KORONER ARTER HASTALIĞININ RİSK FAKTÖRLERİNİN İRDELENMESİNDE ALTERNATİF BİR YAKLAŞIM:

GENETİK ALGORİTMALAR

Hande ŞENOL

Ocak 2014

DENİZLİ

KORONER ARTER HASTALIĞININ RİSK FAKTÖRLERİNİN İRDELENMESİNDE ALTERNATİF BİR YAKLAŞIM:

GENETİK ALGORİTMALAR

Pamukkale Üniversitesi

Sağlık Bilimleri Enstitüsü

Yüksek Lisans Tezi

Biyoistatistik Anabilim Dalı

Hande ŞENOL

Danışman: Prof. Dr. Beyza AKDAĞ

Yardımcı Danışman: Prof. Dr. Handan ANKARALI

Ocak 2014
DENİZLİ

YÜKSEK LİSANS TEZİ ONAY FORMU

Hande ŞENOL tarafından, Prof. Dr. Beyza AKDAĞ yönetiminde hazırlanan "**Koroner Arter Hastalığının Risk Faktörlerinin İrdelenmesinde Alternatif Bir Yaklaşım: Genetik Algoritmalar**" başlıklı tez tarafımızdan okunmuş kapsamı ve niteliği açısından bir Yüksek Lisans Tezi olarak kabul edilmiştir.

Prof. Dr. Handan ANKARALI
Jüri Başkanı

Prof. Dr. Beyza AKDAĞ
Jüri Üyesi(Danışman)

Prof. Dr. Asuman KAFTAN
Jüri Üyesi

Prof. Dr. Erol Ömer ATALAY
Jüri Üyesi

Doç. Dr. A. Gaye TOMATIR
Jüri Üyesi

Pamukkale Üniversitesi Sağlık Bilimleri Enstitüsü Yönetim Kurulu'nun 16.1.15 tarih ve 14-1-11 sayılı kararıyla onaylanmıştır.

Prof. Dr. Z. Melek BOR KÜÇÜKATAY
Müdür

Bu tezin tasarımı, hazırlanması, yürütülmesi, araştırmalarının yapılması ve bulgularının analizlerinde bilimsel etiğe ve akademik kurallara özenle riayet edildiğini; bu çalışmanın doğrudan birincil ürünü olmayan bulguların, verilerin ve materyallerin bilimsel etiğe uygun olarak kaynak gösterildiğini ve alıntı yapılan çalışmalara atfedildiğini beyan ederim.

İmza: 

Öğrenci Adı Soyadı: Hande ŞENOL

TEŞEKKÜR

Tezin planlanmasında, düzenlenmesinde, sonuçların yorumlanmasında ve yüksek lisans eğitimim süresince desteklerini, özverilerini ve bilgilerini esirgemeyen tez danışmanım Prof.Dr. Beyza AKDAĞ'a ve aynı zamanda bu tezin hazırlanmasında emeği geçen yardımcı danışmanım Prof. Dr. Handan ANKARALI'ya

Tezin uygulanması aşamasında yardımlarını esirgemeyen Yard. Doç. Dr. Elif Özge ÖZDAMAR'a ve Araş. Gör. Erdal COŞGUN'a

Bu tez çalışmasında verilerini kullandığımız Uzm. Dr. Işık TEKİN ve Prof. Dr. Asuman KAFTAN'a

Bu tezin hazırlanması esnasında bana her konuda yardımcı ve destek olan arkadaşlarım Fulya SARMAŞIK, Özge GÖKŞAR'a

Beni büyüten, yetiştiren, varlıklarıyla bana güç veren ve her türlü sıkıntımda bana sabırla destek olan sevgili aileme sonsuz teşekkürler ederim.

Hande ŞENOL

Ocak - 2014

ÖZET

KORONER ARTER HASTALIĞININ RİSK FAKTÖRLERİNİN İRDELENMESİNDE ALTERNATİF BİR YAKLAŞIM: GENETİK ALGORİTMALAR

ŞENOL, Hande
Yüksek Lisans Tezi, Biyoistatistik ABD
Tez Yöneticisi: Prof. Dr. Beyza AKDAĞ
Ocak 2014, 73 Sayfa

Bilgisayar uygulamalarının her alanda yaygın olarak kullanıldığı günümüzde, potansiyel olarak depolanan veri hacmi hızla artmaktadır. Özellikle sağlık sektöründe depolanan ancak kullanılmayan bilgilerin hayata geçirilmesi işlemi, büyük önem taşımaktadır. Veri tabanlarında toplanan çok büyük hacimli verilerden, anlamlı bilgilerin elde edilmesi için geliştirilen ve özellikle son yıllarda yaygın kullanım alanları bulan yöntemler veri madenciliği yöntemleri olarak adlandırılırlar. Bu alanda birçok algoritmadan faydalanılmaktadır. Genetik Algoritmalar (GA) yönteminde problemler sanal olarak evrimsel süreçten geçirilir ve bu süreç sonunda en iyi sonucu veren çözüme ulaşılmaya çalışılır.

Bu çalışmada, GA ile optimize edilerek sınıflama yapılmış olan 20, 50, 100, 500 ve 1000 ağaçlı Random Forest (RF) sonuçları; tek bir Karar Ağacından elde edilen sonuçlarla, GA ile optimizasyon yapılmamış olan 20, 50, 100, 500 ve 1000 ağaçlı RF sonuçlarıyla aynı zamanda da sınıflamalara etki eden en önemli 9 değişken temel alınarak sınıflama yapılmış olan 20, 50, 100, 500 ve 1000 ağaçlı RF sonuçlarıyla kıyaslanmıştır.

Yapılan bu çalışmadan elde edilen sonuçlardan yola çıkılarak, GA yöntemiyle optimize edilerek kurulan RF modellerinin, optimize edilmemiş olan RF modellerine oranla çok daha yüksek başarıya sahip olduğu söylenebilir. Aynı şekilde RF modellerinin sınıflama başarısını yükseltmesi açısından, ağaç sayısı kaç olursa olsun, kurulacak olan modellerin GA yöntemiyle optimize edilmesi yapılan bu çalışma sonucunda ulaşılabilen en önemli sonuç olmaktadır.

Anahtar Sözcükler: Genetik Algoritma, Random Forest, Sınıflama ve Regresyon Ağaçları

ABSTRACT

AN ALTERNATIVE APPROACH TO THE EXAMINATION OF CORONARY ARTERY DISEASE RISK FACTORS: GENETIC ALGORITHMS

ŞENOL, Hande

M. SC. Thesis, Biostatistics

Supervisor: Prof. Dr. Beyza AKDAĞ

January 2014, 73 pages

In the present day that computer applications are widely being used in all areas, potentially the data volume being stored is rapidly increasing. Particularly, putting into practice the stored but not used data in the health sector has a great importance. Methods which are developed for obtaining significant informations from huge datas accumulated in data bases, and which are finding wide are of usage particularly in late years, are called data mining. In this area many algorithms are followed up. In Genetic Algorithms (GA) method, problems are virtually undergone an evolutionary process, and at the end of this process it is tried to achieve the solution giving the best result.

In this study, 20, 50, 100, 500 and 1000 trees Random Forest (RF) results classified by optimizing with GA were compared with; the results obtained from only one Decision Tree, nonoptimized 20, 50, 100, 500 ve 1000 ağaçlı RF results, at the same time 20, 50, 100, 500 ve 1000 trees RF results classified based on the most important 9 variables affecting the classifications.

Setting out from the results of this study, it can be speculated that the RF models optimized by GA method have greater succes than nonoptimized RF models. In the same way, the most important result achieved in this study is that, in terms of increasing the succes in classification of RF models, optimizing the established models by GA method irrelevant from the number of trees.

Key Words: Genetic Algorithms, Random Forest, Classification and Regression Trees

İÇİNDEKİLER

ŞEKİLLER DİZİNİ

TABLOLAR DİZİNİ

SİMGELER VE KISALTMALAR DİZİNİ

AID	Automatic Interaction Detection
BYÖY	Bootstrap Yeniden Örnekleme Yöntemi
CART	Classification and Regression Trees
CHAID	Chi Square Automatic Interaction Detection
EYD	Epikardiyal yağ dokusu
GA	Genetik Algoritma
HDL	Yüksek Dansiteli Lipoprotein
KA	Karar Ağacı
KAH	Koroner Arter Hastalığı
LDL	Düşük Dansiteli Lipoprotein
NOO	Negatif Olabilirlik Oranı
NTD	Negatif Tahmin Değeri
POO	Pozitif Olabilirlik Oran
PTD	Pozitif Tahmin Değeri
QUEST	Quick Unbiased Efficient Statistical Tree
RF	Random Forest
VM	Veri Madenciliği
YP	Yanlış Pozitif Oranı
YN	Yanlış Negatif Oranı

1.GİRİŞ

Bilgisayar uygulamalarının her alanda yaygın olarak kullanıldığı günümüzde, potansiyel olarak depolanan veri hacmi hızla artmaktadır. Özellikle sağlık sektöründe depolanan ancak kullanılmayan bilgilerin hayata geçirilmesi işlemi, büyük önem taşımaktadır. Veri tabanlarında toplanan çok büyük hacimli verilerden, anlamlı bilgilerin elde edilmesi için geliştirilen ve özellikle son yıllarda yaygın kullanım alanları bulan yöntemler veri madenciliği yöntemleri (VM) olarak adlandırılırlar (Kahraman 2010). Bu yöntemler üç temel alanda hizmet vermektedirler. Bu alanlar kümeleme, tahmin ve sınıflamadır (Kaya 2010). VM çeşitli değişik branşlardan farklı teknik ve metotları içermektedir. VM alanında birçok algoritmadan faydalanılmaktadır. Genetik Algoritmalar (GA), Karar Ağaçları, Random Forest, Bulanık Mantık ve Yapay Sinir Ağları kullanılan algoritmalardan bazılarıdır (Doğan 2007).

GA, doğal seçim ilkelerine dayanan bir arama ve optimizasyon yöntemidir. Karmaşık çok boyutlu arama uzayında en iyinin hayatta kalması ilkesine göre bütünsel en iyi çözümü arar. Temel ilkeleri John Holland tarafından ortaya atılmıştır. GA'nın, fonksiyon optimizasyonu, mekanik öğrenme, tasarım, hücresel üretim gibi alanlarda başarılı uygulamaları bulunmaktadır. Geleneksel optimizasyon yöntemlerine göre farklılıkları olan GA, parametre kümesini değil kodlanmış biçimlerini kullanırlar. Olasılık kurallarına göre çalışan GA, yalnızca amaç fonksiyonuna gereksinim duyar. Çözüm uzayının tamamını değil belirli bir kısmını tarar. Böylece, etkin arama yaparak çok daha kısa bir sürede çözüme ulaşır (Goldberg 1989).

GA, problemlere tek bir çözüm üretmek yerine farklı çözümlerden oluşan bir çözüm kümesi üretir. Böylelikle, arama uzayında aynı anda birçok nokta değerlendirilmekte ve sonuçta bütünsel çözüme ulaşma olasılığı yükselmektedir. Çözüm kümesindeki çözümler birbirinden tamamen bağımsızdır. GA, problemlerin çözümü için evrimsel süreci bilgisayar ortamında uygular.

Veriler, hacim olarak sayfalarca yer kaplarlar ancak böyle bir durumda işlenmemiş veri oldukları için kullanım değerleri azdır. Sayılar işlenip özetlenirse, harfler düzenlenerek anlamlı cümleler haline dönüştürülürse, uygun grafikler oluşturulursa

ancak o zaman veri bilgiye dönüştürülmüş olacaktır. Bilgi, veriye göre hacim olarak daha az yer kaplar ancak kullanım değeri olarak daha güçlüdür. VM, büyük hacimli veri tabanlarındaki gizli bilgi ve yapıyı açığa çıkarmak için, çok sayıda veri analiz aracını kullanan yöntemler bütünüdür (Gülten ve Doğan 2008, Yıldız 2010).

Tıp alanında kullanılan standart yazılımlar; verinin dağınık yapıda olması, çok farklı türlerinin söz konusu olması, örneklerin karşılaştırılamaması, kritik farkların tespit edilememesi, veritabanları üzerinden sistematik ve tutarlı analizler yapılamaması gibi nedenlerden dolayı, sağlık personeli tarafından yetersiz olarak görülmektedir. Bu bağlamda gerekli bilgileri depolayan, yöneten, bir uzman bilgi sistemi ile temel analizini sağlayan ve karar veren bir yapıyı içeren VM devreye girmektedir. Ayrıca VM yöntemleri sağlık personeline kolaylık ve pratiklik sağlıyor olması nedeniyle tıp alanında kabul gören ve gelecekte de görecek olan bir yöntemdir (Gülten ve Doğan 2008).

Verinin içerdiği ortak özelliklere göre ayrıştırılması işlemi "sınıflandırma" olarak adlandırılmaktadır. Sınıflama; sınıfı tanımlanmış verilerden yola çıkarak, sınıfı belli olmayan verilerin sınıfını belirlemek üzere kullanılan veri madenciliği modelidir (Han ve Kamber 2006). Sınıflandırma işleminde öncelikle tahmin için kullanılacak bir model oluşturulur. Oluşturulan bu model, sınıfı belli olmayan veriler üzerinden uygulanarak sınıflar tahmin edilir (Han ve Kamber 2006). Sınıflandırma yöntemlerinden başlıca kullanılanları; Karar ağaçları (KA), Random Forest (RF), Yapay sinir ağları, Destek Vektör Makinesi, Bayes sınıflandırıcılar şeklinde ifade edilebilir. Sınıflandırma için kullanılan yöntemlerin başında gelen KA; çok sayıda bilgi içeren bir veri kümesini, bazı kuralları uygulayarak daha küçük kümelere bölmek için kullanılan bir yöntemdir. RF ise çok sayıda karar ağacından oluşan ve karar vermek için her ağacın sınıflama performansına göre en iyi sınıflama sonucunu kullanan bir yöntemdir.

Literatürde, sınıflama yöntemleri kullanılarak yapılmış çok sayıda çalışma vardır. Ancak genetik algoritma optimizasyonu kullanılarak yapılmış olan sınıflama çalışması sayısı oldukça azdır. Yapılmış olan bu çalışmadaki amaç; koroner arter hastalığının risk faktörlerini GA yöntemiyle optimize ederek, RF yönteminin GA optimizasyonlu ve optimizasyonsuz sonuçlarını karşılaştırarak daha iyi bir sınıflama başarısı elde etmektir.

Benzer bir çalışma daha önce Jefferson ve arkadaşları tarafından yapılmıştır. Çalışmada kanser hastalarından elde edilen bir veri seti üzerinde genetik algoritma ve lojistik regresyon yöntemleri karşılaştırılmış ve genetik algoritmanın lojistik regresyona göre daha iyi tahmin sonucu verdiği ortaya konulmuştur (Jefferson vd 1997). Başka benzer bir çalışmada Bozcuk ve arkadaşları tarafından yine kanser hastaları üzerinde genetik algoritma ve diğer veri madenciliği metotlarının karşılaştırılmasına yönelik olarak yapılmıştır (Bozcuk vd 2004, Samur vd 2009). (Laurikkala ve Juhola 1998, Fidelis vd 2000, Doğan 2006, Sanz vd 2008, Er vd 2009, Tanwani ve Farooq 2009, Yıldız vd 2012) yapmış oldukları çalışmalarda da karşılaştırmış oldukları yöntemlere göre GA yönteminin üstünlüğünü göstermişlerdir.

Bu çalışmada, genetik algoritma ile optimize edilerek sınıflama yapılmış olan 20, 50, 100, 500 ve 1000 ağaçlı RF sonuçları; tek bir KA'dan elde edilen sonuçlarla, genetik algoritma ile optimizasyon yapılmamış olan 20, 50, 100, 500 ve 1000 ağaçlı RF sonuçlarıyla aynı zamanda da sınıflamalara etki eden en önemli 9 temel değişken alınarak sınıflama yapılmış olan 20, 50, 100, 500 ve 1000 ağaçlı RF sonuçlarıyla kıyaslanmıştır.

2. KURAMSAL BİLGİLER

2.1. Genetik Algoritmalar

Veri madenciliği yöntemlerinin içinde yer alan ve henüz sağlık alanında yaygın olarak kullanılmayan Genetik Algoritmalar (GA), Darwin'in evrim teorisinden esinlenerek ortaya çıkmıştır ve evrim teorisi ilkelerine dayalı olasılık tabanlı araştırma yöntemleridir (Nabiyev, 2010). GA genel anlamda, dizilerden oluşan bir popülasyona seçim (Selection), çaprazlama (Crossover) ve mutasyon (Mutation) işlemlerinin uygulanmasını içerir. Bu işlemlerin uygulanmasından sonra yeni bir popülasyon (yavru popülasyon) meydana gelir. Yeni popülasyon eski (ebeveyn) popülasyon ile yer değiştirir. Her dizinin bir uyum değeri mevcuttur. Diziler uyum değerlerine göre seçilirler. Ortalama uyum değerinin üzerinde uyuma sahip dizilerin gelecek kuşaklarda temsil edilme olasılığı daha yüksektir. Evrim süreci sayesinde, popülasyonun ortalama uyumu giderek artar ve ilerleyen kuşaklarda daha iyi uyum değerleri elde edilir (Taşkın ve Emel 2009).

Bu yöntemde, problemler sanal olarak evrimsel süreçten geçirilir ve bu süreç sonunda en iyi sonucu veren çözüme ulaşılmaya çalışılır (Elmas 2011). Başka bir ifadeyle çözüm de evrimleşmiş olur. Son yıllarda GA'lara ilgi büyüyerek artmaktadır. GA, hem problem çözme hem de modelleme için kullanılmaktadır. Geleneksel yöntemlerden daha etkilidir ve çözümlemede yapılacak küçük değişikliklerle tanımlanabildiğinden dolayı da daha esnektir (Mitchell 1999). Her problemin çözümünde problemin yapısına göre bir genetik algoritma oluşturulabilir (Michalewicz 1992). Genetik algoritma, problem çözümü için kullanılan bir yöntem değildir. Bununla birlikte genetik algoritma bir problemin çözümünü elde etmek için izlenen yol olarak ifade edilebilir. GA'ların önemli özelliklerinden birisi de bir dizi üzerinde çözümü araması ve bu sayede çok sayıda çözüm içerisinden en iyiyi seçebilmesidir (Shah ve Kusiak 2007).

2.1.1. Genetik Algoritmaların Tarihçesi

Michigan Üniversitesinde bulunan John Holland ve arkadaşlarının liderliğinde yapılan çalışmalar sonucu 1970'li yıllarda GA tekniği ortaya çıkmış ve 1975 yılında Holland; "Doğal ve Yapay Sistemlerin Uygulanması" adlı kitabı yayınlamıştır (Sivanandam ve Deepa 2008). Michigan Üniversitesinde psikoloji ve bilgisayar bilimi uzmanı olan John Holland bu konuda ilk çalışmaları yapan kişidir. Mekanik öğrenme (machine learning) alanında çalışan Holland, Darwin'in evrim kuramından etkilenerek canlılarda yaşanan genetik süreci bilgisayar ortamına taşımayı düşünmüştür. Tek bir mekanik yapının öğrenme yeteneğini geliştirmesi yerine böyle yapılardan oluşan bir topluluğun çoğalma, çiftleşme, mutasyon vb. genetik süreçlerden geçerek başarılı (öğrenebilen) yeni bireyler oluşturabildiğini görmüştür. Araştırmalarını arama ve optimumu bulma için doğal seçme ve genetik evrimden yola çıkarak yapmıştır (Elmas 2011).

2.1.2. Genetik Algoritmaların Araştırma Teknikleri İçerisindeki Yeri ve Uygulama Alanları

En iyileme probleminin çözümünde belirli matematiksel ifadelere veya kurallara dayanan algoritmalar kullanılır. Bazı algoritmalar gerçek çözümü bulmayı garanti edemezler. GA bu konuda oldukça başarılıdır. Diğer yandan bir problemin çözümü için farklı araştırma teknikleri de kullanılabilir. (Elmas 2011).

Rastgele araştırma teknikleri, temelde sayma teknikleri gibidir fakat araştırmada kullanılabilmesi için ek bilgiye ihtiyaç duyarlar. Bu araştırma tekniği ile çok karmaşık sistemler çözümlenebilir.

GA'ların fonksiyon optimizasyonu, otomatik programlama ve bilgi sistemleri, mekanik öğrenme gibi alanlarda başarılı uygulamaları bulunmaktadır. Geleneksel optimizasyon yöntemlerine göre farklılıkları olan GA, parametre kümesini değil kodlanmış biçimlerini kullanır. Olasılık kurallarına göre çalışan GA, yalnızca amaç fonksiyonuna gereksinim duyar. Çözüm uzayının tamamını değil belirli bir kısmını tarar. Böylece, etkin arama yaparak çok daha kısa bir sürede çözüme ulaşır (Goldberg 1989).

GA'ların geleneksel optimizasyon yöntemlerine olduğu gibi yapay zeka yöntemlerine göre de çeşitli üstünlükleri bulunmaktadır.
GA'yı diğer yöntemlerden ayıran en belirgin özellikler aşağıda belirtilmiştir:

- GA, parametre kodlarıyla çalışmakta, parametrelerin kendisiyle doğrudan ilgilenmemektedir.
- GA, tek bir alana bağımlı kalarak çözüm aramamakta popülasyonun tamamından çözümü aramaktadır.
- GA, ne yaptığını değil nasıl yaptığını bilmektedir. Yani GA amaç işlevini kullanır, sapma değerleri veya diğer hata faktörlerini kullanmaz.
- GA'nın uygulanmasında kullanılan işlemler rastlantısal yöntemlere dayanmakta, belirli ve kesin yöntemler kullanmamaktadır (Taşkın ve Emel 2009).

2.1.3. Genetik Algoritmaların Temel Kavramları

Gen

Kendi başına anlamlı genetik bilgi taşıyan en küçük genetik yapıdır yani kromozom içerisindeki tek bir özelliktir. GA'da kullanılan gen yapıları araştırmacının tanımlamasına bağlı olarak değişir.

Kromozom

Kromozom, bir bireyin tam bir ifadesidir. Birden fazla genin bir araya gelerek oluşturduğu diziye denir. Kromozomlar, alternatif aday çözümleri gösterirler. Kromozomlar GA yaklaşımında üzerinde durulan en önemli birim olduğu için bilgisayar ortamında iyi bir biçimde ifade edilmeleri gerekir. (Çivril 2009).

Popülasyon (Yığın)

Kromozomlardan oluşan topluluğa denir. Popülasyon, geçerli alternatif çözüm kümesidir. Popülasyondaki birey sayısı (kromozom) genellikle sabit tutulur. GA'da popülasyondaki birey sayısı ile ilgili genel bir kural yoktur. Popülasyondaki kromozom sayısı arttıkça çözüme ulaşma süresi (iterasyon sayısı) azalır. Problemin özelliğine göre seçilecek olan popülasyon sayısı araştırmacı tarafından net bir biçimde belirlenmelidir.

Allel

Bir özelliği temsil eden bir genin alabileceği değişik değerlere denir. Genin değerini gösteren terimdir. Örneğin; yön bir gen olarak gösterilirken, bu genin allelleri "sağ", "sol", "ön", "arka" değerlerini alabilir.

Genotip

Genetik programlar tarafından ortaya çıkarılan çözüm yapısına genotip denir. Yani bir canlının sahip olduğu genlerin toplamıdır.

Fenotip

Genotip özelliklerin bireyin dış yapısında ya da fiziksel görünüşündeki yansımasına fenotip denir. Yani canlının dış görünüşüdür.

2.1.4. Genetik Algoritma İşlemcileri

2.1.4.1. Seçim İşlemcisi

Seçim işlemi, uygunluk değerini temel alarak, topluluktan uygunluk değeri düşük olan bireylerin elenmesi ve yerlerine uygunluk değerleri yüksek olan bireylerin kopyalarının konulmasıdır. Uygunluk değeri; hangi bireyin sonraki topluluğa taşınacağını belirtir. Bir dizinin uygunluk değeri, problemin amaç fonksiyonu değerine eşittir. Bir dizinin gücü uygunluk değerine bağlı olup iyi bir dizi, problemin yapısına göre maksimizasyon problemi ise yüksek, minimizasyon problemi ise düşük uygunluk değerine sahiptir.

Seçim aşamasının önemi, topluluğun boyutu ile ilişkilidir. Seçimde küçük topluluk boyutu ile çalışılması durumunda topluluk çeşitlendirmesi olası iyi alternatiflerin oluşması için yetersiz kalabilir. Bu sebeple seçimde, topluluktaki bireylerin çeşitlendirmesini daraltan bir yöntemin uygulanması iyi sonuç vermeyebilir (Parlak 2007).

Seçim işlemcisi olarak geliştirilmiş birçok yöntem bulunmaktadır. En yaygın kullanılanları roulette wheel selection, tournament selection, elitist selection ve sıralama (ranking selection) yöntemleridir.

2.1.4.1.1. Roulette Wheel Selection (Rulet Tekeri Seçim Yöntemi)

Bu yöntemde her dizi uyum değeriyle orantılı bir olasılık değeri ile seçilmektedir. Rulet tekerleğinin yüzeyi dizilerin uyum değerleri ile orantılı olarak işaretlenir. Tekerlek kaç defa döndürülürse her seferinde bir dizi, eşleme havuzuna atılır. Daha iyi uyum değerine sahip diziler tekerlekte daha fazla yer aldıklarından onların seçilme şansları daha yüksektir. (Yu ve Gen 2010)

2.1.4.1.2. Tournament Selection (Turnuva Seçim Yöntemi)

Turnuva seçiminde bir eşleştirme havuzu vardır ve rastgele seçilen genler bu eşleştirme havuzuna atılır. Popülasyonda bulunan bireyler içerisinden en yüksek uygunluk değerine sahip olanlar eşleştirme havuzuna alınırlar ve geriye kalan bireyler içerisinden yeni bir turnuva seçimi daha yapılır. Turnuvanın boyutu kadar bu işleme devam edilir. Sonuçta havuzda oluşan bireylerin uygunluk değerleri yükseltilmiş olur (Kaya 2010).

2.1.4.1.3. Elitist Selection (Elitist Seçim Yöntemi)

Elitist seçim yönteminde topluluğun en iyi bireyi korunarak, topluluğun geri kalan elemanları uyum orantılı seçim yöntemlerinden birisi kullanılarak yeni bireyler ile değiştirilir. Burada hedef en iyi uyum değerine sahip bireyin, genetik işlemciler kullanıldığında kaybolmasını önlemektir. Elitizm kullanılmadığında, GA rastgele aramaya dönüşür, sonuca ulaşılması zordur ya da ulaşılsa bile bir sonraki nesile aktarılamadığından çözüm elde edilemez. Elitizm, GA için çok önemlidir. Elitizm diğer seçim yöntemleri ile beraber de kullanılabilir (Parlak 2007).

2.1.4.1.4. Ranking Selection (Sıralama Yöntemi)

Bu yöntemde tüm diziler uyum değerlerine göre artan bir şekilde sıralanır. Uygun bir fonksiyon yardımıyla en yüksek uyumlu olanlar havuza daha fazla kopya bırakır, en kötü uyumlu olan dizilerden ise kopya elde edilememiş olur. Bu yöntemde, arama hızı yüksektir.

2.1.4.2. Çaprazlama İşlemcisi

GA içerisindeki çaprazlama işlemi, bireyler arasındaki bilgi değişimini gerçekleştirerek daha iyi bireylerin üretilmesini sağlar. Çaprazlama yapılırken iki bireyin belirli genleri karşılıklı olarak yer değiştirir. Böylelikle iki yeni birey başka bir deyişle iki yeni olası çözüm üretilmiş olur. Çaprazlama işlemini gerçekleştirmek için ilk olarak, üreme işlemi ile oluşturulmuş eşleştirme havuzundaki yeni kopyalanmış dizinin elemanları rastgele eşlenir. İkinci olarak, seçilen dizilerin değerleri, rastgele seçilmiş çaprazlama noktasından itibaren karşılıklı olarak değiştirilirler. Çaprazlamalar, problemlerin türüne göre değişiklikler göstermektedir. Ele alınan probleme bağlı olarak, kullanıcı tarafından seçilen 4 farklı çaprazlama yöntemi bulunmaktadır. Temel olarak kullanılan çaprazlama yöntemleri olan tek nokta çaprazlama ve iki nokta çaprazlama yöntemleri kısaca açıklanmıştır. Bunlara ek olarak, daha seyrek kullanılan diğer çaprazlama yöntemleri ise çok nokta çaprazlama ve uniform çaprazlama yöntemleridir.

- *Tek nokta çaprazlama*

Havuzdaki diziler rastgele eşleşirler. Seçilen her dizi çifti için ilk ve son gen dışında aradaki genlerden rastgele bir yer seçilir. Burası çaprazlama noktasını gösterir. Bu noktadan sonra gelen genler her iki dizide karşılıklı olarak yer değiştirir. Bu işlem için diziler aynı uzunlukta olmalıdır. Aşağıdaki dizilerde 3. konum çaprazlama noktası olacak şekilde örneklenecek olursa, bu noktadan sonraki genler çaprazlandığında yeni diziler elde edilmiş olur.

P1 = 1,0,0,[1,1,1,1,0] => P1’ =1,0,0,[1,0,0,1,0]

P2 = 1,0,1,[1,0,0,1,0] => P2’ =1,0,1,[1,1,1,1,0]

- *İki nokta çaprazlama*

Bu yöntemde dizi üzerinde ilk ve son genler hariç iki tane rastgele nokta seçilir. Çaprazlama işlemi seçilen bu iki nokta arasındaki genlerin yer değişimidir. Aşağıdaki dizilerde 1. ve 5. genden sonra gelen aralık çaprazlama noktaları olacak olursa, bu iki nokta arasında kalan genler çaprazlandığında yeni diziler elde edilmiş olur.

P1 = 1,[0,0,1,1],1,1,0 => P1’ =1,[0,1,1,0],1,1,0

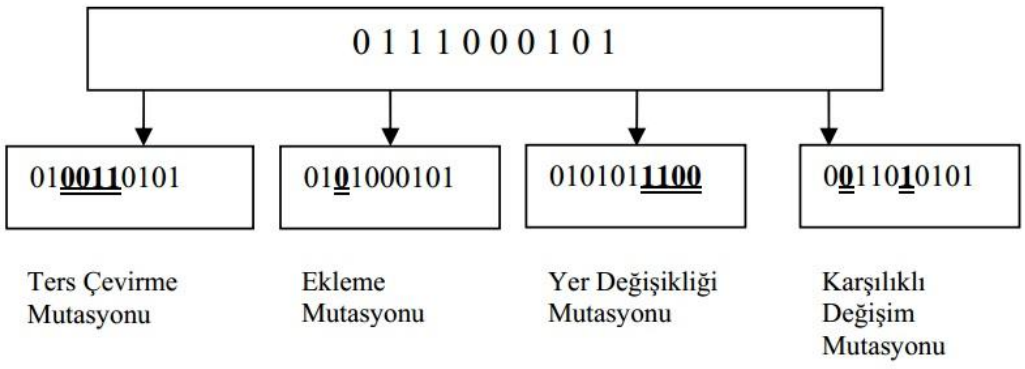
P2 = 1,[0,1,1,0],0,1,0 => P2’ =1,[0,0,1,1],0,1,0

2.1.4.3. Mutasyon İşlemcisi

GA'larda bir diğer işlemci olan mutasyon, oluşan yeni çözümlerin önceki çözümü kopyalamasını önlemek ve sonuca daha hızlı ulaşmak amacıyla kullanılmaktadır. Örneğin; ikili bir kodlamanın kullanıldığı bir dizide, mutasyon işlemcisi ile rastgele seçilen bireyin değeri 1 ise 0 veya 0 ise 1 şeklinde değiştirilerek yeni bir dizi elde edilir. Mutasyon işlemcisi olarak, ele alınan problemin yapısına en uygun olarak aşağıdakilerden birisi seçilir.

- Ters çevirme
- Ekleme
- Yer değişikliği
- Karşılıklı Değişim

Mutasyon işlemcilerinin uygulamaları Şekil 2.1'de görülmektedir. Şekilde altı çizili olarak verilen eleman değerleri mutasyona uğramış elemanları göstermektedir. Ters çevirmede, kromozomdan rastgele iki pozisyon seçilir ve iki ucu arasında ters çevrilir. Eklemede, rastgele bir parça seçilir ve rastgele bir yere yerleştirilir. Yer değişikliğinde rastgele bir alt dizi seçilir ve rastgele bir yere yerleştirilir. Karşılıklı değişimde ise rastgele seçilen iki genin yerleri değiştirilir.



Şekil 2.1 Mutasyon yöntemleri ve etkileri (Bolat vd 2004)

2.1.5. Genetik Algoritmaların Temel Teoremi

GA'ların nasıl arama yaptığı alt dizi kavramıyla açıklanmaktadır. Alt diziler, genetik algoritmaların davranışlarını açıklamak için kullanılan teorik yapılardır. Bir alt dizi, belirli dizi kümeleri arasındaki benzerliği tanımlayan bir dizidir. Alt diziler, {0, 1, *} alfabesi kullanılarak tanımlanır. Örneğin H alt dizisi, ilk konumunda 0, ikinci ve dördüncü konumunda 1 değeri olan kromozomlar dizisi içindir.

H = 0 1 * 1 *

* sembolü, dizinin o konumunun hangi değeri aldığının önemli olmadığı anlamına gelmektedir. Dizi o konumda 0 veya 1 değeri alabilir. Eğer bir x dizisi, alt dizinin kalıbına uyarsa x dizisine "H'nin bir örneğidir" denir. Alt dizilerin iki özelliği mevcuttur. Bu özellikler aşağıda verilmiştir (Goldberg 1989).

- Alt dizi derecesi: Bir H alt dizisinin derecesi o(H) ile gösterilir ve mevcut alt dizi kalıbında bulunan sabit konumların sayısıdır. Bu sayı, ikili alfabede 0 ve 1 değerlerinin sayısının toplamına eşittir.

- Alt dizi uzunluğu: Bir H alt dizisinin uzunluğu δ(H) ile gösterilir ve mevcut alt dizi kalıbında bulunan belirli ilk ve son konumlar arasındaki uzaklıktır.

GA'nın temel teoremi ise şöyle açıklanmaktadır (Yeniay 2001): Popülasyon ortalamasının üstünde uyum gücü gösteren, kısa uzunluğa ve düşük dereceye sahip alt diziler zamanın ilerlemesiyle üssel olarak çoğalırlar. Bu çoğalma, genetik işlemler aracılığı ile gerçekleşmektedir ve sonuçta ana-babadan daha üstün özellikler taşıyan bireyler ortaya çıkmaktadır. Bu çözüm kalitesinin, kuşaktan kuşağa artması iki nedene bağlanmaktadır. Bu nedenler şöyle açıklanabilir (İşlier 2001):

- Başarısız olan bireylerin üreme şansları azaltıldığı için kötüye doğru gidiş engellenmektedir.
- GA'ların yapısı, kötüye gidişi engellemekle kalmamakta, GA'ların temel teoremi uyarınca, zaman içerisinde hızlı bir iyiye gidişi de sağlayabilmektedir.

2.1.6. GA'ların İşlem Adımları

1. Arama uzayındaki tüm gerçekleşmesi muhtemel çözümler dizi olarak kodlanır.
2. Genellikle gelişigüzel olarak bir çözüm kümesi seçilir ve başlangıç topluluğu olarak kabul edilir.
3. Her bir dizi için bir uygunluk değeri hesaplanır, bulunan uygunluk değerleri dizilerin çözüm kalitesini gösterir.
4. Bir grup dizi belirli bir olasılık değerine göre gelişigüzel olarak seçilip çoğalma işlemi gerçekleştirilir.
5. Yeni bireylerin uygunluk değerleri hesaplanarak, çaprazlama ve mutasyon işlemlerine tabi tutulur.
6. Önceden belirlenen kuşak (adım) sayısı boyunca yukarıdaki işlemler devam ettirilir.
7. Yineleme (iterasyon), belirlenen kuşak sayısına ulaşınca işlem sona erdirilir. Uygunluk fonksiyonuna göre en uygun olan dizi seçilir (Engin 2001).

GA'lar bir çözüm uzayındaki her noktayı, kromozom adı verilen ikili bit dizisi ile kodlar. Her noktanın bir uygunluk değeri vardır. Tek bir nokta yerine, GA'lar bir topluluk olarak noktalar kümesini muhafaza eder. Her kuşakta, GA, çaprazlama ve mutasyon gibi genetik işlemcileri kullanarak yeni bir topluluk oluşturur. Birkaç kuşak sonunda, topluluk daha iyi uygunluk değerine sahip üyeleri içerir. GA'lar; çözümlerin kodlanmasını, uygunlukların hesaplanmasını, çoğalma, çaprazlama ve mutasyon işlemcilerinin uygulanmasını içerir.

2.1.6.1. Kodlama

Bir problemin çözümü için GA geliştirmenin ilk adımı, tüm çözümlerin aynı boyutlara sahip bitler dizisi biçiminde gösterilmesidir. Dizilerden her biri, problemin olası çözümler uzayındaki rassal bir noktasını simgeler. Çözümlerin kodlanması, probleme özgü bilgilerin GA'nın kullanacağı şekle çevrilmesidir. Her problem kendine özgü farklı kodlamalara ihtiyaç duyabilir. Tüm problemler için geçerli olabilecek genel bir kodlama tekniği yoktur (Tuğ 2005).

2.1.6.1.1. İkili Kodlama

En yaygın kullanılan kodlama yöntemidir. Problemin olası çözümleri 0 ve 1 sayıları ile kodlanır.

2.1.6.1.2. Sıralı Kodlama

Genellikle birleşik optimizasyon problemlerinde kullanılmaktadır. Dizinin uyum değeri gen değerlerine ve genlerin sıralarına bağlıdır. Bu kodlamada her tamsayı değeri bir olayı gösterir. Bu yöntem sayesinde tekrar eden değerler engellenmiştir. Ayrıca uyum değerlerinin hesaplanması da basit hale getirilmiştir.

Örneğin X=[4 3 0 1 2] dizisi, sırayla dördüncü, üçüncü, merkez olayların, birinci ve ikinci olayların ziyaret edileceğini anlatmaktadır.

2.1.6.1.3. Değer Kodlaması

Değer kodlamasında her dizi, bir değerler kümesinden oluşmaktadır. Değerler, probleme göre herhangi bir sayı ya da karakter olabilir.

Örnek:

Dizi A= 1.2324, 5.3243, 0.4556, 2.3293, 2.4545

Dizi B= ABDJEIFRGT

Dizi C= (geri), (geri), (sağ), (ileri), (sol)

Değer kodlaması bazı özel problemler için oldukça kullanışlıdır. Fakat bu tip kodlama kullanıldığında, probleme özgü çaprazlama ve mutasyon yöntemleri gerekir.

2.1.6.1.4. Ağaç Kodlama Yöntemi

Bu yöntemde her dizi, nesnelerden oluşan bir ağaç yapısı şeklinde ifade edilir. Bu nesneler fonksiyonlar ya da programlama dillerinde geçen komutlar olabilir. Bu yöntem verilen değerlere uygun bir fonksiyon bulmak için kullanılır.

2.1.6.2. Başlangıç Popülasyonunun Oluşturulması

Olası çözümlerin kodlandığı bir çözüm grubu oluşturulur. Çözüm grubu topluluk, çözümlerin kodları da kromozom olarak adlandırılır. İkili alfabenin kullanıldığı kromozomların gösteriminde, ilk topluluğun oluşturulması için rassal sayı üreticileri

kullanılabilir. Rassal sayı üreticisi çağrılır ve değer 0,5'den küçükse konum 0'a değilse 1 değerine ayarlanır. Birey sayısının az ve kromozom uzunluğunun kısa olduğu problemlerde yazı-tura ile de konum değerleri belirlenebilmektedir. GA'larda ikili kodlama yöntemi dışında, çözümü aranan probleme bağlı olarak farklı kodlama yöntemleri de kullanılabilmektedir. Bunlardan biri, 0–9 arasındaki gerçek sayılarla oluşturulan kodlama yöntemidir. Gerçek sayılarla topluluk oluşturulurken, aynı ikili kodlamada olduğu gibi rassal sayılar üretilir fakat burada rassal sayılar 0–10 arasındaki reel sayılardır. Daha sonra bu sayılar bir dizi kurala uygun olarak 0–9 arasındaki tamsayılara yuvarlanarak ilk topluluk oluşturulmuş olur (Parlak 2007).

2.1.6.3. Uygunluk Değerinin Hesaplanması

Bir kuşak oluşturulduktan sonraki ilk adım, topluluktaki her üyenin uygunluk değerini hesaplama adımıdır. Örneğin; bir maksimizasyon problemi için i. üyenin uygunluk değeri f(i), genellikle o noktadaki amaç fonksiyonunun değeridir. Çözümü aranan her problem için bir uygunluk fonksiyonu mevcuttur. Belirli bir kromozom için uygunluk fonksiyonu, o kromozomun temsil ettiği çözümün kullanımıyla veya çözüm niteliğiyle orantılı olan sayısal bir uygunluk değerini verir. Bu bilgi, her kuşaktaki daha uygun çözümlerin seçiminde yol gösterici olur. Bir çözümün uygunluk değeri ne kadar yüksekse, üyelerin yaşama ve çoğalma şansı da o kadar fazla ve bir sonraki kuşakta temsil edilme oranı da o kadar yüksektir.

2.1.6.4. Çoğalma İşleminin Uygulanması

Çoğalma işlemcisinde diziler, amaç fonksiyonuna göre kopyalanır ve iyi nitelikteki kalıtsal özellikleri gelecek kuşağa daha iyi aktaracak bireyler seçilir. Üreme işlemcisi yapay bir süreçtir. Dizileri uygunluk değerlerine göre kopyalama, daha yüksek uygunluk değerine sahip dizilerin, bir sonraki kuşaktaki bir veya daha fazla yavruya daha yüksek bir olasılıkla aktarılması anlamlarına gelmektedir. Çoğalma; bireyleri seçme, seçilmiş bireyleri eşleştirme havuzuna kopyalama ve havuzdaki bireyleri çiftler halinde gruplara ayırma işleminden oluşur.

Uygunluk değerinin hesaplanması adımından sonra mevcut kuşaktan yeni bir topluluk yaratılmalıdır. Seçim işlemi, bir sonraki kuşak için yavru üretmek amacıyla hangi ailelerin yer alması gerektiğine karar vermektir. Bu yöntemin amacı, ortalama uygunluğun üzerindeki değerlere çoğalma fırsatı tanımaktır. Bir dizinin kopyalanma şansı, uygunluk fonksiyonuyla hesaplanmış dizinin uygunluk değerine bağlıdır (Parlak 2007).

2.1.6.5. Çaprazlama İşleminin Uygulanması

Mevcut gen potansiyellerini potansiyelini araştırmak üzere, bir önceki kuşaktan daha iyi nitelik içeren yeni kromozomlar yaratmak amacıyla çaprazlama işlemcisi kullanılmaktadır. Çaprazlama genellikle, verilen çaprazlama oranına eşit olasılıkla seçilen aile türlerine uygulanmaktadır.

GA'nın performansını etkileyen önemli parametrelerden birisi olan çaprazlama işlemcisi; doğal topluluklardaki çaprazlamaya karşılık gelmektedir. Çoğalma işlemi sonucunda elde edilen yeni topluluktan rastgele olarak iki kromozom seçilmekte ve karşılıklı çaprazlama işlemine tabi tutulmaktadır. Çaprazlama işleminde dizi uzunluğu L olmak üzere, $1 <= k <= L-1$ aralığında k tamsayısı seçilmektedir. Bu tamsayı değerine göre dizi çaprazlamaya uğratılmaktadır. En basit çaprazlama yöntemi tek noktalı çaprazlama yöntemidir. Tek noktalı çaprazlama yapılabilmesi için her iki kromozomun da aynı gen uzunluğunda olması gerekir. İki noktalı çaprazlamada ise kromozom iki noktadan kesilir ve karşılıklı olarak pozisyonlar yer değiştirir.

2.1.6.6. Mutasyon İşleminin Uygulanması

Çaprazlama mevcut gen potansiyellerini araştırmak üzere kullanılır. Fakat topluluk gerekli tüm kodlanmış bilgiyi içermezse, çaprazlama tatmin edici bir çözüm üretemez. Bundan dolayı, mevcut kromozomlardan yeni kromozomlar üretme yeteneğine sahip bir işlemci gerekmektedir. Bu görevi mutasyon gerçekleştirir. Yapay genetik sistemlerde mutasyon işlemcisi, bir kez daha elde edilemeyebilir iyi bir çözümün kaybını engellemektedir. İkili kodlama sisteminin kullanıldığı problemlerde mutasyon, düşük bir olasılık değeri altında bir bit değerini (0 veya 1 olabilir) diğer bir bit değerine dönüştürür. İkili kodlama sisteminin kullanılmadığı problemlerde ise daha farklı mutasyon yöntemleri kullanılmaktadır. Hangi yöntem kullanılırsa kullanılsın, mutasyonun genel amacı, genetik çeşitliliği sağlamak veya korumaktır.

2.1.6.7. Yeni Kuşağın Oluşması ve Döngünün Durdurulması

Yeni kuşak; çoğalma, çaprazlama ve mutasyon işlemlerinden sonra tanımlanmakta ve bir sonraki kuşağın ebeveynleri olmaktadır. Süreç, yeni kuşakla çoğalma için belirlenen uygunluk durumu ile devam etmektedir. İstenen hassasiyet derecesine göre de maksimum yineleme (iterasyon) sayısı belirlenebilmekte ve yineleme ilgili sayıya ulaştığında döngü durdurulabilmektedir. Durdurma ölçütü yineleme (adım) sayısı olabileceği gibi hedeflenen uygunluk değeri de olabilmektedir.

2.2. Bootstrap Yeniden Örnekleme Yöntemi

Biyoistatistikte herhangi bir popülasyonun parametresini tahmin etmek için o popülasyona ait gözlemlerden yararlanılır. Üzerinde çalışılan popülasyonun tüm gözlemlerini parametre tahmini için kullanmak hem zaman kaybına yol açacağı hem de maliyeti arttıracağı için popülasyonu en iyi şekilde açıklayacak olan örneklemden elde edilen verilerle çalışmak bu sorunları ortadan kaldırır. İstenilen büyüklük ve miktarda veri setleri oluşturmak için herhangi bir boyuttaki veri setinden gözlemler tesadüfi olarak yer değiştirilerek yeniden örneklenebilir. Bu sayede veri setinden daha fazla bilgi alınabilir. Bu şekilde tanımlanan yöntem "Bootstrap Yeniden Örnekleme Yöntemi" (BYÖY) olarak adlandırılır (Aktükün 2005).

BYÖY literatüre ilk kez Efron 'ın 1979 yılındaki makalesi ile girmiştir. Teorik gelişme Freedman (1981) ve Wu (1986) ile devam etmiştir. Daha sonraki gelişmelerden kitaplaştırılanlar ise tarihsel sırasıyla Beran ve Ducharme (1991), Hall (1992), Mammen (1992), Efron ve Tibshirani (1993), Davison ve Hinkley (1997) ve teorik bir çalışma olan Shao ve Tu (1995) olmuştur (Aktükün 2005).

Günümüzde bilgisayarların da gelişimiyle beraber çok sayıda araştırmaya konu olan BYÖY'de temel düşünce, eldeki örneklemi popülasyon olarak varsayıp buradan belirli sayıda tekrarlı örnekleme yaparak ilgilenilen tahmin edicinin yapay bir örnekleme dağılımını yaratmaktır (Aktütün 2005). Bir kez örnekleme yaparak ardından yeniden örnekleme yöntemleri kullanılarak evren parametrelerine ulaşma konusunda güvenilir sonuçlar elde edildiği yapılan çalışmalarla kanıtlanmıştır. Örneğin; verinin alındığı evrenin ortalaması tahmin edilmek istenirse, elde tahmin olarak örnek ortalaması bulunmaktadır. Evren dağılımı hakkında hiçbir varsayım mevcut değilse, küçük hacimli örneklemede, evren ortalaması için güven aralığı ve aralık tahmini

yapılması çok güvenilir olmayacaktır. Bu gibi sorunların altından kalkmak için eldeki veri üzerinden "yeniden *n* hacimlik örneklemeler" elde edilip, ilgili istatistiğin değeri çok kez gözlenip, istatistiğin dağılımına ilişkin bilgi elde edilebilir. (Takma ve Atıl 2006, Chernick 2008).

Anakütle parametresinin tahmin edicisi olan $\hat{\theta}$'nın örnekleme dağılımının oluşturulmasının amacı, söz konusu anakütle parametresinin tahmin edilmesi ya da test edilmesi içindir. Ancak teorik olarak mümkün olan bu yöntemin uygulanabilirliği konusunda kuşkular bulunmaktadır. Tahmin edicinin örnekleme dağılımını oluşturmak imkansız olmasa da son derece güç ve zaman alıcı bir iştir. Tahmin edicinin deneysel örnekleme dağılımını oluşturmak amacıyla ortaya atılan BYÖY bu sakıncayı ortadan kaldırır. Bu mantık doğrultusunda BYÖY algoritması, aşağıdaki şekilde tanımlanabilir (Özdemir 2011).

- Popülasyondan n hacimlik bir örneğin elde edilmesi
- Anakütle⇒ örnek
- Bu örnek kullanılarak anakütle parametresinin tahmin edicisinin hesaplanması
- Elde edilen bu örnek, anakütle ile ilgili başka hiçbir bilgi olmadığından, anakütlenin tek ve en iyi tahmin edicisi olarak kabul edilir. Bu nedenle bu örnek, anakütle gibi kabul edilerek her defasında iadeli seçimle her bir gözlemin örneğe girme olasılığı 1/n alınarak n hacimlik bir örneğin yeniden elde edilmesi ve bu sürecin B kez tekrarlanması

- Örnek →
  - ▸ 1.Bootstrap örneği
  - ▸ 2.Bootstrap örneği
  - .
  - .
  - ▸ B.Bootstrap örneği

- Her bootstrap örneği için ilgilenilen tahmin edicinin hesaplanması
- B sayıda örnekten hareketle bu tahmin edicilerin örnekleme dağılımının elde edilmesi
- Elde edilen bu dağılımdan, dağılımla ilgili ortalama, standart sapma ve standart hata gibi önemli tahmin ediciler ile parametre tahmin değerlerinin elde edilmesi

- Bu tahminler kullanılarak anakütle hakkında yorumların yapılması (Özdemir 2011)

BYÖY iadeli çekim mantığıyla rastgele örnekleme yaparak eldeki veri kümesini genelleştirme ilkesi üzerine kuruludur. Yani eldeki veri kümesi evren gibi düşünülürse bunun içinden örnek seçmek mümkün olabilmektedir. İadeli rastgele çekim ise şu şekilde örneklenebilir, veri kümesinden rastgele bir örnek çekip geri koyarak tekrar çekim yapılır. Bu yineleme veri kümesindeki gözlem sayısı kadar tekrarlanır. Yani 30 birimlik bir veri kümesinden BYÖY ile parametre tahmini yapılmak istenirse, çekim sonunda iadeli ve rastgele çekim yapılacağı için her türlü kombinasyon mümkün olabilmektedir. Bu yöntemde iadeli yöntem kullanılması sayesinde gözlem sayısı kadar örnek rastgele seçilir. Yerine koyarak örnek çekildiği için verilerin %33'ü örneklemde birden fazla, %67'si ise bir defa yer alacaktır (Karabulut ve Karaağaoğlu 2010).

Bir yeniden örnekleme tekniği olan bu yöntemle, tahmin değerlerini ve standart hataları belirlemek mümkündür. Diğer taraftan Zaman Serileri Analizi, Lineer Olmayan Regresyon Analizi, Kümeleme Analizi, Diskriminant Analizi, Lojistik Regresyon Analizinde ve her türlü hipotez testini sınamak için de kullanılabilmektedir.

Bu yöntemin temel amacı, mevcut veri setinden çok daha büyük veri setleri üretmek için yeniden örnekleme yapmaktır. BYÖY'deki varsayım; gözlenen verilerin, inceleme altındaki ana kütlenin temsilcisi olduğudur. Gözlenen verilerden elde edilen tekrarlı örnekleme gözlemleri, ana kütleden elde edilen tekrarlı örnekleme sürecinin kopyalanmasıdır.

2.3. Karar Ağaçları ve Random Forest Sınıflama Yöntemi

2.3.1. Karar Ağaçları

2.3.1.1. Karar Ağacı Kavramı

Regresyon ağaçları 1960 yılında Morgan ve Sonquist (1963) tarafından AID'in (Automatic Interaction Detection) geliştirilmesiyle literatüre girmiştir. Daha sonra 1970'lerde sınıflandırma ağaçlarını oluşturmak için THAID geliştirilmiştir. Karmaşık bir program olan ve veriye ağaçları uydurmak için 1980'de istatistikçiler (Breiman vd.1984) tarafından CART (Classification and Regression Trees) geliştirilmiştir. Ağaç-tabanlı algoritma olan ve istatistikçiler tarafından geliştirilmiş olan diğer bir metot da QUEST (Quick Unbiased Efficient Statistical Tree)'dir (Loh 1997, Kuzey 2012).

Karar Ağaçları; çok sayıda bilgi içeren bir veri kümesini, bazı kuralları uygulayarak daha küçük kümelere bölmek için kullanılan bir yöntemdir. Verinin içerdiği ortak özelliklere göre ayrıştırılması işlemi "sınıflandırma" olarak adlandırılmaktadır. Sınıflama, sınıfı tanımlanmış verilerden yola çıkarak, sınıfı belli olmayan verilerin sınıfını belirlemek üzere kullanılan veri madenciliği modelidir (Han ve Kamber 2006). Sınıflandırma için öncelikle tahmin için kullanılacak bir model oluşturulur. Oluşturulan bu model, sınıfı belli olmayan veriler üzerinden uygulanarak sınıflar tahmin edilir (Han ve Kamber 2006).

Sınıflandırma ve regresyon ağacı yönteminin kapsayıcı hedefi tahminde kullanılmak üzere bir modelin oluşturulmasıdır. Her bir bilgi birçok değişkenden oluşmaktadır. Tanım kümesi sayısal olan değişkenlere sayısal değişkenler adı verilir, tanım kümesi sayısal olmayan değişkenler ise kategorik değişkenler olarak tanımlanır. Bağımlı değişken adı verilen bir tane seçilmiş değişken mevcuttur. Kalan diğer değişkenler bağımsız değişken olarak adlandırılır; bunlar sayısal ya da kategorik olabilirler. Eğer bağımlı değişken kategorik ise problem bir sınıflandırma ağacı problemi olarak belirtilir ve bağımlı değişken *sınıf etiketi* olarak adlandırılır. Eğer bağımlı değişken sayısal ise, probleme *regresyon ağacı problemi* adı verilir (Gehrke 2003, WEB 1).

KA'lar aşamalı olarak karar vermede kullanılan olası yaklaşımlardan birisidir. Ağaç, dallar ve yapraklardan oluşur. Eğer yaprak artık dallara ayrılmıyorsa o yaprağa "karar düğümü" denir. Tüm yapraklar karar düğümü olana kadar ya da o yaprağa ait veri kalmayana kadar dallara ayrılmaya devam eder (Safavian 1991, Korkem 2013).

KA günümüzde çok çeşitli alanlarda kullanılmaktadır. Belirli bir sınıfın muhtemel üyesi olacak elemanların belirlenmesi, çeşitli değişkenlerin yüksek, orta, düşük risk grupları gibi çeşitli kategorilere ayrılması gibi durumların çözümünde kolay sonuç elde edilebilmesi açısından ön planda olan yöntemlerdir. Tıbbi olarak teşhis koymada, radar sinyal sınıflandırmasında, karakter tanımada, uzaktan algılamada kullanılabilir. KA, diğer sınıflandırma yöntemleriyle karşılaştırıldığında yapılandırılması ve anlaşılmasının daha kolay olması sebebiyle sıkça kullanılan bir yöntemdir (Safavian 1991, Korkem 2013). Yorumlamasının kolay olması, veri tabanı sistemleri ile birleştirilebilir olması nedeniyle tercih edilmektedir (Akdag vd 2006, Akdag vd 2011).

2.3.1.2. Karar Ağaçlarının Avantaj ve Dezavantajları

Sınıflandırma yapmak için KA kullanmak pek çok avantaj sağlar. KA basit olup buradan çıkan sonuçların anlaşılması ve yorumlanması çok kolaydır. Ağacın büyüklüğü veri tabanı büyüklüğünden bağımsızdır. Ağaçlar çok sayıda niteliği olan veriler için de oluşturulabilir. Maddeleyecek olursak;

- KA oluşturmak zahmetsizdir, yorumlamak kolaydır.
- Bağımlı ve bağımsız değişkenler için herhangi bir varsayım aranmaz.
- Bağımsız değişkenlerin hepsi kategorik ya da sayısal ölçüme sahip olabildiği gibi, karışık halde de olabilir.
- Kayıp verilere çözüm bulunur. Sadece değişkenlerin doğrusal kombinasyonu, ayırım kuralı olarak kullanılırsa kayıp veriye çözüm bulunamamaktadır.
- Aşırı değerler, çoklu bağlantı, heterojenlik ya da dağılıma ilişkin hatalardan etkilenmez.
- Veri setindeki değişkenlerin etkileşimleri bulunabilir ve açıklanabilir.
- Bağımsız değişkenlerin dönüştürülmesinden etkilenmez. Yani, bağımsız değişkenlerin logaritmik, karekök gibi dönüşümleri ağaç oluşumunu etkilemez.
- Araştırıcıya rehberlik edecek teorik bilgi ya da kaynak olmasa bile, yöntemin kendisi analiz için yol göstericidir.

- Çok sayıda bağımsız değişken olsa bile birkaç önemli değişkeni kullanarak yararlı sonuçlar üretebilir.
- Her disiplindeki bireyler tarafından kolayca anlaşılabilir (Breiman vd.1984, Sümbüloğlu ve Akdağ 2007).

Dezavantajları ise;

- Sınıf sayısı fazla ve denek sayısı az olduğunda model oluşturma çok başarılı değildir.
- Ağaç oluşturma sonucu ve ağaç budama karmaşası vardır.
- Bir sınıflama ağacının karmaşıklık ölçüsü değeri o ağacın terminal düğüm sayısına eşittir. Terminal düğüm sayıları ve dolayısıyla karmaşıklığı yüksek olan büyük ağacın anlaşılması ve yorumlanması da güçtür.
- Olasılıksal bir modele dayanmamaktadır. Elde edilen sonuçların olasılık düzeyi ya da güven aralığı yoktur. Sonuçların güvenilirliği sadece daha önceki bulguların doğruluğu temeline dayanır.
- Büyük ağacın pratikte ortaya çıkardığı sorunların çözümü için ağacın budanması gereklidir. Büyük ağacın budanması daha küçük ağaçlar dizisi demektir ve oluşturulan bu dizi içerisinden optimum ağaç seçilir. Optimum ağaç büyük ağaçtan daha az karmaşıktır ancak, uygun sonuçla büyük ağaçtan daha az uyumludur ve hatalı sınıflama oranı daha yüksektir (Breiman vd.1984, Temel vd 2005, Sümbüloğlu ve Akdağ 2007).

2.3.1.3. Karar Ağacı Yapısı

KA'ları aşamalı karar vermede kullanılan olası yaklaşımlardan birisidir. Bir ağaç yapısı düğüm, dal ve yaprak olarak adlandırılan üç temel kısımdan oluşur. Bu ağaç yapısında her bir değişken bir düğümle temsil edilir. Dallar ve yapraklar ağaç yapısının diğer elemanlarıdır. Ağaçtaki en son kısım yaprak en üst kısım ise kök olarak adlandırılır. Kök ve yapraklar arasında kalan kısımlar ise dal olarak ifade edilir (Quinlan 1993). Başka bir ifadeyle bir ağaç yapısı; verileri içeren bir kök düğümü, iç düğümler (dallar) ve uç düğümlerden (yapraklar) oluşur. KA'daki her bir düğüm sadece bir ana düğüme ve iki veya daha fazla alt düğüme sahiptir. Veri seti, ağaç tarafından belirlenen karar iskeletine göre aşağı doğru hareket ettirilerek ve sıralı olarak bir yaprağa ulaşılana kadar alt bölümlere ayrılarak sınıflandırılır. Bir KA yapısı oluşturulmasında temel prensip verilere ilişkin bir dizi sorular sorulması ve elde edilen cevaplar doğrultusunda

hareket edilerek en kısa sürede sonuca gidilmesi olarak ifade edilebilir. Bu şekilde KA sorulara karşılık aldığı cevapları toplayarak karar kuralları oluşturur. Ağacın ilk düğümü olan kök düğümünde verilerin sınıflandırılması ve ağaç yapısının oluşturulması için sorular sorulmaya başlanır ve dalları olmayan düğümler ya da yapraklara gelene kadar bu işlem devam eder. Bu noktada belirtilmesi gereken husus yaprak sayısı kadar durum oluşacağıdır. Oluşturulan ağacın işlevselliğinin başka bir ifadeyle yeni bir veri seti için genelleme kabiliyetinin belirlenmesi adına test veri seti kullanılır. İlgili veri seti ile oluşturulan ağaç yapısına yeni gelen bir test verisi, ağacın kökünden girer. Kökte test edilen bu yeni veri test sonucuna göre bir alt düğüme gönderilir. Ağacın belirli bir yaprağına gelene kadar bu işleme devam edilir. Kökten her bir yaprağa giden tek bir yol veya tek bir kararla ilgili durum vardır. Eğer yaprak artık dallara ayrılmıyorsa o yaprağa "karar düğümü" denir. Tüm yapraklar karar düğümü olana kadar ya da o yaprağa ait veri kalmayana kadar dallara ayrılmaya devam ederler (Çölkesen 2009, Kahraman 2010, Kaya 2010).

2.3.1.4. Karar Ağacı Oluşturma Aşamaları

Ağaç oluşturma işlemine ilk noktadan başlanır. Bu noktaya "ana nokta" denir. Ana nokta örneklemdeki tüm bireyleri ya da gözlemleri içerdiğinden heterojen bir yapıya sahiptir. Ana noktada ilk bağımsız değişken için iki sonuçlu (evet – hayır gibi) bir soru ile başlanır. Örneğin, $x \leq d$ burada x, veri setinde bir değişkenin değeri, d ise gerçek bir sayıdır. Yanıt ya evet ya da hayır olacaktır. Evet yanıtı için sola, hayır yanıtı için sağa yönlendirme yapılır. Ayrılan bu iki nokta ana noktaya göre kendi içlerinde daha homojen yapıya sahiptir (Sümbüloğlu ve Akdağ 2007).

1. Aşama: İlk aşamada bağımlı değişken ve söz konusu bağımlı değişkene etki ettiği düşünülen bağımsız değişkenler belirlenir. Bağımlı değişken aynı zamanda başlangıç noktasını oluşturan düğümü, yani ana noktayı ifade etmektedir.

2. Aşama: Bu aşamada bağımlı değişkene etki eden her bağımsız değişken incelenir. Bu bağlamda KA, tekrarlanan bölünmelere dayalı bir süreç izler. Söz konusu işlemler incelenen bölümler arasında en anlamlı bölümlendirme bulunana kadar devam edecektir. Bazı durumlarda bölümlendirme kriteri araştırmacı tarafından da belirlenebilmektedir. KA'da genellikle bölümlendirmenin istatistiksel olarak anlamlılığını tespit etmek için ki-kare analizi kullanılmaktadır. Oluşturulan ikili bölümlendirmeler arasından seçim yapılırken sınıflar arasındaki varyasyonun

maksimize olduğu, sınıf içi varyasyonun ise minimize olduğu bölümlendirme tercih edilmektedir.

3. Aşama: Her değişken için yukarıdaki işlemler gerçekleştirildikten sonra tahmin düzeyi en yüksek olan değişken seçilerek yaprak düğümlerini oluşturmak için kullanılır (Özekeş 2002).

Eldeki verilerden yola çıkılarak birçok farklı KA oluşturmak mümkündür. Amaç en optimum ağacı oluşturmak olsa da, zaman ve farklı kısıtlardan dolayı her zaman bu mümkün olmayabilir. Her ne kadar optimum olmasa da doğruluk derecesi uygun bir ağacı sağlayacak etkin algoritmalar geliştirilmiştir. Bu algoritmalar genellikle, veriyi dallara ayırmak için hangi değişkenden başlanması gerektiği ve bunun yanında ayırma gerçekleştikten sonra oluşan alt veri grubunu tekrar alt dallara ayırmak için hangi değişkenden başlanması gerektiği ile ilgili lokal kararlar alarak bir KA geliştiren, greedy yaklaşımını kullanırlar. Bu algoritmaların birçoğu yukarıdan aşağı (top-down) veya Hunt‟ın algoritması olarak bilinen Böl ve elde et (divide-and-conquer) yönteminin değişik versiyonlarını kullanmaktadırlar (Akman 2010).

2.3.1.5. Karar Ağacı Algoritmaları

2.3.1.5.1. HUNT Algoritması (Böl ve elde et)

Böl ve elde et olarak da adlandırılan Hunt algoritması aşağıdaki adımları içerir.

D*t* , t düğümü ile bağlantılı kayıtlar kümesi ve y={y1,y2,...,yc} sınıf etiketleri olsun;

- Eğer D*t*, Yt ile aynı sınıfa ait kayıtları içerirse; t, Yt olarak etiketlenen yaprak düğümüdür.
- Eğer Dt birden fazla sınıfı içeren kayıtlardan oluşursa, kayıtları daha küçük alt veri kümelerine parçalamak için değişkeni test etme koşulunu uygula
- Aynı şekilde oluşan her alt veri kümesine tekrarlamalı olarak bu testi yaprak düğüme ulaşana kadar yinele (Akman 2010).

KA oluşturulurken ağaca hangi düğümden başlanılacağı büyük önem taşır. Hunt algoritmasının en önemli eksiği ağaca hangi düğümden başlanılacağının rastgele seçilmesidir. Kök düğümün rastgele bir seçimle belirlenmesi olası tüm ağaçları ortaya çıkarmak için büyük bir emek ve zaman kaybına sebep olacaktır. Örneğin, 5 değişken ve 20 elemanlı bir veri kümesinden faydalanarak oluşturabileceğimiz KA sayısı 106'dan daha fazla olacaktır.

2.3.1.5.2. CHAID Algoritması

1980 yılında G.V. Kass tarafından geliştirilen CHAID algoritmasında, bağımlı değişkeni en fazla etkileyen bağımsız değişken, bağımlı değişkenin sürekli olması durumunda F testi, kategorik olması durumunda ise Ki Kare testi kullanılarak belirlenir. Kategorik *(Nominal / Ordinal)* ve sürekli değişkenler üzerinde çalışabilmesi, ağaçta her düğümü ikiden fazla alt gruba ayırabilmesi gibi nedenlerle günümüzde de tercih edilen bir algoritmadır. Kullanılan istatistiksel test, bağımlı değişkenin türüne bağlıdır. Eğer bağımlı değişken sürekli ise, F testi kullanılır. Eğer değişken nominal ise, Pearson Ki Kare testi kullanılır. Eğer değişken sıralı ise, o zaman olabilirlik-oran (likelihood-ratio test) testi kullanılır.

Seçilen her bir çift için, CHAID elde edilen p-değerinin belirli bir birleştirme eşik değerinden büyük olup olmadığını kontrol eder. Eğer cevap pozitif ise, değerleri birleştirir ve ilave potansiyel çiftleri birleştirmek için arama yapar. Bu süreç anlamlı çiftler bulunmayana kadar tekrarlanır. Böylece hâlihazırdaki düğümü bölmek için kullanılan en iyi girdi değişken seçilir, öyle ki her bir alt düğüm seçilen değişkenin değerlerinin homojen bir grubundan oluşur. Diğer yandan, eğer en iyi girdi değişkenin düzeltilmiş p değeri belirli bir bölme eşik değerinden küçük değil ise bölme gerçekleşmez. Aşağıdaki durumlardan herhangi bir tanesi ile karşılaşıldığı zaman işlem de durur:

- Maksimum ağaç derinliğine ulaşıldığında.
- Üst düğüm olmak için, düğümdeki minimum sayıda duruma ulaşıldığında.
- Alt düğüm olmak için düğümdeki minimum sayıda duruma ulaşıldığında (Kuzey 2012).

2.3.1.5.3. ID3 Algoritması

1970'lerin sonunda, J.Ross Quinlan, Hunt'ın 'Böl ve Elde Et' (Divide and Conquer) algoritmasını geliştirerek ID3 KA algoritmasını oluşturmustur. Hunt'ın kullanmış olduğu metotta özellikler rastgele seçilirken Quinlan, değişken seçimi için entropi yöntemini kullanmış ve böylelikle Hunt metodunun en önemli eksikliğini gidermiştir.

ID3, başlangıçta satranç oyununda etkin stratejiler öğrenme şeklindeki amaçlara yönelik olarak kullanılmıştır. Günümüzde ID3 hem akademik, hem de sanayi alanında pek çok sorunu çözme amaçlı kullanılmış, değiştirilmiş, geliştirilmiş ve zaman içerisinde yaygın kullanım alanı bulmuştur (Gülpınar 2008).

2.3.1.5.4. C4.5 Algoritması

ID3'ün bir evrimidir ve aynı yazar tarafından (Quinlan 1993) ortaya sunulmuştur. Bölme kıstası olarak kazanç oranını (gain ratio) kullanır. Bölme işlemi, bölünmesi gereken örneklerin sayısı belirli bir eşik değerinden düşük olduğu zaman sonlanır. Ağacın büyüme işlemi gerçekleştikten sonra hata tabanlı budama (Error –based pruning) işlemi başlar. C4.5 sayısal değişkenleri işleyebilir (Yıldırım 2003).

2.3.1.5.5. CART (Sınıflandırma ve Regresyon Ağaçları) Algoritması

Kısaca CART (Classification and Regression Trees) olarak adlandırılan Sınıflandırma ve Regresyon Ağaçları; Breiman, Freidman, Olshen & Stone tarafından 1984 yılında literatüre girmiştir. İstatistik dünyasında ağaçların kullanılması, 1963 yılında Morgan & Sonquist tarafından yapılan AID ve daha sonra 1970'lerin başında Morgan ve Messenger tarafından yapılan THAID çalışmalarına kadar uzanmaktadır (Gülpınar 2008).

CART algoritması, ikili KA oluşturan, bağımlı değişkenin kategorik veya nümerik olmasına bağlı olarak sınıflandırma ya da regresyon ağaçları üreten parametrik-olmayan bir tekniktir. ID3'te olduğu gibi, en iyi ayrım niteliğinin seçiminde entropi kullanılmaktadır. Ancak, her sayıda düğümün oluşturulabildiği ID3'ten farklı olarak, burada yalnızca iki adet dal oluşturulmaktadır (Gülpınar 2008).

Ağaçtaki her bir düğümde, her bir bağımsız değişken için gelişim skoruna dayalı olarak sürekli değişkenler için en iyi kesim noktası ya da kategorik değişkenler için en iyi kategori grupları oluşturulur. En iyi kestirici değişken seçimi, kategorik bağımlı değişkenler için Gini ya da Twoing, sürekli bağımlı değişkenler için ise En Küçük Kareler Sapması Indeks Hesaplamaları'na göre yapılmaktadır. Burada amaç, bağımlı değişkenle ilgili verinin, mümkün olabildiğince homojen alt kümelerinin meydana getirilmesidir. CART KA tekrarlamalı ve ikili bir bölme prosedürüdür. Veri ham hali ile analiz edilir. Kök düğümünden başlanılarak veri, iki alt düğüme ve daha sonra her bir alt düğüm tekrar iki ayrı alt düğümlere ayrılır. Böylece ağaçlar durdurma olmaksızın en yüksek büyüklüğe ulaşmış olurlar. Daha sonra maksimum büyüklükteki ağaç, budama metodu yolu ile geriye doğru, yani kök' e doğru budanmaya başlanır. Budama işleminden sonraki bölünme ağacın toplam performansına faydası en az olan bölünmedir (Kuzey 2012).

CART mekanizması otomatik sınıf dengelemesi ve eksik değer (missing value) işlemlerini içermektedir. (Steinberg 2009).

2.3.1.6. Karar Ağacı Bölünme Kuralları

En optimum ağacı oluşturmak için dallanmada seçilecek değişken ve bu değişkenin seçilmesi için uygulanacak olan kriterler çok önemlidir. KA sınıflandırıcısıyla veriler belirli kriterler doğrultusunda en iyi özelliğe dayalı olarak bölünürler. Düğümler bölünme şekline göre çoklu ya da ikili olmaktadır. En iyi düğüm bölünmesi için homojen dağılıma sahip düğümler tercih edilir. Homojenliği belirlemek için düğümün impurity değeri hesaplanır. Impurity değerini ölçmek için gini katsayısı, entropi, information gain gibi değerler kullanılır. Tüm veriler KA'da doğru sınıflandırılana kadar KA dallandırılır.

Entropi

Entropi, olayların ortaya çıkma olasılıklarıyla ilişkili olup belirsizliğin ölçülmesi için kullanılan bir ölçüttür. Entropi bilgi ile ilişkilidir ve belirsizlik arttıkça eldeki veriyi tanımlamak için daha fazla bilgi gerekecektir. Entropi 0 ile 1 arası değerler alır ve 1 değerine yaklaştıkça da belirsizliğin arttığın söylenebilir.

ID3, C4.5, CART algoritmaları en iyi ayırıcı özelliğe sahip değişkeni bulmak için entropiden faydalanır. KA'lar kurulurken amaç, veri setinin entropisini örneklerin hepsinin tek sınıf olarak ifade edildiği yaprak düğüm entropi değeri sıfır olana kadar düşürmeye çalışmaktır (Akman 2010).

Ağacın hangi düğümle başlayacağına (kök düğüm) entropi yöntemiyle karar verilir. Bir alanın entropi ölçüsü ne kadar yüksek ise o alan kullanılarak ortaya konulan sonuçlar da o oranda belirsiz ve kararsızdır. Bu nedenle KA'nın kökünde entropi ölçüsü en az olan alanlar kullanılır.

Entropi şu şekilde hesaplanır: Veri tabanından ilgili veri kümesini ele alalım. Veri kümesinin sınıf niteliğinin (bağımlı değişken) alacağı değerlere göre $\{C_1, C_2, \ldots C_k\}$ olmak üzere k tane sınıfa ayrıldığı varsayılsın. Bu sınıflarla ilgili olarak ortalama bilgi miktarına ihtiyaç duyulabilir. Burada T sınıf değerini içeren küme için P_T sınıfların olasılık dağılımları olup aşağıdaki şekilde hesaplanır.

$$P_T = \left(\frac{|C_1|}{|T|}, \frac{|C_2|}{|T|}, \ldots, \frac{|C_k|}{|T|}\right)$$

$|C_i|$ ifadesi C_i kümesindeki elemanların sayısını vermektedir. $\frac{|C_1|}{|T|}$ ise, p_1 olasılığını ifade etmektedir. Bu durumda T için ortalama bilgi miktarı yani entropi aşağıdaki şekilde ifade edilir:

$$H(T) = -\sum_{i=1}^{n} p_i \ log_2(p_i)$$ (Gülpınar 2008).

Information Gain

Information gain ID3, C4.5, CART algoritmalarında veri setindeki değişkenin etkinliğinin ölçüm değeri olarak kullanılır. Bilgi kazancı en yüksek değişken en iyi dallara ayırmayı sağlayacak değişken olarak seçilir ve bölünmeye o değişkenden başlanılır.

Eğer veri seti *D*, *n* tane alt bölüme *X* değişkeninden bölünecekse, *X*'e ait bilgi kazancı aşağıdaki gibi hesaplanır.

$$Information\ Gain\ (D,X) = E(D) - \sum_{i=1}^{n} p(D_i)\ E\ (D_i)$$

E(*D*): Veri setinin X üzerinden bölünmeden önceki entropisi

E (*Di):* i alt bölümünün X üzerinden bölünme olduktan sonraki entropisi

p(*Di*): i alt bölümünün X üzerinden bölünme olduktan sonraki olasılığı

Bilgi kazancı hesaplanırken, öncelikle veri setinin alt bölümlere ayrılmadan önceki durumunun entropisi bulunur, daha sonra her alt bölümün entropisi hesaplanır. Bu iki değer arasındaki farkın en yüksek olduğu değişken dallara ayırma kriterinde en uygun değişken olarak seçilir (Akman 2010).

Gini katsayısı

Gini katsayısı CART için (Breiman vd 1984) geliştirilmiştir. Gini fonksiyonu sınıflara ilişkin değişkenlerin impurity değerlerini ölçer. C_i gibi bazı sınıflara sahip T şeklinde verilen bir eğitim setinin;

$$\mathrm{f}\frac{(C_i, T)}{|T|}$$

şeklinde bir olasılığa sahip olduğu söylenebilir. Genel Gini fonksiyonu veya impurity ölçümü;

$$\sum \sum_{j \neq i} (f(C_i, T)/(f(C_j, T)/|T|)$$

şeklinde ifade edilebilir.

Gini katsayısı basit ve hızlı bir şekilde hesaplanabilir. Bu katsayı rastgele seçilen bir objenin bir düğümden $f(C_i, T)/|T|$ olasılığı ile i sınıfına atanması olarak özetlenebilir (Çölkesen 2009).

2.3.1.7. Budama Metodları

Sınıflandırma ağacının oluşumunda budama (pruning) aşaması; aşırı uyum(overfitting) problemini engellemek ve yanlış sınıflandırma hatasını minimize etmek için ağacın gerçek büyüklüğüne karar vermek için kullanılır. Aşağıdan – yukarı (bottom-up) budamada, ağaç büyüme safhasında ağacın her bir yaprak düğümü, n

değeri kullanıcı tarafından belirlenen c gibi bir eşik değerin altına düşene kadar büyümeye devam eder (Gehrke 2003).

Genel olarak sıkı bir durdurma kriteri uygulanarak küçük ve yetersiz uyumlu (under-fitted) KA oluşturulabilir. Diğer taraftan, gevşek durdurma kriteri kullanıldığı zaman ise, eğitme kümesine aşırı uyumlu (over-fitted) çok büyük bir KA oluşturulabilir. Budama metodu (Breiman vd 1984) tarafından önerilmiş ve yukarıda belirtilen ikilemi çözmek için geliştirilmiştir. Bu yönteme göre, gevşek bir durdurma kriteri kullanılırsa KA'nın veri kümesiyle aşırı uyumlu (overfit) olması sağlanır. Daha sonra aşırı uyumlu olan (overfit) bu ağaç, genellemeye katkısı olmayan alt dallar çıkarılarak daha küçük bir ağaç olacak şekilde kesilir. Birçok çalışmada budama metotlarının uygulanması, KA'nın genelleme performansını iyileştirmektedir (Rokach ve Maimon 2010). Budama metodunun diğer önemli bir tarafı işlem doğruluğu için karmaşıklıktan uzak olmayı sağlamasıdır (Bohanec ve Bratko 1994). Eğer hedef, yeteri kadar doğru ve özlü kavram açıklaması üretmek istemekse, budama bu aşamada çok faydalı olacaktır. Bu süreçte, başlangıç KA doğru bir ağaç olarak kabul edilir. Böylece budanan KA'nın doğruluğunun, başlangıç KA'ya ne kadar yakın olduğu belirlenebilir. Farklı budama yöntemleri mevcuttur. Bunların birçoğu, düğümlerin yukarıdan aşağıya ya da aşağıdan yukarıya gezinmesi şeklinde gerçekleştirilir. Eğer bu işlem esnasında belirli bir kriter sağlanırsa, ilgili düğüm budanır (Kuzey 2012).

2.3.1.8. Karar Ağaçlarının İyileştirilmesinde Kullanılan Yöntemler

Sınıflandırıcıların birleştirilmesi, yeniden örneklenen veri setleri ile oluşan sınıflandırıcıların ayrı ayrı gözden geçirilmesi ve sonuçta ortaya çıkan tahminler ile sınıflandırma işleminin gerçekleştirilmesi süreçlerini içermektedir. Birleştirme sonucu elde edilen sınıflandırıcılar ile yapılan sınıflandırma doğruluğunun, genel olarak her bir sınıflandırıcının tekil olarak kullanılmasına göre daha iyi olduğu ifade edilmektedir Teorik ve deneysel gözlemlerle yapılan araştırmalar, birleştirme yöntemi sonucunda birleştirmede kullanılan sınıflandırma yönteminin doğruluğunda artış olmasının yanında sınıflandırma hatalarında da azalma olduğunu göstermektedir (Çölkesen 2009).

Temel olarak kullanılan yöntemler; Boosting yöntemi, Bagging yöntemi, Multi-Boosting yöntemi ve Random Forest yöntemidir.

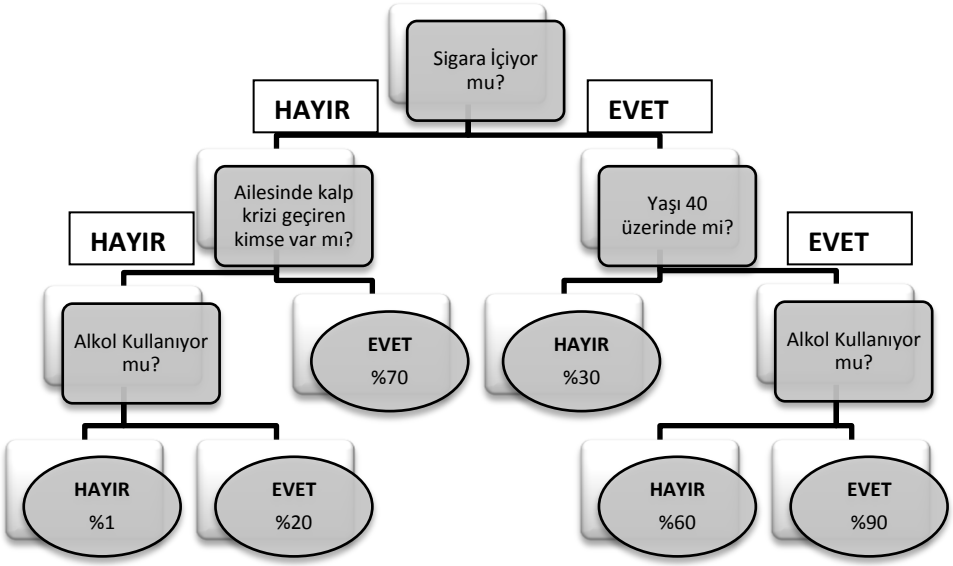
2.3.2. Random Forest Sınıflama Yöntemi

Günümüzde RF algoritması, sınıflandırmada çok iyi performans sergilediği için sıklıkla tercih edilmektedir. Son yıllarda geliştirilen RF sınıflandırıcısı, hem hızlı hem de yüksek doğruluk sağlaması yönünden çok iyi iki yöntem olarak bilinen Boosting (Freund ve Schapire 1996) ve Bagging (Breiman 1994) yöntemlerine göre daha kullanışlıdır.

Farklı veri grupları üzerindeki sınıflandırma başarısının değişkenlik göstermesinden dolayı bazen tek bir sınıflandırıcıdan alınan sonuçlar verimli olmayabilmektedir. Bu durumda başarı oranını ve doğruluğunu arttırmak amacıyla sınıflandırıcılar birleştirilebilmektedir. Bu yöntemler, birbirinden farklı çok sayıda sınıflayıcının yaptığı sınıf tahminlerini oylamaya tabi tutar ve oylama sonucunda en çok oyu alan sınıfı topluluğun sınıf tahmini olarak sunar. Bu yöntemler, çok sayıdaki sınıflayıcının kararını dikkate aldığı için daha güvenilir tahminler ortaya koymaktadır.

RF, birçok KA'dan meydana gelen bir yapıdır. Bu analiz sırasında RF'deki her bir ağaç için BYÖY ile veri kümesinden örneklem seçilir ve seçilen verilerin 2/3'ü ağaç oluşturmak için kullanılır ve bir sınıflama yapılır. Bu sınıflamalara oy verilir. RF algoritması "Forest" içindeki tüm ağaçlardan en çok oy alanı seçer ve onun sınıflamasını kullanır (Türkoğlu 2006, Coşgun ve Karaağaoğlu 2011). Her ağaç, veri setindeki örneklerin rastlantısal bir şekilde, yenisiyle değiştirilmesi mantığından oluşur. Yani her KA'da, veri setindeki tüm değişkenlerden rastgele seçilen az sayıda değişken kullanılmaktadır (Akman vd 2011).

KA'lar çok boyutlu (nitelikli) veriyi, belirlenmiş nitelik üzerindeki bir koşul ile parçalara böler. Her seferinde verinin hangi niteliği üzerinden hangi koşula göre işlem yapılacağına karar vermek ancak çok büyük kombinasyon setleri çözümüyle mümkün olabilir. Örneğin; 5 değişken ve 20 örnekleme sahip bir veri kümesinden, 10^6'dan fazla sayıda farklı KA oluşturulabilir. Bu sebeple her parçalanmanın metodolojik olması gerekmektedir.



Şekil 2.2 Karar Ağacı Yapısı

Şekil 2.2, kalp krizi geçiren kişilerin çeşitli özellikleri ele alınarak verilerin sisteme aktarılması durumunda hayali olarak elde edilebilecek sonuçlardır. Sorulacak sorular ve bu sorulara gelebilecek cevapların yönlendirdiği başka soruların bulunduğu bir ağaç yapısı olarak adlandırılan KA'lar ile değerlendirme yaparken yeni gelen bir veri, ağacın kökünden girer. Kökte test edilen bu yeni veri test sonucuna göre bir alt düğüme gönderilir. Ağacın belirli bir yaprağına gelene kadar bu işlem devam eder. Kökten her bir yaprağa giden tek bir yol vardır ki bu yol kuralı oluşturur. İç düğümlerde (dikdörtgenler) çeşitli girdilere göre ağacın dallanması söz konusudur. Yapraklarda ise sonuç olarak elde edilen değerler gösterilmiştir (WEB 2, Yıldırım 2003, Kayri ve Boysan 2008).

En az örnek sayısıyla çok fazla verinin sınıflandırmasını yapmak için uygun metot bulunması amaçlanmıştır. Random Forest yöntemi bu açıdan bakıldığında en ideal sınıflayıcılar arasında yer almaktadır (Akman 2010).

RF sınıflandırıcısı, özellikle Boosting yöntemine göre, çok daha hızlıdır. Yeterliliği ve doğruluğu ile çok kullanışlı bir sınıflandırıcıdır. Diğer benzer yöntemlerle karşılaştırıldığında ise RF oldukça basit, karmaşık olmayan mekanizmaya sahip bir yöntemdir (Breiman 2001).

2.3.2.1. Random Forest Yönteminin Algoritması

RF, ağaç tipi sınıflandırıcılar topluluğudur. Bagging yönteminin gelişmiş bir şekli olarak kabul edilebilir. (Breiman 2001). Breiman ve Cutler (2005) RF'yi, şuan ki algoritmalar arasından doğruluğu eşsiz olan sınıflandırıcı olarak tanımlamıştır. Ayrıca hızlı ve belirli bir kalıbı olmayan bir yöntem olarak ifade etmiştir. Bu yöntemde ne kadar istenirse o kadar sayıda ağaçla çalışılabilir.

RF içerisindeki her KA, orijinal veri setinden BYÖY tekniği ile farklı örneklemler seçilerek oluşturulmaktadır. Forest, ağaçların yapmış olduğu sınıf tahminlerini bir araya getirerek, en iyi sınıf tahminini ortaya koymaktadır. RF'da ayırıcı özellikteki değişken, tüm değişkenler arasından rastgele olarak belirlenen m tane değişken içerisinden seçilmektedir. Her ağaç için m sayısı sabittir ve genellikle $\sqrt{p}$ (p değişken sayısını ifade etmektedir) şeklinde alınması öngörülmektedir (Breiman 2001, Korkem 2013).

Breiman (2001) RF yönteminde, {h(x, θ_k)k=1,…} şeklinde ağaç tipi sınıflandırıcılar kullanmaktadır. Burada x, girdi verisini; θ_K, rastgele vektörü temsil etmektedir. RF yönteminde Bagging, rastgele özellik seçimi ile birlikte kullanılır. RF'de Bagging yönteminin tercih edilmesinin iki önemli nedeni vardır; birincisi, Bagging işleminde rastgele özellik kullanıldığından doğruluğun artması, ikincisi; genelleştirilmiş hataların (Out-of-bag (OOB)) hesaplanabilmesidir. Rastgele özellik seçimi için öncelikle gerçek veri setinden yer değiştirmeli olarak yeni bir veri seti oluşturulur. Ardından, rastgele özellik seçimi kullanılarak yeni veri setinden bir ağaç geliştirilir. Geliştirilen ağaçlar budanmaz. Pal (2005), budama metodunun seçiminin ve özellik seçim ölçütlerinin olmamasının ağaç tabanlı sınıflandırıcıların performansını etkilediğini belirtmektedir. Budamanın olmaması RF'yi diğer KA yöntemlerinden daha avantajlı hale getirmektedir.

RF sınıflandırıcısı ile bir ağaç üretmek için kullanıcı tarafından tanımlanacak 2 parametre gereklidir. Bu parametreler, en iyi bölünmeyi belirlemek için her bir düğümde kullanılan değişkenlerin sayısı (m) ve geliştirilecek ağaçların sayısı (N)'dir (Pal 2005).

Araştırıcı tarafından başlangıç değeri m rastgele seçilir sonraki m'ler ise genelleştirilmiş hatalara (GH) göre arttırılır ya da azaltılır. m azalınca korelasyon ve güç azalır, m artınca korelasyon ve güç artar. Bu şekilde en uygun m bulunur ve

sınıflandırma duyarlılığı artar, hata azalır. RF' da tüm girdi değişkenleri kullanılmamaktadır. RF algoritmasında hesap karmaşasını sadeleştirmek ve ağaçlar arasındaki korelasyonu azaltmak amacıyla bir m değeri seçimi yapılmaktadır. Bu m değeri, sınıflandırmada önemli olan değişken sayısını ifade etmektedir. Önemli değişkenler, değişken önem ölçümleriyle belirlenmektedir. RF' den elde edilen 3 parametre vardır. Bu parametreler genelleştirilmiş hata, değişken önemi ve yakınlık analizidir.

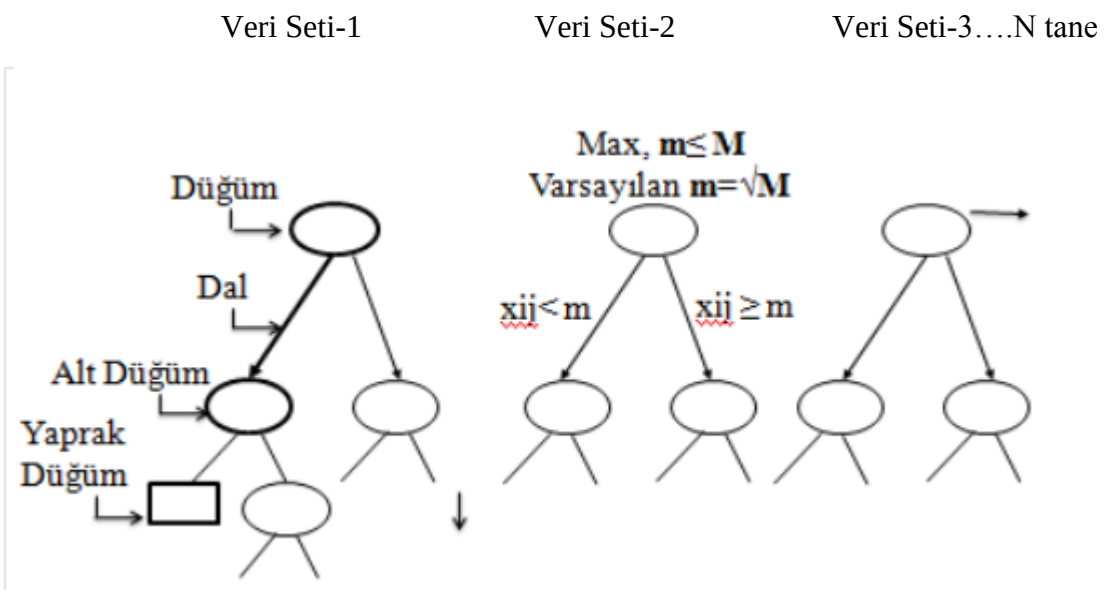
Genelleştirilmiş hata verisi, sınıflandırma doğruluğunun anlaşılmasına yardımcı olur. *T* veri setinden T_k yer değiştirmeli yeni veri seti üretilir. Yeni veri seti kullanılarak $h(x, T_k)$ sınıflandırıcısı oluşturulur. Sınıflandırıcı ile mümkün olan tahminler arasından oylama yapılır. Veri setindeki her (*x, y)* için sadece bu sınıflandırıcı ile oylama gerçekleşir. T_k, *x* ve *y*' yi içermez. (Breiman 2001).

RF algoritması, genelleştirilmiş hata verisindeki verilerin yerleri değiştirildiğinde tahmin hatasının ne kadar olduğunu inceleyerek değişkenlerin önemini, etkilerini hesaplar. Değişken önemi (kullanılan değişkenlerin ne kadar önemli olduğunun belirlenmesi) değişkenlerin yerleri değiştirilerek belirlenir. Değişimler sonucunda oluşan hatalar o değişkenin işlemdeki önemini ortaya koyar. 4 çeşit değişken önemi ölçüm yöntemi vardır. Bunlar; Hata artışı, Ortalama hata payı artışı, Hata payı artışının diferansiyeli ve Gini düşüşüdür.

RF, budama olmaksızın en büyük boyutta ağaç geliştirmek için CART (Classification and Regression Tree) algoritmasını kullanmaktadır (Breiman 2001). CART algoritmasında, bir düğümde belirli bir ölçüt uygulanarak bölünme işlemi gerçekleştirilir. Bunun için önce tüm niteliklerin var olduğu değerler göz önüne alınır ve tüm eşleşmelerden sonra iki bölünme elde edilir. Bu bölünmeler üzerinde seçme işlemi uygulanır. Bölünme işlemlerinde homojen sınıf dağılımına sahip düğümler tercih edilir. Düğüm homojenliğinin ölçümünde; Gini katsayısı, Entropi, Yanlış Sınıflama Hatası, Gain Oranı Kriteri gibi ölçütler kullanılmaktadır. RF yöntemi, Gini katsayısını kullanmaktadır. Verilen bir T veri seti için rastgele bir örnek seçilsin ve bu örnek C_i sınıfına ait olsun. Bu duruma göre Gini katsayısı şöyle ifade edilir;

$$\sum \sum_{j \neq i} \left(\frac{\mathrm{f}(C_i, T)}{|T|} \right) \left(\frac{\mathrm{f}(C_i, T)}{|T|} \right)$$

Eşitlikte, $f(C_i,T)/|T|$ *seçilen* örneğin C_i sınıfına ait olma olasılığını gösterir (Pal 2005). Gini ölçümleriyle, en küçük Gini katsayısına sahip olan bölünme pozisyonu belirlenir. Gini katsayısı büyüdükçe sınıf heterojenliği artarken, Gini katsayısı azaldıkça sınıf homojenliği artar. Bir alt düğümün Gini katsayısı, bir üst düğümün Gini katsayısından daha az olduğunda o dal başarılıdır. Gini katsayısı sıfıra ulaşınca yani her bir yaprak düğümde bir sınıf kalınca ağaç dallanma işlemi sonlanır. Kaç tane ağaç üretmek istenirse her düğüm için en iyi dal belirlenerek o kadar ağaç üretilir (Liaw ve Wiener 2002). Kısaca oluşturulan veriler kullanılarak belirlenen bölünme ölçütlerine göre düğümler dallara ayrılmakta ve ağaç yapıları oluşmaktadır. Şekil 2.3'de RF sınıflandırıcısında belirlenen en uygun bölünme pozisyonlarına göre oluşturulan ağaç yapısı örneği gösterilmiştir. Şekildeki x_{ij}, girdi verilerini temsil etmektedir.

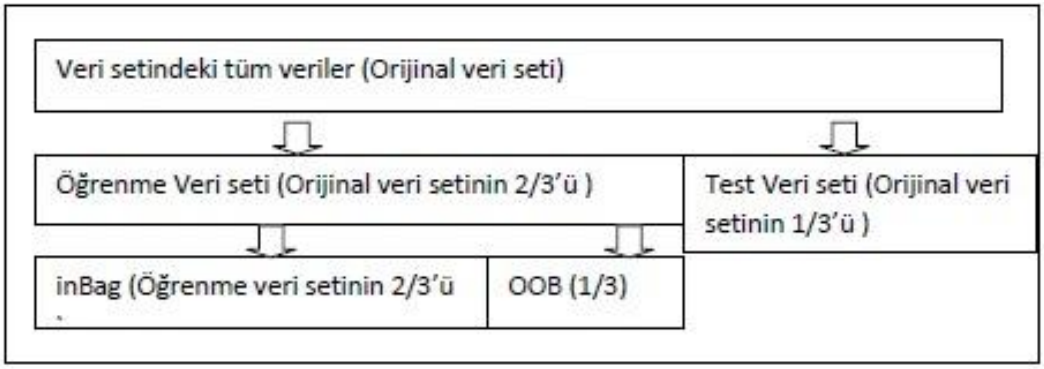


Şekil 2.3 RF yöntemine ait ağaç yapısı

2.3.2.2. Random Forest Sınıflama Yönteminin Aşamaları

1- * Orijinal veri setinin kendine ait bir test seti yoksa orijinal veri setinden BYÖY ile n tane örnekleme seçilir. Her bir örneklemenin 2/3'ü ağaç oluşturmak için kullanılır ve bu verilere inbag veri seti adı verilir. Geriye kalan 1/3'ü ise hata oranını hesaplamak için kullanılır (Tekrarlı Holdout yöntemi) ve bu verilere out-of-bag veri seti adı verilir (Şekil 2.4)

* Orijinal veri setinin kendine ait bir test seti varsa hata oranı bu test setiyle de hesaplanabilir (Holdout hata oranı tahmin yöntemi).
Her iki yolla da elde edilen hata oranı birbirine yakın çıkmaktadır (Korkem 2013).



Şekil 2.4 Veri Setinin Bölünmesi İşlemi

2- Her bootstrap örneği için budanmamış sınıflama veya regresyon ağacı aşağıdaki adımlar kullanılarak oluşturulur.

a. inBag veri setinden her düğümde bütün tahmin değişkenleri içerisinden en iyi değişkeni seçmek yerine rastgele m tane tahmin değişkeni seçilir ve bunların içerisinden en iyi dallara ayıracak (en çok bilgi kazancı sağlayacak) olan belirlenir.

b. Belirlenen tahmin değişkeni için en iyi dallanma kriteri Gini indeksi ile hesaplanır ve hesaplanan değere göre veri seti her düğümde iki alt dala ayrılır

c. Madde a ve b'deki işlemler aşağıya doğru yaprak düğüm elde edilinceye kadar her düğümde tekrar edilir.

Breiman tarafından varsayılan m değeri regresyon ağaçları kurulurken $p/3$, sınıflama ağaçları kurulurken ise $p^{1/2}$ olarak önerilmiştir. Burada p değeri toplam tahmin edici değişkenlerin(bağımsız değişkenlerin) sayısını ifade etmektedir. (Akman vd 2011)

3- n tane ağacın ayrı ayrı yapmış olduğu tahminler bir araya getirilerek yeni bir tahminde bulunulur;

a- Sınıflama ağaçları için en çok oyu alan sınıf en son tahmin durumu olarak seçilir,

b- Regresyon ağaçları için ise yapılan oylamanın ortalaması alınarak nihai tahmin yapılır.

Öğrenme veri setinden hata oranının hesaplaması için ise;

1. Her KA oluşturulurken BYÖY aşamasında; ağaç oluşturulacak veri (inBag) ve ağaç oluşturmak için kullanılmayan veri (out-of-bag veya OOB verisi) olmak üzere ikiye ayrılır. OOB verisiyle ağaç test edilir ve hata oranı tahmini yapılır.
2. Bireysel ağaçların yaptığı OOB tahminleri bir araya getirilir. Bu tahminlerden GH oranı kestirimi yapılır (Akman vd 2011).

Andy (2002), yeteri kadar ağaç oluşturulursa GH oranının oldukça doğru kestirileceğini, yeteri kadar ağaç oluşturulmadığı durumda ise GH oranının olduğundan büyük kestirileceğini belirtmiştir.

2.3.2.3. Modelin Sınıflama Başarısını Test Etme Yöntemleri

Bütün veri madenciliği modellerinin performansını hesaplamak için standart bir ölçütün kullanılması önemlidir. Veri madenciliğinde sınıflama modellerinin karşılaştırılması için en sık kullanılan yöntem hata oranını hesaplamaktır.

Breiman (2001)'a göre RF yönteminde hata oranı ağaç sayısı arttıkça belirli bir limite yakınsamaktadır. Ağaç yapılı sınıflayıcıların oluşturduğu hata oranının düşüklüğü, bireysel oluşturulan ağaçların güçlü olmasına ve aralarındaki korelasyonun az olmasına bağlıdır. Bununla beraber her düğümde rastgele değişkenlerin seçilmesi, hata oranının boosting algoritmasına göre daha düşük olmasını sağlamakta bu durumda modelin daha güçlü olmasına imkan tanımaktadır (Akman 2010).

İki sınıflı model için sınıflama matrisi Şekil 2.5'de verilmiştir.

	Gerçek Durum (Altın Standart tanısı)		
Tanı Testi Sonucu	Pozitif (Hasta)	Negatif (Sağlam)	Toplam
Pozitif (Hasta)	A (DP)	B (YP)	A+B
Negatif (Sağlam)	C (YN)	D (DN)	C+D
Toplam	A+C	B+D	A+B+C+D

Şekil 2.5 İki sınıflı model için sınıflama matrisi

- *Duyarlılık*: Testin, gerçek hastalar içinden hastaları ayırma yeteneğidir.

$$\text{Duyarlılık} = A / (A+C) = DP / (DP + YN)$$

- *Seçicilik*: Testin, gerçek sağlamlar içinden sağlamları ayırma yeteneğidir (Sümbüloğlu ve Akdağ 2010)

$$\text{Seçicilik} = D / (B + D) = DN / (DN + YP)$$

Duyarlılık ve seçicilik oranlarının yanı sıra, testin hatalı yargıları da aşağıdaki gibi hesaplanabilir:

- *Yanlış Pozitif Oranı:* Gerçek sağlamlar içinden testin hatalı olarak hasta dediği olgulardır.

$$YP = (1\text{-Seçicilik}) = B /(B+D) = YP / (YP+DN)$$

- *Yanlış Negatif Oranı:* Gerçek hastalar içinden testin hatalı olarak sağlam dediği olgulardır.

$$YN = (1\text{-Duyarlılık}) = C /(A + C) = YN/ (YN + DP)$$

- *Pozitif Tahmin Değeri:* Tanı testinin sonucuna göre belirlenen pozitifler içerisindeki doğru pozitiflerin oranıdır.

$$PTD = A / (A+B) = DP / (DP + YP)$$

- *Negatif Tahmin Değeri:* Tanı testinin sonucuna göre belirlenen negatifler içerisindeki doğru negatiflerin oranıdır.

$$NTD = D / (C+D) = DN / (DN + YN)$$

- *Pozitif Olabilirlik Oranı:* Hastalık tanısı koymanın doğruluk oranıdır. Testin sonucunun hastalığın varlığında pozitif çıkma olasılığının, hastalığın yokluğunda pozitif çıkma olasılığına oranıdır. Yani hastalığa var dediği zaman doğruyu bildirmesinin, yanılmasına oranıdır.

  POO = L+ = Duyarlılık / (1-Seçicilik)

  POO = A(B+D)/B(A+C)=DP(YP+DN)/YP(DP+YN)

Bu oran ne kadar yüksek olursa, gerçek hastalar o derece iyi ayrımlanabilmektedir.

- *Negatif Olabilirlik Oranı*: Sağlam tanısı koymanın doğruluk oranıdır. Testin sonucunun hastalığın varlığında negatif çıkma olasılığının, hastalığın yokluğunda negatif çıkma olasılığına oranıdır.

  NOO = L- = (1-Duyarlılık) / (Seçicilik) = C(B+D) / D (A+C)

  NOO = YN (YP+DN) / DN (DP+YN)

Bu oran ne kadar düşük olursa, gerçek sağlamlar o derece iyi ayrımlanabilmektedir.

- *Doğruluk:* Duyarlılık ve seçicilik birleştirilerek tek bir ölçü elde edilmek istendiğinde kullanılan ölçülerden birisi de doğru test sonucu olasılığıdır. Gerçekte testin hasta ve sağlam olarak toplam doğru tanı koyma oranına "doğruluk (accuracy)" denir (Dirican 2001).

  Doğruluk = (A+D)/(A+B+C+D) = (DP+DN) / (DP+YP+YN+DN)

3. MATERYAL VE METOT

3.1. Araştırmanın Amacı

Bu çalışmadaki amaç; koroner arter hastalığının risk faktörlerini GA yöntemiyle optimize ederek, RF yönteminin GA optimizasyonlu ve optimizasyonsuz sonuçlarını karşılaştırarak daha iyi bir sınıflama başarısı elde etmektir.

Bu çalışmada, genetik algoritma ile optimize edilerek sınıflama yapılmış olan 20, 50, 100, 500 ve 1000 ağaçlı RF sonuçları; tek bir KA'dan elde edilen sonuçlarla, genetik algoritma ile optimizasyon yapılmamış olan 20, 50, 100, 500 ve 1000 ağaçlı RF sonuçlarıyla aynı zamanda da Gini katsayısı değerlerine göre sınıflamalara etki eden en önemli 9 değişken alınarak sınıflama yapılmış olan 20, 50, 100, 500 ve 1000 ağaçlı RF sonuçlarıyla kıyaslanmıştır.

Uygulaması yapılmış olan tüm ağaçların doğruluk oranları, duyarlılık ve seçicilik değerleri, yanlış pozitif ve yanlış negatif oranları, pozitif ve negatif tahmin değerleri, pozitif ve negatif olabilirlik oranları kıyaslanarak başarılı bir sonuç elde edilip edilmediği incelenmiştir. Aynı zamanda veri sayısı az olduğu durumlarda RF yöntemindeki en başarılı sonuç verecek olan optimum ağaç sayısı da incelenmiştir.

3.2. Araştırmanın Uygulandığı Veri Seti

Yapmış olduğumuz bu çalışmada, 10.01.2011 Tarihinde PAÜ Tıp Fakültesi Girişimsel Olmayan Tıbbi Etik Kurulu'ndan onay almış olan Prof. Dr. Asuman Kaftan'ın yürütücüsü olduğu Arş. Gör. Dr. Işık Tekin'in "Koroner Anjiyografi Endikasyonu Olan Hastalarda Epikardiyal Yağ Dokusunun Transtorasik Ekokardiyografi ile Ölçümü, Koroner Arter Hastlalığı ve Vücut Kitle Endeksi ile İlişkisi" isimli uzmanlık tez çalışmasından elde edilmiş olan veriler kullanılmıştır.

Çalışmaya Pamukkale Üniversitesi Tıp Fakültesi Kardiyoloji polikliniğine göğüs ağrısı nedeniyle başvuran noninvaziv testlerde kesin veya şüpheli iskemi saptanan ve ilk kez koroner anjiografi yapılmış olan 97 hasta dahil edilmiştir. Tüm hastaların koroner anjiografi öncesi biyokimyasal değerleri ve antropometrik ölçümleri alınmış, sol lateral dekübit pozisyonda aynı araştırmacı tarafından transtorasik ekokardiyografileri de yapılmıştır (Tekin 2012). Koroner arter hastalığına (KAH) sahip olan hasta sayısı 48 kişi iken koroner arter hastası olmayan kişi sayısı 49 olarak elde edilmiştir.

Çalışmaya Ocak 2011 ile Ağustos 2012 tarihleri arasında Pamukkale Üniversitesi Kardiyoloji Anabilim Dalı Polikliniği'ne başvuran ve elektif koroner anjiografi endikasyonu alan hastalar dahil edilmişlerdir. Çalışmaya katılmayı kabul eden hastaların koroner anjiografi öncesi ekokardiyografik görüntüleri, ekokardiyografi öncesi antropometrik ölçümleri, son 1 ay içerisindeki lipid, açlık glukoz, hemoglobin değerleri, kardiyak risk faktörleri ve koroner anjiografi sonuçları kaydedilmiştir (Tekin 2012).

Çalışmaya dahil edilme kriterleri: Kardiyoloji polikliniğine başvurarak elektif olarak koroner anjiografi endikasyonu alan hastalar, miyokard perfüzyon sintigrafisi pozitif olan, efor testi pozitif olan, optimal medikal tedaviye rağmen göğüs ağrısı tarifleyen, noninvaziv yöntemlerle şüpheli iskemi saptanan hastalar çalışmaya dahil edilmişlerdir (Tekin 2012).

3.3. Veri Analizinde Kullanılan Programlar ve Paketler

Bu çalışmada R for Windows paket programı ve Salford Systems tarafından geliştirilmiş olan SPM v7.0 - Salford Predictive Modeler software suite paket programı kullanılmıştır.

R for Windows paket programı internet aracılığı ile ücretsiz olarak dağıtılan genel lisanslı bir programdır (WEB-3). Program genel lisans kapsamında, serbest bir şekilde dağıtılabilir ve kullanılabilir ayrıca programı elde eden herkes asıl kaynağını belirterek programı dağıtma ve kullanma hakkına sahiptir ve programın kaynak kodu da açık bir şekilde sunulmaktadır. Dolayısıyla herhangi bir programlama bilgisine sahip olan kişiler bu kod üzerinden değişiklik ve geliştirme yapma hakkına sahiptirler. Programın en büyük üstünlüklerinden birisi de hemen hemen bütün işletim sistemlerinde çalışabiliyor olmasıdır. R for Windows paket programı kullanılarak; istatistiksel analiz, grafik çizme ve veri işleme süreçleri gerçekleştirilebilir.

Bu çalışmada, GA yöntemiyle sınıflama yapabilen, R programı için yazılmış olan "evtree: Evolutionary Learning of Globally Optimal Trees" (WEB-4) paketi kullanıldı. Bu paket Thomas Grubinger, Achim Zeileis ve Karl-Peter Pfeiffer tarafından 26.11.2013 tarihinde yayınlanmıştır.

RF sonuçlarını elde etmek için kullanılan SPM v7.0 - Salford Predictive Modeler software suite paket programı ise Leo Breiman'ın en yeni veri madenciliği teknolojisini içeren bir programdır (WEB-5). Bu çalışmada bu programın deneme sürümü temin edilerek, içerisinde bulunan RandomForest modülü ve CART modülü kullanıldı.

4. BULGULAR

4.1. Üzerinde Çalışılan Değişkenler (Risk Faktörleri) ve Araştırmanın Uygulama Adımları

Bu çalışmada kullanılmış olan değişkenler ve bu değişkenlere ilişkin 97 kişiden elde edilen sonuçlar Tablo4.1'de gösterilmiştir.

Tablo. 4.1 Hastaların klinik ve demografik özellikleri

Değişken Adı	R datası Gösterimi	Değişken Tipi	Değerleri
EYD vertikal (cm) (ortalama ± SS)	x_1	Nicel	0,54 ± 0,16
EYD horizontal (cm) (ortalama ± SS)	x_2	Nicel	2,66 ± 0,71
EYD alan (cm^2) (ortalama ± SS)	x_3	Nicel	1,36 ± 0,67
Koroner Arter Hastalığı	x_4	Kategorik	KAH (+) n= 48 (%49,5) KAH (-) n=49 (%50,5)
Açlık Glukoz (mg/dl) (ortalama ± SS)	x_5	Nicel	124,29 ± 52,36
Ejeksiyon fraksiyonu (ortalama ± SS)	x_6	Nicel	58,85 ± 7,34
Yaş (Ortalama yıl ± SS)	x_7	Nicel	59,13 ± 10,57
Cinsiyet (Erkek(%)/Kadın(%))	x_8	Kategorik	Erkek n= 49 (%50,52) Kadın n= 48 (%49,48)
Diyabetes Mellitus (var) [n(%)]	x_9	Kategorik	Var n= 29 (%29,9) Yok n= 68 (%70,1)
Hipertansiyon (var) [n(%)]	x_{10}	Kategorik	Var n= 50 (%51,55) Yok n= 47 (%48,45)
Sigara kullanımı (var) [n(%)]	x_{11}	Kategorik	Var n= 28 (%28,87) Yok n= 69(%71,13)
Aile öyküsü (var) [n(%)]	x_{12}	Kategorik	Var n= 23 (%23,71) Yok n= 74 (%76,29)
Bel Çevresi/Kalça Çevresi (ortalama ± SS)	x_{13}	Nicel	1 ± 0,07
Vücut kitle indeksi (kg/m^2) (ortalama ± SS)	x_{14}	Nicel	27,42 ± 4,06
LDL Kolesterol (mg/dl) (ortalama ± SS)	x_{15}	Nicel	121,9 ± 32,95
HDL Kolesterol (mg/dl) (ortalama ± SS)	x_{16}	Nicel	46,51 ± 13,19
Total Kolesterol (mg/dl) (ortalama ± SS)	x_{17}	Nicel	201,25 ± 39,35
Trigliserid (mg/dl) (ortalama ± SS)	x_{18}	Nicel	173,55 ± 118,34
Hemoglobin (gr/dl) (ortalama ± SS)	x_{19}	Nicel	14,14 ± 1,39

EYD: Epikardiyal yağ dokusu, LDL: Düşük Dansiteli Lipoprotein, HDL: Yüksek Dansiteli Lipoprotein

4.2. Elde Edilen Klinik Bulguların Sonuçları

Öncelikle, koroner arter hastalığının risk faktörlerini en basit düzeyde inceleyecek olursak; KAH(+) olan grup ile KAH(-) olan grubun yaş ortalamaları başta olmak üzere cinsiyet, bel çevresi/kalça çevresi oranı, sigara kullanımı, HDL kolesterol ve trigliserid değerleri ortalamaları istatistiksel olarak anlamlı farklılık göstermiştir.

Risk faktörleri ile KAH varlığının ilişkisi incelendiğinde; KAH varlığı ile yaşın yüksek olması arasında anlamlı ilişki saptanmıştır. KAH(+) olanların ortalama yaşı 61,6±10,5 iken KAH(-) olan hastaların ortalama yaşı 56,6 ±10,1 olarak saptanmıştır (p=0,021).

Cinsiyet farklılığının KAH üzerine etkisine bakıldığında; erkeklerin %63,3'ünde, kadınların %35,4'ünde KAH varlığı tespit edilmiştir. Erkeklerde KAH(+) oranı kadınlara göre istatistiksel olarak daha fazla saptanmıştır (p=0,006).

Sigara kullanan 28 hastanın 19'unda (% 67,9) KAH(+) saptanmış olup, 9'unda (% 32,1) KAH(-) saptanmıştır. İki grup karşılaştırıldığında sigara kullananlarda istatistiksel olarak anlamlı yükseklikte KAH(+) oranı tespit edilmiştir (p=0,021).

Diğer risk faktörlerinden Diyabetes Mellitus [29 hasta (%29,9)], Hipertansiyon [50 hasta (%51,5)] ve aile öyküsü [23 hasta (%23,7)] ile KAH varlığı arasında istatistiksel olarak anlamlı ilişki saptanmamıştır.

KAH(+) olan hastalarda Bel çevresi/ Kalça çevresi oranının yüksekliği KAH(-) olma durumuna göre anlamlı olarak yüksek bulunmuştur (p=0,009).

KAH varlığı ile biyokimyasal değerler arasındaki ilişki incelendiğinde; KAH varlığı ile HDL düşüklüğü arasında istatistiksel olarak anlamlı ilişki saptanmıştır. KAH(+) olan hastalarda ortalama HDL 43 ±12,9 mg/dl iken KAH(-) olan hastalarda HDL değeri ortalaması 49,9 ± 12,7 mg/dl olarak saptanmıştır(p=0,009).

KAH varlığı ile trigliserid değeri arasında istatistiksel olarak anlamlı ilişki saptanmıştır. KAH(+) olan hastalarda ortalama trigliserid 201,5 ± 141 mg/dl iken KAH(-) olan hastalarda ortalama. trigliserid 145,5 ± 84,2 mg/dl olarak saptanmıştır (p=0,013).

EYD vertikal kalınlığı KAH(+) olan hastalarda ortalama 0,6 ± 0,15 cm, olmayanlarda ortalama 0,46 ± 0,14 cm (p=0,0001), EYD horizontal uzunluğu KAH(+) olan hastalarda ortalama 2,91 ± 0,60 cm, olmayanlarda 2,41 ± 0,73 cm (p=0,001), EYD alan ölçümü KAH(+) olan hastalarda 1,55 ± 0,64 cm^2, olmayanlarda 1,15 ± 0,62 cm^2 (p=0,002) saptanmış olup istatistiksel olarak anlamlı bulunmuştur.

Tablo. 4.2 Koroner arter hastalığı olan ve olmayan grupların karşılaştırılması

	KAH (+) (n=48)	**KAH (-) (n=49)**	**p değeri**
EYD vertikal (cm) (ortalama ± SS)	0,60 ± 0,15	0,46 ± 0,14	0,0001*
EYD horizontal (cm) (ortalama ± SS)	2,91± 0,60	2,41±0,73	0,001*
EYD alan (cm^2) (ortalama ± SS)	1,55 ± 0,64	1,15 ± 0,62	0,002*
Açlık Glukoz (mg/dl) (ortalama ± SS)	130,3 ± 53,5	118,3 ± 50,9	0,09
Ejeksiyon fraksiyonu (ortalama ± SS)	57,88 ± 9,16	59,8 ± 4,88	0,203
Yaş (Ortalama yıl ± SS)	61,6 ± 10,5	56,6 ± 10,1	0,021*
Cinsiyet (Erkek(%)/Kadın(%))	31 (63,3) / 17 (35,4)	18 (36,7) / 31 (64,6)	0,006*
Diyabetes Mellitus (var) [n(%)]	17 (58,6)	12 (41,4)	0,24
Hipertansiyon (var) [n(%)]	29 (58)	21 (42)	0,08
Sigara kullanımı (var) [n(%)]	19 (67,9)	9 (32,1)	0,021*
Aile öyküsü (var) [n(%)]	13 (56,5)	10 (43,5)	0,44
Bel Çevresi/Kalça Çevresi (ortalama ± SS)	1,01 ± 0,06	0,97 ± 0,06	0,009*
Vücut kitle indeksi (kg/m^2) (ortalama ± SS)	27,3 ± 3,8	27,4 ± 4,3	0,98
LDL Kolesterol (mg/dl) (ortalama ± SS)	121,6 ± 35	122,1 ± 32,6	0,65
HDL Kolesterol (mg/dl) (ortalama ± SS)	43,0 ± 12,9	49,9 ± 12,7	0,009*
Total Kolesterol (mg/dl) (ortalama ± SS)	201,9 ± 42,6	200,5 ± 37	0,90
Trigliserid (mg/dl) (ortalama ± SS)	201,5 ± 141	145,5 ± 84,2	0,013*
Hemoglobin (gr/dl) (ortalama ± SS)	14,4 ± 1,3	13,8 ± 1,4	0,08

KAH: Koroner arter hastalığı, LDL: Düşük Dansiteli Lipoprotein, HDL: Yüksek Dansiteli Lipoprotein, EYD: Epikardiyal yağ dokusu, SS: Standart sapma .*p<0,05 İstatistiksel olarak anlamlı kabul edilmiştir.

4.3. Elde Edilen Random Forest Sınıflama Sonuçları

Tablo 4.3'de 97 kişilik veri seti üzerinden KA ile elde edilmiş olan sınıflama sonuçları gösterilmiştir. KA sonucu incelendiğinde; doğruluk oranının %68 olarak elde edildiği görülmektedir. Aynı zamanda duyarlılık oranı %60,42, seçicilik oranı %75,51, yanlış pozitif oranı %24,49, yanlış negatif oranı %39,58, negatif tahmin değeri %66,07, pozitif tahmin değeri %70,73, pozitif olabilirlik oranı 2,47, negatif olabilirlik oranı ise 0,52 olarak elde edilmiştir. Aynı zamanda KA modelinin sınıflamasına etki eden değişkenlerin önem sıralaması Tablo 4.8'de gösterilmiştir. Burada EYD horizontal ölçümünün ilk sırada yer aldığı görülmektedir. Daha sonra sırasıyla yaş, LDL kolesterol, EYD vertikal ölçümü ve EYD alan ölçümü yer almaktadır.

Tablo 4.4'te 20 ağaç için oluşturulmuş olan RF sonuçları görülmektedir. İlk olarak herhangi bir optimizasyon yöntemi uygulanmadan 20 ağaçlık bir RF modeli kurulmuştur. Burada elde edilen doğruluk oranının %64,95 olduğu görülmektedir. Daha sonrasında veri setine GA yöntemiyle optimizasyon uygulanmış ve 20 ağaçlık bir RF sonucu daha elde edilmiştir. Optimizasyon sonrası RF modelinin doğruluk oranı %79,38'e çıkmıştır. Aynı zamanda optimize edilmemiş olan RF modeline etki eden Tablo 4.8'deki değişkenlerin önem sırasına göre ilk 9 tanesi kullanılarak tekrar 20 ağaçlık bir RF modeli elde edilerek sonuçlar incelenmiştir. Bu modelin doğruluk oranının ise % 68,04 olduğu görülmektedir. Duyarlılık değerleri incelendiğinde 3 model arasında ciddi bir farklılık görülmemektedir. Ancak seçicilik değerlerine bakıldığında GA yöntemiyle optimize edilmiş olan modelin seçiciliğinin diğer RF modellerine oranla oldukça yüksek olduğu görülmektedir. Aynı şekilde yanlış pozitif oranının da belirgin bir şekilde düşük çıktığı görülmektedir. Yanlış negatif oranında çok büyük bir fark olmamasına karşın GA yöntemiyle kurulan model diğer modellere oranla başarısız olmuştur. Negatif tahmin değerleri arasında da büyük bir farklılık gözlenmemiş olmasına rağmen pozitif tahmin değeri diğer modellere göre oldukça yüksek çıkmıştır. Negatif olabilirlik oranı açısında GA yönteminin diğer modellere göre daha düşük bir sonuç verdiği görülmektedir. Pozitif olabilirlik oranında ise diğer modellere oranla oldukça yüksek bir sonuç elde edilmektedir.

Tablo 4.3 KA sınıflaması sonuçları

Karar Ağacı	*Gerçek KAH (+)*	*Gerçek KAH (-)*	*Toplam*	*Duyarlılık (%)*	*Seçicilik (%)*	*YP (%)*	*YN (%)*	*PTD (%)*	*NTD (%)*	*POO*	*NOO*	*Doğruluk (%)*
KAH (+)	29	12	41	60,42	75,51	24,49	39,58	70,73	66,07	2,47	0,52	68,04
KAH (-)	19	37	56									
Toplam	48	49	97									

Tablo 4.4 20 Ağaç için test edilmiş RF sonuçları

Random Forest (20 ağaç)	*Gerçek KAH (+)*	*Gerçek KAH (-)*	*Toplam*	*Duyarlılık (%)*	*Seçicilik (%)*	*YP (%)*	*YN (%)*	*PTD (%)*	*NTD (%)*	*POO*	*NOO*	*Doğruluk (%)*
KAH (+)	37	23	60	77,08	53,06	46,94	22,92	61,67	70,27	1,64	0,43	64,95
KAH (-)	11	26	37									
Toplam	48	49	97									
GA (20 ağaç)												
KAH (+)	35	7	42	72,92	85,71	14,29	27,08	83,33	76,36	5,1	0,32	79,38
KAH (-)	13	42	55									
Toplam	48	49	97									
En etkili 9 değişken için RF (20 ağaç)												
KAH (+)	38	21	59	79,17	57,14	42,86	20,83	64,41	73,68	1,85	0,36	68,04
KAH (-)	10	28	38									
Toplam	48	49	97									

Tablo 4.5'te 50 ağaç için oluşturulmuş olan RF sonuçları görülmektedir. Optimize edilmeden kurulmuş olan 50 ağaçlık RF'den elde edilen doğruluk oranının %64,95 olduğu görülmektedir. Optimize edilerek kurulmuş olan 50 ağaçlık RF sonucu modelinin doğruluk oranı %81,44'e çıkmaktadır. Tablo 4.9'de gösterilmiş olan RF modelini etkileyen en önemli 9 değişkenle elde edilmiş olan 50 ağaçlık modelin doğruluk oranının ise % 61,86 olduğu görülmektedir. Duyarlılık değerleri incelendiğinde 3 model arasında ciddi bir farklılık görülmektedir. Ancak seçicilik değerlerine bakıldığında GA yöntemiyle optimize edilmiş olan modelin seçiciliğinin diğer RF modellerine oranla oldukça yüksek olduğu görülmektedir. Aynı şekilde yanlış pozitif oranının da belirgin bir şekilde düşük çıktığı görülmektedir. Yanlış negatif oranında ise büyük bir başarı elde ettiği söylenememektedir. Negatif tahmin değerleri arasında da büyük bir farklılık gözlenmemiş olmasına rağmen pozitif tahmin değeri diğer modellere göre oldukça yüksek çıkmıştır. Negatif olabilirlik oranı açısından da GA yönteminin diğer modellere göre daha düşük bir sonuç verdiği görülmektedir. Pozitif olabilirlik oranında ise diğer modellere oranla oldukça yüksek bir sonuç elde edilmektedir.

Tablo 4.5 50 Ağaç için test edilmiş RF sonuçları

Random Forest (50 ağaç)	***Gerçek KAH (+)***	***Gerçek KAH (-)***	***Toplam***	***Duyarlılık (%)***	***Seçicilik (%)***	***YP (%)***	***YN (%)***	***PTD (%)***	***NTD (%)***	***POO***	***NOO***	***Doğruluk (%)***
KAH (+)	35	21	56	72,92	57,14	42,86	27,08	62,5	68,29	1,7	0,47	64,95
KAH (-)	13	28	41									
Toplam	48	49	97									
GA (50 ağaç)												
KAH (+)	35	5	40	72,92	89,8	10,2	27,08	87,5	77,19	7,15	0,3	81,44
KAH (-)	13	44	57									
Toplam	48	49	97									
En etkili 9 değişken için RF (50 ağaç)												
KAH (+)	30	19	49	62,5	61,22	38,78	37,5	61,22	62,5	1,61	0,61	61,86
KAH (-)	18	30	48									
Toplam	48	49	97									

Tablo 4.6'de 100 ağaç için oluşturulmuş olan RF sonuçları görülmektedir. Optimize edilmeden kurulmuş olan 100 ağaçlık RF'den elde edilen doğruluk oranının %63,92 olduğu görülmektedir. Optimize edilerek kurulmuş olan 100 ağaçlık RF sonuç modelinin doğruluk oranı %83,51'e çıkmaktadır. Tablo 4.9'de gösterilmiş olan RF modelini etkileyen en önemli 9 değişkenle elde edilmiş olan 100 ağaçlık modelin doğruluk oranının ise % 67,01 olduğu görülmektedir. Duyarlılık değerleri incelendiğinde 3 model arasında ciddi bir farklılık görülmemiştir ancak GA yöntemiyle elde edilen duyarlılık değeri diğerlerine oranla daha başarılıdır. Seçicilik değerlerine bakıldığında GA yöntemiyle optimize edilmiş olan modelin seçiciliğinin diğer RF modellerine oranla oldukça yüksek olduğu görülmektedir. Aynı şekilde yanlış pozitif oranının da belirgin bir şekilde düşük çıktığı görülmektedir. Diğer ağaç sayılarından daha fazla oranda, GA negatif tahmin değerinde de diğer modellere göre farklılık göstermektedir. Aynı şekilde pozitif tahmin değeri diğer modellere göre oldukça yüksek çıkmıştır. Negatif olabilirlik oranı açısından GA yönteminin diğer modellere göre daha düşük bir sonuç verdiği görülmektedir. Pozitif olabilirlik oranında ise diğer modellere oranla oldukça yüksek bir sonuç elde edilmektedir.

Tablo 4.6 100 Ağaç için test edilmiş RF sonuçları

Random Forest (100 ağaç)	***Gerçek KAH (+)***	***Gerçek KAH (-)***	***Toplam***	***Duyarlılık (%)***	***Seçicilik (%)***	***YP (%)***	***YN (%)***	***PTD (%)***	***NTD (%)***	***POO***	***NOO***	***Doğruluk (%)***
KAH (+)	34	21	55	70,83	57,14	42,86	29,17	61,82	66,67	1,65	0,51	63,92
KAH (-)	14	28	42									
Toplam	48	49	97									
GA (100 ağaç)												
KAH (+)	38	6	44	79,17	87,76	12,24	20,83	86,36	81,13	6,47	0,24	83,51
KAH (-)	10	43	53									
Toplam	48	49	97									
En etkili 9 değişken için RF (100 ağaç)												
KAH (+)	32	16	48	66,67	67,35	32,65	33,33	66,67	67,35	2,04	0,49	67,01
KAH (-)	16	33	49									
Toplam	48	49	97									

Tablo 4.7'de 500 ağaç için oluşturulmuş olan RF sonuçları görülmektedir. Optimize edilmeden kurulmuş olan 500 ağaçlık RF'den elde edilen doğruluk oranının %64,95 olduğu görülmektedir. Optimize edilerek kurulmuş olan 500 ağaçlık RF sonuç modelinin doğruluk oranı %81.44'e çıkmıştır. Tablo 4.9'da gösterilmiş olan RF modelini etkileyen en önemli 9 değişkenle elde edilmiş olan 500 ağaçlık modelin doğruluk oranının ise % 69,07 olduğu görülmektedir. Duyarlılık değerleri incelendiğinde 3 model arasında ciddi bir farklılık görülmemiştir. Seçicilik değerlerine bakıldığında GA yöntemiyle optimize edilmiş olan modelin seçiciliğinin diğer RF modellerine oranla oldukça yüksek olduğu görülmektedir. Aynı şekilde yanlış pozitif oranının da belirgin bir şekilde düşük çıktığı görülmektedir. Negatif tahmin değerinde ise diğer modellere göre farklılık göstermemektedir. Pozitif tahmin değeri diğer modellere göre oldukça yüksek çıkmıştır. Negatif olabilirlik oranı açısında GA yönteminin diğer modellere göre daha düşük bir sonuç verdiği görülmektedir. Pozitif olabilirlik oranında ise diğer modellere oranla oldukça yüksek bir sonuç elde edilmektedir.

Tablo 4.7 500 Ağaç için test edilmiş RF sonuçları

Random Forest (500 ağaç)	***Gerçek KAH (+)***	***Gerçek KAH (-)***	***Toplam***	***Duyarlılık (%)***	***Seçicilik (%)***	***YP (%)***	***YN (%)***	***PTD (%)***	***NTD (%)***	***POO***	***NOO***	***Doğruluk (%)***
KAH (+)	33	19	52	68,75	61,22	38,78	31,25	63,46	66,67	1,77	0,51	64,95
KAH (-)	15	30	45									
Toplam	48	49	97									
GA (500 ağaç)												
KAH (+)	35	5	40	72,92	89,8	10,2	27,08	87,5	77,19	7,15	0,3	81,44
KAH (-)	13	44	57									
Toplam	48	49	97									
En etkili 9 değişken için RF (500 ağaç)												
KAH (+)	36	18	54	75	63,27	36,73	25	66,67	72,09	2,04	0,4	69,07
KAH (-)	12	31	43									
Toplam	48	49	97									

Tablo 4.8'de 1000 ağaç için oluşturulmuş olan RF sonuçları görülmektedir. Optimize edilmeden kurulmuş olan 1000 ağaçlık RF'den elde edilen doğruluk oranının %64,95 olduğu görülmektedir. Optimize edilerek kurulmuş olan 1000 ağaçlık RF sonucu modelinin doğruluk oranı %81,44'e çıkmaktadır. Tablo 4.9'da gösterilmiş olan RF modelini etkileyen en önemli 9 değişkenle elde edilmiş olan 1000 ağaçlık modelin doğruluk oranının ise % 70,1 olduğu görülmektedir. Duyarlılık değerleri incelendiğinde 3 model arasında ciddi bir farklılık görülmemektir ancak GA yöntemiyle kurulan model en iyi sonucu almamaktadır. Seçicilik değerlerine bakıldığında GA yöntemiyle optimize edilmiş olan modelin seçiciliğinin diğer RF modellerine oranla oldukça yüksek olduğu görülmektedir. Aynı şekilde yanlış pozitif oranının da belirgin bir şekilde düşük çıktığı görülmektedir. Negatif tahmin değerinde ise diğer modellere göre farklılık gözlenmemektedir. Pozitif tahmin değeri ise diğer modellere göre oldukça yüksek çıkmıştır. Negatif olabilirlik oranı açısından GA yönteminin diğer modellere göre daha düşük bir sonuç verdiği görülmektedir. Pozitif olabilirlik oranında ise diğer modellere oranla oldukça yüksek bir sonuç elde edilmektedir.

Tablo 4.8 1000 Ağaç için test edilmiş RF sonuçları

Random Forest (1000 ağaç)	***Gerçek KAH (+)***	***Gerçek KAH (-)***	***Toplam***	***Duyarlılık (%)***	***Seçicilik (%)***	***YP (%)***	***YN (%)***	***PTD (%)***	***NTD (%)***	***POO***	***NOO***	***Doğruluk (%)***
KAH (+)	34	20	54	70,83	59,18	40,82	29,17	62,96	67,44	1,74	0,49	64,95
KAH (-)	14	29	43									
Toplam	48	49	97									
GA (1000 ağaç)												
KAH (+)	35	5	40	72,92	89,8	10,2	27,08	87,5	77,19	7,15	0,3	81,44
KAH (-)	13	44	57									
Toplam	48	49	97									
En etkili 9 değişken için RF (1000 ağaç)												
KAH (+)	37	18	55	77,08	63,27	36,73	22,92	67,27	73,81	2,1	0,36	70,1
KAH (-)	11	31	42									
Toplam	48	49	97									

Tablo 4.9'da modellerde etkili olan değişkenlerin önemlilik düzeyleri görülmektedir. KA'da etkili olan en önemli ilk 5 değişken sırasıyla; EYD horizontal (%100), yaş (%99,81), LDL kolesterol (%74,.84), EYD vertikal (%74,73) ve EYD alan (%56,50) şeklindedir.

20 Ağaç ile kurulmuş olan RF modelinde etkili olan en önemli ilk 5 değişken sırasıyla; LDL kolesterol (%100), Trigliserid (% 57,48), Bel Çevresi/Kalça Çevresi oranı (%55,29), EYD vertikal (%50,37) ve EYD horizontal (%50,03) şeklindedir.

50 Ağaç ile kurulmuş olan RF modelinde etkili olan en önemli ilk 5 değişken sırasıyla; Trigliserid (% 100), LDL kolesterol (%98,28), EYD vertikal (%86,62), Bel Çevresi/Kalça Çevresi oranı (%85,76) ve EYD horizontal (%75,73) şeklindedir.

100 Ağaç ile kurulmuş olan olduğumuz RF modelinde etkili olan en önemli ilk 5 değişken sırasıyla; EYD horizontal (%100), LDL kolesterol (%92,03), Trigliserid (% 88,31), Bel Çevresi/Kalça Çevresi oranı (%74,79) ve EYD vertikal (%70,16) şeklindedir.

500 Ağaç ile kurulmuş olan RF modelinde etkili olan en önemli ilk 5 değişken sırasıyla; EYD horizontal (%100), LDL kolesterol (%83,65), Bel Çevresi/Kalça Çevresi oranı (%79,12), Trigliserid (% 77,85), ve EYD alan (%63,15) şeklindedir.

1000 Ağaç ile kurulmuş olan RF modelinde etkili olan en önemli ilk 5 değişken sırasıyla; EYD horizontal (%100), LDL kolesterol (%86,69), Bel Çevresi/Kalça Çevresi oranı (%82,03), Trigliserid (% 73,24), ve EYD vertikal (%65,29) şeklindedir.

Tablo 4.9 Değişkenlerin Modellere Göre Normalize Edilmiş Gini Katsayıları

Önem Sırası	Karar Ağacı	Normalize Edilmiş Gini Katsayısı değerleri (%)	20 Ağaç	Normalize Edilmiş Gini Katsayısı değerleri (%)	50 Ağaç	Normalize Edilmiş Gini Katsayısı değerleri (%)
1	EYD horizontal	100	LDL Kolesterol	100	Trigliserid	100
2	Yaş	99,81105	Trigliserid	57,4808	LDL Kolesterol	98,28457
3	LDL Kolesterol	74,84077	Bel Çevresi/Kalça Çevresi	55,29057	EYD vertikal	86,62225
4	EYD vertikal	74,73229	EYD vertikal	50,37168	Bel Çevresi/Kalça Çevresi	85,76624
5	EYD alan	56,50545	EYD horizontal	50,02669	EYD horizontal	75,72934
6	Sigara kullanımı	43,22274	Açlık Glukoz	36,73376	EYD alan	58,22758
7	Ejeksiyon fraksiyonu	29,17418	Hemoglobin	33,68714	Hemoglobin	49,40129
8	Vücut kitle indeksi	22,07499	EYD alan	31,30188	Açlık Glukoz	43,89714
9	Trigliserid	18,13765	Total Kolesterol	22,39936	Total Kolesterol	41,71646
10	Açlık Glukoz	16,70769	Vücut kitle indeksi	20,00865	Yaş	39,12256
11	Bel Çevresi/Kalça Çevresi	5,84975	Yaş	17,66244	Vücut kitle indeksi	38,86213
12	Total Kolesterol	3,09243	HDL Kolesterol	8,9957	Sigara kullanımı	29,10614
13	HDL Kolesterol	2,27315	Ejeksiyon fraksiyonu	8,05558	Cinsiyet	29,04172
14	Hipertansiyon	0	Hipertansiyon	7,87034	HDL Kolesterol	24,06964
15	Hemoglobin	0	Sigara kullanımı	7,43483	Hipertansiyon	21,16713
16	Diyabetes Mellitus	0	Diyabetes Mellitus	7,26875	Ejeksiyon fraksiyonu	15,99397
17	Cinsiyet	0	Cinsiyet	4,11934	Diyabetes Mellitus	6,52151
18	Aile öyküsü	0	Aile öyküsü	1,97154	Aile öyküsü	1,91663

Tablo 4.9 Değişkenlerin Modellere Göre Normalize Edilmiş Gini Katsayıları (Devamı)

Önem sırası	100 Ağaç	Normalize Edilmiş Gini Katsayısı değerleri (%)	500 Ağaç	Normalize Edilmiş Gini Katsayısı değerleri (%)	1000 Ağaç	Normalize Edilmiş Gini Katsayısı değerleri (%)
1	EYD horizontal	100	EYD horizontal	100	EYD horizontal	100
2	LDL Kolesterol	92,0335	LDL Kolesterol	83,64658	LDL Kolesterol	86,69217
3	Trigliserid	88,30896	Bel Çevresi/Kalça Çevresi	79,1208	Bel Çevresi/Kalça Çevresi	82,03202
4	Bel Çevresi/Kalça Çevresi	74,78626	Trigliserid	77,85522	Trigliserid	73,24198
5	EYD vertikal	70,15684	EYD alan	63,15164	EYD vertikal	65,2948
6	Hemoglobin	56,80676	EYD vertikal	62,399	EYD alan	59,76909
7	Açlık Glukoz	56,74006	Açlık Glukoz	52,28101	Yaş	52,51531
8	EYD alan	45,73151	Yaş	45,89645	Açlık Glukoz	51,53077
9	Yaş	39,2931	Hemoglobin	41,12826	Hemoglobin	41,59889
10	Vücut kitle indeksi	31,44343	Vücut kitle indeksi	30,52723	Ejeksiyon fraksiyonu	31,67625
11	Total Kolesterol	27,37642	Ejeksiyon fraksiyonu	29,42954	Vücut kitle indeksi	30,32198
12	HDL Kolesterol	23,44744	Total Kolesterol	25,40029	HDL Kolesterol	26,83639
13	Sigara kullanımı	23,39547	HDL Kolesterol	25,36185	Total Kolesterol	23,86272
14	Cinsiyet	23,01739	Cinsiyet	24,61961	Cinsiyet	23,59253
15	Ejeksiyon fraksiyonu	22,9583	Sigara kullanımı	18,92446	Sigara kullanımı	20,23043
16	Hipertansiyon	15,29217	Hipertansiyon	8,95347	Hipertansiyon	7,48536
17	Diyabetes Mellitus	4,30131	Diyabetes Mellitus	4,16518	Diyabetes Mellitus	5,02555
18	Aile öyküsü	3,05012	Aile öyküsü	2,43529	Aile öyküsü	2,70653

5. TARTIŞMA

Bu çalışmada, veri madenciliği alanında sıkça kullanılan ağaç tabanlı sınıflandırma yöntemlerinden birisi olan RF yöntemi ve henüz sağlık alanında yaygın olarak kullanılmaya başlanmamış olan bir optimizasyon yöntemi olan GA'ların, RF yöntemi üzerindeki başarısı koroner arter hastalarına ait 97 kişilik bir veri seti üzerinde uygulanmıştır.

Bu çalışmaya benzer şekillerde daha önceden yapılmış olan bazı çalışmalarda GA yönteminin sınıflandırma problemlerinde başarılı bir şekilde optimizasyon sağlayarak sınıflama başarısını arttırdığı görülmüştür. 2009 yılında yapılmış bir çalışmada astım hastalığı teşhisinde GA yönteminin kullanılmasıyla önemli sayılabilecek doğruluk oranı ile sınıflandırma işlemi başarısı elde edilmiş olduğu görülmektedir. Gerçekleştirmiş oldukları testler sonucunda % 91,31 doğruluk oranı elde etmişlerdir (Er vd. 2009).

2012 yılında yapılmış olan bir başka çalışmada ise meme kanseri sınıflandırması için veri füzyonu ve genetik algoritma tabanlı gen seçimi yapılmış ve yapılan çalışmada belirlenen 10 gen ile sınıflandırma doğruluk oranı %94,65 olarak elde edilmiştir (Yıldız vd. 2012).

Doğan tarafından 2006 yılında yapılan çalışmada, gerçekleştirmiş oldukları karar destek sistemiyle kalp krizi için 61 hasta verisi kullanarak 50 hasta için kalp krizi, 11 hasta için sağlam tanısı; 106 hasta verisi kullanarak 86 hasta verisi için hiperlipidemi, 20 hasta verisi için sağlam tanısı; 96 hasta verisi kullanarak 76 hasta verisi için anemi, 20 hasta verisi için sağlam tanısı; 120 hasta verisi kullanılarak 35 hasta verisi için hipertiroidi, 35 hasta için hipotiroidi, 50 hasta içinse sağlam tanısı koymuşlardır. Daha sonra bu hastalarda sistemin ulaştığı sonuç, sağlık personelinin verdiği kararla %100 doğrulukla örtüşmüştür (Doğan 2006).

Laurikkala ve Juhola (1998) tarafından yapılmış olan çalışmada, kadınlarda idrar kaçırma verileri üzerine çalışılmış ve ileri istatistiksel bir sınıflama tekniği olan diskriminant analizi sonuçlarıyla karşılaştırılmıştır. GA yöntemiyle yapmış oldukları sınıflama sonuçlarının diskriminant analizi sonuçlarına göre daha belirleyici olduğunu göstermişlerdir.

Fidelis ve arkadaşlarının 2000 yılında yapılmış oldukları bir çalışmada, hem dermatoloji verileri hem de meme kanseri verileri incelenmiştir. Dermatoloji verilerinde yapmış oldukları GA optimizasyonu ile sınıflandırma sonucunda %95 doğruluk oranına ulaşmışlardır. Aynı şekilde meme kanseri için uyguladıkları GA optimizasyonu ile sınıflandırmada ise %67 doğruluk oranı elde etmişlerdir.

Tanwani ve Farooq'un 2009 yılında yapmış oldukları çalışmada, biyomedikal veri setlerinde GA sınıflama performansları üzerinde durulmuş ve 31 farklı veri setini 6 farklı evrimsel öğrenme algoritması ile karşılaştırmışlar ve veri setlerinin çoğunda GA yöntemiyle yapmış oldukları sınıflama sonuçlarının daha başarılı doğruluk oranlarına sahip olduğunu göstermişlerdir.

Sanz ve arkadaşlarının, 2008 yılında 7 farklı sağlık merkezinden elde ettikleri kardiyovasküler hastalıklara ait verilerle yapmış oldukları çalışmada, bulanık mantık sınıflandırıcısının GA yöntemiyle optimize edilerek sınıflandırma başarısını yükseltip yükseltmediğini incelemişler ve sonuç olarak GA yöntemiyle yapılmış olan sınıflandırma sonuçlarının doğruluk oranlarının daha yüksek olduğu sonucuna varmışlardır.

Bu çalışmadan elde edilen sonuçlarda GA ile optimize edilmiş olan RF modelinde doğru sınıflama yapma oranlarının optimize edilmeden ele alınan RF modellerine göre çok daha yüksek olduğu sonucuna ulaşılmıştır. Aynı şekilde bağımsız değişken sayısının azaltılarak kurulmuş olan modellerin oranlarına ve KA sonucunda elde edilmiş olan oranlara göre de GA sonuçlarının daha başarılı olduğu görülmektedir.

Kurulmuş olan modellerde 20, 50, 100, 500 ve 1000 ağaçtan oluşan ve GA ile optimize edilmiş olan RF sonuçları elde edilmiştir. Daha sonra aynı ağaç sayılarından oluşan modellerde optimize etmeden RF yöntemi kullanılmış ayrıca bu RF modellerini etkileyen değişkenlerin önem dereceleri bulunarak ve bunlardan yola çıkıp tahmin edici değişken sayısı yarı yarıya azaltılarak aynı sayıda ağaçlardan RF modelleri de kurulmuştur. Burada, sınıflandırmaya en çok etki eden değişkenler ile daha basit ve daha güçlü bir model kurularak, bu modelin GA ile optimize edilmiş olan 18 değişkenlik modele üstünlük kurup kurmadığının da incelenmesi düşünülmüştür. Aynı zamanda KA sonucunda elde edilen sınıflama sonucunun da diğer modellerle karşılaştırılması işlemi gerçekleştirilmiştir.

Genel olarak elde edilmiş olan sonuçlar incelendiğinde;

Gerçekte KAH (+) olan hastalar içinden kurulan model sonucunda doğru sınıflanmış olan hastaların oranını ifade eden duyarlılık değerleri incelendiğinde; kurulan modeller içerisinde hastaları en yüksek oranda belirleyen model GA ile optimize edilmiş olan 100 ağaçlık RF modelidir. Bu modelde duyarlılık değeri %79,17 olarak bulunmuştur. Aynı zamanda 20 ağaçlık RF modelinden elde edilen en etkili 9 değişkenle kurulmuş olan 20 ağaçlık RF modelinin duyarlılık değeri de %79,17 olarak bulunmuştur. Performans ölçüsü açısından bu iki model, en ideal değere ulaşan modeller olmuşlardır.

Gerçekte KAH (-) olan bireyler içinden kurulan model sonucunda doğru sınıflanmış olan bireylerin oranını ifade eden seçicilik değerleri incelendiğinde; kurulan modeller içerisinde KAH (-) olan bireyleri en yüksek oranda tespit eden modellerin GA ile optimize edilmiş olan modeller olduğu görülmektedir. Bu modellerin içerisinde 50, 500 ve 1000 ağaçlık RF sonuçlarının seçicilik değeri %89,8 olarak bulunmuştur. Seçicilik değerlerinde; GA optimizasyonu yapılmış olan tüm RF modelleri, diğer modellere göre açık üstünlük göstermektedir.

Gerçekte KAH (-) olan bireyler içinden kurulan modellerin hatalı bir biçimde KAH (+) olarak sınıflamış olduğu oranı ifade eden, yanlış pozitif oranları incelendiğinde GA ile optimize edilmiş olan modellerde en düşük değere ulaşıldığı görülmektedir (%10,2). Aynı zamanda KA sonucu elde edilmiş olan yanlış pozitif oranının, optimize edilmeden kurulmuş olan tüm RF modellerinin sonuçlarına ve en etkili 9 değişken alınarak kurulmuş olan tüm RF sonuçlarına oranla daha iyi bir performans gösterdiği görülmektedir. Ancak buna rağmen %28,57 olarak gözlenmektedir. Bu performans ölçüsü açısından da GA optimizasyonu yapılmış olan tüm RF modelleri, diğer modellere göre açık ara üstünlük göstermektedir.

Gerçekte KAH (+) olan hastalar içinden kurulan modellerin hatalı olarak KAH (-) olarak sınıflamış olduğu oranı ifade eden yanlış negatif oranları incelendiğinde en başarılı model GA ile optimize edilmiş olan 100 ağaçlık RF modelidir. Bu modelde yanlış negatif oranı %20,83 olarak bulunmuştur. Aynı zamanda 20 ağaçlık RF modelinden elde edilen en etkili 9 değişkenle kurulmuş olan 20 ağaçlık RF modelinin duyarlılık değeri de %20,83 olarak bulunmuştur. Bu performans ölçüsü açısından en

ideal modelin sadece GA optimizasyonu yapılmış olan RF modeli olduğu söylenemez. Ancak yine de GA modeli, optimize edilmeden kurulmuş olan tüm RF modellerine göre daha yüksek bir başarı göstermiştir.

Pozitif olabilirlik oranının yüksekliği gerçek KAH (+) olan hastaların sınıflandırılmasının ne derece doğru olduğunu ifade etmektedir. Oranın yüksekliği daha doğru bir sınıflama yapıldığını göstermektedir. Negatif olabilirlik oranı ise gerçekte KAH (-) olan kişilerin sınıflandırmalarının doğruluğunu ifade etmektedir. Burada ise oranın düşüklüğü daha doğru bir sınıflama yapıldığını göstermektedir. Bu değerler için GA ve KA karşılaştırıldığında, GA'nın çok daha doğru bir sınıflama yaptığı görülmektedir. Pozitif olabilirlik oranının en başarılı olduğu modeller, GA optimizasyonu yapılmış olan RF modelleridir. Aralarında en iyi değere sahip olan ise 50 ağaçlık RF modelidir. Aynı şekilde negatif olabilirlik oranının en başarılı olduğu modeller de GA optimizasyonu yapılmış olan RF modelleridir. Aralarında en iyi değere sahip olan ise 100 ağaçlık RF modelidir.

İlgilenilen tüm performans ölçülerinin içerisinde modelin başarısını en objektif verebilecek olan performans ölçüsü doğruluk oranıdır. Çünkü bu değer; duyarlılık ve seçicilik değerlerinin birlikte hesaplanması sonucu elde edilen bir değerdir. Gerçekte testin, hasta ve sağlam olarak toplamda doğru tanı koyma oranı anlamına gelmektedir. Doğruluk oranları incelendiğinde, ağaç sayılarına göre kurulmuş olan tüm modellerin kendi aralarında karşılaştırılması, bulgular bölümünde yapılmıştır. Tüm modeller içerisinde doğruluk oranı en yüksek olan model %83,51 oranında başarılı sınıflama yapabilmiş olan GA optimizasyonu yapılmış 100 ağaçlık RF modelidir.

GA optimizasyonu yapılmış olan tüm modeller içerisinde tüm performans ölçülerinin sonuçları birlikte düşünüldüğünde en ideal sonuca 100 ağaçlık RF modeli ile ulaşıldığı görülmektedir. 50 ağaçlık olan model de diğer modellere göre daha iyi bir performans göstermektedir. Bunun sebebinin denek sayısının az olmasından olabileceği düşünülmektedir. GA ya da RF yöntemleri iki bağımsız veri madenciliği tekniğidir ve çok büyük veri setlerinde, sonuçlara daha kolay ulaşılması amacıyla geliştirilmişlerdir. Ancak buna rağmen çalışma sonucunda, veri seti ne kadar küçük olsa da başarılı sınıflama oranları elde edilmiştir. Özellikle GA optimizasyonu yapılmış olan RF modellerinin sonuçlarının oldukça başarılı olduğu söylenebilir.

Random Forest yöntemiyle oluşturulmuş olan 20, 50, 100, 500 ve 1000 ağaçlık sınıflandırma sonuçları ise kendi içlerinde incelenecek olursa; ağaç sayılarının artmasıyla incelenen performans ölçülerinde çok büyük değişiklikler olmadığı görülmektedir. Buna karşın, en etkili 9 değişkenle kurulmuş olan modellerde ise "50 ağaçlık RF modeli haricinde" normal RF modellerinden daha yüksek başarı elde edilmektedir.

6. SONUÇ VE ÖNERİLER

RF yöntemi, en çok kullanılan veri madenciliği yöntemlerinden birisi olduğu halde veri sayısı az olan setlerde çok yüksek başarı oranları yakalanamamaktadır. Modeldeki ağaç sayısı yükseltilmiş olsa da doğruluk oranları yükselememiştir.

Yapılan bu çalışmadan elde edilen sonuçlardan yola çıkılarak, GA yöntemiyle optimize edilerek kurulan RF modellerinin, optimize edilmemiş olan RF modellerine oranla çok daha yüksek başarıya sahip olduğu söylenebilir. Aynı şekilde RF modellerinin sınıflama başarısını yükseltmesi açısından, ağaç sayısı kaç olursa olsun, kurulacak olan modellerin GA yöntemiyle optimize edilmesi yapılan bu çalışma sonucunda varılabilecek olan en önemli sonuç olmaktadır.

KAYNAKLAR

Akdag, B., Fenkci, S., Degirmencioglu, S, Rota S., Sermez, Y., Camdeviren, H. (2006) Determination of risk factors for hypertension through the classification tree method. ***Adv Ther***, 23(6): 885 -892.

Akdag, B., Cavlak, U., Cimbiz, A., Camdeviren, H. (2011) Determination of pain intensity risk factors among school children with nonspecific low back pain. ***Med. Sci. Monit.***, 17(2): 12 - 15.

Akman, M. (2010) Veri Madenciliği Yöntemlerine Genel Bakış ve Random Forests Yönteminin İncelenmesi: Sağlık Alanında bir Uygulama., Yüksek Lisans Tezi, Ankara Üniversitesi Sağlık Bilimleri Enstitüsü, Ankara, 78s

Akman, M., Genç, Y., Ankaralı, H. (2011) Random Forests Yöntemi ve Sağlık Alanında Bir Uygulama. ***Turkiye Klinikleri J Biostat.***, 3(1):36-48.

Aktükün, A. (2005) Asal Bileşenler Analizine Bootstrap Yaklaşımı. ***Ekonometri ve İstatistik.***, 1: 5 – 15.

Bolat, B., Erol, K., ve İmrak, C., (2004) Genetic Algorithms in Engineering Aplications and the Function of Operators. ***Journal of Engineering and Natural Sciences***, 4: 264-265.

Bohanec, M. and Bratko, I. (1994) Trading Accuracy for Simplicity in Decision Trees, Machine Learning, (Minton, S.), Kluwer Academic Publishers, Boston, 223-250s

Bozcuk, H., Bilge, U., Koyuncu, E., Gulkesen H. (2004) An application of a genetic algorithm in conjunction with other data mining methods for estimating outcome after hospitalization in cancer patients., ***Med Sci Monit***,10(6): 246-251

Breiman, L. (2001) Random forests., ***Machine Learning***, 45: 5–32.

Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A. (1994) Classification and Regression Trees, Chapman& Hall, 358s

Breiman, L. , 2002. Manual On Setting Up, Using, And Understanding Random Forests V3.1, http://oz.berkeley.edu/users/breiman/Using_random_forests_V3.1.pdf (06.11.2009)

Chernick, M.R. (2008) Bootstrap Methods: A Guide for Practitioners and Researchers, A John Wiley & Sons Inc. Publication, 359s.

Coşgun, E. ve Karaağaoğlu, E. (2011) Veri Madenciliği Yöntemleriyle Mikrodizilim Gen İfade Analizi. ***Hacettepe Tıp Derg.***, 42: 180-189.

Çivril, H. (2009) Hemşire Çizelgeleme Problemlerinin Genetik Algoritma İle Çözümü., Yüksek Lisans Tezi, Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü, Isparta, 87s

Çölkesen, İ. (2009) Uzaktan Algılamada İleri Sınıflandırma Tekniklerinin Karşılaştırılması ve Analizi., Yüksek Lisans Tezi, Gebze Yüksek Teknoloji Enstitüsü Mühendislik ve Fen Bilimleri Enstitüsü, Gebze, 151s

Deb, K. (1998) Genetic Algorithm in Search and Optimization: The Technique and Applications. Proceedings of International Workshop on Soft Computing and Intelligent Systems, 58 – 87.

Dirican, A. (2001) Tanı Testi Performanslarının Değerlendirilmesi ve Kıyaslanması. ***Cerrahpaşa J. Med.,*** 32: 25 -30.

Doğan, S. (2006). Türkçe Dökümanlar İçin N-Gram Tabanlı Sınıflandırma: Yazar, Tür ve Cinsiyet., Yüksek Lisans Tezi, Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü, İstanbul, 61s

Doğan, Ş. (2007) Veri Madenciliği Kullanılarak Biyokimya Verilerinden Hastalık Teşhisi., Yüksek Lisans Tezi, Fırat Üniversitesi Fen Bilimleri Enstitüsü, Elazığ, 91s

Elmas, Ç. (2011) Yapay Zeka Uygulamaları, Seçkin Yayıncılık, Ankara,424s.

Engin, O. (2001) Akış Tipi Çizelgeleme Problemlerinin Genetik Algoritma ile Çözüm Performansının Arttırılmasında Parametre Optimizasyonu., Doktora Tezi, İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü, İstanbul, 214s

Fidelis, M.V., Lopes, H.S., Freitas, A.A. (2000) Discovering Comprehensible Classification Rules with a Genetic Algorithm," Congress on Evolutionary Computation, USA, 805 - 810

Freund, Y., ve Schapire, R.E. (1996) Experiments with a New Boosting Algorithm", Machine Learning: Proceedings of the Thirteenth International Conference, Italy , s.1-15.

Gehrke, J. (2003) Decision Trees, The Handbook of Data Mining, (Ye, N.), Lawrence Erlbaum Associates, London, s. 3-23.

Goldberg, D.E. (1989) Genetic Algorithms in Search, Optimization and Machine Learning, Addison-Wesley Longman Publishing, Boston, 412s.

Gülpınar, V. (2008) Avrupa Birliği Ülkeleri İle Türkiye'nin Ekonomik Göstergelerinin Karar Ağacı Yöntemi İle Karşılaştırılması., Yüksek Lisans Tezi, Marmara Üniversitesi Sosyal Bilimler Enstitüsü, İstanbul, 140s

Gülten, A. Ve Doğan, Ş. (2008) Genetik Algoritmalar Yönteminin Biyomedikal Verileri Üzerindeki Uygulamaları. ***DAUM.*** 7: 12-16.

Han, J. and Kamber, M. (2006) Data Mining: Concepts and Techniques, Elsevier, San Francisco, 743s

İşlier, A. (2001) Üretim Hücrelerinin Bir Genetik Algoritma Kullanılarak Oluşturulması. ***AUJST***., 2: 137 -157.

Jefferson, M.F., Pendleton, N., Lucas, S.B., Horan, M.A. (1997) Comparison of a Genetic Algorithm Neural Network with Logistic Regression for Predicting Outcome after Surgery for Patients with Nonsmall Cell Lung Carcinoma., ***Cancer***, 73(7):1338-1342

Kahraman, M. (2010). Çok Amaçlı Genetik Algoritma Kullanarak DNA Mikrodizi Verilerinin Kümelenmesi., Yüksek Lisans Tezi, Fırat Üniversitesi Fen Bilimleri Enstitüsü, Elazığ, 66s

Karabulut, E. ve Karaağaoğlu, E. (2010) Biyoinformatik ve Biyoistatistik. ***Hacettepe Tıp Derg.***, 41: 162-170.

Kaya, B.(2010). Tıbbi Veri Kümeleri Arasındaki Birliktelik Kurallarının Çok Amaçlı Genetik Algoritma İle Çıkarılması., Yüksek Lisans Tezi, Fırat Üniversitesi Fen Bilimleri Enstitüsü, Elazığ,50s

Kayri, M. ve Boysan, M. (2008) Bilişsel Yatkınlık ile Depresyon Düzeyleri İlişkisinin Sınıflandırma ve Regresyon Ağacı Analizi ile İncelenmesi. ***H.U. Journal of Education***, 34: 168-177.

Korkem, E. (2013) Random Forests ve Naive Bayes Sınıflama Yöntemleri Yaklaşımı., Yüksek Lisans Tezi, Hacettepe Üniversitesi Sağlık Bilimleri Enstitüsü, Ankara, 52s

Kuzey, C. (2012) Veri Madenciliğinde Destek Vektör Makinaları ve Karar Ağaçları Yöntemlerini Kullanarak Bilgi Çalışanlarının Kurum Performansı Üzerine Etkisinin Ölçülmesi ve Bir Uygulama., Doktora Tezi, İstanbul Üniversitesi Sosyal Bilimler Enstitüsü, İstanbul, 318s

Laurikkala, J. and Juhola, M. (1998) A genetic-based machine learning system to discover the diagnostic rules for female urinary incontinence., ***Comput Methods Programs Biomed.***, 55 (1998) 217–228

Liaw, A. and Wiener, M.(2002) Classification and Regression by randomForest., ***R news***, 3: 18 -22s

Loh, W.Y. and Shih, X. (1997) Split selection methods for classification trees., ***Stat Sin***, 7: 815-840.

Michalewics, Z. (1992) Genetic Algorithms + Data Structures =Evolution Programs, Springer, New York, 375s.

Mitchell, M. (1999) An Introduction To Genetic Algorithms, A Bradford Book The MIT Press, London, 158s.

Nabiyev, V.V. (2010). Yapay Zeka, Seçkin Yayıncılık, Ankara,752s.

Özdemir, A. (2011) Doğrusal Olmayan Regresyonda Asimptotik Yöntemle Bootstrap Örneklemesi., Yüksek Lisans Tezi, Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü, İstanbul, 53s

Özekes, S.(2002) Veri Madenciliği Uygulaması., Yüksek Lisans Tezi, Marmara Üniversitesi Fen Bilimleri Enstitüsü, İstanbul, 97s

Pal, M. (2005) Random forest classifier for remote sensing classification. ***Int J Remote Sens.***, 26:1, 217-222s.

Parlak, M. (2007) Genetik Algoritmaların Hesapsal ve Yapısal Olarak İncelenmesi., Yüksek Lisans Tezi, Ondokuz Mayıs Üniversitesi Fen Bilimleri Enstitüsü, Samsun, 116s

Quinlan, J.R. (1993) C4.5:Programs For Machine Learning, Morgan Kaufmann, California, 302s

Rokach, L. And Maimon, O. (2010) Decision Trees, Data Mining And Knowledge Discovery Handbook, (Rokach, L. And Maimon) , Springer, New York, 149 – 175s.

Safavian, S.R., and Landgrebe, D. (1991) A Survey of Decision Tree Classifier Methodology***. IEEE Transactions.***, 21: 660 – 674.

Samur, M.K., Bilge, U., Samur, A.A., Bircan, İ. ve Saka, O. (2009) Klinik Bilgi Sistemi ile Elde Edilen Ailesel Boy Kısalığı Verisinde Lojistik Regresyon ve Genetik Algoritma Sonuçlarının Karşılaştırılması. ***VI. Ulusal Tıp Bilişimi Kongresi Bildirileri.***, *1: 356 – 363.*

Sanz, J., Pagola, M., Bustince, H., Brugos, A., Fernandez, A., Herrera F. (2011) Case Study on Medical Diagnosis of Cardiovascular Diseases Using a Genetic Algorithm for Tuning Fuzzy Rule-Based Classification Systems with Interval-Valued Fuzzy Sets", ***IEEE Symposium on Advances in Type-2 Fuzzy Logic Systems***, Paris, 9-15.

Serhatlıoğlu, S. ve Hardalaç, F. (2009) Yapay Zeka Teknikleri ve Radyolojiye Uygulanması. ***Fırat Tıp Derg.***,14,1:1-6.

Shah, S. and Kusiak, A. (2007) Cancer gene search with data-mining and genetic algorithms. ***Comput Biol Med.***, 37: 251 – 261.

Sivanandam, S.N. and Deepa, S.N. (2008) Introduction to Genetic Algorithms, Springer, Heidelberg, 453s.

Steinberg, D. (2009) CART: Classification and Regression Trees, The Top Ten Algorithms in Data Mining, (Wu, X., Kumar, V.), Taylor & Francis Group, s179 - 203.

Sümbüloğlu, K. ve Akdağ, B. (2007) Regresyon Yöntemleri ve Korelasyon Analizi, Hatiboğlu Yayınları, Ankara, 139s.

Sümbüloğlu, K. ve Akdağ, B. (2010) İyi Klinik Uygulamalar, Pamukkale Üniversitesi Yayınları, Denizli, 179s.

Takma, Ç. Ve Atıl, H. (2006) Bootstrap Metodu ve Uygulanışı Üzerine Bir Çalışma 2. Güven Aralıkları, Hipotez Testi ve Regresyon Analizinde Bootstrap Metodu., ***Ege Üniv. Ziraat Fak. Derg***.,43(2):63-72.

Tanwani, A. K. and Farooq, M. (2009) Performance evaluation of evolutionary algorithms in classification of biomedical datasets," ***11th Annual Conference Companion on Genetic and Evolutionary Computation Conference: Late Breaking Papers***, New York, 2617-2624

Taşkın, Ç. ve Emel, G.G. (2009) Sayısal Yöntemlerde Genetik Algoritmalar, Alfa Aktüel Yayınları, Bursa, 154s.

Temel, G.O. ; Çamdeviren, H. ; Akkuş, Z. (2005) Sınıflama Ağaçları Yardımıyla Restless Legs Syndrome (RLS) Hastalarına Tanı Koyma. ***İnönü Üniv Tıp Fak Derg.***, 12(2):111-117.

Tekin, I. (2012) Epikardiyal Yağ Dokusunun Transtorasik Ekokardiyografi ile Vertikal, Horizontal ve Alansal Değerlerinin Koroner Arter Hastalığı ve Antropometrik Ölçümlerle İlişkisi., Uzmanlık Tezi, Pamukkale Üniversitesi Tıp Fakültesi, Denizli, 58s

Tuğ, E. (2005). Genetik Algoritmalar İle Tıbbi Veri Madenciliği., Yüksek Lisans Tezi, Selçuk Üniversitesi Fen Bilimleri Enstitüsü, Konya,74s

Türkoğlu, F. (2006) Melez Yaklaşımlarla Türkçe Dökümanlarda Yazar Tanıma., Yüksek Lisans Tezi, Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü, İstanbul, 47s

WEB-1 Random Forests, Statistical Methods for Prediction and Understanding. http://www.stat.berkeley.edu/~breiman/RandomForests/

WEB-2 Karar Ağacı Öğrenmesi http://www.bilgisayarkavramlari.com/2012/04/11/karar-agaci-ogrenmesi-decision-tree-learning/

WEB-3 The R Project for Statistical Computing http://www.r-project.org/

WEB-4 evtree: Evolutionary Learning of Globally Optimal Trees

http://cran.r-project.org/web/packages/evtree/index.html

WEB-5 http://www.salford-systems.com/products/spm

Ye, N. (2003) The Handbook of Data Mining, Lawrence Erlbaum Associates, London, 677s

Yeniay,Ö. (2001) An Overview Of Genetic Algorithms. ***AUJST.***, 2: 37 - 49.

Yıldız, E. (2010) Online Çevirmen - Müşteri Buluşturma ve Metin Madenciliği Yöntemlerini Kullanarak Metin Sınıflandırma Sistemi. Dönem Projesi, Kocaeli Üniversitesi Mühendislik Fakültesi Bilgisayar Mühendisliği Bölümü, Kocaeli, 67s

Yıldız, M., Yüksel, A., Korürek, M., Aykan, A.Ç., Yıldız, B.Ş., Şahin, A., Hasdemir, H. Ve Özkan, M. (2012) Kalp Kateterizasyonu ile Hemodinamik Ölçümleri Saptanmış Atriyal Septal Defekt ve Ventriküler Septal Defektli Olguların Genetik Algoritmalar ve Çok Katmanlı Yapay Sinir Ağı ile Sınıflandırılması. ***Koşuyolu Heart J.***, 15: 45 – 50.

Yıldız, O., Tez, M., Bilge, H. Ş., Akcayol, M.A. ve Güler, İ. (2012) Meme Kanseri Sınıflandırması İçin Veri Füzyonu ve Genetik Algoritma Tabanlı Gen Seçimi**.** ***Gazi Üniv. Müh. Mim. Fak. Der.***, 27: 659 – 668.

Yıldırım, S. (2003), Tümevarım Öğrenme Tekniklerinden C4.5'in İncelenmesi., Yüksek Lisans Tezi, İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü, İstanbul,79s

Yu, X. and Gen, M. (2010) Introduction to Evolutionary Algorithms, Springer, London, 417s

EK-1

T.C.
PAMUKKALE ÜNİVERSİTESİ
GİRİŞİMSEL OLMAYAN KLİNİK ARAŞTIRMALAR ETİK KURULU

Sayı : 2012/ 48 29.11.2012
Konu :

Sayın;
Doç.Dr.Beyza AKDAĞ
Tıp Fakültesi
Biyoistatistik Anabilim Dalı
Öğretim Üyesi

İlgi: 12.11.2012 tarihli dilekçeniz.

İlgi dilekçe ile başvurmuş olduğunuz "Koroner Arter Hastalığının Risk Faktörlerinin İrdelenmesinde Alternatif Bir Yaklaşım: Genetik Algoritmalar" konulu çalışmanız da Prof.Dr.Asuman KAFTAN'ın yürütücüsü olduğu "Koroner Anjiyografi Endikasyonu Olan Hastalarda Epikardiyal Yağ Dokusunun Transtorasik Ekokardiyografi ile Ölçümü, Koroner Arter Hastalığı ve Vücut Kitle Endeksi ile İlişkisi" konulu çalışmanın verilerinin kullanılarak yapılması hakkındaki dilekçeniz **27.11.2012 tarih ve 06 sayılı** kurul toplantımızda görüşülmüş olup,

Yapılan görüşmelerden sonra, söz konusu çalışmanın yapılmasında **ETİK AÇIDAN SAKINCA OLMADIĞINA**, altı ayda bir çalışma hakkında Kurulumuza bilgi verilmesine oy birliği ile karar verilmiştir.

Bilgilerinizi rica ederim.

Prof. Dr. Kemalettin ACAR
Başkan

ÖZGEÇMİŞ

1983 yılında Kocaeli ili İzmit ilçesinde doğdu. İlk, orta ve lise öğrenimini İzmit'te tamamladı. 2007 yılında Mimar Sinan Üniversitesi İstatistik bölümünden mezun oldu. 2011 yılında Pamukkale Üniversitesi Sağlık Bilimleri Enstitüsü, Biyoistatistik Yüksek Lisans programında eğitime başladı. Halen Pamukkale Üniversitesi Biyoistatistik Anabilim Dalı'nda Araştırma Görevlisi olarak çalışmaktadır.