MÜHENDİSLİK FAKÜLTESİ

Uluslararası Mühendislik Araştırma ve Geliştirme Dergisi

International Journal of Engineering Research and Development

Cilt/Volume:10 Sayı/Issue:1 Ocak/January 2018 **https://doi.org/10.29137/umagd.419663**

Araştırma Makalesi / Research Article

Koroner Arter Hastalığı Riskinin Veri Madenciliği Yöntemleri İle İncelenmesi

Identification of Coronary Artery Disease Risk Using Data Mining Techniques

Şeyma CİHAN[1], Bergen KARABULUT*[1], Güvenç ARSLAN[2], Gökhan CİHAN[3]

[1]Kırıkkale Üniversitesi, Mühendislik Fakültesi, 71450 Kırıkkale, TÜRKİYE
[2]Kırıkkale Üniversitesi, Fen Edebiyat Fakültesi, 71450 Kırıkkale, TÜRKİYE
[3]Kırıkkale Yüksek İhtisas Hastanesi, Kardiyoloji Bölümü, 71400 Kirikkale, TÜRKİYE

Başvuru/Received: 10/02/2017 **Kabul/Accepted**: 25/07/2017 **Son Versiyon/Final Version**: 29/01/2018

Öz

Günümüzde Kardiyovasküler Hastalıklar oldukça yaygındır ve ölüm nedenlerinin başında gelmektedir. Kardiyovasküler Hastalıkların bir tipi olan Koroner Arter Hastalığının doğru ve zamanında teşhisi çok önemlidir. Koroner arter hastalığının kesin tanısı ve hastalık şiddetinin saptanmasında invaziv bir yöntem olan anjiyografi altın standart olarak kullanılmaktadır. Anjiyografi, maliyeti yüksek ve ileri seviyede uzmanlık gerektiren bir yöntem olmasının yanında ciddi komplikasyonlara da sebep olabilmektedir. Bu nedenlerle daha ucuz ve etkili bir yaklaşım sağlayabilecek olan veri madenciliğinin kullanımı üzerinde çalışmalar yapılmaktadır. Bu çalışmada Koroner Arter Hastalığı riskinin tespitinde bir sınıflama modeli geliştirmek için veri madenciliği yaklaşımı uygulanmıştır. Çalışma kapsamında sınıflandırma yöntemleri ile elde edilen sonuçlar ve doğru sınıflandırma oranları karşılaştırılmıştır. Bunun için Cleveland kliniğine ait, 303 kayıt ve 14 değişken içeren kalp hastalığı veri kümesi kullanılmıştır. Gerekli hesaplamalar ve modelleri elde etmek için Weka paket programında 1R, J48 Karar Ağacı, Naive Bayes ve Çok katmanlı yapay sinir ağı (YSA) sınıflandırma yöntemleri uygulanmıştır. Uygulama sonucunda Koroner Arter Hastalığının tespitinde en iyi sonucun %83,498 doğruluk oranı ile Çok katmanlı YSA sınıflandırma yöntemi ile elde edildiği görülmüştür. Çok katmanlı YSA algoritmasını Naive Bayes ve Düzenlenmiş J48 Karar Ağacı algoritmaları izlemiştir.

Anahtar Kelimeler

"Antideprasan, ilaç etken maddesi, adsorpsiyon, izoterm"

Abstract

Cardiovascular Diseases are quite common nowadays and are one of the leading causes of death. The correct and timely diagnosis of Coronary Artery Disease, a type of Cardiovascular Disease, is very important for further treatment of the patients. For accurate diagnosis of coronary artery disease and determination of disease severity, angiography, which is an invasive and gold standard diagnosis tool, is used. Angiography is a costly and advanced method that requires clinical expertise and may cause serious complications. For these reasons, research on using data mining techniques, which is a cheaper and more effective approach, for diagnosis is one of today's research topics. In this study, classification-based data mining methods were used to determine the risk of coronary artery disease and these methods were compared in terms of accuracy. A data set consisting of 303 patient records and 14 attributes of Cleveland clinic were used. In particular, 1R, J48 Decision Tree, Naive Bayes and Multilayer Artificial Neural Network classification methods were applied on this data set with the help of WEKA program. The best result (in terms of correct diagnosis ratio) in determining risk of Coronary Artery Disease was obtained with Artificial Neural Network classification method with an accuracy of 83.498%. The multi-layer ANN algorithm was followed by Naive Bayes and the J48 Decision Tree algorithms.

Key Words

"Data Mining, Classification, Coronary Artery Disease"

brgnkarabulut@gmail.com

1. GİRİŞ

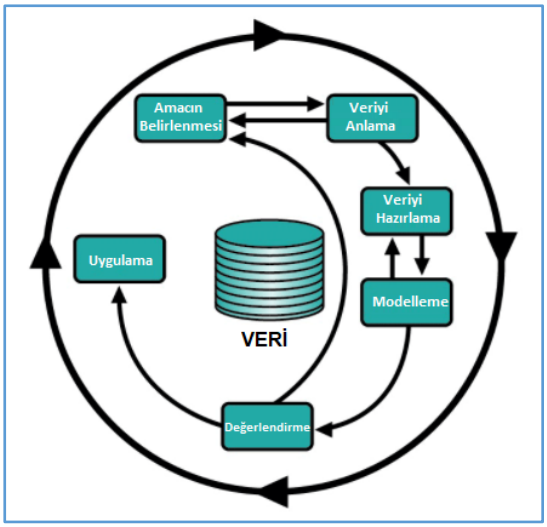
Kardiyovasküler hastalıklar (KVH), kalp ve kan damarlarındaki patolojilerden kaynaklanmakta ve koroner arter hastalığı (KAH), kalp yetmezlikleri, kardiyak arrest, ventriküler aritmiler, ani kalp ölümü, iskemik inme, geçici iskemik atak, subaraknoid ve intraserebral hemoraji, abdominal aort anevrizması, periferal arter hastalıkları ve konjenital kalp hastalıkları ile sonuçlanabilmektedir (Wong, 2014). Dünya Sağlık Örgütü raporlarına (WHO) göre KVH küresel ölüm nedenleri arasında birinci sırada yer almaktadır. 2012 yılında tüm dünyada 17,5 milyon insan KVH nedeniyle yaşamını yitirmiştir. Bu sayı küresel ölümlerin %31'ini oluşturmaktadır (WHO, 2011). Kardiyovasküler hastalıkların bir tipi olan KAH, büyük ve orta çaplı arterlerin damar duvarında kalınlaşma ve esneklik kaybının söz konusu olduğu ve bu damarların intima tabakasında aterom plaklarıyla karakterize bir patoloji olan ateroskleroz sonucunda gelişmektedir (Avşar vd., 2011). KAH, miyokard infarktüsü, kalp yetmezliği ve ani kalp ölümüne neden olabilmektedir.

Koroner arter hastalığının tanısı, nükleer tarama, ekokardiyografi, elektrokardiyogram (EKG), egzersiz stres testi gibi non-invaziv (girişimsel olmayan) ve anjiyografi gibi invaziv (girişimsel) medikal tanı işlemleri gerekmektedir (Verma vd., 2016). Non-invaziv teknikler koroner arter hastalıklarının kesin tanısında yetersiz kalabilmektedir. Bu nedenle invaziv bir yöntem olan anjiyografi tanı yöntemi koroner arter hastalıklarının kesin tanısı ve hastalık şiddetinin saptanmasında altın standart olarak kullanılmaktadır. Ancak anjiyografi işlemi, maliyeti oldukça yüksek ve ileri seviyede teknik uzmanlık isteyen bir tanı yöntemidir (Alizadehsani vd., 2012). Ayrıca anjiyografi sırasında; işlem ile ilgili nadir olmakla birlikte hastanın durumuna, uzman hekimin deneyimine ve işlemin tipine göre değişen oranlarda ölüm, myokard infarktüsü, serebrovasküler olaylar, ritim bozuklukları, damarsal komplikasyonlar, işlem sırasında kullanılan kontrast maddeye bağlı böbrek yetmezliği gibi komplikasyonlar ortaya çıkabilmektedir (Ökçün ve Gürmen, 2007). Bu nedenle, araştırmacılar KAH tanısında, veri madenciliği gibi daha az maliyeti olan ve etkili yöntemler üzerinde çalışmalar yürütmektedirler (Soni vd., 2011; Alizadehsani vd., 2013; El-Bialy vd.,2015; Sharan ve Sathees, 2016).

Klinik karar verme süreçleri sıklıkla doktorun sezgilerine ve tecrübesine dayanmaktadır. Bu yaklaşım istenmeyen ön yargıları, klinik hataları ve medikal maliyeti artırarak hastaya verilen sağlık bakım kalitesini olumsuz yönde etkileyebilmektedir. Veri madenciliği gibi veri modelleme ve analizinin yapıldığı yöntemlerin kullanıldığı klinik karar destek sistemlerinin oluşturulması ile medikal hatalar, istenmeyen uygulama çeşitlilikleri, medikal maliyet azaltılabilir ve hasta güvenliği ve klinik karar kalitesi anlamlı ölçüde artırılabilir (Srinivas vd., 2010).

Veri madenciliği, çalışılan alanla ilgili sorunların veri madenciliği görevlerine dönüştürülmesinde, uygun veri dönüşümü, hazırlığı, veri madenciliği modelinin seçimi, sonuçların etkililiğinin değerlendirilmesi ve deneyimin raporlanmasında standart bir yaklaşıma gereksinim duymaktadır. CRISP-DM (CRoss Industry Standard Process for Data Mining) hem iş sektörü hem de kullanılan teknolojiden bağımsız olarak veri madenciliği projeleri yürütmek için sistematik bir çerçeve sağlayan bir süreç modeli tanımlamaktadır. CRISP-DM süreç modeli, büyük veri madenciliği projelerini, daha az maliyetli, daha güvenilir, daha tekrarlanabilir, daha yönetilebilir ve daha hızlı hale getirmektedir (Wirth ve Hipp, 2000; Palaniappan ve Awang, 2008). CRISP-DM altı ana aşamadan oluşmaktadır. Bunlar; amacın/hedefin belirlenmesi, veriyi anlama, veriyi hazırlama, modelleme, değerlendirme, sonuçları kullanma/uygulamadır (Çınar ve Arslan, 2008). Şekil 1 CRISP-DM süreç modeli aşamalarını göstermektedir.

Bu çalışmada en yaygın kullanılan veri madenciliği yöntemlerinden biri olan sınıflandırma üzerinde durulmuştur. Farklı sınıflandırma yöntemleri kullanılarak koroner arter hastalığı riskinin tespitinde en uygun teşhis aracı tespit edilmeye çalışılmıştır. Yöntemler veri kümesi üzerinde denenmiş ve elde edilen sonuçlar karşılaştırmalı olarak sunulmuştur. Veri madenciliği modelinin uygulanmasında CRISP-DM süreç modelinin adımları izlenmiştir.



Şekil 1. CRISP-DM Süreç Modeli Aşamaları (Çınar ve Arslan, 2008).

1. BENZER ÇALIŞMALAR

Son yıllarda koroner arter hastalığı riskinin ve hastalık şiddetinin veri madenciliği algoritmaları kullanılarak incelenmesi ile ilgili çalışmalar yapılmaktadır. Anbarasi ve ark (2010) 13 değişkenli 909 hasta kaydı üzerinde yaptıkları çalışmada, öncelikle Genetik Algoritma kullanılarak ve değişken sayısı 6 'ya indirgenerek koroner arter hastalığı riskinin belirlenmesine daha çok katkıda bulunduğu düşünülen değişkenler belirlenmiştir. Daha sonra veri kümesine Naive Bayes, Clustering ve Karar Ağacı sınıflandırma yöntemleri uygulanmıştır. En yüksek doğruluk %99,2 oranı ile Karar Ağacı sınıflandırma yöntemi ile elde edilmiştir. Karar Ağacı algoritmasını Naive Bayes ve Clustering algoritmaları izlemiştir.

Chen ve ark (2011) çalışmalarında uzman hekimlere klinik karar almalarında destek olması amacıyla koroner arter hastalığı tahmin sistemi geliştirmişlerdir. Sistem iki aşamada geliştirilmiştir. Öncelikle UCI Machine Learning Repository' den alınan 14 değişken ve 303 hasta kaydından oluşan veri kümesine Yapay Sinir Ağları algoritması uygulanarak değişkenlere dayalı sınıflandırma yapılmıştır. Sınıflandırma algoritmasının doğruluk oranı yaklaşık %80 olarak saptanmıştır. İkinci aşamada, C ve C# programlama dilleri kullanılarak kullanıcı dostu bir arayüz ile Kalp Hastalığı Tahmin Sistemi Programı geliştirilmiştir. Program; klinik veri girişi, ROC (Alıcı işlem karakteristikleri, Receiver Operating Characteristic) eğrisi ve tahmin performans (işlem zamanı, doğruluk, duyarlılık, özgüllük) ve tahmin sonucu bölümlerinden oluşmaktadır.

Abdullah (2012) çalışmasında koroner arter hastalığı riskini incelemek için Cleveland kliniğinden **derlenmiş** 303 kayıt ve hedef değişkenle birlikte 14 değişkenden oluşan kalp hastalığı veri kümesini kullanılmıştır. Çalışmada öncelikle Parçacık Sürü Optimizasyon (PSO) algoritması kullanılarak değişken sayısı 14'den 9'a indirgenmiştir. Bu değişkenler; yaş, cinsiyet, göğüs ağrısı tipi, kolesterol seviyesi, açlık kan şekeri, istirahat ECG, maksimum kalp hızı, floeoskopi sonucu ve defekt tipidir. Çalışmada ayrıca indirgenmiş veri kümesi üzerinde sınıflandırma yapmak amacıyla J48 Karar Ağacı algoritması uygulanmış ve %60.74 doğruluk oranı elde edilmiştir.

Alizadehsani ve ark (2013), çalışmalarında değişken seçme ve veri kümesinden yeni değişkenler oluşturma yöntemleriyle koroner arter hastalığı riskinin belirlenmesinde kullanılan veri madenciliği algoritmalarının doğruluk oranını artırmayı amaçlamışlardır. Çalışma koroner arter hastalığı şüphesi ile kliniğe başvuran 303 hasta ve bunlara ait 54 değişken üzerinde yapılmıştır. Sıralı Minimum Optimizasyon (SMO), Naive Bayes, Bagging ve Sinir Ağları sınıflandırma algoritmaları veri kümesinin analizi için kullanılmıştır. En yüksek doğruluk oranı değişken seçimi ve değişken oluşturma yöntemleri ile birlikte Sıralı Minimum Optimizasyon algoritması ile %94.08 olarak elde edilmiştir.

Pandey ve ark (2013) çalışmalarında J48 Karar Ağacı algoritmasını kullanarak Kalp Hastalığı Tahmin Modelini geliştirmişlerdir. Geliştirilen modelde, J48 algoritması budanmamış, basit budama ve azaltılmış hata budaması yaklaşımı kullanılarak üç farklı yolla uygulanmış ve sonuçlar karşılaştırılmıştır. Çalışmada %75.73 doğruluk oranı ile en iyi sonuç azaltılmış hata budaması yaklaşımı uygulanan J48 Karar Ağacı algoritmasından elde edilmiştir.

Shafique ve ark (2015), yaptıkları çalışmada UCI Machine Learning Repository' den aldıkları 597 hasta kaydından oluşan veri kümesi üzerinde; Karar Ağacı, Yapay Sinir Ağları ve Naive Bayes algoritmalarını kullanarak kalp hastalığı riski açısından sınıflandırma yapmışlardır. Çalışmada, Naive Bayes sınıflandırma algoritmasının %82.914 ile en yüksek doğruluk oranına sahip olduğu saptanmıştır.

Verma ve ark (2016), koroner arter hastalığı riskini belirlemek amacıyla kullandıkları hibrid veri madenciliği modelinde değişken alt kümesi seçimi için korelasyon tabanlı Parçacık Sürü Optimizasyonu ve K-means Kümeleme algoritmalarını kullanmışlardır. Sonrasında modeli oluşturmak için Yapay Sinir Ağları (MLP), çoklu nominal lojistik regresyon (MLR), Bulanık Sırasız Kural Azaltma (FURIA) ve C4.5 algoritmaları kullanılmıştır. Çalışmacılar, geliştirdikleri hibrid modeli, 26 değişkenli 335 kardiyoloji hastasının kayıtları üzerinde test etmişlerdir. MLR algoritmasının kalp hastalığı riskinin belirlenmesinde %88,4 ile en yüksek orana sahip olduğu saptanmıştır.

Sharan ve Sathees (2016) yaptıkları çalışmada veri kümesi üzerinde kalp hastalığı riskinin tahmin edilmesi amacıyla 3 farklı karar ağacı algoritması uygulamışlardır. Bunlar; Sınıflandırma-Karar Ağacı (simple CART), J48 ve Naive Bayes (NB Tree) algoritmalarıdır. Analizler WEKA programı aracılığıyla yapılmıştır. Çalışmada işlem zamanı açısından en iyi performans J48 algoritması ile 0.08 saniye, sınıflandırma doğruluk oranı açısından en iyi performans ise Simple CART algoritması ile %92,2 olarak saptanmıştır.

2. UYGULANAN YÖNTEM

2.1 Veri Kümesi

Çalışmada, UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/index.html)' den alınan kalp hastalığı veri kümesi kullanılmıştır. Bu veri kümesi 4 ayrı veri tabanından oluşmaktadır. Bunlar; Cleveland, Hungary, Switzerland, ve the VA Long Beach' dir. Tüm veri setleri aynı değişkenleri içermektedir ve aynı formatta oluşturulmuştur. Bu proje kapsamında Cleveland Klinik veri kümesi kullanılacaktır. Veri kümesi 76 değişkeni içermektedir; ancak, bu veri kümesi ile yapılan tüm çalışmalar, bu değişkenlerden, hedef değişkeninde dahil olduğu 14 tanesinin en önemli değişkenler olduğunu göstermektedir. Cleveland Klinik veri kümesi 303 hasta kaydından oluşmaktadır. Veri kümesindeki değişkenler ve özellikleri Tablo 1'de verilmiştir.

Tablo 1. Kalp hastalıkları veri kümesi (heart-c.csv) değişkenleri

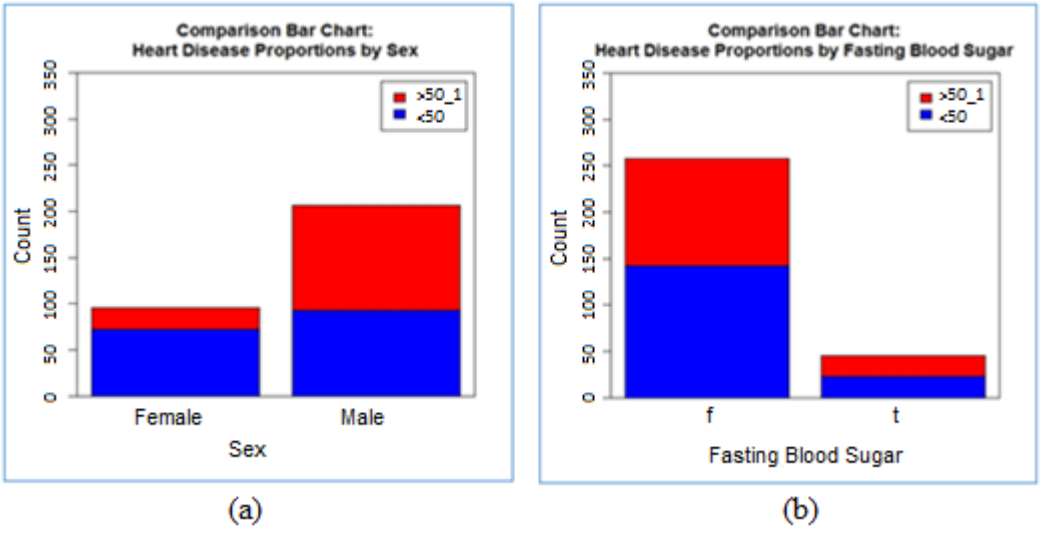
Değişken #	Değişkenler	Değişken Tipi
1	age	Sayısal
2	sex	Kategorik
3	cp	Kategorik
4	trestbps	Sayısal
5	chol	Sayısal
6	fbs	Kategorik
7	restecg	Kategorik
8	thalach	Sayısal
9	exang	Kategorik
10	oldpeak	Sayısal
11	slope	Kategorik
12	ca	Sayısal
13	thal	Kategorik
14	Num	Kategorik (Hedef Değişken)

2.2 Verinin Hazırlanması ve Analizi

Veri madenciliği sürecinin standardı olarak kabul edilen CRISP (DM)' nin 3. aşaması olan verilerin hazırlanması aşamasında, analiz yöntemi öncesinde veri kalitesinin yükseltilmesi amacıyla yapılan işlemlerden biri de eksik veya yanlış verilerin dikkate alınmasıdır. Çalışmada kullanılan veri kümesinin ca ve thal isimli 2 kategorik değişkeninde eksik veriler tespit edilmiştir. R programlama dili yardımıyla yapılan incelemede "ca" değişkende yaklaşık %2 ve "thal" değişkeninde yaklaşık %1 oranında kayıp veri tespit edilmiştir. Eksik verilerin temizlenmesinde mod işleminden faydalanılmıştır. Bu işlemde her bir kayıp veri yerine değişkende en çok tekrar eden değer atanarak eksik veriler tamamlanmaktadır. R programlama dili kullanılarak yazılan bir kod ile mod işlemi uygulanmış ve eksik veriler temizlenmiştir.

Verilerin hazırlanmasından sonra önemli ön bulgular R programı yardımıyla oluşturulan histogram ve çubuk (bar) grafikleri aracılığıyla analiz edilmiştir. Analizler literatür incelemesi ve kardiyoloji alanında uzman bir hekimin görüşleri alınarak yapılmıştır.

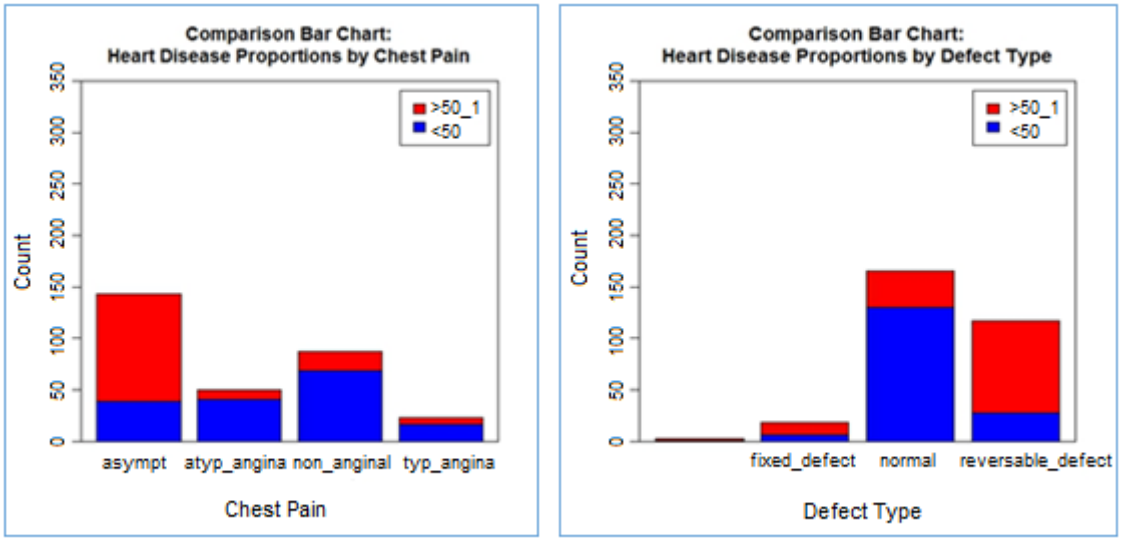
3.2.1. Kategorik Değişkenler İçin Hedef Değişkene Göre Çubuk (Bar) Grafiği Analizi



Şekil 2. (a) Cinsiyet ve koroner arterlerdeki daralma **(b)** Açlık kan şekeri ve koroner arterlerdeki daralma

Şekil 2 (a) cinsiyet ve koroner arterlerdeki ciddi daralma **(>%50)** ile karakterize kalp hastalığı arasındaki ilişkiyi göstermektedir. Erkeklerde kadınlara göre koroner arterlerdeki ciddi daralma oranı daha fazladır. Kadınlarda menopoz öncesi dönemde iskemik kalp hastalığının daha seyrek görülmesi büyük ölçüde östrojenin plazma lipit profili üzerine olumlu etkilerine bağlanmaktadır. Menopoz sonrası dönemde ise koroner arter hastalığı görülme sıklığı erkek ve kadınlarda eşitlenir (De Flines, Scheen,2009; Griffin vd., 2012).

Şekil 2 (b)' deki açlık kan şekeri ve koroner arterlerdeki ciddi daralma **(>%50**) ile karakterize kalp hastalığı arasındaki ilişkiye bakıldığında, açlık kan şekeri 120 mg/dl' den fazla olan grupta ciddi koroner arter daralma oranı daha fazladır. Diyabet hastalığı koroner kalp hastalığı için önemli bir risk faktörüdür. Diyabeti olan hastalarda kardiyovasküler olay (kalp krizi, göğüs ağrısı vb.) gelişme riski diyabeti olmayanlara göre 2-8 kat daha fazladır. Diyabetin neden olduğu damarsal komplikasyonlar sadece diyabet tanı kriterlerinin karşılandığı bireylerde değil, diyabetin gelişme sürecinde yer alan bozulmuş açlık glikozu safhasındaki bireylerde de görülebilmektedir (Onat vd., 2003).

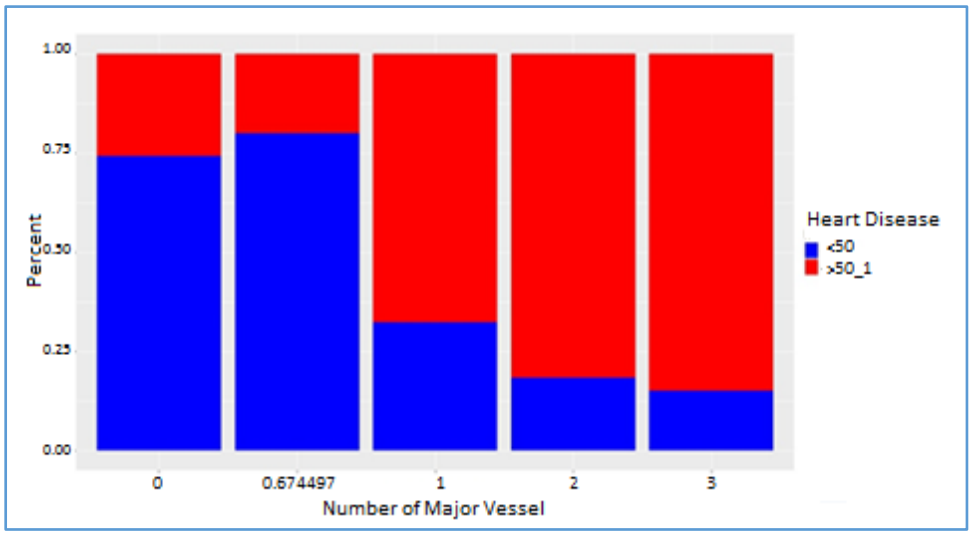


(a) (b)

Şekil 3. (a) Göğüs ağrısı tipi ve koroner arterlerdeki daralma **(b)** Defekt tipi ve koroner arterlerdeki daralma

Şekil 3 (a) göğüs ağrısı tipi ve koroner arterlerdeki ciddi daralma **(>%50**) ile karakterize kalp hastalığı arasındaki ilişkiyi göstermektedir. Asemptomatik bireylerde koroner arterlerdeki ciddi daralma oranının göğüs ağrısı olan hastalara oranla daha fazla olduğu görülmektedir. Bu sonuç asemptomatik hastalar için beklenin aksi yöndedir. Göğüs ağrısı olan 3 grup birlikte değerlendirildiğinde, tipik anjinası olan grupta beklendiği gibi koroner arterlerdeki daralma oranı daha fazla olduğu ve non- anjinal grupta ciddi daralma oranının en az olduğu görülmektedir (Griffin vd., 2012).
Şekil 3 (b)' deki Myokard Perfüzyon Sintigrafisinde (MPS) defekt tipi ve koroner arterlerdeki ciddi daralma **(>%50**) ile karakterize kalp hastalığı arasındaki ilişkiye bakıldığında MPS ile reversible defect tespit edilen grupta ciddi koroner arter hastalığı varlığı diğer iki gruba göre belirgin yüksek bulunmuştur. MPS' de fixed-defect tespit edilen kişi sayısı diğerlerinden az olmakla birlikte bu grupta koroner arter hastalığı oranı yarıdan fazladır. MPS sonucu normal olan grupta koroner arter hastalığı oranı düşüktür (Mann vd., 2014).

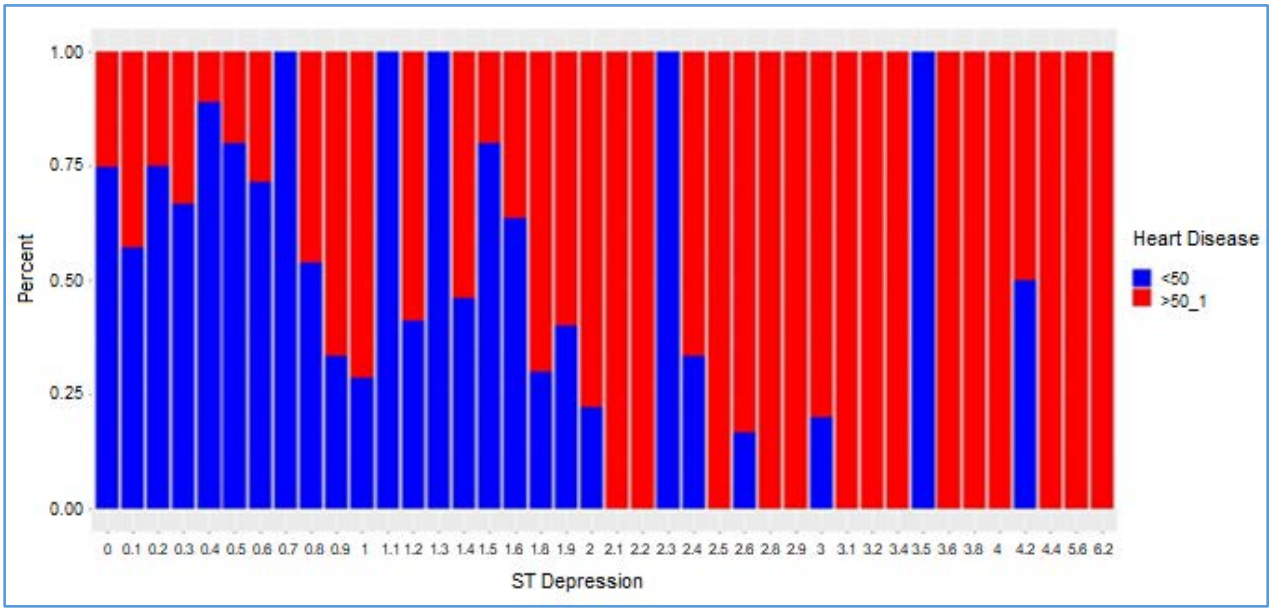
3.2.2. Sayısal Değişkenler için Histogram Analizi



Şekil 4. Koroner kalsifikasyon tespit edilen büyük damar sayısı değişkeninin histogram dağılımı

Şekil 4, Fluoroskopi ile koroner kalsifikasyon tespit edilen büyük damar sayısı değişkeninin histogram dağılımını göstermektedir. Floroskopi ile kalsifikasyon görüntülenen büyük damar sayısı arttıkça, koroner arterlerdeki ciddi daralma oranının da arttığı

görülmektedir. Koroner kalsiyum skorlaması koroner kalsiyumun koroner atherosklerotik plak için belirteç olarak kullanılabilmesi gözlemine dayanır. Çalışmalara göre koroner arter kalsifikasyonunun tamamen yokluğu önemli koroner darlığı olmadığını ve ileride koroner olay riskinin düşük olduğunu gösterir. Erkeklerde ki kalsifikasyon oranı kadınlara göre daha yüksek olduğu bilinmektedir. Herhangi bir yaşta kadınlarda erkeklere göre 5-7 kat daha düşüktür. Genellikle koroner arter kalsifikasyonu kadınlarda erkeklere göre 10 ila 15 yıl sonra gelişmektedir (Erdoğan vd., 2002).



Şekil 5. Koroner kalsifikasyon tespit edilen büyük damar sayısı değişkeninin histogram dağılımı

Şekil 5, ST depresyon değişkeninin histogram dağılımını göstermektedir. İstirahat EKG'sine göre ST depresyonun miktarı arttıkça, koroner arterlerdeki ciddi daralma oranının da arttığı görülmektedir. EKG'nin iskemiye en duyarlı kısmı ST segmentidir. Düşük efor düzeylerinde 2mm ve daha fazla ST depresyonu yaygın koroner arter hastalığı ve kötü prognozu düşündüren egzersiz testi bulgularıdır. Bu durum histogram verileri ile uyumludur (Mann vd., 2014).

3. ARAŞTIRMA BULGULARI

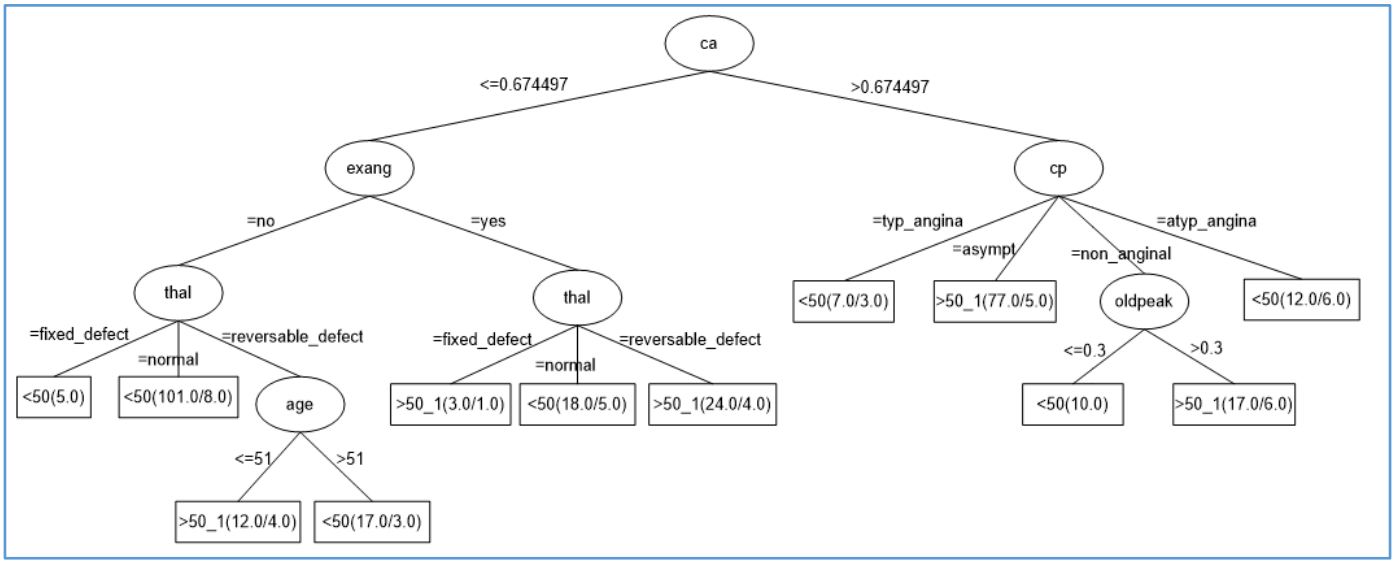
Sınıflandırma yöntemlerinin kıyaslanmasında sık kullanılan ölçülerden biri doğruluk ölçütüdür. Doğruluk ölçütünün gözlemlenmesinde genellikle karışıklık matrisi(confusion matrix) kullanılmaktadır. Karışıklık matrisi bileşenleri yapılan çalışmaya göre anlamlandırılmaktadır. Bu çalışma kapsamında incelenen kalp hastalığı sınıflandırma işlemi için karışıklık matrisi bileşenleri şu şekilde tanımlanmaktadır;

- **True Pozitif - TP:** Koroner arter hastalığına sahip kişiler koroner arter hastalığı var şeklinde doğru bir biçimde sınıflandırılmıştır.
- **False Pozitif - FP:** Sağlıklı kişiler yanlış bir biçimde koroner arter hastalığı var şeklinde sınıflandırılmıştır.
- **True Negatif - TN:** Sağlıklı kişiler doğru bir biçimde sağlıklı şeklinde sınıflandırılmıştır.
- **False Negatif - FN:** Koroner arter hastalığına sahip kişiler yanlış bir biçimde sağlıklı şeklinde sınıflandırılmıştır.

Çalışmada; 1R, Naive Bayes, J48 karar ağacı ve yapay sinir ağları olmak üzere 4 farklı sınıflandırma yöntemi üzerinde durulmuştur. Bu yöntemlerin her biri kalp hastalığı veri kümesine uygulanmıştır. Uygulama işlemi Weka kullanılarak 10 folds cross-validation değerine göre yapılmıştır. 1R, Naive Bayes ve Çok Katmanlı YSA doğrudan uygulanırken J48 karar ağacı üzerinde budama ile sadeleştirme ve ilişkili değişkenlerin çıkarılması ile yeniden düzenleme işlemleri yapılmıştır.

4.1. Budanmış J48 Karar Ağacı

Veri kümesi üzerine J48 Karar Ağacı doğrudan uygulandığında yaprak sayısı 32 ve boyutu 51 olan bir ağaç elde edilmiştir. Bu haliyle geniş ve karmaşık bir yapıya sahip olan karar ağacının yorumlanması zordur. Daha basit ve kolay yorumlanan bir karar ağacı elde edebilmek için budama işlemi gerekmektedir. Weka üzerinde Confidence Factor: 0.30 ve minimum number of object (M): 10 değerleri kullanılarak J48 karar ağacı budanmıştır. Bu işlem sonucunda oluşan ağacın yaprak sayısı 12 ve boyutu 19 olmuştur. Elde edilen budanmış J48 ağaç yapısı Şekil 6'da verilmiştir.



Şekil 6. Budanmış J48 Karar Ağacı

4.2. Düzenlenmiş J48 Karar Ağacı

Veri kümesi incelendiğinde "thalach" değişkeni az ve orta düzeyde hem "age" hem de "oldpeak" değişkenleri ile doğrusal ilişkiye sahip gibi görülmektedir (Tablo 2). Bu nedenle "thalach" değişkeni çıkartılarak tekrar analiz yapılmıştır. Buna göre J48 karar ağacında doğru sınıflama oranı biraz artmıştır.

Tablo 2. Sayısal değişkenler arasındaki korelasyonlar

	age	trestbps	chol	thalach	oldpeak
age	1.000	0.279	0.213	-0.399	0.210
trestbps		1.000	0.123	-0.047	0.193
chol			1.00	-0.010	0.054
thalach				1.000	0.344
oldpeak					1.000

4.3. Karşılaştırmalı Analiz

Sınıflandırma algoritmalarının karşılaştırılmasında Doğruluk (Accuracy), TP, FP, Kesinlik (Precision), F-ölçütü (F-measure), ROC ve zaman ölçütleri sıklıkla kullanılmıştır. Bu çalışmada uygulanan sınıflandırma algoritmaları da bu ölçütler açısından değerlendirilmiş ve elde edilen sonuçlar Tablo 3'te verilmiştir.

Tablo 3. Sınıflandırma algoritmalarının uygulama sonuçları

Algoritma	**Doğruluk (%)**	**TP**	**FP**	**Kesinlik**	**F-ölçüt**	**ROC**	**Zaman (saniye)**
1R	71,947	0,719	0,281	0,721	0,720	0,719	0,00
Naive Bayes	82,838	0,828	0,178	0,828	0,828	0,899	0,00
Budanmış J48 Karar Ağacı	77,557	0,776	0,233	0,775	0,775	0,834	0,00
Düzenlenmiş* J48 Karar Ağacı	78,548	0,785	0,223	0,785	0,785	0,837	0,02
Çok katmanlı YSA	83,498	0,835	0,171	0,835	0,835	0,893	0,78

*Age ve slope ile ilişkili thalach çıkartıldıktan sonraki analiz

Tablo 3' de yer alan uygulama sonuçları incelendiğinde, doğruluk ölçütü açısından en iyi sonuç % 83.498 değeri ile Çok Katmanlı YSA ile elde edilmiştir. Yapay sinir ağı modeli en yüksek doğruluk oranını verse de yorumlama ve uygulama açısından kapalı bir kutu gibidir. Naive Bayes algoritması ise basit olmasına rağmen en iyi doğruluk oranına sahip algoritmalardan biri olmuştur. J48 karar ağacı ise orta düzeyde doğruluk oranına sahip olmasına karşılık uzman hekimler ve araştırmacılara yorumlama imkânı sağlaması açısından önem arz etmektedir.

4. SONUÇ

Veri madenciliği algoritmaları koroner arter hastalığının tanımlanmasında ve risk faktörlerinin belirlenmesinde önemli rol oynamaktadır. Bu çalışmada koroner arter hastalığı riskinin belirlenmesinde 1R, Budanmış, Budanmamış ve Düzenlenmiş J48 Karar Ağacı, Naive Bayes ve Çok katmanlı YSA sınıflandırma yöntemleri kullanılmıştır. Sınıflandırma algoritmaları; doğruluk, TP, FP, Kesinlik, F-ölçüt, ROC ve zaman açısından karşılaştırılmıştır. Veri madenciliği sınıflandırma algoritmaları doğruluk açısından incelendiğinde, en iyi sonuç % 83.498 doğruluk oranı ile Çok katmanlı YSA sınıflandırma yönteminden elde edilmiştir. Yapay sinir ağı modeli en yüksek doğruluk oranı verse de yorumlama ve uygulama açısından kapalı bir kutu gibidir. Basit olmasına karşın Naive Bayes algoritması en iyi doğruluk oranına sahip algoritmalardan birisidir. J48 karar ağacı ise orta düzeyde doğruluk oranına sahip olmasına karşılık uzman hekimler ve araştırmacılara yorumlama imkânı sunabilmektedir. Bu nedenlerle uygulamalarda hangi modelin tercih edileceği uygulamanın özel durumları da dikkate alınarak belirlenebilir.

Bu çalışmanın sonuçlarının koroner arter hastalığı şüphesi ile kliniğe başvuran hastaların tanı ve tedavi sürecinde ve invaziv prosedür uygulanacak doğru hasta grubunun seçilmesinde kardiyoloji alanında çalışan uzmanların klinik kararlarına rehberlik edeceği düşünülmektedir. Ayrıca, geliştirilen veri madenciliği modeli ile medikal hatalar, istenmeyen uygulama çeşitlilikleri, medikal maliyet azaltılabilmekte, dolayısıyla hasta güvenliği ve yaşam kalitesi artırılmaktadır. Bu çalışmada sınıflandırma algoritmaları veri kümesindeki 14 değişken üzerinde uygulanmıştır. Gelecek çalışmalarda optimizasyon algoritmaları kullanılarak veri kümesindeki değişkenler daha detaylı incelenerek sınıflandırma algoritmaları uygulanabilir. Ayrıca modellerin uygulanmasında TP ve FN oranlarının doğurabileceği sonuçlar da dikkate alınarak risk yönetimi açısından değerlendirmeler yapılabilir. Diğer yandan ileriki çalışmalar için her iki oranı, yani TP ve FN, birlikte optimize edecek algoritmalar geliştirilebilir.

REFERANSLAR

Abdullah, A. S. (2012). A Data Mining Model to Predict and Analyze the Events Related to Coronary Heart Disease using Decision Trees with Particle Swarm Optimization for Feature Selection. International Journal of Computer Applications, 55(8).

Alizadehsani, R., Habibi, J., Hosseini, M. J., Mashayekhi, H., Boghrati, R., Ghandeharioun, A., Sani, Z. A. (2013). A Data Mining Approach for Diagnosis of Coronary Artery Disease. Computer Methods and Programs in Biomedicine, 111(1), 52-61.

Alizadehsani, R., Hosseini, M. J., Sani, Z. A., Ghandeharioun, A., & Boghrati, R. (2012). Diagnosis of Coronary Artery Disease Using Cost-Sensitive Algorithms. In Data Mining Workshops (ICDMW), 2012 IEEE 12th International Conference on (pp. 9-16). IEEE.

Anbarasi, M., Anupriya, E., & Iyengar, N. C. S. N. (2010). Enhanced Prediction of Heart Disease with Feature Subset Selection Using Genetic Algorithm. International Journal of Engineering Science and Technology, 2(10), 5370-5376.

Avşar, A., Önder, Akçı., Beyter, M. E. (2011). Aterosklerozun Patogenezi (Aterogenez). Turkiye Klinikleri Journal of Cardiology Special Topics, 4(2), 1-15.

Cardiovascular diseases (CVDs),
http://www.who.int/mediacentre/factsheets/fs317/en/ (Erişim tarihi; Ekim, 2016).

Ceylan, Y., Kaya, Y., & Tuncer, M. (2011). Akut Koroner Sendrom Kliniği ile Başvuran Hastalarda Koroner Arter Hastalığı Risk Faktörleri. Van Tıp Dergisi, 18(3), 147-54.

Chen, A. H., Huang, S. Y., Hong, P. S., Cheng, C. H., & Lin, E. J. (2011, September). HDPS: Heart Disease Prediction System. In Computing in Cardiology, 2011 (pp. 557-560). IEEE.

Çınar, H. ve Arslan, G., 2008. "Veri madenciliği ve CRISP-DM yaklaşımı", XVII. İstatistik Araştırma Sempozyumu, 304-314, Ankara.

De Flines, J., & Scheen, A. J. (2009). Management Of Metabolic Syndrome And Associated Cardiovascular Risk Factors. Acta Gastro-Enterologica Belgica, 73(2), 261-266.

El-Bialy, R., Salamay, M. A., Karam, O. H., & Khalifa, M. E. (2015). Feature Analysis of Coronary Artery Heart Disease Data Sets. Procedia Computer Science, 65, 459-468.

Erdoğan, N., Altın, L., Altunkan, Ş. (2002). Elektron Beam Tomografi ile Koroner Arterlerdeki Kalsiyum Miktar›n›n Saptanması. Tanısal ve Girişimsel Radyoloji, 8, 533-537.

Griffin, B. P., Callahan T.D., Menon, V.(Eds.). (2012). Manual of Cardiovascular Medicine. Lippincott Williams & Wilkins.

Mann, D. L., Zipes, D. P., Libby, P., & Bonow, R. O. (2014). Braunwald's Heart Disease: a Textbook of Cardiovascular Medicine. Elsevier Health Sciences.

Ökçün, B., Gürmen, T. (2007). Koroner Anjiyografi Komplikasyonları ve Tedavisi. Turkiye Klinikleri Journal of Internal Medical Sciences, 3(42), 48-72.

Palaniappan, S., & Awang, R. (2008). Intelligent Heart Disease Prediction System Using Data Mining Techniques. In Computer Systems and Applications, 2008. AICCSA 2008. IEEE/ACS International Conference on (pp. 108-115). IEEE.

Pandey, A. K., Pandey, P., & Jaiswal, K. L. (2013). A Heart Disease Prediction Model Using Decision Tree. IUP Journal of Computer Sciences, 7(3), 43.

Shafique, U., Majeed, F., Qaiser, H., & Mustafa, I. U. (2015). Data Mining in Healthcare for Heart Diseases. International Journal of Innovation and Applied Studies, 10(4), 1312.

Sharan M.L, Sathees, K.B. (2016). Analysis of Cardiovascular Heart Disease Prediction Using Data Mining Techniques. Analysis, 4(1), 55-58.

Soni, J., Ansari, U., Sharma, D., Soni, S. (2011). Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction. International Journal of Computer Applications, 17(8), 43-48.

Srinivas, K., Rani, B. K., & Govrdhan, A. (2010). Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks. International Journal on Computer Science and Engineering (IJCSE), 2(02), 250-255.

Onat, A., Sansoy, V., Soydan, İ., Tokgözoğlu, L., & Adalet, K. (2003). TEKHARF, Oniki Yıllık İzleme Deneyimine Göre Türk Erişkinlerinde Kalp Sağlığı. İstanbul Türkiye, 12-4.

Verma, L., Srivastava, S., Negi, P. C. (2016). A Hybrid Data Mining Model to Predict Coronary Artery Disease Cases Using Non-Invasive Clinical Data. Journal of Medical Systems, 40(7), 1-7.

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a Standard Process Model for Data Mining. In Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining, 29-39.

Wong, N. D. (2014). Epidemiological Studies of CHD and the Evolution of Preventive Cardiology. Nature Reviews. Cardiology, 11(5), 276.