Journal of Information Systems and Management Research
Bilişim Sistemleri ve Yönetim Araştırmaları Dergisi

http://dergipark.gov.tr/jismar

Araştırma Makalesi / Research Article

Veri Bilimi Uygulamalarının Hastalık Teşhisinde Kullanılması: Kalp Krizi Örneği

Mahmut Ensar GÖKTAŞ[*,a], Mete YAĞANOĞLU[b]

[a,*] *Atatürk Üniversitesi, Yönetim Bilişim Sistemleri Bölümü, ERZURUM, 25240, TÜRKİYE*
[b] *Atatürk Üniversitesi, Bilgisayar Mühendisliği Bölümü, ERZURUM, 25240, TÜRKİYE*

MAKALE BİLGİSİ

Alınma: 22.12.2020
Kabul: 29.12.2020

Anahtar Kelimeler:
Kalp Krizi, Veri Bilimi, C4.5 Karar Ağacı, K-En Yakın Komşu, Rastgele Orman, Destek Vektör Makineleri
Sorumlu Yazar
e-posta: mensargoktas@gmail.com

ÖZET

Dünyanın en yaygın hastalıklarından olan kalp hastalıklarının, uzun süre daha bir numaralı ölüm sebebi olmaya devam edeceği öngörülmektedir. Kalp hastalıkları faktörlerinin birçoğunun önlenebilir ya da tedavi edilebilir olması hastalık sonucu can kaybının azalması için bir fırsattır. Bu nedenle, vaka verilerine algoritmik ve istatistiksel yöntemler uygulayarak hastalığın tespitini amaçlayan birçok çalışma yapılmıştır. Bu çalışmanın amacı, belirlenen niteliklerin kalp krizi tanısıyla ilişkisini incelemek ve kalp krizi tanısını maksimum düzeyde doğru tahmin etmektir. 12 nitelik ve 303 olguya sahip veri setindeki değerler temizlenmiş ve ardından analizler yapılmıştır. Göğüs ağrısı tipi, egzersize bağlı anjin ve egzersize bağlı ST depresyonu niteliklerinin kalp krizi tanısıyla yüksek oranda kolerasyona sahip olduğu bulgulanmıştır. Analiz sonuçları görselleştirilerek çalışmaya eklenmiştir. Kalp krizi teşhisi için diğer çalışmalarda başarı gösteren C4.5 karar ağacı, k-en yakın komşu, rastgele orman ve DVM olmak üzere 4 farklı algoritma uygulanmış ve sonuçları karşılaştırılmıştır. Kalp krizi tanısını en doğru tahmin eden algoritma C4.5 karar ağacı algoritması olmuştur.

Using Data Science Applications in Diagnosis of Disease: A Heart Attack Example

ARTICLE INFO

Received: 22.12.2020
Accepted: 29.12.2020

Keywords:
Heart Attack, Data Science, C4.5 Decision Tree, K-Nearest Neighbor, Random Forest, Support Vector Machines
Corresponding Authors
e-mail: mensargoktas@gmail.com

ABSTRACT

It is predicted that heart disease, one of the most common diseases in the world, will continue to be the number one cause of death for a long time. The fact that many of the factors of heart disease can be prevented or treated is an opportunity to decrease the loss of life as a result of the disease. For this reason, many studies aimed at detecting the disease by applying algorithmic and statistical methods to case data have been conducted. The aim of this study is to examine the relationship between the characteristics determined and the diagnosis of heart attack and to predict the diagnosis of a heart attack at the maximum level. The values in the data set with 12 qualities and 303 cases were cleared and then analyzes were made. It has been found that the characteristics of chest pain type, exercise-induced angina, and exercise-induced ST depression have a high rate of correlation with the diagnosis of a heart attack. Analysis results were visualized and added to the study. For the diagnosis of heart attack, 4 different algorithms were applied, namely the C4.5 decision tree, k-nearest neighbor, random forest, and SVM, which were successful in other studies, and their results were compared. The C4.5 decision tree algorithm has been the most accurate algorithm for predicting heart attack diagnosis.

1. GİRİŞ *(INTRODUCTION)*

Dolaşım sisteminin ve insan vücudunun önemli bileşenlerinden birisi olan kalp, küçük yapısına rağmen vücuttaki en güçlü kastır. Anne karnında bir fetüs iken atmaya başlayan kalp, sağlıklı ve yetişkin bir bireyde günde ortalama 100000 atış gerçekleştirmektedir. Kalp, ortalama bir insan ömründe yaklaşık iki buçuk milyar kere atmakta ve her atışında vücudun her tarafına gerekli olan temiz kanı pompalamaktadır [1]. Kalbin içerisinden dakikada yaklaşık beş litre kan geçmekte ve vücuda dağılmaktadır. Ancak kalp içerisinden geçen kan ile beslenmemektedir. Yoğun ve durmadan çalışan kalbi besleyen damarlar bulunmaktadır. Bu damarlar koroner damar olarak adlandırılmaktadır. Koroner damarlarda meydana gelen daralma veya tıkanma sonucunda ortaya çıkan hastalığa ise koroner arter hastalığı ismi verilmektedir [2]. Kalp ve damar hastalıklarının büyük bir kısmının kökeninde koroner arter hastalığı bulunmaktadır [3]. Dünya Sağlık Örgütü(DSÖ) verilerine göre kalp ve damar hastalıkları sonucu ölümler yüzde 30 ile tüm ölümler arasında ilk sırayı almaktadır. Ülkemizde bu oran yüzde 47 civarındadır [4].

Koroner arter hastalıklarından birisi olan ve halk arasında kalp krizi olarak adlandırılan Miyokard infarktüsü (MI), kalbin bir bölümünün kanlanma veya oksijenlenmesinin bozulması sonucu ortaya çıkan hastalıktır. Daha açık bir tabirle kalbe yeterli miktarda kanın ve dolayısıyla oksijenin gidememesinden kaynaklanmaktadır. Yeterli miktarda oksijen gitmediğinde kalpte hasar, uzun süre oksijen gitmediğinde ise kalpte ölüm meydana gelmektedir [5]. Ölüm riski oluşturan veya insan sağlığını olumsuz yönde etkileyen bu hastalığın başlıca nedenleri şu şekilde sıralanabilir [6]:

- Obezite
- Hipertansiyon
- Diyabet
- Kolesterol
- Cinsiyet
- Yaş
- Tütün kullanımı
- Alkol kullanımı
- Aile öyküsü

Bu maddelere ek olarak, uyuşturucu kullanımı, stres ve yoğun yaşam temposu da hastalığı tetikleyen nedenler arasındadır [7]. Kalp krizi tanısı, fiziksel muayene ile kreatinin kinaz, troponin, miyoglobin ve elektokardiyografi(EKG) gibi testlerin sonuçlarına göre koyulabilmektedir [6]. Hastalık tanısı günümüzde yalnızca doktorlar tarafından konmaktadır. Sağlık bilişimi kavramı ile tıp dünyasında daha fazla yer edinmeye başlayan bilişim sistemlerinin bu konuda hekimlere yardım etmesi beklenebilir. Tıp dünyası veri olarak çok zengindir. Bir benzetme yapmak gerekirse, tıp dünyasındaki veriler bir okyanusa, ilişkilerin keşfedildiği ve ortaya çıkan bilgi bir su damlacığına benzetilebilir. Son yıllarda kalp ve damar hastalıklarına yönelik algoritmik ve istatistiksel birçok çalışma yapılmıştır.

Doğan [8] yaptığı çalışmada 50 kalp krizi tanısı almış ve 11 kalp krizi tanısı almamış hasta verileri üzerinden karar ağacı yöntemi ile sınıflandırma yapmış ve %100 oranında duyarlılık ve özgüllük elde etmiştir.

Anbarasi ve arkadaşları [9] yaptıkları çalışmada kalp rahatsızlıklarına neden olan faktörleri Genetik Algoritmalar yardımıyla tespit etmeye çalışmışlardır. Çalışma sonucunda faktörler arasından en önemli 6 tanesi tespit edilmiştir. Çalışmada karar ağacı tekniğinin daha iyi performans gösterdiği gözlemlenmiştir.

Çakmakçı ve Kahyaoğlu [10] yaptıkları çalışmada Cleveland klinik kurumundan alınan 300 hasta verisi üzerinde aşırı yapay öğrenme yöntemi kullanarak kalp krizi tespiti yapmaya çalışmışlardır. Çalışma sonucunda %80 başarı oranı elde etmişlerdir.

Bulut [1] yaptığı çalışmada yapay sınırlandırma yöntemlerini kullanarak kalp krizi risk oranını belirlemeyi amaçlamıştır. Çalışmada Manisa Merkezefendi Devlet Hastanesi ve İzmir Ege Üniversitesi Hastanesinde tedavi gören 62 hastanın verisi kullanılmıştır. Verilerin analizinde karar ağacı yöntemi ve AdaBoost algoritması kullanılmıştır.

Çalışma sonucunda kalp krizini tetikleyen en önemli beş faktör, büyük tansiyon değeri, diyabet, obezite, sigara ve küçük tansiyon değeri olarak bulunmuştur.
Taşçı ve Şamlı [11] yaptıkları çalışmada yüzlerce veri bulunan veri setine çeşitli algoritmalar uygulayarak kalp hastalığı teşhisi koyma çalışması yapmıştır. 9 farklı veri madenciliği yöntemini WEKA yazılımında veri seti üzerinde uygulamışlardır. K-NN yönteminin en başarılı yöntem olduğunu tespit etmişlerdir.

Sunulan çalışmalar verinin işlenip bilgi çıkarılmasının tıp alanında kullanılabilecek yararlı sonuçlar ortaya çıkarılabileceğine birer örnek olarak gösterilebilir. Kalp ve damar hastalıkları üzerine yapılan çalışma sayısı arttıkça kalp hastalıklarının önlenmesi konusunda daha fazla destek sağlayacaktır.

Bu çalışmanın amacı, yaş, cinsiyet, göğüs ağrısı tipi, istirahat kan basıncı, kolesterol değeri, açlık kan şekeri, istirahat elektrokardiyografik sonuç, maksimum nabız, egzersize bağlı anjin, egzersize bağlı ST depresyonu, ST segmenti eğimi ve defekt tipi niteliklerinin kalp krizi tanısıyla ilişkisini incelemek ve kalp krizi tanısını maksimum düzeyde doğru tahmin etmektir. Çalışma kapsamında veri ön işleme yapılmış ve veriler düzenlenerek analiz edilmiştir. Analiz sonuçları görselleştirilerek çalışmaya eklenmiştir. Tanı tahmini için C4.5 karar ağacı, k-en yakın komşu, rastgele orman, destek vektör makineleri algoritmaları kullanılmıştır. Çalışma

Kaggle veri tabanından alınan 12 özniteliğe ve 303 olguya sahip veri kümesi üzerinde ilerlemiştir.

Çalışmanın ikinci bölümünde kullanılan veri seti ve analiz aşamasında kullanılan yöntemler açıklanmıştır. Üçüncü bölümde analiz sonucu elde edilen sonuçlar sunulmuş ve tartışmalar yapılmıştır. Çalışmanın dördüncü ve son bölümünde ise çalışmanın sonuçları açıklanmıştır.

2. MATERYAL ve YÖNTEM *(MATERIAL and METHODS)*

2.1. Veri Seti *(Data Set)*

Çalışmada Kaggle veri tabanında yer alan 85000 kalp krizi vakasından oluşan, kalp krizi tanısı konmuş ya da sağlıklı bireylerin değerlerinden oluşan veri seti kullanılmıştır. Değerler çeşitli yöntemlerle elimine edilmiş ve sonuç olarak 12 nitelik ve 303 vaka ile çalışmaya başlanmıştır. Kullanılan öznitelikler sırasıyla yaş, cinsiyet, göğüs ağrısı tipi, istirahat kan basıncı, kolesterol, açlık kan şekeri, istirahat elektrokardiyografik sonuç, maksimum nabız, egzersize bağlı anjin, egzersize bağlı ST depresyonu, ST segmenti eğimi ve defekt tipidir. Tablo 1'de özniteliklerin ayrıntılı açıklaması verilmiştir.

Tablo 1: Öznitelikler ve Açıklamaları
(Attributes and Descriptions)

Kod	Nitelik Açıklaması	Değer	Değer Açıklaması
age	Hastanın yaşı	Nümerik	
sex	Hastanın cinsiyeti	0,1	0: Kadın
			1: Erkek
cp	Göğüs ağrısı tipi	0,1,2,3	0: Tipik anjin ağrı
			1: Atipik anjin ağrı
			2: Anjinal olmayan ağrı
			3:Asemptomatik ağrı
trestbps	İstirahat kan basıncı	Nümerik	
chol	Kolesterol değeri	Nümerik	
fbs	Açlık kan şekeri	0,1	0: 120 veya daha küçük şeker değeri
			1: 120'den büyük şeker değeri
restecg	İstirahat elektrokardiyografik sonuç	0,1,2	0: Normal değer
			1: ST-T dalga anormalliği
			2: muhtemel ya da kesin sol ventrikül hipertrofisi
thalac	Maksimum nabız	Nümerik	
exang	Egzersize bağlı anjin	0,1	0: Yok
			1: Var
oldpeak	Egzersize bağlı ST depresyonu	Nümerik	
slope	ST segmentinin eğimi	0,1,2	0: Yukarı doğru
			1: Düz
			2: Aşağı doğru
thal	Defekt tipi	3,6,7	3: Normal
			6: Belirlenmiş defekt
			7: Tersinir defekt

2.2. Kullanılan Teknik ve Yöntemler *(Techniques and Methods)*

Çalışma 3 aşama uygulanarak tamamlanmıştır. Veriler analiz edilmeden önce ön işleme tekniklerinden veri temizleme yapılarak düzenlenmiştir. Daha sonra veriler analiz edilerek görselleştirme yapılmıştır. Son olarak karar ağacı, k-en yakın komşu, rastgele orman ve DVM(SVM) algoritmaları kullanılarak veriler makineye öğretilerek tahminde bulunması istenmiş ve başarı sonuçları listelenmiştir.

2.2.1. Veri temizleme *(Data cleansing)*

Veri temizleme, veri analiz işlemleri yapılmadan önce kesinlikle yapılması gerekilen bir işlemdir. Eksik verilerin tamamlaması, tutarsızlıkların giderilmesi ve aykırı değerlerin saptanması için gürültünün giderilmesi gibi işlemlerden oluşmaktadır [12]. Eksik verilerin tamamlanmasında şu yöntemler kullanılabilir [13]:

- Eksik değer içeren kayıtlar silinebilir.
- Eksik değerler yerine ortalama kullanılabilir.
- Eksik değerler yerine medyan kullanılabilir.
- Eksik değerler yerine içinde bulunduğu sınıfın ortalaması kullanılabilir.
- Eksik değerler yerine regresyon gibi yöntemler kullanarak en uygun değer kullanılabilir.

Aykırı değerlerin saptanması için ise kullanılan 3 yöntem bulunmaktadır. Bunlar, binning, kümeleme ve regresyon yöntemleridir. Olabilecek veri tutarsızlıkları ise dışsal referans kullanarak giderilebilir [12].

2.2.2. C4.5 karar ağacı algoritması *(C4.5 decision tree algorithm)*

C4.5 karar ağacı algoritması, son yıllarda sık kullanılan sınıflandırma yöntemlerinden birisidir [14]. Algoritma öncelikle her hedef ve tahmin değişkeni için bilgi değeri hesaplar. Daha sonra en fazla bilgi toplayan sınıfı belirlemek amacıyla her tahmin değişkeni için bilgi kazancı hesaplar. En fazla bilgi kazanımı sağlayan tahmin değişkeni tespit edilir ve ağaç oluşmaya başlar. Veriler her dalda dengeli bir biçimde dağılır. İlk tahmin edici değişken belirlendikten sonra geriye kalan tahmin edici değişkenlerin hangisiyle bölümlenmesinin en yüksek

bilgi kazanımı sağlayacağı hesaplanır. Bu işlem tahmin edici değişkenlerin tümü ağaca yerleşene kadar devam eder [11].

2.2.3. K-en yakın komşu algoritması *(K-nearest neighbor algorithm)*

K-En yakın komşu algoritması, eğitim setindeki verilere dayalı olarak öğrenmeyi gerçekleştiren örnek tabanlı bir öğrenme algoritmasıdır. Gruptaki yeni veriyi, eğitim setindeki verilerle karşılaştırarak sınıflandırma işlemini yapmaktadır [15]. K-En yakın komşu algoritmasında, eğitim setindeki her örnek uzayda bir noktayı temsil edecek şekilde tutulmaktadır. Yeni bir örnek uzaya katıldığında, yeni örneğe en yakın eğitim setindeki örneklerden k tanesi belirlenerek yeni örneğin sınıfı belirlenmektedir [16].

2.2.4. Rastgele orman algoritması *(Random forest algorithm)*

2001 yılında geliştirilen rastgele orman algoritmasının amacı, tek bir karar ağacı yerine farklı eğitim kümelerinde eğitilmiş ağaç kararlarını birleştirip sunmaktır. Her seviyedeki nitelik belirlenirken, bütün ağaçlarda birtakım hesaplar yapılarak nitelik belirlenmektedir. Daha sonra ise diğer ağaçlardaki nitelikler birleştirilerek en fazla kullanılan nitelik seçilmektedir. Seçilen nitelik ağaca dahil edilip bütün seviyelerde işlem tekrarlanmaktadır. Algoritmanın başlatılabilmesi için her düğümde kullanılacak değişken sayısı ve ağaç sayısı kullanıcı tarafından belirlenmelidir [17].

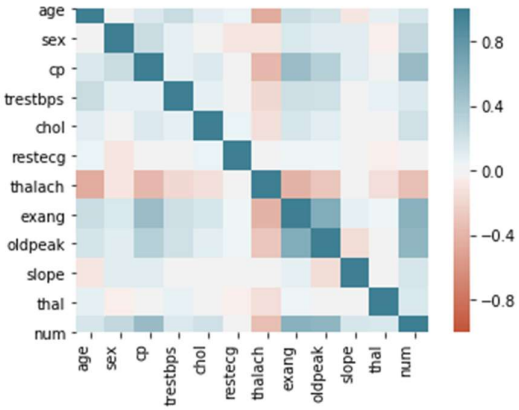
2.2.5. Destek vektör makineleri *(Support vector machine)*

Destek vektör makineleri, bilinen adıyla SVM, riski minimize etmeyi amaçlayan, dış bükey optimizasyona dayalı makine öğrenim algoritmalarıdır. Dağılım bilgisine ihtiyaç duymadığı için bağımsız öğrenme yapabilmektedir. Algoritmaların amacı, sınıfları ayıracak en uygun hiper düzlemi elde etmektir. Diğer bir ifadeyle, farklı sınıflar arasındaki uzaklığı maksimize etmeyi amaçlamaktadır [18].

Çalışmada kullanılan veri seti 12 nitelik ve 303 olguya sahiptir. 12 nitelik bağımsız değişken olarak ele alınmış ve 0 ve 1 değerlerini içeren tanı adında bir bağımlı değişken tahmin edilmeye çalışılmıştır. Çalışmada kullanılan veriler öncelikle temizlenmiştir. Temizleme işleminde eksik veri barındıran thalac değişkeni nümerik değerler taşıdığından eksik verilerin yerine ortalama değer atanmıştır. Boş değerler barındıran nümerik olmayan değişkenler için ise medyan değeri eksik veriler yerine kullanılmıştır. Uygulama aşamasında veriler Python ortamında, Python dili kullanılarak analiz edilmiştir. Veriler öncelikle korelasyon analizine tabii tutulmuştur. Aralarında korelasyon ilişkisi çıkan nitelikler analiz edilmiş ve sonuçlar görselleştirilmiştir. Tahmin aşamasında Scikit-learn kütüphanesi kullanılmıştır. Bu kütüphane Python dilinde makine öğrenimi konusunda en kullanışlı kütüphanelerden olduğu için tercih edilmiştir.

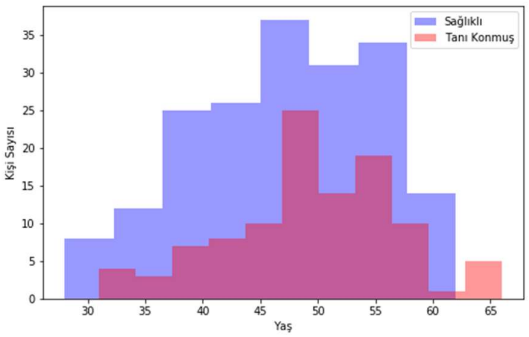
3. BULGULAR *(FINDINGS)*

Çalışmada kullanılan veri setinin niteliklerinin tanı durumu ile olan korelasyonunu belirlemek için korelosyon analizi yapılmıştır. Test sonuçları Şekil 1'de gösterilmiştir.



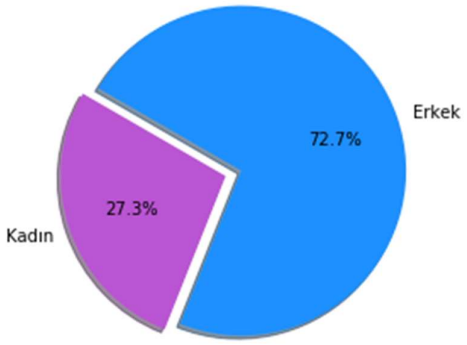
Şekil 1. Korelasyon analizi sonuçları
(Correlation analysis results)

Şekil 1 incelendiğinde, niteliklerin tanı durumu ile olan korelasyonunda göğüs ağrısı tipi, egzersize bağlı anjin ve egzersize bağlı ST depresyonu arasında pozitif yüksek bir korelasyon olduğu görülebilmektedir. Bu durumda göğüs ağrısı tipi, egzersize bağlı anjin ve egzersize bağlı ST depresyonunun tanı durumunu etkilediği söylenebilmektedir. Korelasyon analizinde yüksek bir korelasyon çıkmasa da yapılmış çalışmalarda yaşın kalp krizi vakasını etkilediği sıklıkla geçmektedir. Yaş-tanı grafiği Şekil 2'de gösterilmiştir.



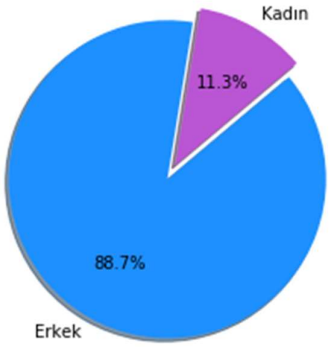
Şekil 2.Yaş-tanı grafiği
(Age-diagnosis chart)

Şekil 2 incelendiğinde yaş ve tanı arasında bir bağlantı olmadığı rahatlıkla görülebilmektedir. Bu durum korelasyon analizini de desteklemektedir. Literatürde bahsedilen bir diğer değişken ise cinsiyettir. Vakaların cinsiyete göre dağılımına ilişkin grafik Şekil 3'te gösterilmiştir.



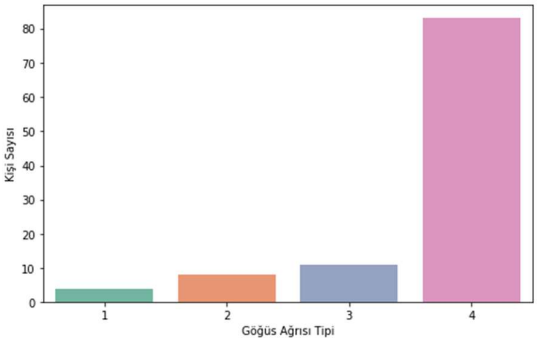
Şekil 3.Vakaların cinsiyet dağılımı
(Gender distribution of the cases)

Kaydedilen 303 vakanın %72.7'sini erkek hastalar, %27.3'ünü ise kadın hastalar oluşturmaktadır. Bu grafik tek başına bir yorumlanamayacağından kalp krizi tanısı konmuş hastaların cinsiyete göre dağılımına bakıp birlikte yorumlanmalıdır. Tanı konmuş vakaların cinsiyet dağılımına ilişkin grafik Şekil 4'te gösterilmiştir.



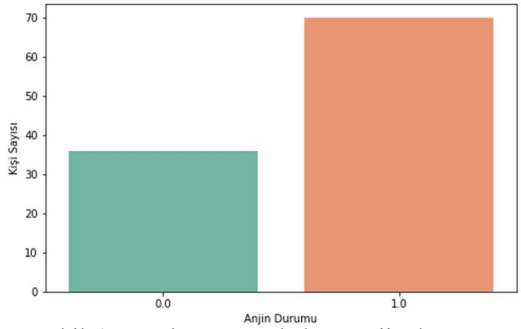
Şekil 4. Tanı konmuş vakaların cinsiyet dağılımı
(Gender distribution of diagnosed cases)

Kalp krizi tanısı konmuş 106 hastanın %88.7'sini erkek hastalar, %11.3'ünü ise kadın hastalar oluşturmaktadır. Şekil 3 ve Şekil 4 birlikte incelendiğinde kalp krizi ile cinsiyet arasında bir ilişki olduğu söylenebilmektedir. Bu durum kolerasyon testi sonucunda ortaya çıkan normal pozitif korelasyon sonucunu desteklemektedir. Korelasyon testi sonucunda güçlü pozitif ilişki çıkan niteliklerden birisi olan göğüs ağrısı tipine ilişkin grafik Şekil 5'te gösterilmiştir.



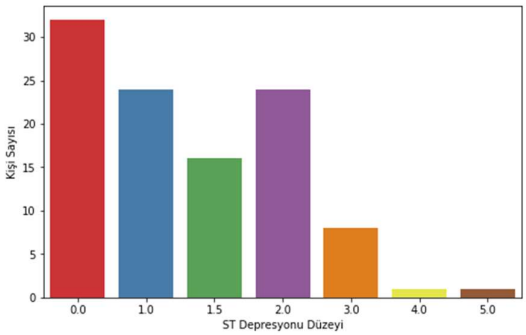
Şekil 5. Tanı konmuş vakaların göğüs ağrısı tipi dağılımı
(Chest pain type distribution of diagnosed cases)

Şekil 5 incelendiğinde, kalp krizi tanısı konmuş vakalarda hastalarda büyük oranda asemptomatik ağrı tipi olduğu görülebilmektedir. Bu durum hasta kayıt altına alınırken herhangi bir göğüs ağrısı yaşamadığını ifade etmektedir. Asemptomatik ağrıyı sırasıyla, anjinal olmayan ağrı, atipik anjin ağrı ve tipik anjin ağrı takip etmektedir. Kalp krizi tanısı ile yüksek korelasyona sahip bir diğer nitelik egzersize bağlı anjin durumudur. Tanı konmuş vakalardaki anjin durumu dağılımı Şekil 6'da gösterilmiştir.



Şekil 6. Tanı konmuş vakaların anjin durumu dağılımı
(Angina status distribution of diagnosed cases)

Şekil 6 incelendiğinde, kalp krizi tanısı konmuş hastalarda büyük oranda egzersize bağlı anjin olduğu görülebilmektedir. Anjin, kalbi besleyen damarların kalbi besleyecek yeterli kanın geçemeyeceği şekilde daralması halinde kalbin yaptığı ağrıdır. Bu durumda egzersiz sırasında oluşan anjinin kalp krizi belirtilerinden olduğu söylenilebilir. Tanı ile yüksek pozitif ilişkisi bulunan son nitelik ise egzersize bağlı ST depresyonudur. Tanı konmuş vakalardaki ST depresyonuna ilişkin grafik Şekil 7'de gösterilmiştir.



Şekil 7. Tanı konmuş hastalardaki egzersize bağlı ST depresyonu durumu dağılımı
(Distribution of exercise-induced ST depression status in diagnosed patients)

Egzersize bağlı ST depresyonu, kalp kansızlığının en sık rastlanan bulgularından bir tanesidir. Genellikle ST çökmesi yaşanmadan önce ST segmentinde düzleşme(depresyon) meydana gelmektedir [19]. Şekil 7 incelendiğinde, tanı konmuş hastaların yaklaşık üçte ikisinde egzersize bağlı ST segment düzleşmesinin en az 1 defa olduğu görülmektedir. Bu durum dikkate alınarak egzersize bağlı ST depresyonunun kalp krizi için önemli bir bulgu olduğu söylenebilir.

Çalışmada dört farklı algoritma kullanılarak sınıflandırma işlemi yapılmış ve algoritmaların yeni vaka tahminindeki başarı oranları incelenmiştir. Bu aşamada age, sex, cp, trestbps, fbs, restecg, thalach, exang ve oldpeak nitelikleri test verisi olarak kullanılmıştır. Test sınfının büyüklüğü %25 olarak belirlenmiştir. Karar ağacı algoritmasında kriter olarak entropi belirlenmiştir. K-en yakın komşu algoritmasında, test verisi büyüklüğü %33 olarak belirlenmiştir. Metrik olarak minkowski tekniği uygulanmıştır. Rastgele orman algoritmasında test verisi büyüklüğü %30 olarak ele alınmıştır. Son olarak Destek Vektör Makinelerinde ise test boyutu %33 olarak ele alınmış ve lineer yöntem kullanılmıştır.

Algoritmaların başarı oranları Tablo 2'de gösterilmiştir.

Tablo 2. Algoritmaların tahmin başarı yüzdesi
(The prediction success percentage of algorithms)

Algoritma	**Başarı Oranı(Yüzde)**
C4.5 Karar Ağacı Algoritması	83,099
K-En Yakın Komşu	68,085
Rastgele Orman	81,456
Destek Vektör Makineleri (Doğrusal)	75,532

C4.5 karar ağacı algoritması vaka tahmini konusunda %83,099 oranında başarı elde etmiştir. C4.5 karar ağacı algoritmasını %81,456 başarı oranıyla rastgele orman, %75,532 başarı oranıyla destek vektör makineleri ve %68,085 başarı oranıyla k-en yakın komşu algoritmaları takip etmiştir.

4. SONUÇ *(RESULTS)*

Bu çalışmanın amacı, yaş, cinsiyet, göğüs ağrısı tipi, istirahat kan basıncı, kolesterol değeri, açlık kan şekeri, istirahat elektrokardiyografik sonuç, maksimum nabız, egzersize bağlı anjin, egzersize bağlı ST depresyonu, ST segmenti eğimi ve defekt tipi niteliklerinin kalp krizi tanısıyla ilişkisini incelemek ve kalp krizi tanı değerini(0,1) maksimum düzeyde doğru tahmin etmektir. Veri setinde yer alan değerler temizlenerek analiz aşamasına uygun hale getirilmiştir. Sonrasında veriler analiz edilmiş ve görselleştirme yapılarak çalışmaya eklenmiştir. Analiz sonucunda göğüs ağrısı tipi, egzersize bağlı anjin durumu ve egzersize bağlı ST depresyonu niteliklerinin kalp krizi tanısı ile yüksek pozitif kolerasyona sahip olduğu görülmüştür. Bu durum, göğüs ağrısı türü, egzersize bağlı anjin durumu ve egzersize bağlı ST depresyonu değişkenlerinin kalp krizi hastalığı için birer işaret olabildiğini göstermektedir. Cinsiyet ve yaş niteliklerinin ise kalp krizi tanısı ile normal pozitif kolerasyona sahip olduğu saptanmıştır. Bu durum ise cinsiyet ve yaşın kalp krizinde etkili olduğunu göstermiştir. Kadınlarda erkeklerden, gençlerde ise yaşlılardan daha az kalp krizi riski olduğunu göstermektedir.Kalp krizi tanısını tahmin etmek için C4.5 karar ağacı, k-en yakın komşu, rastgele orman ve destek vektör makineleri algoritmaları veri setine uygulanmış ve başarı puanları kaydedilmiştir. Yapılan testler sonucunda C4.5 karar ağacı algoritmasının en iyi sonucu verdiği görülmektedir. Karar ağacı algoritmasını rastgele orman, Destek Vektör Makineleri ve K-En yakın komşu algoritmaları takip etmiştir.

KAYNAKLAR *(REFERENCES)*

[1] F. Bulut, "AdaBoost ile Kalp Krizi Risk Tespiti," *CBÜ Fen Bilimleri Dergisi,* pp. 459-472, 2016.

[2] N. Bilir, Sigara ve Kalp-Damar Hastalıkları, Ankara: T.C. Sağlık Bakanlığı Yayınları, 2008.

[3] WHO, Noncommunicable Diseases Country Profiles, Geneva: World Health Organization, 2014.

[4] F. Bulut, "Heart Attack Risk Detection Using Bagging Classifier," %1 içinde *24th Signal Processing and Communication Application Conference (SIU)*, Zonguldak, 2016.

[5] A. B. Storrow ve W. B. Gibler, "Chest Pain Centers: Diagnosis of Acute Coronary Syndromes," *Annals of Emergency Medicine,* cilt 35, no. 5, pp. 449-461, 2000.

[6] İ. Buçan, Kalp Krizi Karar Destek Sistemi, Trabzon: Karadeniz Teknik Üniversitesi Sağlık Bilimleri Enstitüsü, 2017.

[7] S. Güleç, "Kalp Damar Hastalıklarında Global Risk ve Hedefler," *Türk Kardiyoloji Derneği Araştırmaları,* cilt 37, no. 2, pp. 1-10, 2009.

[8] Ş. Doğan, "Heart attack detection from cardiac by using decision trees," *Engineering Sciences,* pp. 39-50, 2007.

[9] M. Anbarasi, E. Anupriya ve N. C. S. N. Iyengar, "Enhanced Prediction of Heart Disease with Feature Subset Selection Using Genetic Algorithm," *International Journal of Engineering Science and Technology,* pp. 5370-5376, 2010.

[10] S. Çakmakçı ve D. T. Kahyaoğlu, "Yağ Asitlerinin Sağlık ve Beslenme Üzerine Etkileri," *Türk Bilimsel Derlemeler Dergisi,* pp. 133-137, 2012.

[11] M. E. Taşçı ve R. Şamlı, "Veri Madenciliği İle Kalp Hastalığı Teşhisi," *Avrupa Bilim ve Teknoloji Dergisi,* pp. 88-95, 2020.

[12] A. Oğuzlar, "Veri Ön İşleme," *Erciyes Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi,* pp. 67-76, 2003.

[13] R. J. Roiger, Data Mining:A Tutorial-Based Primer(Second Edition), CRC Press, 2016.

[14] H. Hong, J. Liu, D. T. Bui, B. Pradhan, T. D. Acharya, B. T. Pham, A.-X. Zhu, W. Chen ve B. B. Ahmad, "Landslide susceptibility mapping using C4.5 Decision Tree with AdaBoost,Bagging and Rotation Forest ensembles in the Guangchang area (China)," *Catena,* pp. 399-413, 2018.

[15] E. Taşçı ve A. Onan, "K-En Yakın Komşu Algoritması Parametrelerinin Sınıflandırma Performansı Üzerine Etkisinin İncelenmesi," %1 içinde *18. Akademik Bilişim Konferansı*, Aydın, 2016.

[16] J. Han, J. Pei ve M. Kamber, Data Mining: Concepts and Techniques, Burlington: Elseiver, 2011.

[17] B. Daş ve İ. Türkoğlu, "DNA Dizilimlerinin Sınıflandırılmasında Karar Ağacı Algoritmalarının Karşılaştırılması," %1 içinde *Eleco 2014 Elektrik – Elektronik – Bilgisayar ve Biyomedikal Mühendisliği Sempozyumu*, Bursa, 2014.

[18] S. Ayhan ve Ş. Erdoğmuş, "Destek Vektör Makineleriyle Sınıflandırma Problemlerinin Çözümü İçin Çekirdek Fonksiyonu Seçimi," *Eskişehir Osmangazi Üniversitesi İktisadi ve İdari Bilimler Dergisi,* cilt 9, no. 1, pp. 175-201, 2014.

[19] M. Ertürk, "Rose Angina Anketi"nin Türk Toplumunda Uygulanabilirliği ve Efor Testi Sonuçları ile Karşılaştırılması, İstanbul: Prof. Dr. Siyami Ersek Göğüs Kalp ve Damar Cerrahisi Merkezi, 2005.