir. Klasik teori, LDA'nın sabit veri boyutsallığı p ve büyük bir eğitim örneği boyutu n için tutarlı Bayes olduğunu göstermektedir. Bununla birlikte, $p \gg n$ olduğunda yüksek boyutlu ortamlarda, kovaryans matrisinin tutarsız tahmini ve popülasyonların ortalama vektörleri nedeniyle LDA zordur [28].

2.1.5. Gaussian naive Bayes

Naive Bayes, veri madenciliğinde ilk 10 algoritmadan biri olup birçok uygulamada yaygın olarak kullanılan etkin bir sınıflandırıcı türüdür [29]. Algoritma, hızlı öğrenme ve test etme sürecine sahip üretken model tabanlı bir sınıflandırıcıdır [30]. Bayes sınıflandırıcılar, Bayes kuralı ve olasılık teoremine göre çalışırlar. Naive Bayes, Bayes sınıflandırıcısının basitleştirilmiş bir sürümüdür.

Gaussian Naive Bayes sınıflandırıcısı ise sınıf etiketi verilen özellik değerleri üzerinde bir Gauss dağılımı olduğu varsayılarak oluşturulmuş Naive Bayes yöntemidir [31].

2.1.6. Gradient boosting

İleri aşamalı optimizasyon algoritması olan gradient boosting (GB), güçlü bir sınıflandırıcı oluşturmak için her iterasyonda öğrenilen her zayıf sınıflandırıcının kararını kullanır. Bu yöntemde, zayıf sınıflandırıcılar olarak regresyon ağacı modellerini kullanır ve bir kayıp fonksiyonu en aza indirilecek şekilde gradyanlar kavramına dayalı güçlü bir model oluşturur [32]. GB sınıflandırıcı, regresyon ağaçlarının kayıp fonksiyonunun negatif gradyanına uyduğu ilave bir modeldir [33].

2.1.7. Random forest

Random forest (RF) sınıflandırma yönteminin ikili bağımlı değişken için bir sınıflandırmanın olasılığını tahmin etmede etkili olduğu kanıtlanmıştır [34].

Random forest birden fazla ağaç yetiştirir ve etiketleri tüm ağaçların kararlarına göre sınıflandırır. Tüm özellikler arasında, her düğüm için rastgele bir özellik alt kümesi seçilir ve özellik alt kümesindeki en iyi bölünme düğümü bölmek için seçilir [32].

Sınıflandırıcının temel fikri, zayıf öğrenenlerin bir araya gelip daha güçlü bir öğrenen oluşturması, bir kök ile başlaması, dallarını büyütmeye devam etmesi ve sonuçta yapraklar adı verilen terminal düğümüne ulaşmasıdır. Ağaca aktarılan dallar, bu özelliklere dayanan özellikler veya işlenmiş bilgilerdir. Diğer algoritmalarla karşılaştırıldığında, RF sınıflandırıcıları, düşük nispeten yüksek bir doğrulukla büyük bir veri tabanında verimli bir şekilde çalışır [35].

Sınıflandırıcı karşılaştırılmasında her algoritmanın o parametre için kendi varsayılan (default) ayarları kullanılmıştır.

2.2. Değerlendirme kriterleri

Uygulanan veri setinin %80'i eğitim, %20'u test amaçlı ayrılmıştır. Bu yukarıda kullanılan tüm sınıflandırıcılar için geçerlidir.

Modellerin performansını karşılaştırmak için karışıklık matrisi (confusion matrix) kullanılmıştır. Matris, sınıflandırma doğruluğunu değerlendirmek için kullanılan bir ölçüm yöntemidir. Yapıyı oluşturan doğru pozitif (TP), doğru negatif (TN), yanlış pozitif (FP) ve yanlış negatif (FN) ifadelerinin matris üzerinde gösterimi Tablo 2'de verilmiştir.

Tablo 2. İkili sınıflandırma için karışıklık matrisi

Karışıklık Matrisi				
		Doğru sınıf		
		KHOA	KHOF	Toplam
Tahmin sınıfı	KHOA	**TP**	FP	TP+FP
	KHOF	FN	**TN**	FN+TN
	Toplam	TP+FN	FP+TN	

KHOA (kalp hastalığı olasılığı az) ve KHOF (kalp hastalığı olasılığı fazla) etiketli örnekler, sırasıyla pozitif sınıf ve negatif sınıf olarak kabul edilmiştir. Burada;

TP, KHOA olarak sınıflandırılmış KHOA örneklerini; FP, KHOA olarak sınıflandırılmış KHOF örneklerini; TN, KHOF olarak sınıflandırılmış KHOF örneklerini, FN ise KHOF olarak öngörülen KHOA örneklerini belirtmektedir.

Sınıflandırma algoritmaların performansını değerlendirmek için karışıklık matrisinden elde edilen duyarlılık (TPR), özgüllük (TNR), Matthews korelasyon katsayısı (MCC), Fowlkes-Mallows (FM) endeks, F1 skoru ve doğruluk (ACC) parametreleri kullanılmıştır. Bu puanlar Denklemler (1)-(6) 'de verildiği şekilde tanımlanmıştır.

$$TPR = \frac{TP}{TP + FN} \quad (1)$$

$$TNR = \frac{TN}{TN + FP} \quad (2)$$

$$MCC = \frac{TP*TN - FP*FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (3)$$

$$FM = \sqrt{\frac{TP}{TP+FP} * TPR} \quad (4)$$

$$F1 = \frac{2TP}{2TP + FP + FN} \quad (5)$$

$$ACC = \frac{TP + TN}{TP + FP + FN + TN} \quad (6)$$

Tablo 3. 165 KHOF ve 138 KHOA olan kişi kayıtlarıyla ilgili tanımlayıcı istatistikler.

	KHOF	KHOA	p-değeri
Yaş	52.50 (0.74)	56.60 (0.68)	<0.01
Cinsiyet	0.56 (0.04)	0.83 (0.03)	<0.01
Göğüs ağrısı türü	1.38 (0.07)	0.48 (0.08)	<0.01
Dinlenme hali kan basıncı	129.30 (1.26)	134.40 (1.59)	0.035
Serum kolestrolü	242.23 (4.17)	251.09 (4.21)	0.036
Açlık kan şekeri	0.14 (0.03)	0.16 (0.03)	0.626
Dinlenme hali EKG sonuçları	0.59 (0.04)	0.45 (0.05)	0.010
Maksimum kalp atış hızı	158.47 (1.49)	139.10 (1.92)	<0.01
Egzersize bağlı anjin	0.14 (0.03)	0.55 (0.04)	<0.01
Dinlenmeye göre egzersizle indüklenen ST depresyonu	0.58 (0.06)	1.59 (0.11)	<0.01
Pik egzersiz ST segmentin eğimi	1.59 (0.05)	1.17 (0.05)	<0.01
Floroskopi ile renklendirilen büyük damar sayısı	0.36 (0.07)	1.17 (0.09)	<0.01
Thal	2.12 (0.04)	2.54 (0.06)	<0.01

3. Bulgular ve tartışma

Yapılan işlemler, Intel CoreI5 3.40 GHz işlemci, 8 GB RAM ve Windows 10 Professional Edition'ın 64 bit sürümünü çalıştıran bir laptop bilgisayarda gerçekleştirilmiştir.

SPSS programı sayesinde kalp hastalığına yakalanma olasılığı fazla ve az olan kişi kayıtlarıyla elde edilen tanımlayıcı istatistikler elde edilmiş ve sonuçları Tablo 3'de sunulmuştur. Tablo 3'deki değerler ortalama (standart sapma) olarak verilmiştir. Tabloya dahil edilen p değerleri, her girdi için normallik varsayımları bir Kolmogorov-Smirnov testi ile değerlendirildikten sonra Mann-Whitney U testleri kullanılarak elde edilmiştir.

KHOA ve KHOF olan kişi kayıtlarının yaş, cinsiyet, göğüs ağrısı türü, dinlenme hali EKG sonuçları, maksimum kalp atış hızı, egzersize bağlı anjin, dinlenmeye göre egzersizle indüklenen ST depresyonu, pik egzersiz ST segmentin eğimi, floroskopi ile renklendirilen büyük damar sayısı ve thal bağımsız değişkenlerinin ortalamaları karşılaştırıldığında istatiksel olarak yüksek düzeyde olarak anlamlı fark saptanmıştır ($p<0.01$).

Kişi kayıtlarından dinlenme hali kan basıncı, serum kolestrolü ve dinlenme hali EKG sonuçları bağımsız değişkenlerinin ortalamaları karşılaştırıldığında ise istatiksel olarak anlamlı fark saptanmıştır ($p<0.05$).

13 özellik arasında açlık kan şekeri için bulunan değerin (0.626) 0.05'ten büyük olmasından dolayı istatiksel olarak anlamlı bir fark saptanmamıştır.

Karışıklık matrisini oluşturan parametrelerden elde edilen performans kriterlerine göre, kullanılan sınıflandırıcı yöntemlerinin karşılaştırılması Tablo 4'de gösterilmiştir.

Doğru olarak tanımlanan "gerçek kalp hastalığı olasılığı fazla" olanların oranını ölçen özgüllük değeri, k-NN'de %82.8 ile en düşük oran yakalanırken; gradient boosting ve random forest ile %100'lük mükemmel bir başarı oranı elde edilmiştir

Destek vektör makinelerinin sınıflandırma performansını gösteren kriterlerden duyarlılık, Matthews korelasyon katsayısı, Fowlkes-Mallows indeks, F1 skoru ve doğruluk değerleri sırasıyla %43.8, 0.62, 0.42, %58.3 ve %67.2 bulunmuştur. Bu değerlere göre, DVM, çalışmanın en kötü tahminini yapan algoritmasıdır.

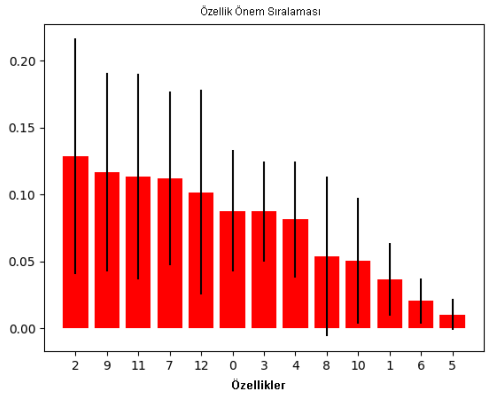
Random forest algoritmasının sınıflandırma performansına ait özgüllük, Matthews korelasyon katsayısı, Fowlkes-Mallows indeks, F1 skoru ve doğruluk değerleri sırasıyla %100, 0.90, 0.82, %89.7 ve %90.2'dir. Bu sonuçlara göre, RF, çalışmanın en iyi tahminini yapan algoritması olmuştur.

İşlem için sınıflandırıcılardan en az süre harcayan 0.008487 sn ile Gauss naive Bayes iken; en başarılı tahmin de bulunan Random forest ise sınıflandırma işlemini en fazla zaman harcayarak 0.147559 sn'de bitirmiştir.

Tablo 4. Sınıflandırıcı modellerin değerlendirme kriterlerine göre kıyaslanması.

Yöntemler	TP	FP	TN	FN	TPR	TNR	FM	MCC	F_1 Skor	ACC
Destek vektör makineleri	14	2	27	18	43.8	93.1	0.62	0.42	58.3	67.2
k-NN	19	5	24	13	59.4	82.8	0.69	0.43	67.9	70.5
Karar ağaçları	23	1	28	9	71.9	96.6	0.83	0.70	82.1	83.6
Lineer diskriminant	25	1	28	7	78.1	96.6	0.87	0.75	86.2	86.9
Gaussian naive Bayes	28	3	26	4	87.5	89.7	0.89	0.77	88.9	88.5
Gradient boosting	25	0	29	7	78.1	100	0.88	0.79	87.7	88.5
Random forest	26	0	29	6	81.3	100	0.90	0.82	89.7	90.2

Şekil 1, kalp hastalığı olasılık tahmini için kullanılan özelliklerin etkisini gösteren ağırlıkları ve önem sıralaması özetlemektedir.



Şekil 1. Kalp hastalık olasılık tahmini için kullanılan özelliklerin ağırlıkları ve önem sıralaması.

Bağımsız değişkenlerin, en başarılı tahmin yapan Random forest sınıflandırıcısına etkisi açısından önem sırası aşağıdaki gibi bulunmuştur.

- ✓ Göğüs ağrısı türü (f2)
- ✓ Dinlenmeye göre egzersizle indüklenen ST depresyonu (f9)
- ✓ Floroskopi ile renklendirilen büyük damar sayısı (f11)
- ✓ Maksimum kalp atış hızı (f7)
- ✓ Thal (f12)
- ✓ Yaş (f0)
- ✓ Dinlenme hali kan basıncı (f3)
- ✓ Serum kolestrolü (f4)
- ✓ Egzersize bağlı anjin (f8)
- ✓ Pik egzersiz ST segmentin eğimi (f10)
- ✓ Cinsiyet (f1)
- ✓ Dinlenme hali EKG sonuçları (f6)
- ✓ Açlık kan şekeri (f5)

En önemli üç özelliğin sınıflandırmaya etkisi sırayla %13.43 (göğüs ağrısı türü) , %13.39 (dinlenmeye göre egzersizle indüklenen ST depresyonu) ve %12.51 (floroskopi ile renklendirilen büyük damar sayısı) iken açlık kan şekeri %2.43 ile en az etkili değişken olduğu görülmüştür.

Tablo 3'e bakıldığında, Mann-Whitney U testine göre ortalama açlık kan şekeri değerleri için p değerinin 0.05'ten büyük olduğu bulunmuş olup iki grup arasında istatiksel olarak anlamlı bir fark bulunamamıştır. Aynı zamanda, açlık kan şeker değerinin değişkenler arasında en az etkili olmasından dolayı sınıflandırma algoritmaları, giriş olarak kullandıkları bağımsız değişkenlerden bir tanesini çıkartılarak (açlık kan şekeri olmadan) tekrar çalıştırılmıştır.

Sınıflandırma algoritmaların girişlerine açlık kan şekeri haricinde 12 bağımsız değişken uygulandığında karar ağaçları haricinde tüm sınıflandırıcıların performanslarında önemli bir değişiklik olmamıştır. Karar ağaçlarında ise Tablo 5'de gösterildiği üzere özgüllük haricinde sınıflandırma performansının geliştiği görülmüştür.

Tablo 5. 12 bağımsız değişken girişli karar ağaçlarının sınıflandırma performansı.

Yöntem	TPR	TNR	FM	MCC	F_1 Skor	ACC
Karar Ağ.	81.3	93.1	0.87	0.75	86.7	86.9

12 bağımsız değişken kullanılarak yapılan sınıflandırma işlemi için algoritmalar arasında en az zaman harcayan 0.005984 sn ile yine Gauss naive Bayes olmuştur. En başarılı tahminde bulunan Random forest ise sınıflandırma işlemini 0.144610 sn'de bitirmiştir. Açlık kan şekerini olmadan yapılan 12 girişli sınıflandırıcılar için işlemi bitirme sürelerinin kısaldığı görülmüştür.

Tablo 6, önceki çalışmalar ile mevcut çalışma arasındaki karşılaştırmayı göstermektedir. Bu çalışma ile elde edilen sonuçlar tabloda belirtilen önceki çalışmalara göre makul düzey doğrulukta kabul edilebilir.

Tablo 6. Önceki çalışmalarla mevcut çalışma arasında bir karşılaştırma.

Çalışmalar	Doğruluk	Duyarlılık	Özgüllük
Chen vd. [12]	≈80	≈85	≈70
Patel vd. [13]	56.8		
Jarad vd. [14]	52.3		
Enriko vd. [15]	81.9		
Shao vd. [16]	83.9		
Taşçı ve Şamlı [17]	88.5		
Tunç vd.[18]	81.4	80.4	82.3
Akalın vd. [19]	82.5		
Khan ve Mondal [20]	88.9		
Tougui vd. [21]	85.9	83.9	87.5
Bu çalışma	90.2	81.3	100

**Değerler % olarak verilmiştir.*

Random Forest, verinin alt kümesinde olabildiğince çok sayıda ağaç oluşturur ve çıktılarını birleştirir. Bunu yaparak karar ağaçlarındaki aşırı uyum problemi ve varyans azaltarak sınıflandırıcı doğruluğunu arttırmaktadır. Bu güçlü yönünü, çalışmada gösterdiği yüksek doğruluk performansı ile ispatlamıştır. Algoritmanın zayıf yönleri ise, sınıflandırma tahminini doğru yapmak için çok daha fazla hesaplama gücü ve kaynak gerektirmesi ve ayrıca eğitim için fazla zaman harcaması denilebilir. Çalışmada sınıflandırma işlemi için 0.147559 saniyelik süre ile en fazla zaman harcayan algoritma olması bunu doğrulamıştır.

Destek vektör makinelerinde eğitimin görece olarak daha kolay olması nedeniyle kısa sürede sınıflandırma işlemini bitirmesi avantajlı yönüdür. Zayıf yönü ise, çakışan hedef sınıflar olduğunda iyi bir performans göstermediği için çekirdek işlevi ve model parametre kombinasyonlarının test edilmesinin gerekliliğidir. Çalışmada, destek vektör makinelerinin sınıflandırmayı diğerlerine göre kısa süre içerisinde (0.031914 saniye) fakat kötü bir doğruluk oranıyla bitirmesi olumlu ve olumsuz yönlerini göstermektedir.

4. Sonuç

Bu çalışma, tıp doktorlarına kalp hastalığını tahmin etmede yardımcı olabilmeyi amaçlamıştır. 303 kişiye ait yaş, cinsiyet, göğüs ağrısı türü, dinlenme hali kan basıncı, serum kolestoral, açlık kan şekeri, dinlenme hali EKG sonuçları, maksimum kalp atış hızı, egzersize bağlı anjin, dinlenmeye göre egzersizle indüklenen ST depresyonu, pik egzersiz ST segmentin eğimi, floroskopi ile renklendirilen büyük damar sayısı, thal gibi demografik ve klinik veriler kullanılmıştır. En iyi kalp hastalık tahmini için 7 sınıflandırma algoritması karşılaştırılmıştır. Performans kriterlere göre kullanılan sınıflandırıcılar arasında en iyisi random forest olmuştur. 13 özellik arasında rastgele ormanın sınıflandırma performansını en az etkileyen özellik % 2,43 ile açlık kan şekeri olarak bulunmuştur. Bu çalışma, daha güçlü kalp hastalığı tahmini için daha fazla çabaya katkıda bulunacaktır.

Kaynakça

[1] World Health Organization, Cardiovascular Diseases, https://www.who.int/health-topics/cardiovascular-diseases/#tab=tab_1 (Erişim Tarihi: 12.07.2020).

[2] Anbarasi, M., Anupriya, E., Iyengar, N.C.H.S.N. 2010. Enhanced prediction of heart Disease with feature subset selection using genetic algorithm, International Journal of Engineering Science and Technology, Cilt. 2, s. 5370-5376.

[3] Palaniappan, S., Awang, R. 2008. Intelligent heart disease prediction system using data mining techniques, International Journal of Computer Science and Network Security, Cilt. 8, s. 343-350.

[4] Nahar, J., Imam, T., Tickle, K.S., Chen,Y.-P.P. 2013. Computational intelligence for heart disease diagnosis: A medical knowledge driven approach, Expert Systems with Applications, Cilt. 40, s. 96-104. DOI: 10.1016/j.eswa.2012.07.032.

[5] Soman, K.P., Shyam, D.M., Madhavdas, P. 2003. Efficient classification and analysis of ischemic heart disease using proximal support vector machines based decision trees. Conference on convergent technologies for AsiaPacific region, 15-17 Oct., Bangalore, India, 214-217. DOI: 10.1109/TENCON.2003.1273317

[6] Kim, B.-H., Lee, S.-H., Cho, D.-U., Oh, S.-Y. 2008. A proposal of heart diseases diagnosis method using analysis of face color. International conference on advanced language processing and web information technology, 23-25 July, Dalian Liaoning, China, 220-225. DOI: 10.1109/ALPIT.2008.27

[7] Gamboa, A.L.G., Mendoza, M.G., Orozco, R.E.I., VARGAS, J.M., Gress, N.H. 2006. Hybrid Fuzzy-SV clustering for heart disease identification, Computational intelligence for modelling International conference on control and automation, and international conference on intelligent agents, web technologies and internet commerce, 28 Nov.-1 Dec., Sydney, NSW, Australia. DOI: 10.1109/CIMCA.2006.114

[8] Zheng, J., Jiang, Y., Yan, H. 2006. Committee machines with ensembles of multilayer perceptron for the support of diagnosis of heart diseases. Proceedings of the international conference on, communications, circuits and systems, 25-28 June, Guilin, China, 2046-2050. DOI: 10.1109/ICCCAS.2006.285080

[9] Obayya, M., Abou-chadi, F. 2008. Data fusion for heart diseases classification using multi-layer feed forward neural network. International conference on computer engineering&systems, ICCES, 25-27 Nov., Cairo, Egypt, 67-70. DOI: 10.1109/ICCES.2008.4772968

[10] Avci, E. 2009. A new intelligent diagnosis system for the heart valve diseases by using genetic-SVM classifier, Expert Systems with Applications, Cilt. 36, s. 10618–10626. DOI: 10.1016/j.eswa.2009.02.053

[11] Maglogiannis, I., Loukis, E., Zafiropoulos, E., Stasis, S. 2009. Support vectors machine-based identification of heart valve diseases using heart sounds, Computer Methods and Programs in Biomedicine, Cilt. 95, s. 47–61. DOI: 10.1016/j.cmpb.2009.01.003

[12] Chen, A.H., Huang, S.Y., Hong, P.S., Cheng, C.H., Lin, E.J. 2011. HDPS: Heart Disease Prediction System, Computing in Cardiology, Cilt. 38, s. 557-560.

[13] Patel, J., Tejalupadhyay, S., Patel, S.B. 2016. Heart Disease Prediction Using Machine learning and Data Mining Technique, International Journal of Computer Science & Communication, Cilt. 7, s. 129-137. DOI: 10.090592/IJCSC.2016.018

[14] Jarad, A., Katkar, R., Shaikh, A.R., Salve, A. 2015. Intelligent Heart Disease Prediction System with MongoDB, International Journal of Emerging Trends & Technology in Computer Science, Cilt. 4, s. 236-239.

[15] Enriko, I.K.A., Suryanegara, M., Gunawan, D. 2016. Heart Disease Prediction System using k-Nearest Neighbor Algorithm with Simplified Patient's Health Parameters, Journal of Telecommunication, Electronic and Computer Engineering (JTEC), Cilt. 8, s. 59-65.

[16] Shao, Y.E., Hou, C.-D., Chiu, C.-C. 2014. Hybrid intelligent modeling schemes for heart disease classification, Applied Soft Computing, Cilt. 14, s. 47-52. DOI: 10.1016/j.asoc.2013.09.020

[17] Taşçı, M.E., Şamlı, R. 2020. Veri Madenciliği İle Kalp Hastalığı Teşhisi, Avrupa Bilim ve Teknoloji Dergisi, Özel sayi:88-95. DOI: 10.31590/ejosat.araconf12

[18] Tunç, Z., Cicek, I.B., Guldogan E. 2020. Performance evaluation of the deep learning models in the classification of heart attack and determination of related factors, The Journal of Cognitive Systems, Cilt. 5 s. 99-103. https://dergipark.org.tr/tr/pub/jcs/issue/59409/801555

[19] Akalın, B., Veranyurt, Ü., Veranyurt, O. 2020. Classification of individuals at risk of heart disease using machine learning, Cumhuriyet Medical Journal, Cilt 42, s. 283-289. DOI: 10.7197/cmj.vi.742161

[20] Khan, I.H., Mondal, R.H.M. 2020. Data-Driven Diagnosis of Heart Disease, International Journal of Computer Applications, Cilt 176, s. 46-54. DOI: 10.5120/ijca2020920549

[21] Tougui, I., Jilbab, A., El Mhamdi, J. 2020.Heart disease classification using data mining tools and machine learning techniques, Health Technol., Cilt.10, s.1137–1144. DOI: 10.1007/s12553-020-00438-1

[22] Datasets. "Health care: Data set on heart attack possibility". https://www.kaggle.com/nareshbhat/health-care-data-set-on-heart-attack possibility?select=heart.csv (Erişim Tarihi: 10.07.2020).

[23] Aburomma, A.A., Reaz M.B.I. 2017. A novel weighted support vector machines multiclass classifier based on differential evolution for intrusion detection systems, Information Sciences, Cilt. 414, s. 225-246. DOI: 10.1016/j.ins.2017.06.007

[24] Kausar, N., Samir, B.B., Sulaiman, S.B., Ahmad, I., Hussain, M. 2012. An approach towards intrusion detection using pca feature subsets and svm. 2012 International Conference on Computer & Information Science (ICCIS), 12-14 June, Kuala Lumpeu, Malaysia, 569-574. DOI: 10.1109/ICCISci.2012.6297095

[25] Taharwat, A., Mahdi, H., Elhoseny, M., Hassanien, A.E. 2018. Recognizing human activity in mobile crowd sensing environment using optimized k-NN algorithm, Expert Systems With Applications, Cilt. 107, s. 32-44. DOI: 10.1016/j.eswa.2018.04.017

[26] Venkatasubramaniam, A., Wolfson, J., Mitchel, N., Barnes, T., JaKa, M., French, S. 2017. Decision trees in epidemiological research, Emerg Themes Epidemiol., Cilt. 14, s. 1-12. DOI: 10.1186/s12982-017-0064-4

[27] Sifaou, H., Kammoun, A., Alouini, M.-S. 2020. High-dimensional Linear Discriminant Analysis Classifier for Spiked Covariance Model, Journal of Machine Learning Research, Cilt. 21, s. 1-24.

[28] Zhang, Z., Wang, S., Bian, W. 2020. Sign consistency for the linear programming discriminant rule, Pattern recognition, Cilt. 100, s. 1-11. DOI: 10.1016/j.patcog.2019.107083

[29] Wu, X., Kumar, V., Quinlan, J.R., Ghosh, J., Yang, Q., et al. 2008. Top 10 algorithms in data mining, Knowledge and information systems, Cilt. 14, s. 1–37. DOI: 10.1007/s10115-007-0114-2

[30] Ng, A.Y., Jordan, M.I., 2002. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes, Advances in neural information processing systems, Cilt. 2,s. 841–848.

[31] Jahromi, A.H., Taheri, M. 2017. A non-parametric mixture of Gaussian naive Bayes classifiers based on local independent features. 2017 Artificial Intelligence and Signal Processing Conference (AISP), 25-27 Oct. Shiraz, Iran, 209-212. DOI: 10.1109/AISP.2017.8324083

[32] Lee, K., Kwan, M.-P.. 2018. Physical activity classification in free-living conditions using smartphone accelerometer data and exploration of predicted results, Computers, Environment and Urban Systems, Cilt. 67, s. 124-131. DOI: 10.1016/j.compenvurbsys.2017.09.012

[33] Friedman, J.H. 2001. Greedy Function Approximation: A Gradient Boosting Machine, The Annals of Statistics, Cilt. 29, s. 1189–1232. DOI: 10.1214/aos/1013203451

[34] Cutler, D.R., Edwards, T.C., Beard, K.H., Cutler, A., Hess, K.T., et al. 2007. Random forests for classification in ecology, Ecology, Cilt. 88, s. 2783–2792. DOI: 10.1890/07-0539.1

[35] Fu, Y. 2017. Combination of Random Forests and Neural Networks in Social Lending, Journal of Financial Risk Management, Cilt. 6, s. 418-426. DOI: 10.4236/jfrm.2017.64030