AdaBoost ile Kalp Krizi Risk Tespiti

Faruk Bulut

Katip Çelebi Üniversitesi, Mühendislik Mimarlık Fakültesi, Bilgisayar Müh. Bölümü, Çiğli, İzmir,
Tel: +90 232 329 3535
faruk.bulut@ikc.edu.tr

Geliş / Received: 23 Nisan (April) 2016
Kabul / Accepted: 7 Aralık (December) 2016
DOI: 10.18466/cbayarfbe.280652

Özet

Dünyada genelinde en çok ölüme sebep olan hastalık türü Kalp ve Damar hastalıklardır. Çalışmamız, kolektif yapay sınıflandırma yöntemlerini kullanarak bireyde kalp krizi risk oranını belirlenmesi üzerinedir. Bu amaçla kalp krizi veya bir kalp rahatsızlığı geçirmiş hastalara ait gerekli bir takım tıbbi veriler resmi izinlerle hastanelerden alınmıştır. Bu veriler bir veri setine aktarılarak sınıflandırıcı algoritmalarda kullanılmıştır. Yapılan deneysel uygulamalarda başarı oranı yüksek AdaBoost kolektif sınıflandırıcısı ile bireyde olası kalp hastalıkları riski erkenden tespit edilebilmiştir. Çapraz geçerleme işlemleri ile önerilen model başka sınıflandırıcılara kıyaslanmış ve daha yüksek sınıflandırma performansı elde edildiği gözlemlenmiştir. Tasarlanan erken uyarı sistemi ile bu sayede muhtemel kalp krizine karşı önceden önlem alınabilmesi sağlanmış olmaktadır.

Anahtar Kelimeler — Kalp krizi tespiti, Regresyon, AdaBoost, Klinik Karar Destek Sistemleri (KKDS).

Determining Heart Attack Risk Ration Through AdaBoost

Abstract

Cardiovascular diseases are the most common cause of death all over the world. This study is based on to predict the heart attack risk of an individual by using Artificial Ensemble Classification algorithms. For this reason, relevant clinical data has been obtained by the official permissions from the hospitals where there are some patients who have had heart attacks before. A dataset has been constructed using this collected data in order to use in the classifiers. In the practical applications, heart attack risks can be recognized for an individual by using a powerful ensemble classifier, AdaBoost. Furthermore, it is detected that proposed technique explicitly shows a high-performance in validating when compared with other classifiers. Therefore, this suggested early warning system allows taking required precautions before a possible heart attack.

Keywords — Heart attack detection, regression, AdaBoost, Clinical Decision Support System (CDSS).

1 Giriş

Kalp sadece bir yumruk büyüklüğünde olmasına rağmen insan vücudundaki en güçlü kastır. Yaklaşık olarak gebeliğin dördüncü haftasında atmaya başlayan kalp, yetişkin ve sağlıklı bir bireyde ortalama olarak günde 100000 kere atar. 70 yıllık bir ömürde iki buçuk milyar kez çarpan kalp, her defasında vücudun tüm bölgelerine kan pompalar. Bu olağanüstü sistem önlenebilir ve tedavi edilebilir birçok faktöre bağlı olarak, çeşitli bozulma ve deformasyonlara açık bir yapıdadır. Kalp krizi koroner arterlerde yani kalp kasına kan ile oksijen taşıyan atardamarların birinde oluşan tıkanıklık ve tıkanıklığın oluşturduğu komplikasyonlara denir. Oluşan oksijen kesintisi kalp dokusunda hasara ya da dokunun kalıcı olarak zarar görmesine neden olabilmektedir ve bu durum tıbbi müdahale gerektirmektedir [1].

Kalp ve Damar Hastalıkları (KDH), Dünya Sağlık Örgütü (WHO) verilerine bakıldığında ölüm nedenleri içerisinde yaklaşık %30'luk oranla birinci sıradadır. Türkiye'de ise bu oran yaklaşık % 47 civarındadır. Ülkemizde yetişkin bireylerin yaklaşık olarak yarıya yakını kalp damar hastalıkları tehlikesi ile karşı karşıyadır [2].

Birinci ölüm nedeni olan ve yaşam kalitesini düşüren kalp damar hastalıklarının başlıca nedenleri şu şekilde sıralanabilir [3]:

- Düşük iyi kolesterol (HDL),
- Yüksek kötü kolesterol (LDL ve trigliserit),
- Hipertansiyon,
- Damar tıkanıklığı ve buna benzer hastalıklar,
- Ritim bozukluğu ve senkop olayı gibi çeşitli kalp hastalıkları,
- Daha önceden geçirilmiş kalp krizi,
- İleri yaş (kadınlarda 50, erkelerde ise 40 yaş üstü),
- Diyabet,
- Obezite,
- Aşırı alkol tüketimi,
- Sigara,
- Uyuşturucu,
- Stres,
- Yoğun yaşam temposu.

Kalp hastalıklarına ait teşhis işlemi günümüzde sadece doktorlar tarafından yapılmaktadır. Sağlık bilişimi alanında yapılan yapay karar sistemleri ile hekimler üzerindeki bu yük kısmen de olsa kaldırılmaktadır. Her ne kadar yapay karar destek sistemleri %100 güvenilir sonuçlar vermese de sağlık personelini bilgilendirici ve uyarıcı bir yapıda olması bu tür çalışmaları önemli kılmaktadır. Aynı zamanda hekimler tarafından verilen yanlış tıbbi yorumlar ve teşhisler, korkunç sonuçlar ortaya çıkarabilmektedir. Karar destek sistemleri ise bu noktada tıbbi teşhislerin yeniden değerlendirilmesini mümkün kılmaktadır.

Bilgisayar ve istatistik bilimleri açısından incelendiğinde, sağlık sektörü genellikle veri zengini ancak bilgi yoksunu olarak tanımlanabilir. Sağlık sistemleri içinde mevcut verilerde bir zenginlik olduğunu apaçık ortadadır. Ancak, veriler içerisinde gizli kalmış bilgi ve ilişkileri keşfetmek için etkili analiz yöntemlerinin yeterince kullanılmadığı da bir gerçektir.

Son yıllarda, örüntü tanıma (Pattern Recognition), veri madenciliği (Data Mining) ve yapay öğrenme (Machine Learning) teknikleri sağlık bilişiminde birçok alana uygulanmıştır. Bilgi keşfi ve veri madenciliği birçok sektöre ve iş alanına uygulanabilmektedir. Değerli bilgi, sağlık sisteminde veri madenciliği tekniklerinin uygulanması yolu ile tespit edilebilir. Yapay karar destek sistemlerinin sağlık alanındaki uygulamaları günden güne artmaktadır. Hekimler tarafından yapılan teşhis ve tedavilere yardımcı olmak ve insan kaynaklı hataları engellemek için yapay zekâ tabanlı karar destek sistemlerinin kullanımı da yaygınlaşmaktadır.

Kalp ve damar hastalıklarını önceden tespit etmeye yönelik algoritmik çalışmaların yanında değişik istatiksel çalışmalar da gerçekleştirilmiştir. Yapay zekâ ve örüntü tanıma disiplini altında yapılan birbirinden farklı oldukça fazla çalışma vardır. M. Anbarasi ve arkadaşları yaptıkları bir çalışmada Genetik Algoritmalar (GA) yardımıyla kalp rahatsızlıklarına neden olan en önemli faktörleri tespit etmeye çalışmışlardır. Çalışmalarında kalp hastalıklarına neden olan 13 faktörün içinden GA ile en etkenleri tespit edilmiştir. Ayrıca bu faktörler değişik sınıflandırma algoritmalarında kullanılmıştır [4].

K. Srinavas ve arkadaşları ile S. Palaniappan isimli bir araştırmacı veri madenciliği tekniklerini kullanarak sağlık durumunu ve kalp krizini tahmin etmeye yönelik iki ayrı çalışma gerçekleştirmişlerdir. Hazır ve ortak kullanıma açık olan UCI [5] veri setlerinden Heart-c, Heart-h ve Heart-statlog seçilmiş ve Weka yazılımı kullanılarak bu veri setleri üzerinde bir takım sınıflandırma işlemleri yapılmıştır. Çalışmalarda kullanılan verilerin özgün olmamasına ve klasik sınıflandırıcıların kullanılmış olmasına rağmen, bu çalışmalar alanında bir ilk olma niteliği kazanmıştır [6], [7]. Bu çalışmalara benzeyen başka bir çalışma da D. Chandna tarafından yapılmıştır. UCI veri setleri ile yapılan deneylerde Bilgi Kazancı (Information Gain) ile ANFIS yöntemi (Adaptive Neural Fuzzy Inference System) kullanılarak yüksek performans elde edildiği belirtilmiştir [8].

Yapılan başka bir çalışmada cep telefonu tabanlı giyilebilir bir düzenek ile herhangi bir zamanda ve yerde, gerçek zamanlı olarak, EKG sinyalleri sürekli olarak izlenmiş ve oluşturulan sinyal kayıtları ile anormal durumlar tespit edilmeye ve önlenmeye çalışılmıştır. Çalışmada kapsamlı bir literatür

incelemesi yapılmış ve gelişmiş makine öğrenmesi algoritmaları Android ortamında kullanılmıştır. Gömülü bir elektronik sistem ve cep telefonu yardımıyla ile bir EKG izleme ve ritim sınıflandırma sistemi inşa edilerek sisteme taşınabilir bir özellik kazandırılmıştır [9].

M. Karakoyun ve M. Hacıbeyoğlu isimli araştırmacılar ise, biyomedikal veri kümelerini sınıflandırma metotlarında kullanarak istatistiksel olarak karşılaştırmalar yapmışlardır. Ayrıca çalışmada makine öğrenmesi algoritmalarından birçok öğrenme algoritması denenmiştir. Yaptıkları deneysel ve istatistiksel çalışmalarda Yapay Sinir Ağları algoritmasının daha yüksek oranda başarılı sonuçlar verdiği görülmüştür. Ayrıca *Big Data* grubuna girmeyen küçük ve orta büyüklükteki veri kümeleri için *k* En Yapın Komşuluk (*k* Nearest Neighbors, *k*-NN) metodunun daha yüksek performansta çalıştığı görülmüştür [10].

2015 yılında yayınlanan bir makalede Ateş Böceği optimizasyon algoritması kullanılarak kalp hastalıklarının tespiti üzerinde teorik ve deneysel bir çalışmalardan bahsedilmektedir. Ortak kullanıma açık veri setleri kullanılarak yapılan çalışmada boyut indirgeme işlemi IT2FLS (Interval Type-2 Fuzzy Logic System) ve Ateş Böceği (Fire Fly) optimizasyonu ile gerçekleştirilmiştir. Tekil öğrenici olan Destek Vektör Makineleri, Naive Bayes ve Yapay Sinir Ağları ile yapılan sınıflandırma işlemlerine göre önerilen boyut indirgemeli sistemin daha az karmaşıklık ile daha yüksek tahmin başarısı elde edildiği belirtilmiştir [11].

Aşırı Yapay Öğrenme yöntemi (Extreme Learning Machine, ELM) kullanarak kalp krizi tespitini yapan bir çalışmada ise 300 adet hastanın verisi Cleveland klinik kurumundan (Cleveland Clinic Foundation) alınmıştır. Hastalara ait cinsiyet, yaş, kolesterol ve kan şekeri düzeyleri gibi faktörler ele alınarak yapılan deneysel çalışmalarda %80 başarı oranı elde edildiği belirtilmiştir [12].

Vinitha Dominic ve çalışma arkadaşları tarafından 2015 yılında yapılan detaylı bir çalışmada popüler denetimli öğrenicilerin kalp rahatsızlıkları üzerinde performanslarının analiz edilmesi ve kalp hastalıklarına etki eden faktörlerin sınıflandırılması üzerine bir çalışma gerçekleştirilmiştir. Çalışmada kalp hastalıklarını tetikleyen olası 75 anatomik faktör ele alınmıştır. Her bir kalp hastalığı türüne etki eden anatomik faktörlerin belirlenmesi bilgi kazancı (Information Gain) ölçütüyle ve bu faktörlerin gruplandırılması denetimli öğreniciler tarafından yapılmıştır [13].

R. Kavitha ve T. Christopher tarafından 2016 yılı içerisinde yayınlanan bir çalışmalarında denetimli ve denetimsiz öğreniciler yardımıyla tıbbi cihazlardan alınan verileri kullanarak kalp hastalıklarının tespiti ve sınıflandırılması üzerine bir çalışma gerçekleştirilmiştir [14]. Elektrokardiyogram ya da EKG olarak isimlendirilen cihaz, cilde yapıştırılan elektrotlar aracılığı ile kalbin elektriksel aktivitesini (kalp atışlarının ritmini ve frekansını) ve HRV (Heart Rate Variability) sinyalleri grafiksel değerlerle vermektedir. Elde edilen etiketli sinyaller, Destek Vektör Makineleri gibi genel kurallar üreten bir sınıflandırıcılardan ziyade daha yüksek sınıflandırma başarısı için Geliştirilmiş Aşırı Yapay Öğrenme (Improved Extreme Learning Machine, IELM) metotlarına aktarılmıştır. Kullanılan IELM mekanizmasında BFO (Bacterial Foraging Optimization) en iyileştirme yöntemi kullanılarak sınıflandırıcı için en uygun parametrelerin belirlenmesi ve veri setindeki en iyi özniteliklerin seçilme işlemi yapılmıştır. Diğer bir deyişle EKG cihazından gelen lineer ve lineer olmayan HRV sinyalleri ve öznitelik seçimi BFO sayesinde yapılmıştır.

Ayrıca bu çalışmada KFCM (Kernel Fuzzy C-Means) ile başarılı bir kümeleme ve sınıflandırma gerçekleştirilmiştir.

Son aylarda aynı alanda yapılan başka bir çalışmada ise diğerlerinden farkı olarak Gri Kurt Optimizasyon (Gray Wolf Optimization, GWO) algoritması ile Yapay Sinir Ağlarının hibritleştirilmesinden bahsedilmektedir [15]. Stokastik bir arama algoritması olan GWO'nun çalışma prensibi, gradient tabanlı bir geri yayılım metodu üzerine dayalıdır. Önerilen bu hibrit model Yapay Sinir Ağlarının daha az iterasyon ile en iyi şekilde eğitilmesini sağlamıştır. Bu sayede standart geri yayılımlı Yapay Sinir Ağlarına göre hastalık teşhisinde daha az Ortalama Karekök Hata ile daha yüksek sınıflandırma başarısı elde edilmiştir.

Literatürde var olan çalışmaların bazıları hazır veri setleri üzerinde belirli sınıflandırıcı yöntemlerinin uygulanması ve elde edilen değerlerin

karşılaştırılması üzerinedir. Bazı çalışmalar ise veri setinde bulunan özniteliklerden en etkili olanların seçilmesi ve daha az karmaşıklık ile sınıflandırma yapılabilmesi üzerindedir. Aynı şekilde EKG sinyallerinin analiz edilerek kalp hastalıkların tespit edilmesi üzerine değişik çalışmalar da mevcuttur. Bazı makalelerde ise farklı kalp hastalıklarını etkileyen faktörlerin gruplanması üzerine yöntemler sunulmuştur. Önerilen bu çalışma, diğer çalışmalarla kıyaslandığında iki noktada farklılık göstermektedir. Birincisi eğitim setinin Anadolu insanının genetik, biyolojik, beslenme ve yaşam alışkanlıklarını temel alarak hazırlanan bir anket ile toplanan verilerden oluşmasıdır. İkincisi ise regresyon yöntemi kullanılarak yüzdelik olarak kalp krizi risk oranın hesaplanması üzerindir.

Kalp krizi riskinin oransal olarak bir yazılım tarafından tespit edilmesine yönelik olan çalışmamızın özgünlük ve güvenilirlik kazanabilmesi için verilerin güncel, gerçek ve güvenilir olması amaçlanmıştır. Bu nedenle ilgili kamu hastanelerine resmi müracaat ile başvurulmuştur. Alınan resmi izin belgeleriyle devlet ve üniversite hastanelerinde tedavi gören hastalara hazırlanan anketler uygulanmıştır. Hastalardan elde edilen veriler bir veri setine reel sayı değerlerine dönüştürülerek aktarılmıştır. Oluşturulan veri seti ile kolektif sınıflandırıcılar eğitilmiştir. Deneysel uygulamalarda ve çapraz geçerleme işlemlerinde kullanılan yöntemler ile güvenilir sonuçların elde edilmiştir.

Çalışmamızın geriye kalan kısmında dört bölüm daha vardır. İkinci bölümde model oluşturmak için kalp krizine neden olan faktörlerin belirlenmesine, anket sorularının hazırlanmasına ve veri setinin sayısal değerler ile ifade edilme biçimine yer verilmiştir. Üçüncü bölümde ise kullanılan kolektif sınıflandırma ve regresyon yöntemlerine, dördüncü bölümde deneysel sonuçlara ve son bölümde de çalışmanın değerlendirilmesine yer verilmiştir.

2 Model Oluşturma ve Veri Toplama

Önerilen modelin çalışması temel olarak dört ana adımdan oluşmaktadır. Bunlar:

1. Kalp krizine neden olan risk faktörlerinin belirlenmesi ve anket sorularının hazırlanması,
2. Anketin hastanedeki hastalara uygulanarak verilerin toplanması,
3. Kolektif sınıflandırma ve regresyon algoritmaların uygulanması,
4. Sonuçların elde edilmesi ve yorumlanması.

2.1 Anket Oluşturma

Anadolu insanının yaşamsal alışkanlıklarına ve psikolojik durumuna özgü hazırlanan ankette kalp hastalıklarına neden olan faktörler altı ana risk grubuna ayrılmıştır:

1. Temel Demografik özellikler
2. Bireysel ve Yaşamsal Alışkanlıklar
3. Genel Sağlık Durumu
4. Spor ve Hareketlilik
5. Beslenme Alışkanlıkları
6. Psikolojik Durumu

2.1.2 Demografik özellikler

Bu alanda bireye ait cinsiyet, yaş, kilo, boy ve medeni durum gibi temel sorular sorulmaktadır. Vücut Kitle İndeksi (VKİ), boy, kilo ve cinsiyet bilgisi kullanılarak elde edilir ve bireydeki obezite durumunu en iyi belirleyen en iyi ölçüttür. VKİ, esasen bireydeki kalp hastalıklarını tetikleyen önemli bir faktördür. VKİ, aşağıdaki 1 numaralı formül ile hesaplanır.

$$VKI = \frac{Ağırlık}{boy^2} \quad (1)$$

VKİ sonucuna göre bireyin obezlik durumu Çizelge 1'deki verilere göre kategorilere ayrılır.

Çizelge 1. VKİ'ye göre obezite.

Vücut Kitle İndeksi (VKİ)	Sonuç
18.50 kg/m²'den düşük	Zayıf
18.50-24.99 kg/m² arası	Normal kilolu
25-29.99 kg/m² arası	Fazla kilolu
30-39.99 kg/m² arası	Obez (şişman)
40 kg/m²'den büyük	İleri derecede obez

2.1.2 Yaşamsal Alışkanlıklar

Bu bölümde bireyin sigara alışkanlığı, alınan uyku saati, haftada tüketilen alkol miktarı gibi sorular ile bireyin günlük yaşamsal alışkanlıkları tespit edilmeye çalışılmıştır. Kalp krizine geçirilmesine neden olan en önemli faktörlerin bu bölümde sorulan sorular olduğu literatürde yayınlanan çalışmalarda belirtilmiştir [3]. Sigara bağımlılığı kalp damar hastalarının yaklaşık %80'inde görülmektedir. Sigara bağımlılığı yüksek kan basıncına ve yüksek kolesterol seviyesine neden olduğu için önemli bir risk faktörüdür. Bu faktörler koroner kalp hastalığının gelişmesinde büyük rol oynamaktadır.

Soru 1. Sigara içiyor musunuz?
a) Sürekli olarak kullanıyorum
b) Yeni bıraktım
c) Yıllar önce bıraktım
d) Ara sıra içiyorum
e) İçmiyorum

Soru 2. Günde kaç saat uyuyorsunuz?
a) 10 saatten fazla
b) 8-10
c) 6-8
d) 4-6

Soru 3. Haftada ne kadar alkol tüketiyorsunuz? (1 bardak=200 ml)
a) 600 ml den fazla
b) 400-600 ml arası
c) 200-400 ml arası
d) Tüketmiyorum

2.1.3 Sağlık Durumu

Bu bölümde bireye ait nabız değeri, ailede kalp hastalığından dolayı erken yaşta ölüm olup olmaması, diyabet durumu, iyi ve kötü huylu kolesterol seviyeleri, genel tansiyon değerleri, doğum kontrol hapı kullanım durumu ve menopoz durumu sorulmuştur. Bazı bilimsel çalışmalarda, nabız değerlerindeki değişik zamanlardaki farklılıkların, kolesterol ve tansiyondaki düzensizliğin bir atardamar hastalığı olan aterosklerozun bir nedeni olabileceğini göstermektedir. Bu rahatsızlık zaman içerisinde kalp krizini tetikleyici bir etkene dönüşebilmektedir [3]. Ayrıca kandaki kolesterol seviyesinin yükselmesi koroner arter hastalığı riskinin de artması anlamına gelmektedir. Kolesterol seviyesi fiziksel aktivite, beslenmede düzenlilik ve ilaçlar ile düşürülebilmektedir [16].

Soru 4. Nabız değeriniz nedir?
a) 120 den fazla
b) 100-120 arası
c) 80-100 arası
d) 60-80 arası

Soru 5. Menopozda mısınız?
a) Evet
b) Hayır değilim
c) Genç yaşta menopoza girdim
d) Hormon tedavisi kullandığım için menopoza giremedim
e) Ben erkeğim

Soru 6. Kontraseptif (doğum kontrol) hap kullanıyor musunuz?
a) Evet
b) Hayır
c) Ben erkeğim

Soru 7. Ailenizde kalp hastalıklarından ötürü erken yaşta ölen var mı? (45 yaş altı)
a) Evet
b) Hayır

Soru 8. İyi huylu kolesterol (HDL) seviyesi nedir?
a) 40'dan az
b) 40-60 arası
c) 60'dan fazla

Soru 9. Kötü huylu kolesterol (LDL) seviyesi nedir?
a) 100'den az
b) 100-150 arası
c) 150'den fazla

Soru 10. Diyabetiniz var mı?
a) Evet
b) Hayır

Soru 11. Tansiyon değerleriniz nedir?
Büyük ……...
Küçük ……...

2.1.4 Spor ve Hareketlilik

Anketin bu bölümdeki sorular, bireyin bedensel hareketliliğine ve düzenli olarak yaptığı sportif faaliyetlere göre kalp hastalığına olan yatkınlığını

incelemek ve araştırmak için sorulmuştur. Yapılan bilimsel çalışmalarda, bireyin sıkça yürüyüş yapması veya düzenli olarak bir spor dalı ile ilgilenmesi kalp krizi riskini düşürmektedir [3], [16].

Soru 12. Günde ortalama kaç dk. spor yapıyorsunuz?
a) 45 dk.'den fazla
b) 30-45 dk.
c) 15-30 dk.
d) 15 dk.'den az

Soru 13. Gün içinde ne kadar yürüyorsunuz?
a) 3 saatten fazla
b) 2-3 saat
c) 1-2 saat
d) 1 saatten az

Soru 14. Bel çevreniz kadınsanız 80cm ya da erkekseniz 94cm'den fazla mı?
a) Evet
b) Hayır

Soru 15. Günde ortalama ne kadar oturuyorsunuz?
a) 12 saatten fazla
b) 8-12 saat
c) 4-8 saat
d) 4 saatten az

2.1.5 Beslenme Alışkanlıkları

Çalışmanın bu bölümünde var olan bilimsel çalışmalardan elde edilen bilgilere göre sorular hazırlanmıştır.

Bu bölümde bulunan "Haftada kaç öğün balık yersiniz?" sorusu, ω3 yağ asitlerinin ne sıklıkta alındığını anlamak için sorulmuştur. Bilindiği üzere ω3, kalp dostu bir yağ asidi türüdür.

Ayrıca kan kolesterol miktarını sanıldığı kadar yükseltmeyen yumurta, her yaştaki bireylerin tüketmesi gereken, besleyici değeri çok yüksek bir hayvansal gıda kaynağı olduğu belirtilmiştir. Bu nedenle haftada ne kadar yumurta yenildiği sorulmuştur.

"Pişirilen yemeklerde en çok hangi yağ türünü kullanıyorsunuz?" sorusu ilgili kişinin margarin tüketimini tespit etmek için sorulmuştur. Bilindiği üzere margarinlerde kalp damar hastalıklarını tetikleyen oldukça yüksek oranda trans yağ asidi vardır. Trans yağ asitleri, LDL kolesterolünü artırırken HDL kolesterolünü de azaltır. Ticari mutfaklarda, hazır yemekler kullanılan bu yağların tüketimi azaltılmalıdır [17]. Bunun yanında zeytinyağı tüketimimin kalp sağlığı açısından oldukça faydalı bir besin türü olduğu bilinmektedir [18].

Ayrıca "Besinleri kavurma ve kızartma yöntemleri ile mi pişirirsiniz?" sorusu sorulmuştur. Yapılan bazı çalışmalarda besinlerin kızartma yerine haşlama, fırınlama ve buhar şeklinde pişirilmesinin kalp krizini önleyebileceği ifade edilmektedir [19].

Bu bölümde sorulan soruların tamamı şu şekildedir:

Soru 16. Sakatat ve şarküteri ürünlerini ne sıklıkta tüketirsiniz?
a) Haftada birden fazla
b) Haftada bir
c) Ayda bir
d) Daha az

Soru 17. Haftada kaç yumurta yersiniz?
a) 4'den fazla
b) 2-3 arası
c) 1 ya da hiç

Soru 18. Haftada ne kadar kırmızı et tüketirsiniz?
a) 3'den fazla
b) 1-3
c) Hiç tüketmiyorum

Soru 19. Haftada kaç öğün balık tüketiyorsunuz?
a) 2'den çok
b) 1-2
c) Hiç tüketmiyorum

Soru 20. Her gün meyve sebze tüketiyor musunuz?
a) Tüketiyorum
b) Tüketmiyorum

Soru 21. Kızartma ve kavurma yöntemleri ile mi besinleri pişirirsiniz?
a) Evet
b) Hayır

Soru 22. Yemeklerde en çok hangi tür yağ kullanıyorsunuz?
a) Zeytinyağı
b) Ayçiçek, mısırözü yağı
c) Tereyağı
d) Margarin

Soru 23. Kilo vermek için ilaç kullanır mısınız ya da şok diyet yapar mısınız?
a) Evet
b) Hayır

Soru 24. Kuru baklagillerden haftada ne kadar tüketirsiniz?
a) 3 tabaktan fazla
b) 2-3 tabak
c) 1-2 tabak
d) Hiç

Soru 25. Günlük ortalama kuruyemiş (fındık, ceviz, fıstık vb.) tüketim miktarı ne kadardır?
a) Bir kâse
b) Bir avuç
c) Çok az
d) Tüketmiyorum

Soru 26. Ayda kaç kez ayaküstü tarzı yiyecek ve kızartma tüketiyorsunuz?
a) 6'dan fazla
b) 4-6
c) 1-3
d) Hiç

Soru 27. Her gün soğan, sarımsak tüketiyor musunuz?
a) Tüketiyorum
b) Tüketmiyorum

Soru 28. Günlük çay tüketiminiz ne kadar?
a) 5' ten fazla
b) 3-5
c) 3'ten az
d) Tüketmiyorum

Soru 29. Bitter çikolata tüketiyor musunuz?
a) Tüketiyorum
b) Tüketmiyorum

Soru 30. Günlük kaç fincan kahve içersiniz?
a) 5' ten fazla
b) 3-5
c) 3'ten az
d) Tüketmiyorum

2.1.6 Psikolojik Durumu

Kalp hastalıkları psikolojik durum ile iki yönlü bir ilişkiye sahiptir. Değişik hastalıkların baş göstermesi, ani ölümler, depresyon, kişiye mobbing uygulanması, panik atak hastalığı, tükenmişlik sendromu, ekonomik kriz gibi çeşitli psikososyal etmenler kalp krizinde önemli bir etkiye sahiptir. Bu nedenlerden dolayı kalp rahatsızlıklarını etkileyen psikolojik faktörler cevabı evet ya da hayır olacak şekilde şu sorularla tespit edilmeye çalışılmıştır:

Soru 31. Psikolojik bir rahatsızlık geçirdiniz mi?
Soru 32. Sizi çok üzen bir olay yaşadınız mı?
Soru 33. Aile içinde sorun yaşıyor musunuz?
Soru 34. Ani duygu değişimleri yaşar mısınız?
Soru 35. Bulunduğunuz çevrede hava kirli mi?
Soru 36. Bulunduğunuz çevrede yoğun bir trafik var mı?
Soru 37. Sık sık öfkelenir veya kızar mısınız?
Soru 38. Maddi durumunuzdan memnun musunuz?
Soru 39. İşyerinde size mobbing uygulanıyor mu?
Soru 40. Stresli biri olduğunuzu düşünüyor musunuz?
Soru 41. Daha önceden kalp krizi geçirdiniz mi?
Soru 42. Spor yaparken vücudunuza çok yüklenir misiniz?
Soru 43. Öfkelenince hemen sakinleşebilir misiniz?

2.2 Veri Toplama ve Eğitim Seti Oluşturma

Alınan resmi izin belgeleriyle Manisa Merkezefendi Devlet Hastanesi ve İzmir Ege Üniversitesi Hastanesindeki Kardiyoloji bölümlerine gidilerek tedavi gören toplam 62 hastaya anketler yapılmıştır. Anketteki sorulara hastalar tarafından verilen cevaplar gerçel sayısal değerler olarak veri setine kaydedilmiştir. Bazı sorulara verilen cevap "hayır" ise veri setine 0 olarak; "evet" ise 1 olarak kaydedilmiştir. Bazı çoktan seçmeli sorularda bulunan a, b, c ve d şıkları ise veri setine sırasıyla 1, 2, 3 ve 4 tam sayısı olarak kaydedilmiştir. Ayrıca bazı çoktan seçmeli sorular ise harf olarak kaydedilmiştir.

Anket çalışmasına katılan hastalardan bir kısmı, bazı soruları cevaplamamıştır. Bundan dolayı yeterince cevap alınamayan 24., 25., 26., 27. ve 28. sorular eğitim setine alınmamıştır.

Anket sorularında yer alan bazı faktörlerin kalp hastalıklarına olan etkisinin beklenenden daha az olduğu proje üzerinde çalışma yapılırken gözlemlenmiştir.

Eğitim setindeki belirli bazı sorulara ait verilen cevapların sayısal değerleri arasında büyük farklılıklar vardır. Aynı şekilde bazı verilerin sayısal değerlerin belirli bir skala arasında olmayışı, normalleştirme gereksinimin ortaya çıkarmıştır. Verilerin sınıflandırıcılar tarafından kullanılabilmesi ve güvenilir sonuçların elde edilebilmesi için Min-Max normalizasyon tekniği kullanılmıştır. Tüm özniteliklere ait verileri [0,1] aralığına çeken bu normalleştirme yönteminin formülü şu şekildedir:

$$x_{yeni} = \frac{x - Min}{Max - Min} \quad (2)$$

Bu formülde gözlemlenen x değeri, normalleştirilmiş olan x_{yeni}'ye dönüştürülmüştür. Max ve Min değerleri sırasıyla öznitelikteki en büyük ve en küçük değerlerdir.

Anket çalışması hasta ile sözlü olarak görüşülüp yapıldığı için elde edilen veriler güvenilirdir ve geçerlilik analizi gibi testlerin yapılmasına gerek yoktur. Ayrıca belirtmek isteriz ki resmi izinlerin alınması ve hastane ortamında hastalarla yüz yüze görüşerek anketlerin doldurulması oldukça zor bir iştir. Doğru ve hassas sonuçlar veren bir yapay karar destek sisteminin oluşturulması amaçladığı için bu tür zorluklara katlanılmıştır.

3 Yöntemler

Yapay öğrenme yöntemleri insana karar verme aşamasında yardımcı olan bir sistemdir. Denetimli yapay öğrenme yöntemleri tahmin (prediction), sınıflandırma (classification) ve regresyon (regression) şeklindedir. Denetimli öğrenme, türü yani etiketi bilinen verilerden faydalanarak etiketsiz yani sınıfı belli olmayan verileri etiketleme yöntemidir. Denetimli öğrenme için etiketli verilerin hazırlanması insan desteği gerektiren, vakit alan, zor ve masraflı bir iştir. Denetimli öğrenme, temel olarak sınıflandırma ya da regresyon yöntemleri ile yapılır. Sınıflandırma ve regresyon işlemleri için yapay sınıflandırıcı kendisi için hazırlanmış olan veri seti ile eğitir. Sınıflandırmada işlem sonucu sınıf türü yani bir etiket iken regresyonda işlem sonucu bir reel değerdir.

Sınıflandırıcılar (classifiers), temel öğrenici (base learner) olarak da adlandırılırlar. Denetimli öğrenme alanında karar ağaçları (Decision Tree, DT), En yakın k komşulukları (k-NN) tekniği, Destek Vektör Makineleri (Support Vector Machines, SVM), Çok Katmanlı Algılayıcılar (Multi-Layer Perceptrons, MLP), Naive Bayes (NB) gibi sınıflandırma algoritmaları mevcuttur. Regresyon alanında ise CART (Classification and Regression Tree), Bayesian Logistion Regression ve Linear Regression gibi yöntemler vardır. Bu çalışmada temel öğrenici olarak kural tabanlı bir sınıflandırıcı olan CART karar ağaçları tercih edilmiştir. Çünkü kolektif öğrenme metotlarında kullanılan temel öğreniciler genellikle karar ağaçlarıdır. Eğitilmelerinin kolay olması, şeffaf yapıları, hızlı karar verebilmeleri ve veri setindeki gürültüye karşı gürbüz olmaları, veri setindeki lokal özellikleri öğrenebilmesi gibi nedenlerden dolayı sıklıkla tercih edilmektedir [20].

3.1 Karar Ağacı Sınıflandırıcısı

Karar ağaçları, kolektif karar sistemlerinde genelde temel öğrenici olarak kullanılır. Karar ağaçlarının inşa edilmesi ve kullanımı kolay olduğu için sıklıkla kullanılmaktadır. Karar ağacı sınıflandırıcısında, ters bir ağaç yapısı inşa edilir. Gövdeden yani ana düğümden yapraklara doğru giden düğümler üzerinde karşılaştırmalar ve yapraklar üzerinde de sınıf türleri yer almaktadır. Ağacın inşa edilmesi sırasında, her bir özniteliğin bilgi kazancı değerine bakılarak bölütleme işlemi yapılır. Özyinelemeli olarak önemini yitirene kadar bu işlem devam sürekli tekrar eder. Karar ağacı öğrenmesi hem sınıflandırma hem de regresyon amaçlı kullanılmaktadır. Bilindiği üzere sınıflandırma işleminde test verisi için ortaya çıkan sonuç bir sınıf etiketi; regresyon işleminde ise sayısal bir değerdir. ID3, C4.5 ve CART algoritmaları en bilinen karar ağacı yöntemleridir.

3.2 AdaBoost (Adaptive Boosting)

Tekil öğrencililerin ve doğal olarak kararlarının birleştirilmesi ile oluşan öğrenci topluluğuna (ensemble) kolektif öğrenme denir. Genel olarak kolektif öğrenme uygulamalarındaki sınıflandırma başarısı tekil öğrenmeye göre daha yüksektir. AdaBoost en çok kullanılan boosting algoritmaları arasındadır ve ilk olarak Freund ve Schapire tarafından önerilmiştir [21]. Tahmin hızının diğerlerine göre yüksek oluşu, az hafıza kullanması, uygulanabilir olması gibi özelliklerinden dolayı diğer kolektif yöntemlere göre tercih edilmektedir.

Çalışmamızda kullanılan veri seti örnek tabanlı tekil öğrenicilere uygunluğundan dolayı ardışık topluluklarla öğrenme (Boosting) metodu tercih edilmiştir. Öğrenicilerin eğitileceği örneklerin

seçiminde önceki temel öğrenicilerin hata yaptıkları örneklere öncelik verilmektedir. Diğer bir kolektif öğrenme metodu olan Bagging yönteminde her bir iterasyonda tüm örneklerin eğitim kümesine seçilme şansları ve olasılıkları aynıdır. Fakat Boosting'de her iterasyonda örneklere ait seçilme olasılıkları güncellenmektedir. Bu da doğru verilen kararlardan çok, yanlış verilen kararlar üzerine odaklanılmasını sağlayan bir yöntem olmasını sağlamaktadır [22]. Bu sayede sistemin daha doğru tahmin yapması sağlanmış olmaktadır.

AdaBoost algoritmanın çalışma prensibi, her bir öznitelikten zayıf bir sınıflandırıcı oluşturularak bu zayıf sınıflandırıcılardan kurulu bir topluluk oluşturulmasına dayanır. Zayıf sınıflandırıcıların karar sınırları her bir öznitelik için negatif ve pozitif örneklerin ağırlıklı ortalaması alınarak bulunur. Daha sonra hata oranı en düşük zayıf sınıflandırıcılar kullanılarak güçlü bir sınıflandırıcı meydana getirilir. Güçlü sınıflandırıcı içerisinde yer almayan zayıf sınıflandırıcılara ilişkin öznitelikler silinmiş olur. Algoritmanın kaba kodu (pseudo code) şu şekildedir [23].

Adım 1: Veri setindeki *N* adet eğitim örneği {(x_1,y_1), ... ,(x_N,y_{nN})} şeklinde verilmiş olsun. $x_i, y_i \in [0, 1]$ tanımlamasında x_i, her bir örneğin sınıf etiketi; y_i ise regresyon algoritmasının verdiği karardır. Yapılan risk tahmini için y_i değeri 0 ile 1 aralığında reel değerler alabilmektedir.

Adım 2: I toplam iterasyon sayısı olmak üzere, her bir iterasyon *t*=1, 2, … , *I* için:

a) Her bir iterasyon için tüm örnekler ele alınarak ağırlıklar normalize edilir: $w_i \leftarrow \frac{w_i}{\sum_{a=1}^{N} t,a}$

b) Her bir *j* özniteliği için, sadece *j* özniteliğini kullanan her bir h_j sınıflandırıcısı eğitilir. Hata oranı w_t ağırlığına göre şu şekilde ölçülür:

$$\varepsilon_j = \sum_i^N w_i \left| h_j(x_i) - y_i \right|$$

c) En az ε_j hatasına sahip h_j sınıflandırıcısı seçilir.

d) Ağırlıklar güncellenir: $w_{t+1,i} = w_{t,i}\beta_t^{1-e_i}$

Burada h_i düşük hata oranı ile sınıflandırma yaptıysa $e_i = 0$, aksi halde $e_i = 1$ olur.

$\beta_t = \frac{\varepsilon_t}{1-\varepsilon_t}$ olarak hesaplanır.

Adım 3: $\alpha_t = \log\frac{1}{1-\beta_t}$ olarak alındığında $h(x)$ sınıflandırıcısının son durumu şu şekilde olur:

$$h(x) = \begin{cases} 1, & \sum_{t=1}^{I} \alpha_t h_t(x) \geq \frac{1}{2}\sum_{t=1}^{I} \alpha_t \\ 0, & diğer\ durmlar \end{cases} \quad (3)$$

AdaBoost algoritmasında, en güçlü zayıf sınıflandırıcıların bir araya getirilmesiyle güçlü bir sınıflandırıcının oluşturulması amaçlanmaktadır. Bunun için bu yöntemde işlemler her bir eğitim örneği için eşit bir *D* dağılımıyla başlar. Her iterasyonda sınıflama başarısına göre en iyi zayıf sınıflandırıcı tespit edilir ve ağırlıklar güncellenerek bir olasılık dağılım fonksiyonu oluşturulur. İlerleyen adımlarda bu işlemler tekrar edilir. Belirlenmiş bir iterasyon sonucunda en güçlü zayıf sınıflandırıcıların birleştirilmesiyle yüksek performanslı bir sınıflandırıcı oluşturulmuş olur.

Bu çalışmada neden AdaBoost algoritmasının tercih edildiği merak konusu olabilir. Elbette ki diğer birçok sınıflandırıcı veri setine uygun genel bir model oluşturduğu bilinmektedir. Bu tarz yöntemlerde veri setinde hata yapılan bölgeye özgü ayrı bir çözüm sunulması söz konusu değildir. Bu nedenle genelleştirilmiş modellerin hata oranları beklenen düzeyin üzerinde olmaktadır. Karar ağaçları sınıflandırıcısında ise sistem aşırı öğrenme (overfitting, overlearning) gerçekleştirdiğinde bazı problemlerin oluşmasına engel olamamaktadır. Bu durumda test veri kümesine en uygun karar ağacı oluşturulduğu için uygulama safhasında başarılı sonuçlar elde edilememektedir. Ayrıca aşırı öğrenme yapmış ağaç budandığında ise başarı oranı düşmektedir. AdaBoost yöntemi tekil karar ağacı sınıflandırıcısına göre daha uygulanabilir bir model ortaya sunarak yüksek sınıflandırma başarısı sağlamaktadır.

4 Deneysel Sonuçlar

Sınıflandırıcı yöntemlerin sınıflandırma başarılarını (performansını) belirlemek için çapraz geçerleme işlemi yapılmıştır. Literatürde önerilen ve sıklıkla tercih edilen çapraz geçerleme yöntemleri 10-Fold, Birini Devre Dışı Bırak (LOO-CV, Leave One Out Cross Validation) ve 5x2 yöntemlerdir. Etiketli yani sınıfı belirli tıbbi verilerin elde edilmesi yukarıda

belirtildiği üzere zor ve zahmetli bir iştir. Bu alanda oluşturulan veri setleri genellikle seyrek (*sparse*) bir yapıdadır yani fazla sayıda örnek içermemektedir. Literatürde seyrek yapıdaki veri kümeleri için tavsiye edilen en uygun teknik, LOO-CV'dur [24]. Ayrıca önerilen kolektif öğrenici ve temel öğrenicilerin verdikleri rasyonel sonuçların doğruluklarını test etmek amacıyla da Ortalama Karekök Hata yöntemi ve korelasyon katsayısı kullanılmıştır.

4.1. Çapraz Geçerleme

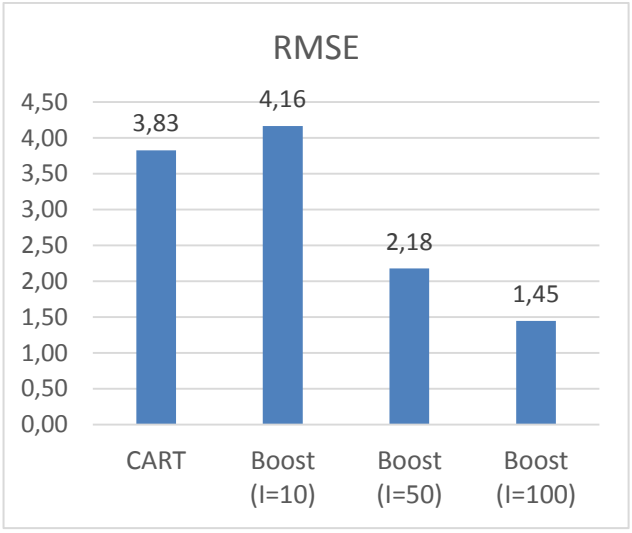
Çalışmamızda oluşturulan veri seti için en uygun ve en güvenilir sonucu geçerleme tekniği LOO-CV'dir. Birini Dışarda Bırakarak çapraz geçerleme yöntemi, veri seti ile eğitilen sınıflandırıcının toplam sınıflandırma başarısını belirlemek için kullanılır. LOO-CV'de, veri setindeki örnekler önce tek tek ele alınıp test örneği olarak seçilir. Arta kalan diğer örnekler ise eğitim seti olarak kullanılır ve ilgili sınıflandırıcıya aktarılarak eğitim işlemi yapılır. Her bir test noktası eğitilen sınıflandırıcıya sorulur. Her çevrimde sınıflandırıcının verdiği cevapların hatalı olup olmadığı kaydedilir. İşlem sonunda ortamala hata oranı ve ortalama karekök hatasının bulunur.

4.2 RMSE

RMSE (Root Mean Square Error), ortalama karekök hatadır. Kuadratik ortalama olarak da bilinen karekök ortalama değişen miktarların büyüklüğünün ölçümünde kullanılan istatistiksel bir yöntemdir. Regresyon çalışmaları için uygun bir başarı ölçüm kriteridir. RMSE tahmin edilmiş değerler ve orijinal değerler arasındaki farklarının karesini alınarak hesaplar. RMSE ölçütü, bir veri setinde gürültülü (noise) örneklerin çok olduğu ve etiketli verilerin az olduğu durumlarda ortalama hata (MAE) yöntemine göre daha tutarlı ve güvenilir sonuçlar verir. Formül şu şekildedir:

$$RMSE = \sqrt{\sum_{t=1}^{n} \frac{(a_i - p_i)^2}{n}} \quad (4)$$

Burada n veri setinde bulunan örnek yani hasta sayısını, a gerçekte var olan (actual) kalp krizi yaşını, p ise KKDS tarafından yapılan tahmini kalp krizi geçirme yaşını (predicted) göstermektedir. RMSE değerlerinin 0'a çok yakın çıkması bireylerin geçirdiği gerçek kalp krizi yaşları ile sistemin tahmin ettiği olası kalp krizi yaşlarının hemen hemen aynı çıkması anlamına gelir. Bu da önerilen çalışmanın oldukça başarılı sonuçlar verdiğini ve güvenilir olduğunu gösterir.



Şekil 1. Yöntemlerin RMSE Değerleri.

Deneysel sonuçlar MATLAB ve C++'da yazılan kodlar ile elde edilmiştir. RMSE değerleri Şekil 1'de verildiği gibidir. Çizelgeye göre hata oranı ne kadar az ise sınıflandırıcı o kadar başarılıdır.

Karar ağacı sınıflandırıcısı olan CART algoritması tekil bir öğrenici olarak gösterdiği performansın hata oranı oldukça yüksektir. Bu sınıflandırıcı kolektif öğrenici içerisinde tekrar tekrar kullanıldığında sistemin genel performansı artmaktadır.

İterasyon sayısı arttıkça AdaBoost algoritmasının hata oranı da azalmaktadır. Görüldüğü üzere genel olarak 50 ile 100 arasında seçilen iterasyon sayısında sınıflandırma performansları sırasıyla 2.18 ve 1.45 oranında çıkmıştır. Bu da AdaBoost metodunun çok az bir hata oranı ile ne kadar doğru tahminler verdiğini göstermektedir. Diğer bir deyişle I değeri 100'e eşitlendiğinde AdaBoost metodu ile bireye ait kalp krizi geçirme yaşı ±1.45 hata payı ile tahmin edilebilmektedir.

4.3 Korelasyon Katsayısı

Korelasyon iki değişken arasındaki ilişki durumunu tespit etmekte kullanılan istatistiksel bir yöntemdir. Bir olayın veya değişkenin diğer bir olaya veya değişkene olan etki oranını bulmak için kullanılır. +1 değerindeki korelasyon kat sayısı, iki değişken arasında tam bir lineer düz ilişkiyi; 0, ilişkisizliği; -1

ise tam ters ilişkiyi ölçeklemektedir. Pearson korelasyon katsayısı şu şekilde hesaplanır:

$$C = \frac{Cov(P, A)}{\sigma_p . \sigma_a} \quad (5)$$

Cov(P, A) fonksiyonu tahmini kalp krizi geçirme yaşı (P) ile gerçek kalp krizi yaşı (A) arasındaki kovaryansı hesaplayan formüldür. σ_p ve σ_a ise sırayla tahmin edilen değer ile gerçek değerlerin standart sapmalarıdır.

Kovaryans, iki değişkenin birlikte ne kadar değiştiklerini tespit eden istatistiksel bir ölçüdür. Kovaryansı 0 olan iki rassal değişkene korelasyonsuz değişkenler adı verilir. Kovaryansın 1'e yakın çıkması, aralarında doğru orantılı bir ilişki olduğu gösterir. Veri setinde bulunan gerçek kalp krizi geçirme yaşları olan A={A_1,..., A_n} vektörü ile tahmini kalp krizi geçirme yaşları olan P={P_1,..., P_n} vektörleri arasındaki kovaryans şu şekilde ifade edilebilir:

$$Cov(P, A) = \frac{1}{n-1}\sum_{i=1}^{n}(P_i - \mu_P) * (A_i - \mu_A) \quad (6)$$

μ_P ile μ_A değerleri sırasıyla P ve A vektörlerinin aritmetik ortalamalarıdır. * işlemi ise karmaşık eşleniği (complex conjugate) göstermektedir.

Yapılan uygulamada gerçek kalp krizinin geçirildiği yaş bilgilerini barındıran A vektörü ile hesaplamalar sonucunda tahmini kalp krizi geçirme yaş bilgilerini barındıran P vektörü arasındaki korelasyon katsayısı 0.942 çıkarak bu iki değişken arasında lineer ve tama yakın bir ilişki olduğu gözlemlenmiştir. Vektörlerde bulunan her bir hasta için KKDS sisteminin verdiği tahmini risk değerlerin tek tek incelenmesiyle performansın anlaşılması zordur. Bunun yerine veri setindeki gerçek değerler ile hesaplanan tahmini değerlere bütüncül bir yaklaşımla bakıldığında 0.942'lik bir korelasyon katsayının sınıflandırma başarısı açısından iyi bir sonuç olduğu söylenebilir

4.4 Sınıflandırma Başarısı

Ayrıca veri setinde kaydı bulunan her bir hastanın gerçek kalp krizi geçirme yaşı ile AdaBoost (I=100 olarak seçildiğinde) algoritmasının hesaplamalar sonucu bulduğu tahmini kalp krizi yaşı Şekil 2'de verilmektedir. Şekildeki her bir mavi nokta veri setinde bulunan hastaları simgelemektedir. Her bir noktanın yatay eksendeki değeri gerçek kalp krizi geçirme yaşını; dikey eksendeki değeri ise KKDS sistemi tarafından tahmin edilen kalp krizi geçirme yaşını göstermektedir. Bu durumda yatay lineer çizgiye yakın olan noktalarda algoritmanın ne kadar başarılı olduğu görülmektedir.

Şekil 2. Sınıflandırma başarısı.

4.5 Öznitelik Seçimi (Feature Selection)

Özellik seçimi, çok boyutlu bir veri kümesini daha küçük boyutta temsil etmek için tercih edilen bir yöntemdir. Amaç veri kümesindeki boyut sayısını azaltarak karmaşıklığı önlemek ve sınıflandırmadaki hesaplama süresini azaltmaktır. Occam's Razor yaklaşımında da belirtildiği üzere daha az özellik sayısıyla birim zaman dilimi içerisinde daha başarılı bir sınıflandırma performansı elde edilebilir [25].

Veri setindeki her bir öznitelik için hesaplanan Bilgi Kazancı değerlerinin yüksek olması ilgili özniteliğin sınıflandırmada etken bir rol oynadığını göstermektedir. Şu durumda çizelgede yer alan ilk üç faktörün kalp krizine neden olan en önemli faktörler olduğu görülmektedir.

Her bir sınıf etiketi üzerindeki IG değerlerinin hesaplanabilmesi için entropi değerlerinin öncelikle hesaplanması gerekir. Entropi, sistemdeki belirsizlik ölçüsüdür ve şu şekilde hesaplanır:

$$H(T) = -\sum_{i=1}^{n} p(t_i)\log_2 p(t_i) \quad (7)$$

T öznitelik vektörü $\{t_1, \dots, t_n\}$ şeklinde tanımlansın ve bu vektörde n tane sınıf etiketi bulunuyor olsun. Her bir sınıf etiketinin bulunma olasılığı $p(t_i)$ şeklindedir. 0 ile 1 arasında değişen Bilgi Kazancı entropinin basitçe tersidir ve şu şekilde bulunur:

$$H(t,T) = H(T) - \sum_{i=1}^{n} \frac{|T_i|}{|T|} H(T_i) \quad (8)$$

T veri kümesi, t ise hesaplanması istenen sınıf türüdür.

Çizelge 2. Bazı özniteliklerin IG değerleri.

Önem Sırası	Soru	IG
1	Büyük tansiyon değeri nedir?	0.8789
2	Diyabet hastalığı var mı?	0.7512
3	Vücut Kitle İndeksi (VKİ)	0.6124
4	Sigara kullanımı	0.4841
5	Küçük tansiyon değeri	0.4124
…		
43	Ailevi sorunlarınız var mı?	0.0117
44	Maddi durumunuzdan memnun musunuz?	0.0018
45	Kaç saat uyuyorsunuz?	0.0015
46	Bulunduğunuz çevrede yoğun bir trafik var mı?	0.0010
47	Menopozda mısınız?	0.0002

Kalp krizine neden olan 47 faktör, veri setini oluşturan her bir örneğin özniteliği olmuştur. Çalışmamızın başında her bir olası faktör dikkate alınmış ve veri seti bu doğrultuda oluşturulmuştur. Fakat bilindiği üzere kalp krizini tetikleyen bazı faktörler diğerlerine göre oldukça fazla bir etkiye sahiptir. Çalışmamızın bu bölümünde kalp krizine neden olan en çok öneme sahip 5 faktör ile en az öneme sahip 5 faktör Bilgi Kazancı (*Information Gain, IG*) yöntemi ile Çizelge 2'de sunulmuştur.

Hazırlanan eğitim setinde elde edilen bu değerler ışığında kalp krizini tetikleyen en etken beş faktör görüldüğü üzere büyük tansiyon değeri, diyabet hastalığı, obezite durumu, sigara tiryakiliği ve küçük tansiyon değeridir. Elbette ki değişik tıbbi çalışmalar ile kalp krizini tetikleyen genel faktörler belirlenmiştir. Fakat burada oluşturulan veri seti sayesinde kullanılan yapay öğrenme algoritmaları ile bu sıralama hesaplanmıştır.

5 İleri Çalışmalar ve Değerlendirme

Tıbbi vakıalarda bir hastalığı tetikleyen faktörler için genelleştirilmiş bir kural yoktur ve bu hastalıklar kişiye göre farklılıklar arz edebilmektedir. Ayrıca bir hastalığın nedeni birden fazla olabilmektedir. AdaBoost yönteminde, yapay karar destek sistemi eğitilirken veri uzayında hata yaptığı bölgeye odaklamaya çalışılarak daha doğru sonuçların elde edilmesini sağlamaktadır. Hastalığı tetikleyen her bir faktör yani öznitelik ayrı ayrı ele alınıp genel bir modelin çıkarılması çalışılmaktadır. Bu da doğal olarak genel sınıflandırma başarısının artmasına neden olmaktadır.

Üzerinde çalışılan konunun geliştirilmesi ve pratikte kullanılabilmesi için yapılabilecek bazı ek çalışmalar olabilir. Kolektif sınıflandırma yöntemleri için en uygun tekil öğrenicinin Karar Ağaçları olduğu bilinmektedir. Fakat veri setine uygun örnek tabanlı öğreniciler, destek vektör makineleri, yapay sinir ağları ve karmaşık modeller gibi başka tekil sınıflandırıcılar da seçilebilir. Ayrıca bireyde kalp hastalıklarına neden olan değişik faktörler de analiz edilip çalışmaya ilave edilebilir. Kişilere ait kan grupları, kan ve idrar tahlilleri, genetik hastalık türleri, protein yapıları ve elde edilebilecek bazı laboratuvar ve tahlil sonuçları da kalp hastalıklarını tetikleyebilecek risk faktörleri grubuna girebileceği için bunlar da çalışmaya dâhil edilebilir. Bunun yanında kalp hastalığı geçirmiş hastalardan alınan tıbbi verilerle sınıflandırıcıların kalp hastalıkları ile ilgili her olasılığa karşı önceden eğitilmiş olabilir. Bu nedenle veri zengini hastanelerdeki var olan hasta kayıtları bu çalışmaya ilave edilebilir. Bu sayede olası tüm faktörlere ait durumlar veri setine dâhil edildiği için daha güvenilir yapay tahminler yapılabilir. Veri madenciliği disiplininde bulunan değişik algoritmalarla coğrafi bölgelerdeki kalp krizi nedenleri tespit edilebilir. Bilgi keşfi gibi diğer veri madenciliği algoritmaları da bu alana uygulanabilir.

Çalışmamızdaki amaç bireysel gayretlerle oluşturulmuş özgün bir veri kümesi üzerinde, var olan sınıflandırıcı modellerinin uygulanması ve sınıflandırma başarılarının test edilmesi ve bazı karşılaştırılmaların yapılması değildir. Amaç, prototip bir yapay klinik karar destek sisteminin sınıflandırma ve regresyon alanında önerilmesidir. Burada çalışmamızı özgün ve faklı kılan iki unsurun olduğunu düşünüyoruz. Birincisi kalp krizini tetikleyebilen olası tüm faktörleri içeren bir anket ile kalp krizi geçirmiş hastalardan ilgili verilen elde edilmesi ve özgün bir veri setinin oluşturulmasıdır. İkincisi ise tıbbi verilere uygun, yüksek doğruluk oranına sahip, hızlı ve kolektif bir tahmin mekanizmasının inşa edilmesidir.

Bilişim ve sağlık bilim dallarını içeren çok disiplinli çalışmamızda, yüksek performanslı kolektif sınıflandırıcılarla bireye ait kalp krizi risk oranı belirlenmeye çalışılmıştır. Gerçek, özgün, ve güvenilir bir veri seti oluşturularak ve kolektif yöntemler kullanılarak çağın ölümcül hastalığına karşı suni bir klinik karar destek sistemi geliştirilmiştir. Geliştirilen bu sistem sayesinde bireyin olası kalp krizi geçirme yaşı önceden az bir hata oranıyla tahmin edilebilmektedir. Çıkan deneysel sonuçlara göre daha yüksek yüksek risk yüzdesine sahip bireyler önceden bilgilendirilip gerekli önlemlerin vaktinde alınması sağlanmıştır. Sağlık bilişimi alanında önerilen bu özgün erken uyarı sistemi, prototip bir çalışma olmasının yanı sıra, insan sağlığını ve hayatını korumayı amaçladığı için ayrı bir öneme sahiptir diye düşünüyoruz.

Teşekkürler

Hasta kayıtlarına ulaşmada ve ilgili verilerin elde edilmesinde bize resmi izni veren ve destekleyen Ege Üniversitesi Rektörlüğü'ne ve Manisa İl Sağlık Müdürlüğü'ne teşekkürlerimizi borç biliriz. Yardımlarını esirgemeyen, verileri toplayıp tasnif eden ve büyük gayret sarf eden Mustafa Ege ŞEKER'e ile Yusuf Miraç UYAR'a teşekkürlerimizi sunarız.

Kaynaklar

[1] Coşkun, M.Z.; Tarı. E., Ateş. S., Kırma. C., Kılıçgedik. A., İzgi. A., Durduran. K. İstanbul'da Akut Kalp Krizi Haritalarının Coğrafi Bilgi Sistemleri İle Üretilmesi ve Geoistatiksel Olarak İncelenmesi, Türkiye Harita Bilimsel ve Teknik Kurultayı, Ankara, 11-15 Mayıs 2009.

[2] Haney, M. Ö.; Bahar, Z. Çoçuk Kalp Sağlığını Geliştirme Tutum Ölçeği'nin Geçerlik ve Güvenirliği, Dokuz Eylül Üniversitesi Hemşirelik Fakültesi Elektronik Dergisi. 2014; 2.

[3] Güleç, S. Kalp Damar Hastalıklarında Global Risk Ve Hedefler, Türk Kardiyol Derneği Arştırmaları, 2009.

[4] Anbarasi, I. N. C. S. N. Enhanced prediction of heart disease with feature subset selection using genetic algorithm, International Journal of Engineering Science and Technology. 2010; 2, 5370-5376.

[5] Lichman. M. UCI Machine Learning Repository, Irvine, CA: University of California, School of Information and Computer Science., 2015. [http://archive.ics.uci.edu/ml]. Erişildi: 01/10/2016.

[6] Srinivas, K.; B, Rani, K., Govrdhan, A. Applications of data mining techniques in healthcare and prediction of heart attacks. International Journal on Computer Science and Engineering (IJCSE). 2010; 2, 250-255.

[7] Palaniappan, S.; & Awang, R. Intelligent heart disease prediction system using data mining techniques. IEEE/ACS International Conference on Computer Systems and Applications. 2008; 108-115.

[8] Chandna, D. Diagnosis Of Heart Disease Using Data Mining Algorithm. Int. J. Comput. Sci. Inf. Technol.(IJCSIT). 2014; 5, 1678-1680.

[9] Jin, Z.; Sun, Y., & Cheng, A.C. Predicting cardiovascular disease from real-time electrocardiographic monitoring: An adaptive machine learning approach on a cell phone. Annual International Conference of the IEEE Engineering in Medicine and Biology Society. 2009; 6889-6892.

[10] Karakoyun, H.M. Biyomedikal Veri Kümeleri İle Makine Öğrenmesi Sınıflandırma Algoritmalarının İstatistiksel Olarak Karşılaştırılması, DEÜ Mühendislik Fakültesi Mühendislik Bilimleri Dergisi. 2014; 16, 30-41.

[11] Öztürk, S. Kardiyovasküler Risk Faktörü Olarak Dislipidemilere Yaklaşım, Abant Medical Journal. 2012; 2, 89-93.

[12] Çakmakçı, S.; Kahyaoğlu, D. Yağ Asitlerinin Sağlık Ve Beslenme Üzerine Etkileri, Türk Bilimsel Derlemeler Dergisi. 2012; 2, 133-137.

[13] Buckland. G.; Carlos. A. G. The role of olive oil in disease prevention: a focus on the recent epidemiological evidence from cohort studies and dietary intervention trials., British Journal of Nutrition. 2015; 113, 94-101.

[14] Nehir, S.; Çam, O. Miyokard İnfarktüsü Geçiren Hastalarda Psikososyal Sağlık ve Hastalık Uyumu, Ege Üniversitesi Hemşirelik Yüksek Okulu Dergisi. 2010; 26, 73-84.

[15] Zhou, Z.-H. Ensemble Methods: Foundations and Algorithms, Press: CRC, New York, 2012.

[16] Freund, Y.; Schapire, R., Abe, N. A short introduction to boosting. Journal-Japanese Society For Artificial Intelligence. 1999; 14, 771-780.

[17] Breiman, L. Bias, Variance, and Arcing Classifiers, Technical Report, STATISTICS Department, Unıversity Of California, Berkeley, 1996.

[18] Tetik, Y.E.; Bolat, B. Gürültülü Ortamlarda Konuşma Tespiti İçin Yeni Bir Öznitelik Çıkarım Yöntemi. Elektrik-Elektronik ve Bilgisayar Sempozyumu, Fırat Üniversitesi-Elazığ, 2011.

[19] Alpaydın, E. Yapay Öğrenme, Press: Boğaziçi Üniversitesi Yayınevi, İstanbul, 2010; 417.

[20] Domingos, P. A Few Useful Things To Know About Machine Learning, Communications of the ACM. 2012; 55, 78-87.

[21] Şimşek, H.; Demiral. Y., Aslan. Ö, Toğrul. B. Ü. Bir Üniversite Hastanesinde Koroner Kalp Hastalarına Uygulanan Tedavi Oranları, DEÜ Tıp Fakültesi Dergisi. 2012; 2, 111-117.

[22] Estevez, P.A.; Tesmer, M., Perez, C.A., Zurada, J.M. Normalized Mutual Information Feature Selection, IEEE Neural Networks. 2009; 20, 189-201.