Random Forests Yöntemi ve Sağlık Alanında Bir Uygulama

Random Forests Methods and an Application in Health Science

Muhammet AKMAN,[a]
Yasemin GENÇ,[b]
Handan ANKARALI[c]

[a]Savunma Sanayii Müsteşarlığı, Ankara
[b]Biyoistatistik AD,
Ankara Üniversitesi Tıp Fakültesi,
Ankara
[c]Biyoistatistik AD,
Düzce Üniversitesi Tıp Fakültesi,
Düzce

Geliş Tarihi/*Received:* 11.05.2010
Kabul Tarihi/*Accepted:* 26.07.2010

Yazışma Adresi/*Correspondence:*
Muhammet AKMAN
Savunma Sanayii Müsteşarlığı, Ankara,
TÜRKİYE/TURKEY
makman@ssm.gov.tr

ÖZET Amaç: Veri madenciliği, genel olarak tanımlayıcı ve tahmin edici olmak üzere iki ana başlıkta incelenmektedir. Özellikle tıp alanında veri madenciliği daha çok tahmin edici yönüyle kullanılmaktadır. Bu çalışmada, ağaç tabanlı veri madenciliği yöntemlerinden birisi olan Random Forests (RF) yönteminin incelenmesi ve sağlık alanından elde edilen bir veri seti üzerine uygulaması yapılarak sonuçlarının tartışılması amaçlanmıştır. **Gereç ve Yöntemler:** RF yönteminde, karar ormanını oluşturan karar ağaçları orijinal veri setinden bootstrap yöntemiyle seçilen farklı örneklerden oluşturulmaktadır. Her karar ağacında veri setindeki tüm değişkenlerden rastgele seçilen az sayıda değişken kullanılmaktadır. Her ağaç bir sınıf için oy vermektedir ve orman sınıflayıcısı bütün ağaçların verdiği oyları toplayarak bir sınıf için son tahminini yapmaktadır. Yöntemin uygulanması amacıyla Diş hekimliği alanından elde edilen bir veri seti kullanılmıştır. **Bulgular:** Toplam 43 tane demografik, dental ve serolojik özelliklere ait veriler kullanılarak RF yöntemi ile %95.4 oranında başarılı bir sınıflandırma yapılmıştır. Bu karar ormanının hata oranı ise %3.33 olarak bulunmuştur. Aynı veri seti için Bagging ve CART yöntemi ile de sınıflama yapılmış ve Bagging yöntemi ile hata oranı %5.4, CART yöntemi ile %8.75 olarak bulunmuştur. **Sonuç:** RF yöntemi ile veri setindeki değişken sayısı ve örnek sayısı ne kadar çok olursa olsun genellikle hata oranı düşük sınıflamalar yapılmaktadır. Hata oranının düşüklüğü ise bir topluluk yöntemi olmasından kaynaklanmaktadır. Özelikle çok sayıda değişkenin olduğu DNA veri seti gibi binlerce gen arasından önemli olanları tespit etmek için kullanılabilir.

Anahtar Kelimeler: Veri madenciliği; sınıflandırma; random forests; karar ağaçları; karar ormanı

ABSTRACT Objective: In general, data mining has descriptive and predictive perspectives. In medicine, especially its predictive aspects are used. This study is aimed to analyze tree-based data mining method Random Forests (RF) and apply it on the health science dataset and discuss the results. **Material and Methods:** In RF method, decision trees which form decision forest are created with different datasets. These datasets are bootstrapped samples from original dataset. Also each decision tree is created with less randomly selected parameters from all of the predictors. Each decision tree votes for one class and forest aggregates votes from all trees, and makes final decision for the class. In order to apply this method, a dentistry dataset is used. **Results:** Analyzing 43 demographic, dental and serological features of the data, 95.4 % of succesful classification rate was achieved by using RF method. The error rate of the decision forest was found 3.33 %. Also same data was analyzed by using CART and Bagging methods. The error rate was found 5.4 % by using Bagging method and 8.75 % by using CART method. **Conclusion:** Using RF method, even there exists many predictors and large amount of data, generally lower error rate of classification is achieved. As RF is an ensemble method it gives better results. It can be used for determining important genes from large amount of DNA data set which has thousands of predictors (genes).

Key Words: Data mining; classification; random forests; decision trees; decision forest

Turkiye Klinikleri J Biostat 2011;3(1):36-48

Veri madenciliği, büyük boyutlardaki veriyi değişik perspektiflerden ele alarak analiz etme ve veriyi işe yarar bilgiye dönüştürme işlevleri olarak tanımlanabilir. Veri madenciliği yöntemleri genel olarak tanımlayıcı ve tahmin edici modeller olmak üzere iki ana başlık altında incelenebilir. Ancak bazı yöntemler hem tanımlayıcı hem de tahmin edici özelliğe sahip olabilmektedir. Tahmin edici modellerde, sonuçları bilinen verilerden yola çıkarak bir model kurulması ve kurulan bu modelden sonuçları bilinmeyen verilerin sonuç değerlerinin tahmin edilmesi amaçlanmaktadır.[1] Tanımlayıcı modellerde, eldeki verilerden strateji geliştirme ve karar verme süreçlerinde kullanılabilecek bilgiler sağlanmaktadır. Tanımlayıcı modeller daha çok veriler arasındaki gizli kalmış ilişkiyi ortaya çıkarırlar.[2]

Sınıflama genel anlamda, geçmişte toplanan verilerin hangi sınıfa ait olduğu bilindiğinde, yeni gelen verinin hangi sınıfa ait olduğunu bulma işlemidir ve tahminsel bir modeldir. Sınıfı bilinen nesneler ile (öğrenme veri seti) bir model kurulur. Kurulan model öğrenme kümesinde yer almayan nesneler ile (test veri seti) test edilerek performansı ölçülür.

Karar ağaçları, sınıflama, karar teorisi, kümeleme ve tahminsel fonksiyonlarda kullanılmaktadır. Karar ağaçlarını oluşturacak verideki değişkenler kategorik veya sürekli olabilirler. Eldeki verilerden yola çıkarak birçok farklı karar ağacı 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. Son yıllarda karar ağacı temelli olan ve çok sayıda karar ağacı ile oluşturulan topluluk yöntemler (ensemble learning) ve algoritmalar önem kazanmıştır. Bu yöntemlerden en çok öne çıkan ise Breiman tarafından ortaya konulan Random Forests (RF) yöntemidir. Bu yöntem Bagging, CART ve Random Subspace yöntemlerini temel almaktadır.[3]

RF yönteminde birden çok sınıflayıcının ortaya koyduğu sonuçlar bir araya getirilerek, topluluk adına tek bir karar verilmekte böylece daha güvenilir tahminler yapılmaktadır. RF yönteminde sonradan gelen veriye ait tahmin yapılmasının yanında, değişkenlerin önem derecelerinin hesaplaması, model indirgemesi ve aşırı değerlerin tespiti de yapılmaktadır.

Bu çalışmada, Random Forests yönteminin incelenmesi ve sağlık alanından elde edilen bir veri seti üzerine uygulaması yapılarak sonuçlarının tartışılması amaçlanmıştır.

GEREÇ VE YÖNTEMLER

RANDOM FORESTS YÖNTEMİ

RF yöntemi çok sayıda karar ağacından oluşan orman sınıflayıcısıdır ve bu yöntemle sınıflama veya regresyon ağaçları kurulabilmekte ve kümeleme yapılabilmektedir. Veri setindeki "sınıf değişkeni" kategorik ise sınıflama, sürekli ise regresyon ağaçları kurulmaktadır. Ormanda yer alan her bir karar ağacı, bootstrap tekniği ile orijinal veri setinden örneklem seçilmesi ve her karar düğümünde tüm değişkenler içinden belirlenen sayıda rastgele değişkenin seçilmesi ile oluşturulmaktadır. Bu yöntemde CART algoritması ile ağaçlar oluşturulur ve ağaçlar budanmaz. CART algoritması veri setinin hangi değişkenden başlayarak dallara ayrılacağına "bilgi kazancını" kullanarak karar verir. Ayrıca dallara ayrılmak için seçilen değişkenin uygun test kriteri (cut-off değeri) "gini katsayısı" ile belirlenir. Veri seti, test ve öğrenme olmak üzere iki bölüme ayrılmamış olsa da, veri setinin bütünü ele alınarak model kurulabilmekte ve modelin performansı modelin iç hatası olarak da bilinen Out-Of-Bag (OOB) hata oranı ile ölçülebilmektedir. Oluşturulan her ağaca önceden hesaplanan OOB hata oranına göre bir ağırlık verilir. En düşük hata oranına sahip ağaç en yüksek ağırlığı, en yüksek hata oranına sahip ağaç en düşük ağırlığı alır. Her ağaç belirlenen ağırlığa göre yaptığı sınıf tahmini için bir oy verme işlemine tabi tutulur. Ağaçların oyu oluştuktan sonra RF algoritmasında bu ağırlıklı oylar toplanır.

Random Forests Algortiması

Random Forests algoritması aşağıdaki adımlardan oluşur.

1) Orijinal veri setinden n tane bootstrap örnekleme yap. Her örneklemin 2/3'ünü ağacı oluşturmak için öğrenme verisi olarak kullan (inBag).

2) Her bootstrap örnekleme için budanmamış sınıflama veya regresyon ağacı aşağıdaki adımlar kullanılarak oluşturulur.

- inBag veri setinden her düğümde bütün tahmin değişkenleri içerisinden en iyi değişkeni seçmek yerine rastgele ***m*** tane tahmin değişkeni seç ve bunların içerisinden en iyi dallara ayıracak (en çok bilgi kazancı sağlayacak) olanı belirle.
- Belirlenen tahmin değişkeni için en iyi dallanma kriterini gini indeksi ile hesapla ve hesaplanan değere göre veri setini her düğümde iki alt dala ayır
- Yukarıdaki verilen işlemleri aşağıya doğru yaprak düğüm elde edilinceye kadar her düğümde tekrar et.

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 değişken sayısını ifade etmektedir.

3) ***n*** tane ağacın ayrı ayrı yapmış olduğu tahminleri bir araya getirerek yeni bir tahminde bulun;

- Sınıflama ağaçları için en çok oyu alan sınıfı final tahmin olarak seç,
- Regresyon ağaçları için ise yapılan oylamanın ortalamasını alarak nihai tahmini yap

Veri setinden hata oranını hesaplamak için;

- Her karar ağacı oluşturulurken, bootstrap aşamasında, bootstrap örneklemeyi, ağaç oluşturulacak veri (inBag) ve ağaç oluşturmak için kullanılmayan veri (OOB verisi) olmak üzere ikiye ayır. OOB verisiyle ağacı test et ve hata oranı tahmini yap.
- Bireysel ağaçların yaptığı OOB tahminlerini bir araya getir. Bu tahminlerden ormanın hata oranı kestirimi yap.

RF yönteminde Değişken Önem Derecesinin hesaplanması

Bir değişkenin önem derecesi diğer değişkenlerle olan ilişkisine bağlı olabileceğinden değişkenlerin önem derecesinin hesaplanması oldukça zordur. Değişken önem derecesi aşağıda detayları verilen iki farklı yöntemle bulunabilir.

Standart Yöntem: RF yönteminde, m. değişkenin önem derecesi şu şekilde bulunur. Karar ağacı oluşturulduktan sonra, OOB test verisi ağaçta yukarıdan aşağıya doğru yerleştirilir ve doğru sınıflama sayısı kaydedilir (c_i). Daha sonra, OOB test verisindeki m. değişkenin değerleri kendi içinde karıştırılır yani tüm değerlerin yeri değiştirilir. Değiştirilmiş OOB test verisi daha önce oluşturulmuş karar ağacı üzerine yukarıdan aşağıya doğru yerleştirilir ve doğru sınıflama sayısı kaydedilir (c_i^*). Değiştirilmemiş OOB verisi ile yapılan doğru sınıflama sayısından, değiştirilmiş OOB verisi ile yapılan doğru sınıflama sayısı çıkartılır ($d_i=c_i-c_i^*$). m. değişken için bu işlem, ormanı oluşturan tüm ağaçlarda tekrar edilir. Oluşan d_i'lerin ortalaması alınarak m. değişkenin kabaca önem derecesi belirlenmiş olur ($\bar{d}$). Orijinal veri setindeki tüm değişkenler için aynı işlem tekrar edilir ve tüm değişkenlerin ayrı ayrı kabaca önem derecesi bulunur. Bu aşamadan sonra, oluşturulan tüm ağaçların birbirinden bağımsız olduğu ve d_i'lerin normal dağıldığı varsayılarak d_i'lerin standart hatası hesaplanır. m.değişken için hesaplanan kaba değişken önem derecesi, hesaplanan standart hataya bölündüğünde (SE_{di}), bulunan değer ilgili değişkenin önem derecesi skorunu oluşturur (Eşitlik 1.1).

$$\text{Önem Derecesi Skoru} = \frac{\bar{d}}{SE_{di}} \qquad 1.1$$

Gini Yöntemi: Bu yöntemde, m. değişkenden dallara ayırma gerçekleşmeden önce ve alt dalla ayrıldıktan sonra bölünen veri için gini değerleri hesaplanır. Bölünme olmadan önceki verinin gini değeriyle, bölünme olduktan sonraki verinin gini değeri arasındaki fark alınır. Ormanda m. değişken kullanılarak oluşturulan her ağaç için bölünme olmadan önceki gini değeri ile bölünme olduktan sonraki gini değeri arasındaki farklar hesaplanır ve tüm ağaçlar aradaki farklar toplanır. Bulunan değer m. değişkenin gini önem derecesini verir. Bu işlemler tüm değişkenler için hesaplanır. Bu yöntemle elde edilen değişken önem derecesi, çoğu zaman standart yöntem ile bulunan değişken önem derecesine paralel bir değer olmaktadır.

RF Yönteminde Örnekler Arası Yakınlık (Proximity) Hesaplaması

Proximity değeri, veri setindeki kayıtlı bir verinin diğer satırlarda kayıtlı verilerle olan mesafesini ölçen bir gösterge olduğu için, sınıfından uzakta olan aşırı değerlerin (outlier) tespit edilmesinde kullanılabilmektedir. Her bireysel ağaç oluşturulduktan sonra, OOB ve inBag verisi dahil olmak üzere her veri ağaçta yukarıdan aşağıya doğru yerleştirilir ve aynı yaprak düğümde (terminal düğüm) sonlanan deneklerin sayıları kaydedilerek bir proximity matrisi oluşturulur. Proximity değerleri kullanılarak veri setindeki eksik değerlerin tahmin edilmesi de sağlanmaktadır. Proximity matrisi ile birbirine yakın verileri bir araya toplayarak kümeleme de yapılabilmektedir.

Random Forest Algoritmasının Üstün Yönleri ve Kısıtları

Random Forests yönteminin üstün yönleri ve kısıtlılıkları aşağıda sıralanmıştır.

Üstün Yönleri:

- Aşırı öğrenmeye ve veri setindeki kayıp verilere karşı oldukça güçlüdür. Çok fazla kayıp veri olduğu durumda da başarılı bir şekilde sınıflama yapmaktadır.
- Random Forests yöntemi kategorik, sürekli veya her ikisinin de bulunduğu veri setlerinde uygulanabilmektedir.
- Oluşturulan ağaçlar için budama işlemi yapmaya gerek yoktur.
- Hem büyük hem de küçük boyutlu veri setlerinde doğru sonuçlar vermektedir.
- Orijinal veri setini, öğrenme ve test veri seti olarak ayırmadan da model test edilebilir. Model orijinal veri setini kullanarak, iç hata oranını hesaplamaktadır. Hesaplanan iç hata oranı sapmasızdır.
- Çok büyük sayıda değişken içeren veri setleri için de uygun bir yöntemdir.
- Veri setindeki sınıflama değişkeninde (bağımlı değişken) sınıf sayısı çok fazla olsa da bu yöntem sınıflama yapabilmektedir.
- Random Forests yönteminde sınıflamayı gerçekleştiren değişkenlerin önem derecesi hesaplanabilmektedir.
- Değişkenler arasındaki ilişkiler ve mesafe, proximity matrisi oluşturarak tespit edilebilmektedir. Proximity, kümeleme yaparken, aşırı değerleri (outlier) tespit etmek için kullanılmasının yanında verinin görsel olarak gösterilmesinde de oldukça yararlıdır.
- Sınıflara düşen örnek sayısı (sınıf dağılımları) dengesiz olduğunda, sınıfları dengeli olacak şekilde ele alabilmektedir.
- Uygulamada analistin sadece iki parametre seçmesi gerekir. Bunlar "oluşturulacak ağaç sayısı" ve "her düğümde rastgele seçilecek değişken sayısı"dır. Ancak bu değerler ne olursa olsun sonuç çok fazla değişmemektedir.

Kısıtlılıkları:

- Tek bir karar ağacında olduğu gibi ortaya çıkan sonuç, ağaç yapısıyla görsel olarak görülemez. Model karmaşık olduğu için birçok karar ağacının değerlendirme sonucu, işlemlerin adımları görülemeyecek (black box) şekilde verilir.
- Lojistik regresyon ve yapay sinir ağları yöntemlerindeki gibi oluşan sonuç için bir güven aralığı verilmemektedir.
- Random Forests yöntemi için geliştirilen programlar, oluşan her ağacı sistem belleğinde tuttuğu için çok fazla bellek kullanmakta ve bu nedenle düşük bellekli bilgisayarlarda uygulaması zorlaşmaktadır.

VERİLER

Gazi Üniversitesi Diş Hekimliği Fakültesi Periodontoloji bölümüne periodontal hastalık şikayeti ile gelen 175 kişiye ait veriler kullanılarak Random Forests yöntemi uygulanmıştır.

Hastalar sistemik ve dental muayene değerlerine göre hem sistemik olarak kalp hastalığı hem de dental olarak periodontal hastalığı olan bireyler (PER+AMI, n=80), sistemik olarak sağlıklı ancak dental olarak periodontal hastalığı olan bireyler (PER-AMI, n=80) ve hem sistemik hem dental olarak sağlıklı bireyler (Kontrol Grubu, n=15) olmak üzere 3 gruba ayrılmıştır. Bu bireylerin demografik, dental ve serolojik özelliklerine ilişkin toplam 43 değişkeninden veriler toplanmıştır. Bu değişkenler Tablo 1'de gösterilmiştir.

TABLO 1: Çalışmada kullanılan değişkenlerin tanımlaması.

Sistemik Özellikler			
Değişken Adı	**Ölçü Birimi**	**Değişken Tipi**	**Açıklama**
CRP (Creaktif protein)	mg/l	Sürekli	Kalp hastalığında artar
CHOLmgdl (Total kolesterol)	mg/dl	Sürekli	
LDLmgdl (LDL kolesterol)	mg/dl	Sürekli	
HDLmgdl (HDL kolesterol)	mg/dl	Sürekli	
TGmgdl (Trigliserit)	mg/dl	Sürekli	
GLUCmgdl	mg/dl	Sürekli	Glukoz(açlık kan şekeri)
WBC		Sürekli	Beyaz kan hücresi
Diabetes	0=yok	Kategorik	Şeker hastalığının olup olmadığı sınıflandırılmıştır
	1=var		
Hypertens	0=yok	Kategorik	Yüksek tansiyon hastalığının olup olmadığı sınıflandırılmıştır
	1=var		
Hyperlip	0=yok	Kategorik	Yüksek lipidemi(yağ oranında artış)
	1=var		
Demografik Özellikler			
Değişken Adı	**Ölçü Birimi**	**Değişken Tipi**	**Açıklama**
AGE		Sürekli	Yaş
GENDER	1=erkek	Kategorik	Cinsiyet
	2=bayan		
Smoking	1=hiç içmemiş	Kategorik	Sigara içme durumu
	2=bırakmış		
	3=sigara içen		
BMI	kilo/boy	Sürekli	Vücut kitle indeksi
SosEcSt	1=düşük	Kategorik	Gelir düzeyleri ve eğitim düzeylerine göre hastalar sınıflandırılmıştır
	2=orta		
	3=yüksek		
Education	1=ilkokul 2=lise	Kategorik	Eğitim durumu
	3=üniversite		
Dental Özellikler			
Değişken Adı	**Ölçü Birimi**	**Değişken Tipi**	**Açıklama**
NTeeth		Kesikli	Ağızda mevcut diş sayısı
NExtract		Kesikli	Çekilmiş diş sayısı
NTeethPDless4		Kesikli	Cep derinliği 4 mm'den az olan diş sayısı
NTeethPD4to5		Kesikli	Cep derinliği 4mm ve 5mm ye eşit diş sayısı
NTeethPDover5		Kesikli	Cep derinliği 5 mm'den fazla dolan diş sayısı
NTeethCAL4to5		Kesikli	Klinik ataşman düzeyi 4 ve 5 mm olan diş sayısı
NTeethCALover5		Kesikli	Klinik ataşman düzeyi 5mm'den fazla olan diş sayısı
NSitesPD4to5		Kesikli	Cep derinliği 4 ve 5 mm olan diş bölgesi sayısı
NSitesPDover5		Kesikli	Cep derinliği 5 mm'den fazla olan diş bölgesi sayısı
NSitesCAL4to5		Kesikli	Klinik ataşman düzeyi 4 ve 5 mm olan diş bölgesi sayısı
NSitesCALover5		Kesikli	Klinik ataşman düzeyi 5 mm'den fazla olan diş bölgesi sayısı
PIImean (Plak indeks)	0=plak yok	Kategorik	
	1=gözle görülemeyen ve		Ağızda biriken ve periodontal hastalığa sebep olan
	aletin ucuna gelen plak varlığı		bakteri ürünleri ve yiyecek artıklarının oluşturduğu bir yapıdır.
	2=ince film şeridi şeklinde plak varlığı		Her bir hasta için ortalama değer olarak verilmiştir.
	3=dişin 2/3'ünü kaplayan plak varlığı		
GImean (Gingival indeks)	0=sağlıklı dişeti	Kategorik	Dişeti iltihabını sınıflayan bir indeks sistemidir
	1=iltihap başlangıcı, kanama yok		
	2=ilerlemiş iltihap ve adacıklar		
	şeklinde kanama		
	3=ilerlemiş iltihap ve spontan kanama		
PDmean (Cep derinliği)		Sürekli	Periodontal hastalığa bağlı olarak meydana gelen kemik yıkımını göstermektedir
CALmean (klinik ataşman düzeyi)		Sürekli	Kemik yıkımı ile beraber yumuşak dokulardaki kayıpları gösterir ve her bir hasta
için ortalama değer verilir			
BOPmean (sondlamada kanama)	var ya da yok olarak değerlendirilir	Kategorik	Hastalığın akut durumda olup olmadığını gösterir ve her bir hasta için ortalama
			değer verilir

devamı →

TABLO 1: ***devamı***

Serolojik Özellikler			
Değişken Adı	**Ölçü Birimi**	**Değişken Tipi**	**Açıklama**
LimulusLPS		Sürekli	Serumda belirlenen lipopolisakkarit düzeyi
AaPAL1000ratio		Sürekli	A.a (periodontal hastalıklarda etkili olan bir mikroorganizmadır) bu mikroorganizmanın hastalık yapan yüzey antijenlerinden PAL a karşı antikor cevabınının serumda 1/1000 dilüsyonda ölçülmesi
AaPAL2000ratio		Sürekli	Mikroorganizmanın (A.a) hastalık yapan yüzey antijenlerinden PAL a karşı antikor cevabınının serumda 1/2000 dilüsyonda ölçülmesi
AaLPS1500ratio		Sürekli.	Mikroorganizmanın (A.a) hastalık yapan yüzey antijenlerinden lipopolisakkarite karşı antikor cevabınının serumda 1/1500 dilüsyonda ölçülmesi
AaLPS3000ratio		Sürekli	Mikroorganizmanın (A.a) hastalık yapan yüzey antijenlerinden lipopolisakkarite karşı antikor cevabınının serumda 1/3000 dilüsyonda ölçülmesi
AaOMP1000ratio		Sürekli	Mikroorganizmanın (A.a) hastalık yapan yüzey antijenlerinden outer membrane proteine (dış membran proteini) karşı antikor cevabınının serumda 1/1000 dilüsyonda ölçülmesi
AaOMP2000ratio		Sürekli	Mikroorganizmanın (A.a) hastalık yapan yüzey antijenlerinden outer membrane protein (dış membran proteini) karşı antikor cevabınının serumda 1/2000 dilüsyonda ölçülmesi
PgLPS100ratio		Sürekli	P.g(periodontal hastalıklarda etkili olan bir mikroorganizmadır) bu mikroorganizmanın hastalık yapan yüzey antijenlerinden lipopolisakkarite karşı antikor cevabınının serumda 1/100 dilüsyonda ölçülmesi
PgLPS200ratio		Sürekli	Mikroorganizmanın (P.g) hastalık yapan yüzey antijenlerinden lipopolisakkarite karşı antikor cevabınının serumda 1/200 dilüsyonda ölçülmesi
PgOMP200ratio		Sürekli	Mikroorganizmanın (P.g) hastalık yapan yüzey antijenlerinden outer membrane protein (dış membran proteini) karşı antikor cevabınının serumda 1/200 dilüsyonda ölçülmesi
PgOMP400ratio		Sürekli	hastalık yapan yüzey antijenlerinden outer membrane protein (dış membran proteini) karşı antikor cevabınının serumda 1/200 dilüsyonda ölçülmesi

VERİ ANALİZİNDE KULLANILAN PROGRAM

Çalışmamızda veri analizi için Salford System tarafından, Leo Breiman ve Adele Cutler'ın oluşturmuş oldukları kaynak kodları kullanılarak geliştirilen RandomForests programı kullanılmıştır. Program Salford System tarafından http://salford-systems.com web adresinde yayınlanmakta olup değerlendirme sürümü eğitim amaçlı olarak temin edilebilmektedir.

BULGULAR

Karar ormanını oluşturacak karar ağacı sayısı Breiman tarafından önerilen ve varsayılan değer olan 500 olarak seçilmiştir. Orijinal veri setimiz dışında ayrı bir test veri seti olmadığı için, modelin testi RF algoritması tarafından ayrılan OOB test verisi ile yapılmıştır.

Sınıflara düşen denek sayıları birbirine yakın olmadığından sınıflara düşecek denek sayılarını dengelemek için gerekli ağırlıklar ***RandomForests*** programındaki **"Class Wts"** menüsünden **"BALANCED"** seçeneği işaretlenerek Tablo 2'deki gibi elde edilmiştir.

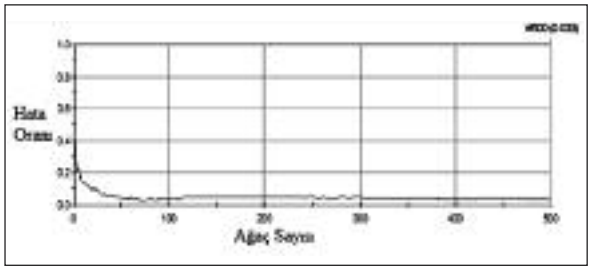
TABLO 2: Sınıf değişkenlerini dengelemek için uygulanacak ağırlıklar.

Hedef sınıf	Ağırlık
PER+AMI	1
PER-AMI	1
Kontrol grubu	5.333

Her düğümde rastgele seçilecek değişken sayısı, toplam değişken sayısının karekökünün tam sayıya yuvarlanmasıyla belirlenmiştir. Bu yönteme göre veri setimizde 43 değişken olduğundan bireysel ağaçlar oluşturulurken her düğümde $\sqrt{43} \cong 7$ değişken kullanılması uygundur.

Ağaçlar oluşturulduktan sonra tüm veriler ağaçlarda yukarıdan aşağıya doğru yerleştirilerek aynı düğümde sonuçlanan denekler belirlenmiş ve proximity matrisi oluşturulmuştur. Örneğimizde proximity matrisi 175 x 175 boyutlarındadır.

Karar ormanını oluşturan her karar ağacının OOB hata oranı hesaplandıktan sonra, her karar ağacına OOB hata oranına göre bir ağırlık verilmiştir.



ŞEKİL 1: Karar ormanındaki ağaç sayısına göre hata oranının değişimi.

Ağaç sayılarına göre, ormanın hata oranı Şekil 1'de verilmiştir. Şekilde görüldüğü gibi orman, bir ağaçla temsil edildiğinde hata oranı 0.4166 (%41.66), 10 ağaçla temsil edildiğinde 0.1375 (%13.75), 150 ağaçla temsil edildiğinde ise 0.04583 (%4.583)'tür. Ormandaki ağaç sayısı arttıkça hata oranı da azalmaktadır. Nihai olarak karar ormanını 500 karar ağacı oluşturduğunda ormanın genel hata oranı 0.0333 (%3.333) olarak hesaplanmıştır.

RF yöntemine göre yapılan sınıflama sonucunda, doğru ve yanlış sınıflandırılan denek sayıları, doğru ve yanlış sınıflama oranları Tablo 3'de gösterilmiştir. Tabloda da gösterildiği üzere sınıf değeri "PER+AMI" olan 80 örneğin 75 tanesi doğru sınıflandırılırken, 5 tanesi yanlış sınıflandırılmıştır. Sınıf değeri "PER-AMI" olan 80 deneğin 3 tanesi yanlış sınıflandırılmıştır. Kontrol grubunda yer alan 15 deneğin ise tümü doğru sınıflandırılmıştır.

$$\text{PER+AMI için hata oranı} = \frac{5}{80} = 0.0625 \qquad (3.1)$$

$$\text{PER-AMI için hata oranı} = \frac{3}{80} = 0.0375 \qquad (3.2)$$

$$\text{Kontrol grubu için hata oranı} = \frac{0}{15} = 0.0 \qquad (3.3)$$

RF algoritması sınıflardaki farklı denek sayılarını dengelemek için Tablo 2'de verildiği gibi ağırlıklar oluşturmaktadır. Bu ağırlıklar ile sınıf değişkenlerine ait eşit sayıda denek varmış gibi analiz yapılmaktadır. RF yöntemi sınıflardaki denek sayılarını dengelediği için ormanın hata oranı bulunurken yanlış sınıflanan denek sayısını Tablo 2'de verilen ağırlıkları da kullanarak Eşitlik 3.4'te ki gibi ağırlıklı toplam denek sayısına bölmektedir.

$$\text{RF yöntemine göre ormanın hata oranı} = \frac{5+3}{(80x1+80x1+15x5,333)} = \frac{8}{240} = 0,0333 \qquad (3.4)$$

Eğer model kurulduktan sonra sınıf değişkenlerine verilen ağırlıklar dikkate alınmadan, yanlış sınıflanan denek sayısını (8 denek) veri setindeki tüm denek sayısına (175 denek) bölünmesi ile sınıflama hata oranı hesaplaması yapılmış olsaydı Eşitlik 3.5' teki gibi bulunacaktı.

$$\text{Sınıflama hata oranı} = \frac{8}{175} = 0.0457 \qquad (3.5)$$

Bulunan bu değer modelin kurulmasındaki varsayımları dikkate almadığından modelin hata oranını doğru yansıtmayabilir. Bu çalışmada yapılan tüm değerlendirmeler ve karşılaştırmalar, Eşitlik 3.4'te verilen hata oranı hesaplama yöntemine göre yapılmıştır.

Yapılan sınıflama sonucunda, Tablo 4'teki sınıflama çizelgesi elde edilmiştir. Modelin doğru sınıflama oranı, tablodaki doğru sınıflanan deneklerin sayısının toplam denek sayısına bölünmesi ile elde edilmiştir.

$$\text{Ormanın doğru sınıflama oranı} = \frac{75+55+15}{175} = 0.9543$$

$$\text{PER+AMI doğru sınıflama oranı} = \frac{75}{80} = 0.9375$$

TABLO 3: Yapılan sınıflama sonucunda elde edilen oranlar.

Sınıf	Toplam Denek Sayısı	Doğru Sınıflanan Denek Sayısı	Yanlış Sınıflanan Denek Sayısı	Doğru Sınıflama Oranı	Yanlış Sınıflama Oranı
PER+AMI	80	75	5	0,9375	0,0625
PER-AMI	80	77	3	0,9625	0,0375
Kontrol Grubu	15	15	0	1	0

$$\text{PER-AMI doğru sınıflama oranı} = \frac{77}{80} = 0{,}9625$$

$$\text{Kontrol grubu doğru sınıflama oranı} = \frac{15}{15} = 1$$

Random Forests programı yardımıyla, değişkenlerin standart yöntemle ve gini yöntemiyle hesaplanan önem dereceleri Tablo 5a'da ve Tablo 5b'de verilmiştir. Modele katkısı fazla olan değişkenler yüksek önem derecesi skoruna sahiptir.

Gini Yöntemine göre CALMEAN, PDMEAN, WBC değişkenleri sınıflara ayırmada en çok bilgi sağlayan ilk 3 değişken olarak görülmektedir.

Standart Yönteme göre CALMEAN, WBC, AAPAL200 değişkenleri sınıflara ayırmada en çok bilgi sağlayan ilk 3 değişken olarak görülmektedir.

Değişkenlerin önem dereceleri belirlendikten sonra önemli görülen değişkenlerle yeniden model kurulabilir. Tablo 5a'da ve Tablo 5b'de önem derecesine göre en önemli bulunan ilk 10 değişken ile

TABLO 4: OOB test verisi ile yapılan test sonucu oluşturulan sınıflandırma çizelgesi.

	Tahmin edilen sınıf			
	PER+AMI	**PER-AMI**	**Kontrol Grubu**	**Toplam**
Gerçek Sınıf				
PER+AMI	75	3	2	80
PER-AMI	2	77	1	80
Kontrol Grubu	0	0	15	15
Tahmin edilen denek sayısı	77	80	18	175
Doğru Sınıflama Oranı		0,9543		

TABLO 5a: Değişkenlerin gini yöntemiyle hesaplanan önem dereceleri.

Değişken	Önem Derecesi Skoru	Değişken	Önem Derecesi Skoru	Değişken	Önem Derecesi Skoru	Değişken	Önem Derecesi Skoru
CALMEAN	16.765	V22_A	2.264	LDLMGDL	0.862	PGLPS100	0.184
PDMEAN	11.646	NSITESPD	1.825	V21_A	0.758	PGOMP200	0.178
WBC	7.972	BOPMEAN	1.804	V26_A	0.666	NTEETHPD	0.162
AALPS300	7.792	GIMEAN	1.344	LIMULUSL	0.618	PGOMP400	0.154
AAPAL200	6.605	V24_A	1.323	TGMGDL	0.567	NTEETH	0.121
CRP	6.098	CHOLMGDL	1.168	SMOKING	0.521	NEXTRACT	0.118
AALPS150	5.522	V28_A	1.154	AAOMP100	0.488	AGE	0.111
GLUCMGDL	5.473	NTEETHCA	1.088	HYPERLIP	0.417	EDUCATIO	0.030
AAPAL100	4.883	NSITESCA	1.034	DIABETES	0.403	SOSECST	0.007
PLIMEAN	2.884	AAOMP200	1.032	PGLPS200	0.328	GENDER	0.004
HDLMGDL	2.520	BMI	0.886	HYPERTEN	0.221		

TABLO 5b: Değişkenlerin standart yöntemle hesaplanan önem dereceleri.

Değişken	Önem Derecesi Skoru	Değişken	Önem Derecesi Skoru	Değişken	Önem Derecesi Skoru	Değişken	Önem Derecesi Skoru
CALMEAN	16.080	HDLMGDL	1.954	HYPERLIP	0.926	PGLPS100	0.062
WBC	14.081	AAOMP200	1.816	V28_A	0.924	PGLPS200	0.037
AAPAL200	10.418	GIMEAN	1.553	CHOLMGDL	0.921	AAOMP100	0.032
CRP	9.279	DIABETES	1.206	V21_A	0.724	AGE	0.025
AAPAL100	7.772	PLIMEAN	1.147	NSITESCA	0.320	GENDER	0.023
GLUCMGDL	6.216	BOPMEAN	1.099	SMOKING	0.299	EDUCATIO	0.010
V22_A	5.469	LIMULUSL	1.025	NTEETHPD	0.290	SOSECST	0.000
PDMEAN	4.412	HYPERTEN	0.998	LDLMGDL	0.191	NTEETH	0.000
AALPS300	2.665	V26_A	0.951	NEXTRACT	0.166	PGOMP400	0.000
AALPS150	2.598	NTEETHCA	0.934	TGMGDL	0.132	BMI	0.077
V24_A	2.157	NSITESPD	0.928	PGOMP200	0.084		

TABLO 6: Ağaç sayılarına ve modele giren değişke sayısına göre hata oranları.

Ağaç Sayısı	Tüm değişkenler modele girdiğinde hata oranı	*Gini yöntemiyle* en önemli bulunan 10 değişken modele girdiğinde hata oranı	*Standart yöntemle* en önemli bulunan 10 değişken modele girdiğinde hata oranı
300	0,046	0,067	0,033
400	0,038	0,075	0,038
500	0,033	0,071	0,033
600	0,038	0,071	0,033
700	0,038	0,058	0,033
800	0,033	0,063	0,033
900	0,038	0,071	0,029
1000	0,042	0,067	0,033
1300	0,033	0,063	0,033
1600	0,038	0,063	0,033
2000	0,033	0,058	0,029

yeniden model kurulmuş ve bu modellere ilişkin hata oranları Tablo 6'da verilmiştir.

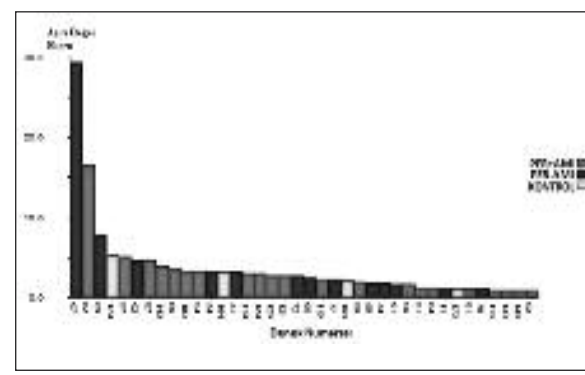
Tablo 6'ya göre gini yöntemi kullanılarak en önemli ilk 10 değişken modele alındığında 700 ağaç ile en iyi sonuç gözlenmiştir. Standart yöntemle belirlenen en önemli ilk 10 değişken modele alındığında ise 900 ağaç ile en iyi sonuç elde edilmiştir. Standart yöntemle belirlenen en önemli ilk 10 değişkenle yeniden model kurulursa, 500 ağaçlık ormanın hata oranı 0.033, 900 ağaçlık ormanın hata oranı 0.029 olarak bulunmaktadır. Standart yöntemle belirlenen en önemli ilk 10 değişken, tüm değişkenlerin modele girdiğinde hesaplanan ormanın hata oranından düşük bir hata oranı vermektedir. Bu sonuçlara göre periodontoloji veri seti için değişken önem derecesinin standart yöntemle hesaplanması önerilebilir.

Tablo 7'de farklı değişken sayılarına göre 500 ve 1000 ağaçlık ormanların hata oranları gösterilmiştir. Tabloda da görüldüğü gibi hata oranları birçok değişken sayısı için aynı kalmaktadır. Bu sonuca göre değişken sayısının model üzerinde önemli bir etkisinin olmadığı söylenebilir.

RF yöntemi ile model kurulduktan sonra veri setindeki hangi değerlerin aşırı değer (outlier) oldukları benzerlik veya farklılık ölçüleri yardımıyla hesaplanır. Bu amaçla her bir bireyin ait olduğu gruptaki diğer deneklerden ortalama uzaklığı hesaplanmış bu uzaklık değerleri ortanca değer ve interquartile range kullanılarak standardize edilmiştir. Bu işlem test veri setindeki tüm deneklere uygulanarak her denek için aşırı değer skoru hesaplanmış olur. Bu skor 10' un üzerinde ise söz konusu gözleme "aşırı değer" denir. 5 ile 10 arasında ise "şüpheli aşırı değer" olarak göz önüne alınmalıdır. Şekil 2'nin düşey eksenindeki değerler standardize edilmiş aşırı değer skorlarını, yatay eksendeki değerler ise denek numaralarını göstermektedir. En yüksek dereceli aşırı değere göre diğer aşırı değerlerde göreceli olarak aşırı değer olma derecesi almaktadırlar. Şekil 2'ye göre "PER-AMI" sınıfına düştüğü modellenen 67. denek ve 69.denekler aşırı değerdir ve aşırı değer olma derecesi yüksektir. "PER+AMI" sınıfına düştüğü modellenen 82. denek bir aşırı değerdir ve aşırı değer olma derecesi yüksektir. Aşırı değer olduğu tespit

TABLO 7: Değişken sayılarına göre ormanın hata oranları.

Ağaç Sayısı	Değişken Sayısı	Hata Oranı	Ağaç Sayısı	Değişken Sayısı	Hata Oranı
500	3	0,033	1000	3	0,033
500	7	0,033	1000	7	0,042
500	14	0,046	1000	14	0,042
500	21	0,033	1000	21	0,033
500	28	0,033	1000	28	0,038
500	35	0,033	1000	35	0,033



ŞEKİL 2: Aşırı değer olan deneklerin gösterilmesi.

edilen deneklere ait veri, analist tarafından yeniden gözden geçirilmelidir. Aşırı değer olmasının sebepleri klinik olarak araştırılmalı ve ilgili deneğe ait verinin çıkartılmasına veya modelde kalmasına karar verilmelidir. Aşırı değerlerin modelden çıkartılmasına karar verilirse, modelin yeni veri seti ile yeniden kurulması gerekir.

RF YÖNTEMİ VE BAGGING YÖNTEMİNE AİT SONUÇLARIN KARŞILAŞTIRILMALI İNCELENMESİ

Bagging, bootstrap yöntemiyle seçilen örneklemlerle oluşturulan çok sayıda karar ağacının yapmış olduğu tahminleri toplayarak nihai sınıf tahmini yapan bir yöntemdir. Kurulan ağaç yapısında orijinal veri setindeki tüm değişkenleri kullanmaktadır. Bootstrap örnekleme seçerken de orijinal veri setindeki toplam veri sayısı kadar örneği sınıf yapısını bozmayacak şekilde seçmektedir.

Bagging yöntemi ile yapılan uygulamada, seçilen bootstrap örneklem büyüklüğü 175, ağaçlarda kullanılacak değişken sayısı 43, oluşturulacak ağaç sayısı 500 olarak seçildiğinde modelin hata oranı Tablo 8'de görüldüğü gibi hesaplanmaktadır. Bagging yöntemi uygulandığında modelin hata oranı 0,054 olarak bulunmaktadır.

RF yönteminde 500 ağaç ile oluşturulan ormanın hata oranı 0,033 idi. Bagging yöntemi ile elde edilen hata oranı 0,054 daha yüksektir. Periodontoloji veri seti için, RF yönteminin Bagging yöntemine göre daha iyi sonuç verdiği söylenebilir.

RF YÖNTEMİ VE CART YÖNTEMİNE AİT SONUÇLARIN KARŞILAŞTIRMALI İNCELENMESİ

CART yöntemi, orijinal veri setinden en çok bilgi kazancı sağlayan değişkenden başlayarak alt dallara ayrılan ve seçilen değişkenin hangi değeri ile iki ayrı dala ayrılacağına ise gini katsayısı kullanarak belirleyen bir yöntemdir. CART yönteminde bütün veri seti tek bir ağaçla modellenir.

TABLO 8: Bagging yöntemi uygulandığında hata oranı.

Ağaç sayısı	Ağaçlarda kullanılan örnek sayısı	Ağaçları oluşturmak için kullanılan değişken sayısı	Modelin hata oranı
500	175	43	0,054

TABLO 9: CART yöntemi uygulandığında hata oranı.

Ağaç sayısı	Ağaçlarda kullanılan denek sayısı	Ağaçları oluşturmak için kullanılan değişken sayısı	Modelin hata oranı
1	175	43	0,0875

Periodontoloji veri seti için CART yöntemi ile sınıflama yapılmış ve modelin hata oranı Tablo 9'da gösterildiği gibi 0,0875 olarak bulunmuştur.

CART yöntemi ile elde edilen hata oranı, RF yöntemiyle bulunan hata oranından yüksektir. Periodontoloji veri seti için, RF yönteminin CART yöntemine göre daha iyi bir sonuç verdiğini söylenebilir.

TARTIŞMA

Bu çalışmada, ağaç tabanlı veri madenciliği yöntemleri ve topluluk sınıflama modelleri arasında yer alan RF yöntemi tanıtılmış ve periodontoloji bilim dalından elde edilen bir veri seti üzerinde uygulanmıştır.

Topluluk yöntemler zayıf öğrenicileri bir araya getirerek güçlü öğreniciler oluşturan yöntemlerdir. Öğrenme bir anda değil yavaş yavaş ve dengeli olmakta ve tüm zayıf öğrenicilerin sonuçları toplanarak bir komite oluşturulmaktadır. Her zaman komitenin sınıflama performansı bireysel sınıflayıcılardan daha yüksek olmaktadır. Yapılan çalışmalar çok sayıda karar ağacının rastgele veri ve değişkenlerle oluşturularak bir araya getirilmesi yoluyla bireysel oluşturulan karar ağaçlarına kıyasla çok daha başarılı ve istikrarlı sonuçlar verdiğini göstermektedir.[4] Çalışmamızda örnek veri setine RF yöntemi uygulandığında modelin genel hata oranı %3.33 bulunurken tek bir ağaçla sınıflama yapan CART yöntemi ile bu oran %8.75 olarak elde edilmiştir. Ayrıca RF yönteminden elde edilen sonuçlar diğer bir topluluk öğrenme yöntemi olan bagging ile de karşılaştırılmış, bagging yöntemi ile genel hata oranı %5.4 olarak bulunmuştur.

Breiman, RF yönteminin, diskriminant analizi, destek vektör makineleri ve yapay sinir ağları gibi sınıflandırıcılar dahil birçok diğer sınıflandırıcılara göre de daha iyi sonuçlar verdiğini bunun ya-

nında yöntemin aşırı öğrenmeye karşı da oldukça güçlü olduğunu belirtmiştir.[3] Son yıllarda RF yöntemi araştırmacılar tarafından, çok farklı veri setleri kullanılarak çeşitli yöntemlerle karşılaştırılmıştır. Clark ve ark. yaptıkları çalışmada RF yönteminin destek vektör makinelerine göre daha iyi sonuçlar verdiğini ve veri setinde çok sayıda kayıp veri bulunsa da RF yönteminin oldukça başarılı olduğunu vurgulamışlardır.[5] Wu ve ark. yaptıkları çalışmada, kanser hastalığını teşhis etmek için Mass spectrometry veri seti için doğrusal diskriminant analizi, quadratik diskriminant analizi, k-komşuluk sınıflayıcısı, bagging, boosting, destek vektör makineleri ve RF yöntemlerinin sınıflama başarılarını karşılaştırmış ve RF yönteminin diğerlerinden daha başarılı sonuçlar verdiği gözlemlemişlerdir.[6] Yoonhee ve ark., Random Forests yöntemini ki-kare ve lojistik regresyon modelleri ile karşılaştırmışlardır. RF yönteminin çok sayıda değişken içeren veri setlerinde ve genler arasındaki ilişkilerin bulunmasında geleneksel istatistiksel yöntemlere göre daha avantajlı olduğunu belirtmişlerdir. Ayrıca önemli değişkenlerin tespit edilmesinde de RF yönteminin ki-kare testinden daha iyi performans ortaya koyduğunu açıklamışlardır.[7] Statnikov ve ark., kanser hastalığına sebep olan en önemli genlerin tespit edilmesi için yaptıkları mikroarray çalışmasında RF ve destek vektör makineleri yöntemlerinin sonuçlarını karşılaştırmışlar ve destek vektör makinelerinin Random Forests yönteminden daha iyi performans gösterdiğini ortaya koymuşlardır.[8] Ruiz-Gazen ve ark., sınıf dağılımları dengesiz olan meteoroloji veri seti için lojistik regresyon ve RF yöntemini birçok açıdan karşılaştırmışlardır. Her iki yönteminde diğer standart yöntemlerden daha iyi sonuçlar verdiğini tespit etmişlerdir. Ancak iki yöntem arasında önemli bir fark bulamadıklarını ve daha sağlıklı bir karşılaştırmanın yapılabilmesi için veri setine yeni ilave verilerin girilmesi gerektiğini belirtmişlerdir. Diğer taraftan lojistik regresyonun RF yöntemine göre daha hızlı olduğunu ve yorumlanmasının daha kolay olduğu için meteorolojistler için tercih edilebileceğini söylemişlerdir.[9] Amasyalı ve ark., Cline ailesi algoritmaları diye adlandırdıkları Cline tabanlı algoritmalar ile içlerinde Random Forests yönteminin de bulunduğu 23 farklı ağaç tabanlı yöntemi sekiz farklı veri seti kullanarak karşılaştırmışlardır. Bu çalışmaya göre RF yöntemi bazı Cline ailesi algoritmalarından daha iyi performans gösterirken, Cline tabanlı olan ve RF yöntemindeki gibi tekil karar ağaçlarının sonuçları birleştiren CLMIX ormanları yöntemi RF yönteminden biraz daha iyi sonuç vermiştir.[10]

RF yöntemi bireysel oluşturulan ağaçların sonuçlarını birleştiren bir topluluk yöntemdir. Periodontoloji veri seti analiz edilirken en düşük hata oranını verecek ağaç sayısına ulaşmak için 300 ila 2000 ağaçtan oluşan ormanlar oluşturulmuş ve hata oranları hesaplanmıştır. Hata oranları %3.33 ile %4.6 arasında değişmekte olup belirgin bir trend gözlemlenmemiştir. En düşük hata oranı 500, 800, 1300 ve 2000 ağaç ile oluşturulan ormanlarda elde edilmiştir. RF yöntemiyle sınıflama yapıldığında en az 500 ağaç oluşturulmalı ancak daha fazla sayıda ağaç oluşturularak hata oranının azalıp azalmayacağı gözlemlenmelidir.

Bu çalışmada optimum sonuca ulaşmak için seçilecek değişken sayısının kaç olması gerektiği araştırılmıştır. Değişken sayısının hata oranına etkisini belirlemek amacıyla ormanı oluşturan ağaç sayıları 500 ve 1000 olarak sabit tutularak farklı sayılarda değişkenle orman oluşturulmuştur. Ağaç sayısı 500 iken değişken sayısı 3, 7, 21, 28, ve 35 olarak belirlendiğinde ormanın hata oranı %3.33 iken değişken sayısı 14 iken ormanın hata oranı %4.6'dır. Ağaç sayısı 1000 seçildiğinde de değişken sayısı arttıkça ormanın hata oranında düşme trendi gözlenmemiştir. Ağaç yapısında kullanılacak değişken sayısının modelin genel hata oranını çok fazla etkilemediği görülmüştür.

RF yönteminde en önemli adımlardan biriside değişkenlerin önem derecesinin hesaplanmasıdır. Bu çalışmada veri setindeki 43 değişkenden en önemli ilk 10 değişken tespit edilmiş ve model tespit edilen 10 değişken ile yeniden kurulmuştur. En önemli 10 değişkenin modele girmesi ile elde edilen ormanın hata oranı standart yöntemle hesaplanan önem derecesinde %2.9, gini yöntemiyle hesaplanan önem derecesine göre ise %5.8 olarak bulunmuştur. Tüm değişkenlerin modele girdiği

durumda ise hata oranı %3.33'tür. Standart yöntemle önem derecesi hesaplandığında en önemli değişkenlerin kullanılması ile elde edilen sonuç, tüm değişkenlerin kullanıldığı sonuçtan biraz daha iyidir. Özellikle binlerce değişkenin olduğu veri setleri için değişken önem derecesinin belirlenmesi çok yararlıdır. Örneğin genetik çalışmalarında, hastalıklara neden olan en önemli genlerin tespit edilmesinde bu özelliği nedeniyle RF yönteminin sıklıkla kullanıldığı gözlenmektedir.

Sınıflama modellerinde, model kurulduktan sonra test edilebilmesi için, öğrenme veri setinden bağımsız test veri seti olmalıdır. Bu çalışmada kurulan modeli test etmek için ayrı bir test veri seti yoktu. RF yönteminde model iç hata oranı (OOB hata oranı) hesaplanarak, modelin genel hata oranı tahmin edilmiş olur. Bu çalışmada, RF yönteminin OOB hata oranı tahmini olan %3.33 ormanın genel hata oranı olarak kabul edildi. Modelin doğru sınıflama oranı ise %95.4 olarak bulunmuştur. Bu değer sınıflama için oldukça başarılı bir sonuçtur. Bu durumda benzer şikâyetlerle gelen yeni kişilerden, bu çalışmada sınıflama başarısı önemli bulunan 10 değişkene ait ölçümler alındığı takdirde o kişilerin gruplarının başarılı bir şekilde belirlenebileceği söylenebilir.

RF yöntemi yardımıyla aşırı değer olan denekler tespit edilebilmektedir. Aşırı değerler, proximity matriste sınıfındaki diğer örneklerle proximity değeri en az olan örneklerdir. Buradan, sınıflandığı gruptaki deneklerle aynı özelikleri taşımadığı sonucu çıkarılabilir. Aşırı değerler RF yöntemiyle tespit edilerek, bu değerle ait ölçümler klinik olarak incelenebilir. Yapılan inceleme sonucunda aşırı değerlerin veri setinden çıkartılmasına veya kalmasına karar verilebilir. Periodontoloji veri setindeki 67., 82. ve 69. denekler aşırı değer olarak bulunmuştur.

SONUÇ VE ÖNERİLER

Veri madenciliğinde en yaygın olarak kullanılan yöntemler ağaç tabanlı yöntemlerdir. Ağaç tabanlı yöntemler, diğer yöntemlere göre daha kolay kurulan, daha kolay yorumlanabilen ve oldukça başarılı sonuçlar veren yöntemlerdir. Son yıllarda yapılan çalışmalar birden çok sınıflayıcının bir araya gelerek ortaya koyduğu sonucun, yani bir sınıflayıcı komitenin sonucunun her zaman tek bir sınıflayıcının ortaya koyduğu sonuçtan daha başarılı olduğunu göstermiştir. Random Forests yöntemi, bir topluluk yöntem olmasına rağmen, topluluk yöntemlerden farklı olarak modele ayrı bir katman olan rastgeleliği de katmıştır. Bu rastgelelik sayesinde sınıflandırıcının daha az sapmasız olması sağlanmıştır.

Tıpta tanı koyma amaçlı olarak veri madenciliği yöntemlerinden karar ağaçları yöntemleri kullanılmasına rağmen, çok yaygın değildir. Tanı koyma çalışmalarında topluluk yöntemler uygulanırsa daha az hata ile tahmin yapılabilir. RF yöntemi, veri setindeki değişken sayısı ve örnek sayısı ne kadar çok olursa olsun sonuçları, makul sayılan bir sürede verebilmektedir. Tanı koyma çalışmalarında hastalığın olup olmadığını öğrenmek için çok sayıda tetkik yapılmaktadır. RF yöntemi, değişkenlerin önem derecesini hesaplayarak, tanı koymada hangi değişkenlerin en önemli rol oynadıklarını tespit ederek, daha az sayıda tetkik yapılmasını önerebilir. RF yöntemi özelikle, çok sayıda değişkenin olduğu mikroarray çalışmalarında ve DNA veri seti gibi binlerce gen arasından önemli olanları tespit etmek için yapılan çalışmalarda kullanılabilir.[11]

Bu çalışmada kullanılan veri setinde RF yöntemiyle %95.4 oranında başarılı bir sınıflama yapılmıştır. Oluşturulan karar ormanının genel hata oranı ise %3.33 olarak bulunmuştur. Bu çalışmada ayrıca, Bagging ve tek bir ağaçla sınıflama yapılan CART yöntemi de aynı veri seti ile modellenmiş ve RF yöntemiyle karşılaştırılmıştır. Bagging yönteminde modelin genel hata oranı %5.4 bulunurken CART yönteminde %8.75 olarak bulunmuştur. Periodontoloji veri seti için RF yöntemi, bu iki yöntemden daha iyi sonuç vermiştir.

Teşekkür

Bu çalışmada kullanılan veriler Gazi Üniversitesi Diş Hekimliği Fakültesi Periodontoloji bölümünden Öğr.Gör. Dr. Ahu Uraz tarafından sağlanmıştır. Kendisine gösterdiği destekten dolayı teşekkür ederiz.

KAYNAKLAR

1. Akpınar H. [Knowledge discovery from databases and data mining]. İÜ İşletme Fakültesi Dergisi 2000;29(1):1.
2. Silahtaroğlu G. [Data mining models]. Kavram ve Algoritmalarıyla Temel Veri Madenciliği. 1. Baskı. İstanbul: Papatya Yayıncılık; 2008. s.29-30.
3. Breiman L. Random forests. Machine Learning 2001;45(1):5-32.
4. Lempitsky V, Verhoek M, Noble JA, Blake A. Random forest classification for automatic delineation of myocardium in real-time 3d echocardiography. In: Ayache N, Delingette H, Sermesant M, eds. Functional Imaging and Modeling of the Heart. 1st ed. Berlin: Springer; 2009. p.447-56.
5. Clarke B, Fokoue E, Zhang H. Tree-based classifiers. Principles and Theory for Data Mining and Machine Learning. 1st ed. London New York: Springer Dordrecht Heidelberg; 2009. p.262.
6. Wu BA, Fishman D, Mcmurray W, Mor G, Stone K, Ward D, et al. Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. Bioinformatics 2003;19(13):1636-43.
7. Yonhee K, Ho K. Application of random forests to association studies using mitochondrial single nucleotide polymorphisms. Genomics&Informatic 2007;5(4):168-73.
8. Statnikov A, Wang LF, Aliferis C. A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. BMC Bioinformatics 2008;9(319):1-10.
9. Ruiz-Gazen A, Villa N. Storms prediction: logistic regression vs random forest for unbalanced data. Case Studies in Business, Industry and Government. Statistics (CSBIGS) 2007;1(2):91-101.
10. Amasyalı MF, Ersoy O. Cline: a new decision-tree family. IEEE Transactions On Neural Networks 2008;19(2):356-63.
11. Archer KJ, Kimes RV. Empirical characterization of random forest variable importance measures. Computational Statistics&Data Analysis 2008;52(4):2249-60.