Avrupa Bilim ve Teknoloji Dergisi
Özel Sayı, S. 88-95, Nisan 2020

Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
Special Issue, pp. 88-95, April 2020

Research Article

Veri Madenciliği İle Kalp Hastalığı Teşhisi*,**

Merve Esra Taşcı[1**], Rüya Şamlı[2]

[1] Istanbul Üniversitesi – Cerrahpaşa, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, İstanbul, Türkiye (ORCID: 0000-0001-7659-0455)
[2] Istanbul Üniversitesi – Cerrahpaşa, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, İstanbul, Türkiye (ORCID: 0000-0002-8723-1228)

(Konferans Tarihi: 5-7 Mart 2020)

(**DOI:** 10.31590/ejosat.araconf12)

ATIF/REFERENCE: Taşçı, M. E. & Şamlı, R. (2020). Veri Madenciliği İle Kalp Hastalığı Teşhisi. *Avrupa Bilim ve Teknoloji Dergisi*, (Özel Sayı), 88-95.

Öz

Gelişen ve değişen çevre koşulları, sınırların kalkması ile küreselleşen dünya, farklı pazarlama ve ar-ge (araştırma geliştirme) yöntemleri "veri"nin değil "bilgi"nin önemini her geçen gün daha da artacak şekilde ortaya koymaktadır. İnternetin yaygınlaşması ve kolaylaşması ar-ge ekiplerinin "bilgi"ye erişmelerini zorlaştırmaktadır. İnternette arama motorları kullanılarak yapılan araştırmalar çoğu zaman istenilenden farklı bir şekilde sonuçlanmaktadır. Büyük bir perakendecinin, fatura bilgilerinden müşteri eğilimlerini belirleyip ona göre pazarlama taktikleri üretebilmesi, rakiplerinin önüne geçmesini sağlayacaktır. Verilen örneklere dikkat edilirse, "veri"nin "bilgi"ye dönüşme işleminin vurgulandığı görülecektir. Veri madenciliği, büyük ölçekli veriler arasından bilgiye ulaşma, bilgiyi madenleme işidir. Ya da bir anlamda büyük veri yığınları içerisinden gelecek ile ilgili tahminde bulunabilmemizi sağlayabilecek bağıntıların bilgisayar programı kullanarak aranmasıdır. Veri madenciliği, eldeki verilerden üstü kapalı, çok net olmayan, önceden bilinmeyen ancak potansiyel olarak kullanışlı bilginin çıkarılmasıdır. Veri madenciliği sürecinin çeşitli aşamalarında; istatistiksel yöntemler, bellek tabanlı yöntemler, genetik algoritmalar, yapay sinir ağları ve karar ağaçları gibi algoritmalar kullanılabilmektedir.

Kalp hastalıkları (kardiyovasküler hastalıklar) bugün dünyanın en yaygın hastalıklarından biridir. Küresel ölçekte kardiyovasküler hastalıkların uzun süre bir numaralı ölüm nedeni olmaya devam edeceği tahmin edilmektedir. Gelişmiş batı ülkelerindeki kardiyovasküler hastalık ölümleri gelişmekte olan ülkelere göre azalma eğilimi göstermektedir. Dünyadaki ölüm oranlarındaki pozitif faktör, kardiyovasküler hastalıklar açısından büyük ölçüde önlenebilir olmalarıdır. Bu nedenle, kalp hastalığı tanısı alan hastaların verilerine dayanarak, metin madenciliği ve algoritmalar kullanılarak kalp-öncesi hastalığı tahmin etmek için çalışma yapılmıştır. Bu çalışma veri madenciliğinin büyük veri setlerinin incelenmesi üzerinde ne kadar büyük bir öneme ve yere sahip olduğunu göstermek için yapılmıştır. Yüzlerce bilginin bulunduğu kalp veri setinden, WEKA programı kullanılarak, çeşitli algoritmalar uygulayarak kişilere kalp hastalığı teşhisi koyma çalışması yapılmıştır.

Kalp hastalığının kesin tanısı ve hastalık şiddetinin saptanması için çeşitli uygulamalar ve yöntemler vardır. Bu çalışmada, daha ucuz ve daha etkili bir yaklaşım sağlayabilecek veri madenciliğinin kullanımı incelenmiştir. Bu çalışmada sınıflandırma yöntemleri ve doğru sınıflandırma oranları ile elde edilen sonuçlar karşılaştırılmıştır. Gerekli hesaplamaları ve modelleri elde etmek için ZeroR, OneR, Naive Bayes, J48 Karar Ağacı, Rastgele Orman, Çoklu Algılayıcılar, k-en yakın komşu (k-Nearest Neighbour - k-NN), Lojistik Regresyon, destek vektör makinesi (Support Vector Machine - SVM) gibi sınıflandırma algoritmaları, Weka programında uygulanmıştır. Uygulamanın sonucu olarak kalp hastalığı teşhisinde en iyi sonuçu veren algoritma tespit edilmeye çalışılmıştır. Veri madenciliği algoritmaları ile kalp hastalığını belirleyen birçok farklı çalışma vardır. Ancak yaptığımız incelemelerde, veri setine 9 farklı algoritma uygulayan bir çalışmaya rastlanmamıştır ve bu çerçevede bu çalışma ilk kez bu kadar çok algoritmanın kullanıldığı çalışma olacaktır.

Anahtar Kelimeler: Kalp Hastalığı, WEKA, Veri Madenciliği, ZeroR, OneR, Naive Bayes, J48 Karar Ağacı, Rastgele Orman, Multiplayer Perceptrons, k-en yakın komşuluk, Logistik Regresyon, Destek Vektör Makinesi

* Bu makale ***International Conference on Access to Recent Advances in Engineering and Digitalization (ARACONF 2020)*** de sunulmuştur.

Abstract

Developing and changing environmental conditions, the globalization of the borders and the globalization of the world, different marketing and R&D (research and development) methods reveal the importance of "information" rather than "data". The widespread and easing of the Internet makes it difficult for R&D teams to access “information”. Research on the internet using search engines often results in a different way than desired. The ability of a large retailer to identify customer trends from the invoice information and produce marketing tactics accordingly will prevent them from getting ahead of their competitors. If attention is given to the examples given, it will be seen that the process of turning “data” into “information” is emphasized.

Data mining is the business of accessing and mining information among large-scale data. Or, in a sense, it is the search for the relations that can enable us to make predictions about the future from large data stacks using a computer program. Data mining is the extraction of implicit, unclear, previously unknown but potentially useful information from the available data. At various stages of the data mining process; Algorithms such as statistical methods, memory-based methods, genetic algorithms, neural networks and decision trees can be used. Heart diseases (cardiovascular diseases) are one of the most common diseases in the world today. It is estimated that cardiovascular diseases will continue to be the number one cause of death for a long time on a global scale. Cardiovascular disease deaths in developed western countries show a decreasing tendency in developing countries. The positive factor in mortality rates in the world is that they are largely preventable in terms of cardiovascular diseases. Therefore, based on the data of patients diagnosed with heart disease, the study was carried out to predict pre-cardiac disease by using text mining and algorithms. This study was conducted to show how much importance and place data mining has on the study of big data sets. From the heart data set containing hundreds of information, by using WEKA program, by applying various algorithms, the study was made to diagnose people with heart disease.

There are various applications and methods for the definitive diagnosis of heart disease and detection of disease severity. In this study, the use of data mining, which could provide a cheaper and more effective approach, was studied. In this study, the results obtained by classification methods and correct classification rates were compared. In order to obtain the necessary calculations and models, classification algorithms such as ZeroR, OneR, Naive Bayes, J48 Decision Tree, Random Forest, Multiplayer Perceptrons, k-nearest neighbor (k-NN), Logistic Regression, support vector machine (SVM), have been applied in Weka packet program. As a result of the application of the best results in the determination of heart disease algorithm has been tried to be determined. There are many different studies that determine heart disease by data mining algorithms. But there is no study that implements 9 different algorithms to the data set and this paper will be the first one.

Keywords: Heart diesease, WEKA, Data Mining, ZeroR, OneR, Naive Bayes, J48 Decision Tree, Random Forest, Multiplayer Perceptrons, k-nearest neighbor, Logistic Regression, support vector machine

1. Giriş

Temel bir tanım olarak, “Veri Madenciliği, veriler hakkında örtük, daha önce bilinmeyen ve potansiyel yararlı bilgilerin önemsiz bir şekilde çıkarılmasıdır”. Başka bir deyişle, veri madenciliği, verileri farklı bir perspektiften analiz etme ve bilgiyi ondan toplama sürecidir.Veri madenciliğinin ilk adımında öncelikle ulaşılmak istenen bilgiye karar verilir. Bu konuda net ve planlı olmak sonuca ulaşmada faydalı olacaktır. Hedefin belirlenmesinin ardından bilgi bir sınıflandırmaya yerleştirilir ve bilginin işleneceği en uygun veri tabanları seçilir. Araştırmanın yapıldığı veri tabanı içerisinde hedef bilgiden uzaklaştıracak olan alakasız verilerin temizlenmesi ve ayrıştırılması gerekir. Ardından hedefe en uygun olarak seçilmiş tabanda toplanan en doğru veriler örüntülü şekilde farklı metotlarla işlenir. Ortaya çıkan sağlıklı, işlenmiş ve alakalı veriler amacına uygun kullanılmak üzere hazırdır. Bu hayli büyük, gizli ve ham veri gruplarını birer faydalı bilgiye dönüştürme ve sonrasında analizlerini yapma işlemi, otomatik olarak programlar ve teknikler ile yapılmakta ve teknolojinin de verdiği yetkiye dayanarak pek çok alana fayda sağlamaktadır. Veri mdenciliğinin pazarlama, bankacılık, e-ticaret, telekominükasyon, tıbbi aratırmalar, taşımacılık ve ulaşım, turizm ve otelcilik, eğitim, bilim ve mühendislik gibi uygulama alanları bulunmaktadır.

Veri madenciliği teknikleri, bilinmeyen kalıpları keşfetmek için karmaşık algoritmalar kullanarak verileri analiz etmek ve ayıklamak için kullanılır. Araştırmacılar kalp hastalığı, diyabet ve kanser gibi birçok hastalığın teşhisi için veri madenciliği tekniklerini kullanmakta ve bu sayede sağlık hizmetlerinde veri madenciliği her geçen gün daha popüler hale gelmektedir. Birçok resmi araştırmaya göre, kalp hastalıkları dünya çapında bir numaralı ölüm nedenidir. Her yıl kalp hastalıklarından ötürü diğer nedenlerden daha fazla insan ölmektedir. Son araştırmalar kalp hastalığının risk faktörlerini tanımlayabilmiştir, ancak birçok araştırmacı bu bilgiyi kalp hastalıkları oluşumunu azaltmak için kullanmak için daha fazla araştırmaya ihtiyaç olduğu konusunda hemfikirdir. Kalp hastalıklarını analiz ederken hastaların diyabet, hipertansiyon, yüksek kan kolesterolü, yeme alışkanlıkları, fiziksel hareketsizlik, sigara içme, alkol alımı, obezite, yaş, cinsiyet gibi aynı özellikleri ele alır. Kalp hastalıkları farklı özelliklere bağlı olabilir. Bazı literatür çalışmaları kalp hastalığı için bu risk faktörlerinin azaltılmasının aslında kalp hastalıklarının önlenmesine yardımcı olabileceğini göstermiştir. Kalp hastalığı riskinin önlenmesi konusunda birçok çalışma ve araştırma vardır. Kalp hastalıkları hakkında daha fazla çalışma, kalp hastalıklarını önlemek için daha fazla fırsat sunacaktır.

Bu çalışma, kalp hastalıklarının veri madenciliği algoritmaları ile teşhis edilmesini amaçlamıştır: ZeroR, OneR, Naive Bayes, J48 Karar Ağacı, Rastgele Orman, Multiple Perceptrons, k-en yakın komşu, Lojistik Regresyon, Destek Vektör Makinesi. Bu teknikler, Kaggle veri tabanından alınan 13 özniteliği olan 303 olgunun veri kümesine uygulanmıştır. Kalp hastalığı veri setinde bu birçok tekniği uygulayacak hiçbir çalışmanın bulunmadığı belirtilmelidir. Doğruluk oranlarına göre, KNN'nin kalp hastalığı tanısında en başarılı yöntem olduğu görülmektedir.

Bu çalışmanın organizasyonu aşağıdaki gibidir: Bölüm 2 kullanılan veri setini ve kullanılan verinin özelliklerini, kullanılan veri madenciliği algoritmalarını ve çalışma mantıklarını açıklar. Bölüm 3, algoritmalar hakkındaki sonuçları tablolarla birlikte sunar ve Bölüm 4 kullanılan algoritmaların sonuçları birbirleriyle karşılaştırılarak en iyi sonucu veren algoritmanın hangisi olduğu açıklar.

2. Materyal ve Metot

2.1. Veri Seti

Bu çalışmada Kaggle veri tabanından 85 000 vakadan oluşan bir veri seti seçilerek kullanılmıştır. Bu veri seti hem hasta hem de sağlıklı vakalardan oluşur. Değerler bazı optimizasyon yöntemlerine göre elimine edilmiştir ve sınıflandırma için 13 özelliğe sahip toplam 303 vaka kullanılmıştır. Bu özellikler yaş, cinsiyet, göğüs ağrısı tipi, istirahat kanbasıncı, serum kolestoral, açlık kan şekeri, istirahat elektrokardiyografik sonuçlar, elde edilen maksimum kalp hızı, egzersize bağlı anjin, istirahate göre egzersize bağlı ST depresyonu, ST segmentinin eğimi, pik egzersiz için floroskopi ile renklendirilmiş büyük damar sayısı, defekt tipidir. Bu verilere toplam 9 farklı veri madenciliği yöntemi uygulanmıştır. Öznitelikler aşağıdaki ayrıntılı şekilde açıklanmıştır:

1) yaş (numerik)

2) cinsiyet (0, 1)
   0: dişi
   1: erkek

3) cp (0,1,2,3): göğüs ağrısı tipi
   0: tipik anjin ağrı
   1: atipik anjin ağrı
   2: anjin olmayan ağrı
   3: asemptomatik ağrı

4) trestbps (numerik): istirahat kanbasıncı

5) chol (numerik): serum kolestoral

6) fbs (0, 1): açlık kan şekeri
   0: yanlış
   1: doğru

7) restecg (0, 1, 2): istirahat elektrokardiyografik sonuçlar
   0: normal
   1: ST-T anormalliği
   2: olası veya belirgin sol ventriküler Estes kriteri
8) thalach (numerik): elde edilen maksimum kalp hızı

9) exang (0, 1): egzersize bağlı anjin
   0: yok
   1: var
10) oldpeak (numerik): istirahate göre egzersize bağlı ST depresyonu
11) slope (0,1,2): ST segmentinin eğimi
   0: yukarı doğru
   1: düz
   2: aşağı doğru
12) ca (numeric - 0,1,2,3,4): pik egzersiz için floroskopi ile renklendirilmiş büyük damar sayısı

13) thal (3,6,7): defekt tipi
   3: normal
   6: belirlenmiş defekt
   7: tersinir defekt

2.2.Yöntemler

2.2.1.ZeroR Algoritması

ZeroR algoritması, veri madenciliğinde en basit ve en temel sınıflandırma algoritması olarak kabul edilir. Önceden en fazla veriye sahip olan sınıftaki tüm verileri kabul eder. Bunu yaparken, gelen verinin sıklığına bakarak verileri tahmin eder. ZeroR algoritması sınıflandırma algoritmalarına dahil edilir, ancak sınıflandırma yerine taban performansını belirlemek için kullanılır. Çünkü zeka içermeyen bir algoritmadır. Taban performansı belirleyerek eğitim verilerinin tahmin sonuçlarında çıkacak oranı görülebilir. Mantığı oldukça basittir. Elimizde bulunan eğitim verilerinde çıkan sonuçların bir biri arasındaki oranına bakar ve en çok olan orandaki sonuç, bundan sonraki verilerde tahmin sonucu olarak kullanılır.

2.2.2.OneR Algoritması

ZeroR algoritmasının genişletilmiş bir versiyonudur ve genellikle ZeroR'dan daha başarılı sonuçlar verir. Algoritma, eğitim verilerindeki tüm sınıflar arasında mümkün olan en iyi sonucu döndürecek olan sınıfı seçer.

2.2.3.Naive Bayes Algoritması

Naive Bayes sınıflandırma algoritması, adını Matematikçi Thomas Bayes'den alan bir sınıflandırma/ kategorilendirme algoritmasıdır. Naive Bayes sınıflandırması olasılık ilkelerine göre tanımlanmış bir dizi hesaplama ile, sisteme sunulan verilerin sınıfını yani kategorisini tespit etmeyi amaçlar.

Naive Bayes sınıflandırmasında sisteme belirli bir oranda öğretilmiş veri sunulur. Öğretim için sunulan verilerin mutlaka bir sınıfı/kategorisi bulunmalıdır. Öğretilmiş veriler üzerinde yapılan olasılık işlemleri ile, sisteme sunulan yeni test verileri, daha önce elde edilmiş olasılık değerlerine göre işletilir ve verilen test verisinin hangi kategoride olduğu tespit edilmeye çalışılır. Elbette öğretilmiş veri sayısı ne kadar çok ise, test verisinin gerçek kategorisini tespit etmek o kadar kesin olabilmektedir.

Naive Bayes sınıflandırma yönteminin birçok kullanım alanı bulunabilir fakat, burada neyin sınıflandırıldığından çok nasıl sınıflandırıldığı önemlidir. Yani öğretilecek veriler ikili veya metin türünde veriler olabilir, burada veri tipinden ve ne olduğundan ziyade, bu veriler arasında nasıl bir oransal ilişki kurulduğu önem kazanmaktadır. Yöntemin matematiksel olarak gösterimi aşağıdaki gibidir:

$$P(C_i|X) = \frac{P(X|C_i) * P(C_i)}{P(X)}$$

2.2.4. J48 Karar Ağacı Algoritması

J48 algoritması değişkenin entropi değerini ve Shannon'un Bilgi Kuramı'ndan (Information Theory) yararlanarak karar ağacını optimize etmeyi amaçlar. Entropi, rastgele bir değişkenin belirsizliğinin ölçüsüdür. Bilgi kazanımı, bir tahmin edici değişken kullanılarak veri bölümlendiğinde, hedef değişkendeki belirsizliğin ne kadar değiştiğinin bir ölçüsüdür.

Bu algoritma önce her hedef değişken/sınıf için entropi değerini ve her tahmin değişkeni/sınıf için bilgi değerini hesaplar. Bundan sonra, her tahmin değişkeni/sınıfı için bilgi kazancını hesaplar. Bu hesaplamaların amacı, en fazla bilgi toplayan sınıfı belirlemektir.

En yüksek bilgi kazanımı sağlayan tahmin edici değişken tespit edilir ve ağaç bu değişkenden itibaren dallandırılmaya başlanır. Böylece her bir dalın altında veriler dengeli bir biçimde dağılacaktır. İlk tahmin edici değişken tespit edildikten sonra aynı işlem bu defa toplam entropi üzerinden değil, bu belirlenen tahmin edici değişkenin bilgi değeri üzerinden tekrarlanarak geriye kalan tahmin edici değişkenlerden hangisiyle bu belirlenen değişkenin bölümlenmesinin daha fazla bilgi kazanımı sağlayacağı hesaplanır. Bu işlem tüm tahmin edici değişkenler ağaca yerleştirilinceye kadar devam eder. J48 algoritmasında dallanmalar belirlenirken bilgi kazancı, kazanç oranı ya da Gini Indeks baz alınabilir.

2.2.5.Rastgele Orman Algoritması

Random Forest, rastgele bir orman oluşturan denetimli bir öğrenme algoritmasıdır. "Orman" genellikle torbalama yöntemi ile eğitilmiş bir ağaç koleksiyonudur. Literatürden bilindiği gibi torbalamanın genel fikri şudur: "öğrenme yöntemlerinin kombinasyonu" doğruluğu arttırmaktadır. Dolayısıyla rastgele ormanın sınıflandırma başarısını artırmak için birden fazla karar ağacını birleştirdiği söylenebilir.

2.2.6. Çoklu Algılayıcılar Algoritması

Multiple Perceptrons'un temeli yapay sinir ağlarına dayanır. Yapay sinir ağları çalışmaları, bilgisayar mühendisliğinin bir alt disiplini olan yapay zekanın çalışma alanına girmektedir. Bu disiplinde insan beyni ve beyinde bulunan nöronların çalışmalarını taklit etmek hedeflenmektedir.

Çok katmanlı perceptrons (Multilayer Perceptron), perceptrons olarak adlandırılan basit sinir hücreleri ağıdır. Temel fikir olan tek perceptron (single perceptron) ilk olarak 1958 yılında Rosenblatt tarafından tanıtılmıştır. Perceptron birçok değeri girdi olarak alıp tek bir çıktı üretir. Çıktıyı oluştururken girdi ağırlıklarına uygun olarak bir lineer kombinasyon oluşturulur ve bazı lineer olmayan etkinleştirme fonksiyonu yoluyla çıktılara yerleştirilir.

2.2.7. k-NN Algoritması

k-NN, tüm vakaları saklayan ve yeni vakaları benzerlik ölçüsüne göre sınıflandıran bir sınıflandırma algoritmasıdır. Algoritma ayrıca vaka temelli akıl yürütme, örnek tabanlı akıl yürütme, örnek tabanlı öğrenme, bellek tabanlı akıl yürütme ve tembel öğrenme olarak da adlandırılır. k-NN algoritması, istatistiksel tahmin ve örüntü tanıma gibi birçok uygulamada 1970'den beri kullanılmaktadır. Gözetimli öğrenme metotlarından sınıflandırma (classification) işleminde kullanılır. Sınıflandırma işleminde k değerine bakılarak

eleman sayısını belirler. Her ne kadar k-NN algoritması k-ortalamalar algoritmasındaki benzer özellikleri taşısa da büyük farklılıklar da içermektedir.
k-NN algoritması bir eğitim verisi içerirken k-means algoritması bir eğitim verisi içermez. Yeni bir değer geldiğinde k değerine mesafeler hesaplanır ve yeni değer bir kümeye ilave edilir. Eğitim kümesinin büyük olması ve k değerini uygun seçilmesi k-NN açısından çok önemlidir. k-NN algoritmasına karar verildiğinde eğitim verisi ve k değerini arttırarak başarıya bakılır ve başarı sabitlenmeye başladığında iyi bir k-NN tahmin sistemi yapılmıştır denilebilir. k-NN algoritmasının çalışma prensibi, nesnelerin birbirleri arasında yakınlık ilişkilerine göre kümeleme işlemi yapmasına dayanır.

2.2.8.Lojistik Regresyon Algoritması

Bağımlı değişkenin kategorik bir değişken olduğu bir regresyon problemidir. Lineer sınıflandırma problemlerinde çok kullanılır. Regresyon olarak adlandırılsa da bir sınıflandırma algoritmasıdır. Lojistik regresyon, bir veya daha fazla bağımsız değişken içeren bir veri kümesini analiz etmek için kullanılan istatistiksel bir yöntemdir. Sonuçlar bir ikili değişken ile ölçülür. Başka bir deyişle, bağımlı değişken sadece 1/0, Doğru/Yanlış, Erkek/Dişi ve benzeri iki değerle tanımlanır. Bu yöntemin amacı, iki bağımlı değişken arasında en uygun modeli bulmaktır.

2.2.9. Destek Vektör Makinesi (Support Vector Machine - SVM) Algoritması

SVM, sınıflandırma işlevini tahmin eden bir tür sınıflandırma yöntemidir ve sınıflandırma için kullanılan verileri analiz eden ve örüntüleri tanıyan bir dizi denetimli öğrenme yöntemidir. Genellikle diğer yöntemlere kıyasla daha iyi sınıflandırma sonuçları ürettiği bildirilen doğrusal olmayan bir sınıflandırma yöntemidir. SVM'nin ana fikri, pozitif ve negatif örnekler arasındaki ayrım marjını en üst düzeye çıkaracak şekilde bir karar yüzeyi olarak hiper bir düzlem oluşturmaktır. Bu işlem, girdi örneği verilerini, verilerin doğrusal olarak ayrılabildiği, böylece daha yüksek sınıflandırma (veya regresyon) doğruluğu sağlayabilen yüksek boyutlu bir alana doğrusal olmayan bir şekilde eşler. SVM'ler, gerçek dünya uygulamalarında, özellikle biyoinformatikte, hem sağlam bir teorik temeli hem de en son başarıyı elde etmeleri bakımından oldukça ilginçtir.

3. Araştırma Sonuçları ve Tartışma

Bu makalede, veri seti kalp hastalığı veri tabanındaki toplam 303 kayıttan oluşmaktadır. Veri madenciliği aracı Weka deney için kullanılır. Tablo 1 algoritmaların performansını göstermektedir.

Tablo 1: Algoritmaların Performansları

Algoritma	Doğruluk Oranı	Geçen Süre (saniye)
ZeroR	%49.1803	0
OneR	%73.7705	0,02
Naive Bayes	%88.5246	0,03
Rastgele Orman	%83.6006	0,3
J48 Karar Ağacı	%78.6885	0,06
Lojistik Regresyon	%85.2459	0,33
k-NN	%81.9672	0
Çoklu Algılayıcılar	%86.8852	1,74
SVM	%86.8852	0,09

Sınıflandırmanın doğruluğunu hesaplamak için bir karmaşıklık (confusion) matrisi elde edilir. Bir karmaşıklık matrisi (Tablo 2) her sınıfa kaç örnek atandığını gösterir.

Tablo 2: Karmaşıklık Matrisi

	a (kalp hastalığı yok)	b (kalp hastalığı var)
a (kalp hastalığı yok)	TP	FN
b (kalp hastalığı var)	FP	TN

TP (True Positive – Doğru Pozitif): Gerçekte doğruyken doğru olarak sınıflandırılan kayıt sayısı
FN (False Negative – Yanlış Negatif): Gerçekte doğruyken yanlış olarak sınıflandırılan kayıt sayısı
FP (False Positive – Yanlış Pozitif): Gerçekte yanlışken doğru olarak sınıflandırılan kayıt sayısı
TN (True Negative – Doğru Negatif): Gerçekte yanlışken yanlış olarak sınıflandırılan kayıt sayısı

Algoritmaların TP, FN, FP ve TN değerleri Tablo 3-11 aşağıda gösterilmiştir:

Tablo 3: ZeroR Algoritması Karmaşıklık Matrisi

	a	b
a	0	31
b	0	30

Tablo 4: OneR algoritması Karmaşıklık Matrisi

	a	b
a	24	7
b	9	21

Tablo 5:Naive Bayes Algoritması Karmaşıklık Matrisi

	a	b
a	25	6
b	1	29

Tablo 6: Rastgele Orman Algoritması Karmaşıklık Matrisi

	a	b
a	24	7
b	3	27

Tablo 7: J48 Karar Ağacı Algoritması Karmaşıklık Matrisi

	a	b
a	23	8
b	5	25

Tablo 8: Lojistik Regresyon Algoritması Karmaşıklık Matrisi

	a	b
a	26	5
b	4	26

Tablo 19: k-NN algoritması Karmaşıklık Matrisi

	a	b
a	24	7
b	4	26

Tablo 10: Çoklu Algılayıcılar Algoritması Karmaşıklık Matrisi

	a	b
a	26	5
b	3	27

Tablo 11: Destek Vektör Makinesi Algoritması Karmaşıklık Matrisi

	a	b
a	25	6
b	2	28

Algoritmaların doğruluk oranları Tablo 12'de formülleri verilmiş olan, Ortalama Mutlak Hata, Kök Ortalama Kare Hata, Bağıl Mutlak Hata, Kök Bağıl Kare Hata gibi bazı kriterlerle verilmiştir.

Tablo 12: Kriter Formülasyonları

Ortalama Mutlak Hata (Mean Absolute Error – MAE)	$MAE = \frac{1}{n}\sum_{1}^{n}\lvert y_i - \hat{y_i}\rvert$
Kök Ortalama Kare Hata (Root Mean Squarred Error – RMSE)	$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y_i})^2}{n}}$
Bağıl Mutlak Hata (Relative Absolute Error – RAE)	$RAE = \sum_{i=1}^{n}\frac{\lvert e_i\rvert}{y_i - \hat{y_i}}$
Kök Bağıl Kare Hata (Root Relative Squarred Error – RRSE)	$RRSE = \sqrt{\sum_{i=1}^{n}\frac{e_i^2}{(y_i - \hat{y_i})^2}}$

Tablo 13, veri madenciliği tekniklerinin doğruluğu için karşılaştırmayı vermektedir.

Table 13: Veri Madenciliği Tekniklerinin Doğruluk Oranlarının Karşılaştırılması

Algoritmalar	MAE	RMSE	RAE	RRSE
ZeroR	0.5009	0.5042	%100	%100
OneR	0.2623	0.5121	%52.3605	%47.5732
Naive Bayes	0.1498	0.3046	%29.9017	%60.4177
J48 Karar Ağacı	0.2788	0.4235	%55.6542	%83.9956
Rastgele Orman	0.2367	0.3338	%47.2473	%66.1964
Çoklu Algılayıcılar	0.1447	0.3422	%28.8800	%67.8654
k-NN	0.2111	0.3115	%42.1435	%61.7824
Lojistik Regresyon	0.1884	0.3161	%37.6183	%62.6842
Destek Vektör Makinesi	0.1311	0.3621	%26.1803	%71.8231

4. Sonuç

Bu çalışmanın amacı, yaş, cinsiyet, göğüs ağrısı tipi, dinlenme kan basıncı, serum kolestoral, açlık kan şekeri, dinlenme elektrokardiyografik sonuçlar, elde edilen maksimum kalp atış hızı, egzersiz anjin, istirahate göre egzersizle indüklenen ST depresyonu, pik egzersiz için ST segmentinin eğimi, floroskopi ile renklendirilen büyük damarların sayısıdan kaynaklı 13 özellik ile kalp hastalığının varlığını daha doğru tahmin etmektir. ZeroR, OneR, Naive Bayes, J48 Karar Ağacı, Rastgele Orman, Çoklu Algılayıcılar, k-NN, Lojistik Regresyon, Destek Vektör Makinesi algoritmaları bu veri setine uygulanmıştır. Algoritmalar Kaggle veri setinden 303 hastaya uygulanmış ve bu veri madenciliği sınıflandırma algoritmalarının doğruluk oranları karşılaştırılmıştır. Sonuçlara göre, farklı metodolojilerin farklı kriterlerle başarılı olduğu ve kriterlerin ortalaması alındığında k-NN'nin en doğru ve en iyi sonucu veren yöntem olduğu görülmektedir.

Kaynakça

[1] W.J. Frawley, G. Piatetsky-Shapiro, Knowledge Discovery in Databases: An Overview. The AAAI/MIT Press, 1996.

[2] D. Wood, G. De Backer, O. Faergeman, I. Graham, G. Mancia, K. Pyörälä, Prevention of coronary heart disease in clinical practice: recommendations of the Second Joint Task Force of European and other Societies on Coronary Prevention. Atherosclerosis, 140 (1998) 199– 270.

[3] J. Soni, U. Ansari, D. Sharma, S. Soni, Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction, International Journal of Computer Applications, 17 (2011) 43-48.

[4] C.S. Dangare, S.S. Apte, Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques, International Journal of Computer Applications, 47 (2012) 44-48.

[5] R. Alizadehsani, J. Habibi, M.J. Hosseini, H. Mashayekhi, R. Boghrati, A. Ghandeharioun, B. Bahadorian, Z.A. Sani, A data mining approach for diagnosis of coronary artery disease, computer methods and programsin biomedicine, 111 (2013) 52-61.

[6] A. Rajkumar, G.S. Reena, Diagnosis Of Heart Disease Using Datamining Algorithm, 10 (2010) 38-43.

[7] J. Nahar, T. Imam, K. S. Tickle, Y.P. Chen, Computational intelligence for heart disease diagnosis: A medical knowledge driven approach, Expert Systems with Applications, 40 (2013) 96-104.

[8] Y. Xing, J. Wang, Z. Zhao, Y. Gao, Combination data mining methods with new medical data to predicting outcome of Coronary Heart Disease, International Conference on Convergence Information Technology, 2007.

[9] N. Bhatla, K. Jyoti, An Analysis of Heart Disease Prediction using Different Data Mining Techniques, International Journal of Engineering Research & Technology, 1 (2012) 1-4.

[10] D. Yeh, C. Cheng, Y. Chen, A predictive model for cerebrovascular disease using data mining, Expert Systems with Applications, 38 (2011) 8970-8977.

[11] S.U. Amin, K. Agarwal, R. Beg, Genetic Neural Network Based Data Mining in Prediction of Heart Disease Using Risk Factors, IEEE Conference on Information and Communication Technologies, 2013.

[12] K. Srinivas, G. Raghavendra Rao, A. Govardhan, Analysis of Coronary Heart Disease and Prediction of Heart Attack in Coal Mining Regions Using Data Mining Techniques, International Conference on Computer Science & Education, 2010.

[13] T.J. Peter, K. Somasundaram, An Empirical Study On Prediction Of Heart Disease Using Classification Data Mining Techniques, IEEE-International Conference On Advances In Engineering, Science And Management, 2012.

[14] H.D. Masethe, M. A. Masethe, Prediction of Heart Disease using Classification Algorithms, World Congress on Engineering and Computer Science, 2014.

[15] B. Bahrami, M.H. Shirvani, Prediction and Diagnosis of Heart Disease by Data Mining Techniques, Journal of Multidisciplinary Engineering Science and Technology, 2 (2015) 164-168.

[16] J. Kim, J. Lee, Y. Lee, Data-Mining-Based Coronary Heart Disease Risk Prediction Model Using Fuzzy Logic and Decision Tree, Healthcare Informatics Research, 21 (2015) 167-174.

[17] K. R. Lakshmi, M. Veera Krishna, S. Prem Kumar, Performance Comparisonof Data Mining Techniques for Predicting of Heart Disease Survivability, International Journal of Scientific and Research Publications, 3 (2013) 1-10.

[18] S. B. Patel, P. K. Yadav, D. P. Shukla, Predict the Diagnosis of Heart Disease Patients Using Classification Mining Techniques, IOSR Journal of Agriculture and Veterinary Science, 4 (2013) 61-64.
[19] N. Bhatla, K. Jyoti, A Novel Approach for Heart Disease Diagnosis using Data Mining and Fuzzy Logic, International Journal of Computer Applications, 54 (17), 2012.
[20] M.G. Tsipouras, D.I. Fotiadis, Automated Diagnosis of Coronary Artery Disease Based on Data Mining and Fuzzy Modeling, IEEE Transactions on Information Technology In Biomedicine, 12(4), 2008.
[21] B. Venkatalakshmi, M.V. Shivsankar, Heart Disease Diagnosis Using Predictive Data mining, International Journal of Innovative Research in Science, Engineering and Technology, 3(3), 2014.
[22] M.A. Jabbar, B.L Deekshatulu, P. Chandra, Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm, International Conference on Computational Intelligence: Modeling Techniques and Applications (CIMTA) 2013.
[23] I.A. Zriqat, A.M. Altamimi, M. Azzeh, A Comparative Study for Predicting Heart Diseases Using Data Mining Classification Methods, International Journal of Computer Science and Information Security (IJCSIS), 14(12), 2016.
[24] S. Sharmila, M.P. Indra Gandhi, Analysis of Heart Disease Prediction Using Data Mining Techniques, International Journal of Advanced Networking & Applications (IJANA), 8(5) (2017), 93-95.
[25] M. Sharma, F. Khan, V. Ravichandran, Comparing Data Mining Techniques Used For Heart Disease Prediction, International Research Journal of Engineering and Technology (IRJET), 4(6) (2017).
[26] M. Abdar, S.R.N. Kalhori, T. Sutikno, I.M.I. Subroto, G. Arji, Comparing Performance of Data Mining Algorithms in Prediction Heart Diseases, International Journal of Electrical and Computer Engineering (IJECE), 5(6) (2015) 1569-1576.
[27] S. Aydin, M. Ahanpanjeh, S. Mohabbatiyan, Comparison And Evaluation Data Mining Techniques In The Diagnosis Of Heart Disease, International Journal on Computational Science & Applications (IJCSA), 6(1) (2016).
[28] T.K. Keerthana, Heart Disease Prediction System using Data Mining Method, International Journal of Engineering Trends and Technology (IJETT), 47(6) (2017).
[29] H.B.F. David, S.A. Belcy, Heart Disease Prediction Using Data Mining Techniques, ICTACT Journal On Soft Computing, 9(1) (2018).
[30] S. Cihan, B. Karabulut, G. Arslan, G. Cihan, Identification of Coronary Artery Disease Risk Using Data Mining Techniques, International Journal of Engineering Research and Development, 10(1), (2018) 85-93.
[31] F. Rabbi, P. Uddin, A. Ali, F. Kibria, M.I. Afjal, S. Islam, A.M. Nitu, Performance Evaluation of Data Mining Classification Techniques for Heart Disease Prediction, American Journal of Engineering Research (AJER), 7(2) (2018), 278-283.
[32] S. Joshi, A. Sasanapuri, S. Anand, S. Nandi, V. Nemade, Predictive Analysis using Data Mining Techniques for Heart Disease Diagnosis, International Journal of Engineering & Technology, 7(3) (2018) 166-170.