TÜRKİYE CUMHURİYETİ
ANKARA ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ

VERİ MADENCİLİĞİNE GENEL BAKIŞ VE RANDOM FORESTS YÖNTEMİNİN İNCELENMESİ: SAĞLIK ALANINDA BİR UYGULAMA

Muhammet AKMAN

BİYOİSTATİSTİK ANABİLİM DALI
YÜKSEK LİSANS TEZİ

DANIŞMAN
Doç.Dr. Yasemin GENÇ

II.DANIŞMAN
Doç.Dr. Handan ANKARALI

2010 – ANKARA

Ankara Üniversitesi Sağlık Bilimleri Enstitüsü

Biyoistatistik Anabilim Dalı Yüksek Lisans Programı

çerçevesinde yürütülmüş olan "Veri Madenciliği Yöntemlerine Genel Bakış ve Random Forests Yönteminin İncelenmesi: Sağlık Alanında bir Uygulama" başlıklı Muhammet AKMAN'a ait bu çalışma, aşağıdaki jüri tarafından Yüksek Lisans Tezi olarak kabul edilmiştir.

Tez Savunma Tarihi: 11 / 02 /2010

Prof.Dr. Ersöz TÜCCAR
Ankara Üniversitesi Tıp Fakültesi
Biyoistatistik Anabilim Dalı
Jüri Başkanı

Prof.Dr.Ergun KARAAĞAOĞLU
Hacettepe Üniversitesi Tıp Fakültesi
Biyoistatistik Anabilim Dalı
Üye

Doç.Dr. Atilla Halil ELHAN
Ankara Üniversitesi Tıp Fakültesi
Biyoistatistik Anabilim Dalı
Üye

Doç.Dr. Yasemin GENÇ
Ankara Üniversitesi Tıp Fakültesi
Biyoistatistik Anabilim Dalı
Tez Danışmanı

Yard.Doç.Dr. S.Kenan KÖSE
Ankara Üniversitesi Tıp Fakültesi
Biyoistatistik Anabilim Dalı
Üye

İÇİNDEKİLER

ÖNSÖZ

Veri madenciliği, son yıllarda oldukça önemli bir konu haline gelmiştir ve hemen hemen her alanda uygulama sahası bulmuştur. Veri madenciliği ülkemizde sağlık alanında tanı koyma çalışmalarında çok yaygın kullanılmamaktadır. Bu tez çalışmasında özellikle tahminleme başarısı oldukça yüksek olan ağaç tabanlı veri madenciliği yöntemi Random Forests kullanılarak periodontoloji alanında elde edilen veri setindeki bireylere tanı koyma çalışması yapılmıştır. Bu çalışma ile sağlık alanında veri madenciliği yöntemlerinin kullanılmasının faydalı olacağı gösterilmeye çalışılmıştır.

Yüksek lisans tez konusunu seçerken ve daha sonrasında önemli katkıları olan danışman hocam Doç. Dr. Yasemin Genç ve Düzce Üniversitesi Tıp Fakültesi Biyoistatistik Anabilim Dalından eş danışmanım Doç. Dr. Handan Ankaralı'ya teşekkür ederim.

Tezin uygulama bölümünde kullanılan veriler Gazi Üniversitesi Diş Hekimliği Fakültesi Periodontoloji bölümünden Öğr. Gör.Dr. Ahu Uraz tarafından sağlanmıştır. Tezin tamamlanmasında gösterdiği destekten dolayı kendisine teşekkürü borç bilirim.

Ayrıca bu tezi yazmamda büyük katkısı olan eşim Derya AKMAN ve manevi destekleriyle katkıda bulunan çocuklarım Ahmet Ethem ve Muammer Taha'ya da teşekkür ediyorum.

SİMGELER VE KISALTMALAR

VTBK	**V**eri **T**abanlarından **B**ilgi **K**eşfi
SQL	**S**tructured **Q**uery **L**anguage(Veritabanı Sorgulama Dili)
CRISP-DM	The **Cr**oss-**I**ndustry **S**tandard **P**rocess for **D**ata **M**ining
SEMMA	SAS Institute Veri Madenciliği Süreci(**S**ample, **E**xplore, **M**odify, **M**odel, **A**ssess)
CART	**C**lassification **A**nd **R**egression **T**ree
RF	**R**andom **F**orests
OOB	**O**ut-**O**f-**B**ag(RF yönteminde iç test için ayrılan veri)
CVA	Cross Validation Average(Çapraz Doğrulama Hata Tahmini)
DVM	**D**estek**V**ektör **M**akinaları

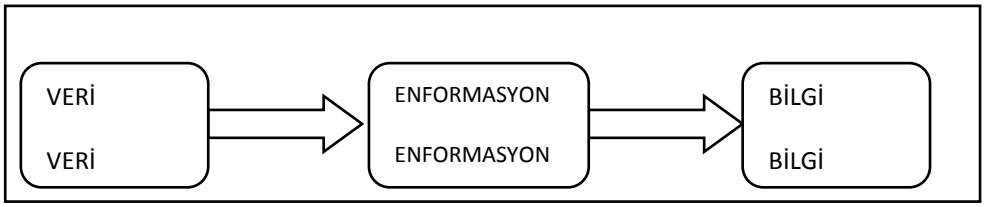
ŞEKİLLER

ÇİZELGELER

1. GİRİŞ

1.1 Temel Kavramlar ve Çalışmanın Amacı

Veri madenciliği kavramını anlamak için önce “veri” ve “madencilik” kavramlarını ele almak gerekir. Veri, istatistik terimleri sözlüğünde “Deneyler ya da gözlemler sonunda elde edilen nicel ya da nitel değerler” olarak tanımlanmaktadır. Türk Dil Kurumuna göre ise “Bir araştırmanın, bir tartışmanın, bir muhakemenin temeli olan ana öğe, done” dir. Genel anlamda ise veri, "kayıt altına alınan işlenmemiş ham olarak duran herhangi bir olay, sayılar veya dokümanlar" olarak tanımlanır. Günümüzde, tüm kuruluşlar çok çeşitli ve günden güne artan verileri çeşitli formatlarda ve farklı veritabanlarında depolamaktadırlar. Verilerden yola çıkarak ilgili alanda enformasyon(malumat) sahibi olunur. Enformasyon, verilerin düzenlenmiş, ilişkilendirilmiş ve anlamlandırılmış halidir. Enformasyondan bilgiye geçiş; geçmiş dönem analizlerinden yola çıkarak gelecek dönemle ilgili yeni trendlerin kavranılması ve yeni trendlere göre davranış ve yönelimlerin belirlenmesi şeklindedir. Madencilik ise, en yaygın bilinen anlamı ile "yer altında saklı olan değerli madenlerin, yeryüzüne çıkartılması için çeşitli işlemlerin yapılması" olarak tanımlanabilir. Veriden bilgiye geçiş Şekil 1.1.’ de gösterilmiştir.



Şekil 1.1. Bilgiye ulaşma süreci

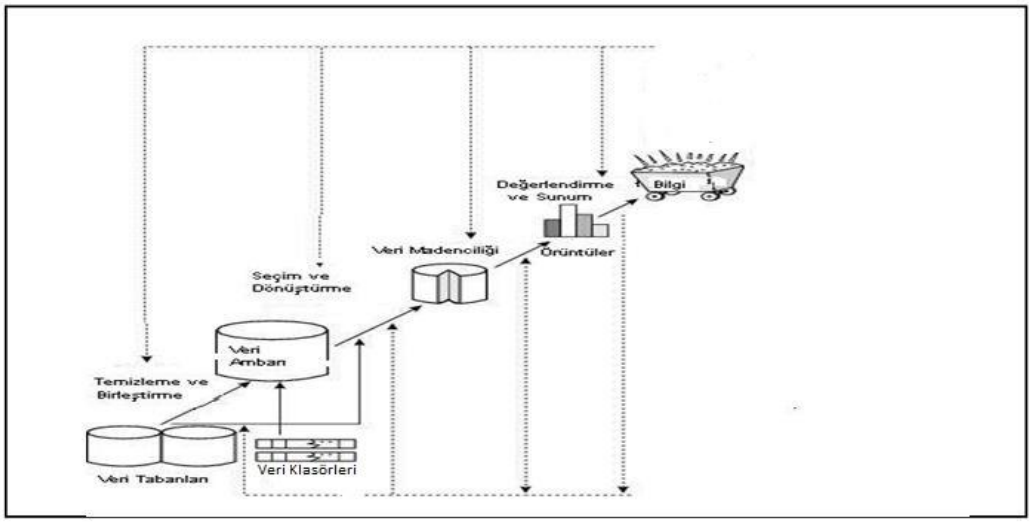
Bu tez çalışmasında, veri madenciliğinde kullanılan temel kavramların tanımlanması, veri madenciliği yöntemlerinin genel özelliklerinin açıklanması, ülkemizde son yıllarda çalışılmaya başlanan ağaç tabanlı Random Forests(RF) yönteminin özelliklerinin anlatılması ve bu yöntemin sağlık alanından elde edilen bir veriye uygulanması, elde edilen sonuçların tartışılarak diğer alternatif veri madenciliği yöntemlerinden farklarının açıklanması amaçlanmıştır.

1.1.1. Veri Madenciliği ve Veri Tabanlarından Bilgi Keşfi

Günümüzde bilgi, kazanılan en önemli değerlerdendir. Bilginin önemi ona olan ihtiyacı arttırmıştır. Bunun sonucu olarak bilişim sistemlerindeki hızlı gelişme ve otomatik veri depolama araçlarındaki teknolojik gelişmeler yoluyla artık yaptığımız her işlem sayısal ortamda kayıt altına alınmaya başlanmıştır. Organizasyonlar, firmalar tüm mali ve operasyon bilgilerini veri tabanlarında, veri ambarlarında depolamaktadırlar. Veritabanlarında depolanan veriler arasındaki ilişkiler, saklı kalmış bilgiler çıkarılmayı beklemektedir. Depolanan veri yığınları arasında saklı kalmış bilgilerin nasıl açığa çıkarılabileceği üzerine yapılan çalışmalar sonucu Veri Tabanlarında Bilgi Keşfi (VTBK) kavramı ortaya çıkmıştır.

VTBK süreci içerisinde, modelin kurulması ve değerlendirilmesi aşamalarından meydana gelen Veri Madenciliği (Data Mining) en önemli kesimi oluşturmaktadır (Şekil 1.2.). Bu önem, bir çok araştırmacı tarafından VTBK ile veri madenciliği terimlerinin eş anlamlı olarak da kullanılmasına neden olmaktadır (Akpınar, H., 2000).

Veri madenciliği, büyük boyutlardaki veriyi değişik perspektiflerden ele alarak analiz etme ve veriyi işe yarar bilgiye dönüştürme işlevleri olarak tanımlanabilir. Burada önemli olan konu veri madenciliği sonucu ortaya çıkarılması hedeflenen bilginin önceden bilinmiyor veya kestirilemiyor olmasıdır. Çünkü bilinen bir gerçeğin veri madenciliğiyle ispat edilmesi veri madenciliğinin anlamına ters düşmektedir. Bunun yanında veri madenciliği oldukça maliyetli olduğundan bu amaçla veri madenciliği yapmak pek de ekonomik olmayacaktır. "Ancak, tıp gibi bazı alanlarda veri madenciliği sonucu ortaya çıkan sonuç üzerinde şüphe varsa, son kararın bir uzman tarafından verilmesi gerekebilir" (Alpaydın, E., 2004).



Şekil 1.2. Veri Madenciliğinin VTBK'ndeki yeri (Han,J.2004)

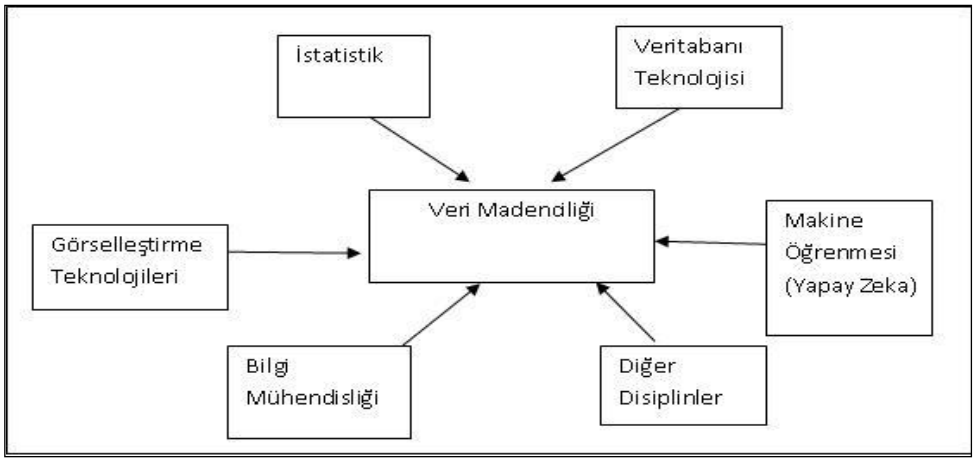
Veri madenciliği veri tabanlarında saklı kalmış bilginin ortaya çıkartılmasının yanında geleceğe dönük kararlar alınmasında da etkin olarak kullanılır.İleriye dönük doğru kararlar alınması için, doğrusal regresyon, lojistik regresyon, zaman serileri analizi ve bayesian yaklaşım gibi istatistiksel yöntemler kullanılarak tahminler yapılmaktadır (Silahtaroğlu, G. 2008).

1.1.2. Veri Madenciliği ile Diğer Disiplinler Arasındaki İlişkiler

Makine öğrenmesi, istatistik ve veri madenciliği birbirleriyle yakından ilişkilidir (Zhou, Z., 2003). Bu üç disiplin veri içindeki bağıntıları ve örüntüleri bulmayı amaçlar. Makine öğrenmesi yöntemleri, veri madenciliği algoritmalarında kullanılan yöntemlerin temelini oluşturur. Makine öğrenmesi ve yapay zeka uygulamalarında kullanılan karar ağaçları, kural çıkartımı, sınıflama ve kümeleme gibi pek çok veri madenciliği algoritmasında da kullanılmaktadır. Ancak makine öğrenmesinde ve istatistiksel yöntemlerde kullanılan örnekleme genişliği, veri madenciliği algoritmalarında kullanılan örneklem boyutuna göre çok daha küçüktür. Veri madenciliği, makine öğrenmesi yöntemlerine göre gürültülü, eksik ve boş değerleri işlemede daha başarılıdır. Veri madenciliği, istatistiksel yöntemleri veri setindeki değişkenler arasındaki bağımlılığın derecesini ölçmek, veriyi tanımlamak, verinin

özetini çıkarmak ve veri setindeki eksik değerlerin tahminlerini yapmak gibi konularda kullanmaktadır (Şekil1.3).

Veri madenciliği ve veritabanı teknolojisi arasında da önemli bir ilişki vardır. Veri madenciliği yöntemlerinin uygulanacağı veriler genellikle büyük boyutlu veritabanlarında tutulmaktadır. Veri tabanları sorgu dili SQL(Structered Query Language) ise veri tabanlarındaki var olan ve bilinen ilişkileri ortaya koymak için kullanılmaktadır. Veri madenciliği ise, veritabanlarında bulunan veriler arasındaki bilinmeyen ilişkileri ortaya çıkarmaktadır.



Şekil 1.3. Veri madenciliği ile diğer disiplinler arasındaki ilişki

1.1.3. Denetimli (Supervised) Öğrenme

Belirli bir amaca ve sonuca yönelik olarak yapılan veri madenciliği yöntemlerine denetimli yöntemler denilebilir. Örnekten öğrenme olarak da isimlendirilen denetimli öğrenimde, ilgili sınıflar önceden belirlenen bir kritere göre ayrılarak, her sınıf için çeşitli örnekler verilir. Sistemin amacı verilen örneklerden hareket ederek her bir sınıfa ilişkin özelliklerin bulunması ve bu özelliklerin kural cümleleri ile ifade edilmesidir. Öğrenme süreci tamamlandığında, tanımlanan kural cümleleri verilen yeni örneklere uygulanır ve yeni örneklerin hangi sınıfa ait olduğu kurulan model tarafından belirlenir.

Denetimli öğrenimde seçilen algoritmaya uygun olarak ilgili veriler hazırlandıktan sonra, ilk aşamada verinin bir kısmı modelin öğrenimi (Learning Dataset), diğer kısmı ise modelin geçerliliğinin test edilmesi (Testing Dataset) için ayrılır. Öğrenim veri seti kullanılarak kurulduktan sonra, test veri seti ile modelin doğruluk derecesi belirlenir. Test sonucunda doğruluk derecesine göre üç durum ortaya çıkabilir. Modelin kabul edilmesi, modelin yeniden ele alınarak bir iyileştirme yapılması veya modelin tamamen reddedilmesi.

1.1.4 Denetimsiz (Unsupervised) Öğrenme

Ulaşılmak istenen sonuç için bir tanımlama yapılmamışsa veya bir belirsizlik varsa denetimsiz öğrenmeden bahsedilebilir. Denetimsiz öğrenme daha çok, veriyi anlamaya, tanımaya ve keşfetmeye yönelik olarak kullanılmaktadır.

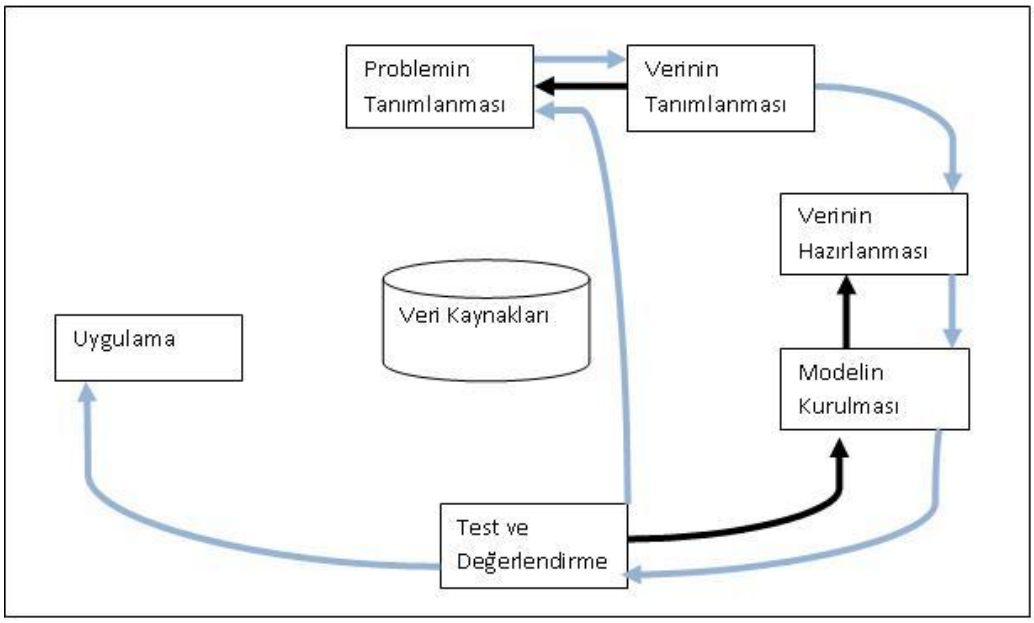
Denetimsiz öğrenmede, ilgili örneklerin gözlenmesi ve bu örneklerin özellikleri arasındaki benzerliklerden hareket ederek sınıfların tanımlanması amaçlanmaktadır

1.2. Veri Tabanlarında Bilgi Keşfi Süreçleri

Sistematik bir veri madenciliği analizi için genellikle standart süreçler kullanılır. Bu süreçlerden en çok kullanılanı CRISP-DM (The Cross-Industry Standard Process for Data Mining) sürecidir ve Şekil1.4'te gösterilmiştir. CRISP-DM süreci 1996 yılında Daimler Chrysler, SPSS ve NCR firmalarından analistler tarafından geliştirilmiştir (Larose,T.D, 2004). CRISP-DM veri madenciliği yaşam döngüsü içerisinde birbirini takip eden altı adımdan oluşmaktadır. Bu adımlar aşağıda sıralanmıştır.

- Problemin Tanımlanması
- Verilerin Tanımlanması

- Verilerin Hazırlanması
- Modelleme
- Değerlendirme
- Uygulama-Kullanma



Şekil 1.4. CRISP-DM Veri Madenciliği Süreci

1.2.1 Problemin Tanımlanması (Business Understanding)

Bu aşamada amacın belirlenmesi ve amacın gerçekleştirilmesi için bir proje planının geliştirilmesi safhaları yer almaktadır. Ayrıca tahminlerdeki fayda-maliyet analizi de bu aşamada yer alır.

1.2.2 Verinin Tanımlanması (Data Understanding)

Bu aşama, uygun verinin nasıl ve hangi kaynaktan toplanacağının belirlenmesi, verinin açıklanması ve veri kalitesinin doğrulanması aşamasıdır. Veriye ait özet istatistiklerin oluşturulması ve gözden geçirilmesi de bu aşamanın son çalışmasıdır.

1.2.3. Verinin Hazırlanması (Data Preparation)

Veri kaynakları belirlendikten sonra, verinin seçilmesi, birleştirilmesi, temizlenmesi, dönüştürülmesi ve istenilen biçime sokulması gerekmektedir. Verinin temizlenmesi, dönüştürülmesi bu aşamada yapılmaktadır. Bu aşama, zaman ve enerji olarak tüm sürecin %60-%95'ini oluşturmaktadır (De Veaux,R.,2009). Modelin kurulması aşamasında ortaya çıkacak sorunlar nedeniyle bu aşamaya sıklıkla geri dönülebilir ve veri yeniden ele alınabilir. Modelin kurulması aşamasında fazla zaman harcamamak için bu aşamada titiz bir çalışma yürütülmelidir.

Verilerin hazırlanması aşaması kendi içerisinde toplama, birleştirme ve temizleme, seçme ve dönüştürme adımlarından meydana gelmektedir.

1.2.3.1. Toplama

Belirlenen amaç için gerekli olduğu düşünülen verilerin ve bu verilerin toplanacağı veri kaynaklarının belirlenmesi adımıdır. Verilerin toplanmasında kuruluşun kendi veri kaynaklarının dışında, konuyla ilgili başka kurumların veri kaynaklarından da faydalanılabilir. Veri madenciliğinde kullanılacak verilerin farklı kaynaklardan toplanması, veri uyumsuzluğu oluşturabilir. Bu uyumsuzluk, kullanılmak istenen verilerin farklı zamanlara ait olmaları, kodlama farklılıkları ve farklı ölçü birimlerine sahip olmalarından kaynaklanabilir. Ayrıca verilerin nerede ve hangi koşullar altında toplandığı da önemlidir.

1.2.3.2. Birleştirme ve Temizleme

Bu adımda toplanan verilerde bulunan farklılıklar giderilmeye çalışılır. Hatalı veya analizin yanlış yönlenmesine sebep olabilecek verilerin temizlenmesine çalışılır. Yanlış veri girişinden kaynaklanan verideki gürültü veya aynı verinin farklı bir şekilde girilmiş olmasından kaynaklanan verideki tekrarlar giderilir. Bunun yanında

kayıp verilerin tespit edilip istatistiksel yöntemlerle tahmin edilmesi, elle yeniden girilmesi veya ilgili kaydın iptal edilmesi gibi işlemler bu aşamada gerçekleştirilir.

1.2.3.3. Seçim

Bu adım, bağımlı ve bağımsız değişkenlerin ve modelin eğitiminde kullanılacak veri kümesinin seçilmesi aşamasıdır. Anlamlı olmayan ve diğer değişkenlerin modeldeki ağırlığını azaltacak değişkenlerin modele girmemesi gerekmektedir. Bazı veri madenciliği algoritmaları değişkenlerin önem derecesini otomatik olarak tespit ederek modelden çıkarılması gereken değişkenleri çıkarmaktadırlar. Ancak yine de modele katkısı olmayan değişkenlerin analist tarafından tespit edilmesi daha anlamlı olacaktır. Verilerin grafiksel olarak görselleştirilmeleri ve bunun sonucu ortaya çıkan ilişkiler bağımsız değişkenlerin seçilmesinde önemli bir rol oynayabilir. Modelde kullanılacak verilerin çok büyük olması, modelin kurulmasını zorlaştırabilir. Bu durumda orijinal verinin yapısını bozmayacak şekilde örnekleme yapılabilir ve bu örneklem üzerinden çok sayıda model oluşturulup en güçlü model seçilebilir.

1.2.3.4. Dönüştürme

Modeli daha güçlü yapmak ve etkinliğini artırmak için, verideki bazı değişkenleri sürekliden sayısal bir aralığa veya kategorik bir veriyi sayısal bir aralığa dönüştürmek gerekebilir.

1.2.4. Modelleme

Bu aşamada, ham veriden bilgiye ulaşmak için ileri çözümleme yöntemleri kullanılmaktadır. Uygun modelleme tekniğinin seçilmesi, test tasarımı (öğrenme veri seti ve test veri seti bu aşamada belirlenmelidir), model geliştirme ve tahminsel işlemleri içermektedir. Veri madenciliği, farklı problemler için farklı yöntemler içermektedir. Bazı yöntemler, veri tipi için uygun olmayabilir bu nedenle veri hazırlama aşamasına geri dönülebilir.

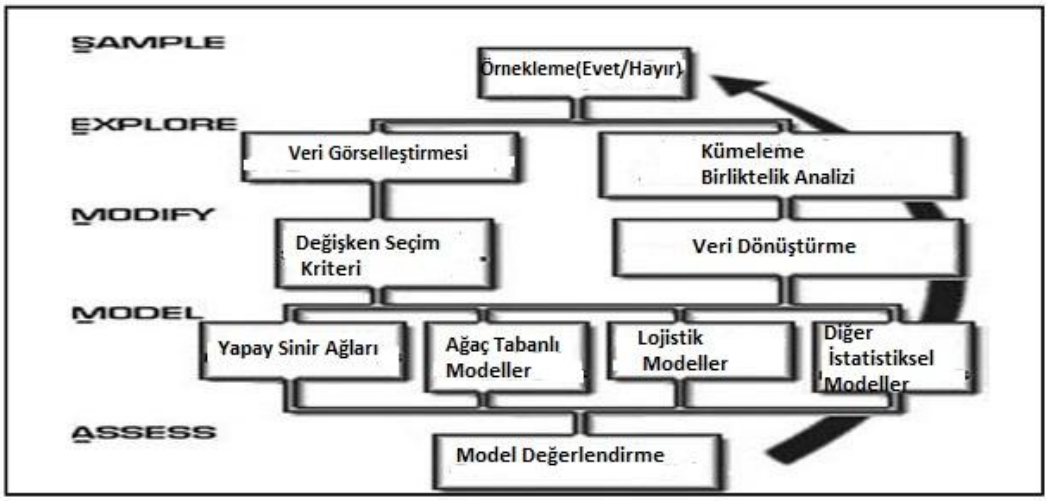
1.2.5. Değerlendirme

Bu aşama sonuçların değerlendirilmesini içermektedir. Ayrıca, veri madenciliği aşamalarının gözden geçirilmesi ve sonraki adımlara karar vermeyi kapsamaktadır. Bu aşama sonucunda yapılan çalışmaların önemli bir sonuç ortaya koyup koymadığı değerlendirilmekte ve sonuçların kullanılıp kullanılmayacağına karar verilmektedir.

1.2.6. Uygulama

Bu safha, ortaya çıkartılan bilgilerin uygulanabilmesine yönelik bir planlama yapma, gözden geçirme ve bakım faaliyetlerini içerir. Araştırma raporu yazılarak proje gözden geçirilir ve süreç tamamlanmış olur.

CRISP-DM standardı dışında uygulanan veri madenciliği süreçleri de vardır. Bunlardan en çok kullanılanı ise SAS Institute tarafından ortaya konulan ve SEMMA (**S**ample, **E**xplore, **M**odify, **M**odel, **A**ssess) adı verilen veri madenciliği sürecidir. Bu süreç Şekil 1.5' te gösterilmiştir.



Şekil 1.5.SEMMA veri madenciliği süreci (SAS Institute orjinal gösterimi)

1.3 Veri Madenciliği Uygulama Alanları

Günümüzde veri madenciliği pek çok alanda kullanılmaktadır. Bu alanlardan bankacılık, pazarlama, sigortacılık ve sağlık gibi sektörler başı çekmektedirler. Genel olarak veri madenciliğinin sektörlere göre hangi amaçla kullanıldığı aşağıda sayılmıştır.

Sağlık ve Farmakoloji: İlaç geliştirme, hastalıkların teşhisi, tedavi sürecinin belirlenmesi alanlarında kullanılmaktadır.

Biyoloji: DNA sıra analizi ile hastalıklara neden olan gen sıralamasını belirlemek amacıyla kullanılmaktadır (Microarray veri analizi).

Pazarlama Yönetimi: Müşterilerin demografik özellikleri arasındaki bağlantıların kurulmasında, satın alma eğilimlerinin belirlenmesinde, pazarlama kampanyalarının planlanmasında, mevcut müşterilerin elde tutulması ve yeni müşterilerin kazanılması için geliştirilecek pazarlama stratejilerinin oluşturulmasında, pazar sepeti ve çapraz satış analizlerinde, müşteri ilişkileri yönetiminde ve satış tahminlerinde kullanılmaktadır.

Bilişim ve Mühendislik: İnternet işlemleri dolandırıcılığının tespit edilmesinde, bilgisayar sistemlerine ve bilgisayar ağlarına yetkisiz girilmesinin tespit edilmesinde, parmak izi ve yüz şeklinden kimlik tespitinde ve yapay zeka uygulamalarında kullanılmaktadır.

Bankacılık: Farklı finansal göstergeler arasındaki gizli korelasyonların bulunmasında, kredi kartı dolandırıcılıklarının tespitinde, kredi taleplerinin değerlendirilmesinde, usulsüzlük tespiti, risk analizleri ve risk yönetiminde kullanılmaktadır.

Meteoroloji ve Atmosfer Bilimleri: Bölgesel iklim ve yağış haritaları oluşturma, hava tahminleri yapma amacıyla kullanılmaktadır.

Sigortacılık: Yeni poliçe talep edecek müşterilerin tahmin edilmesinde, sigorta dolandırıcılıklarının tespitinde ve riskli müşteri tipinin belirlenmesinde kullanılmaktadır.

Perakendecilik: Satış noktası veri analizleri, alış-veriş sepeti analizleri, tedarik ve mağaza yerleşim optimizasyonunda kullanılmaktadır.

Borsa: Hisse senedi fiyat tahmini, genel piyasa analizleri, alım-satım stratejilerinin optimizasyonunda kullanılmaktadır.

Endüstri: Kalite kontrol analizlerinde, lojistik ve üretim süreçlerinin optimizasyonunda kullanılmaktadır.

Telekomünikasyon: Kalite ve iyileştirme analizlerinde, hatların yoğunluk tahminlerinde ve telefon dolandırıcılığının tespit edilmesinde kullanılmaktadır.

Veri madenciliği uygulama alanları burada sayılmayan, eğitim/öğretim, güvenlik gibi daha başka bir çok alanda da yapılmaktadır. www.kdnuggets.com sitesinin veri madenciliğinin uygulama alanlarını ve oranlarını belirlemek için, web sitesi üzerinden analistlere yönelik olarak yaptığı anket sonuçları Şekil 1.6'da yer almaktadır.

2008 yılında veri madenciliğini uyguladığınız endüstri/alanlar nelerdir?	
CRM/ Müşteri Analizleri (41)	% 38.3
Bankacılık (34)	% 31.8
Dolandırıcılık Tespiti (21)	% 19.6
Finans (18)	% 16.8
Doğrudan Pazarlama (15)	% 14.0
Diğer (14)	% 13.1
Yatırırım/Borsa kararları (14)	% 13.1
Kredi kartı Skorlama (14)	% 13.1
Telekomünikasyon13)	% 12.1
Perekandecilik(13)	% 12.1
Reklam (13)	% 12.1
Biyoteknoloji/Genetik (12)	% 11.2
Bilim (11)	% 10.3
Sigortacılık (11)	% 10.3
Sağlık (10)	% 9.3
İmalat (9)	% 8.4
E-ticaret (8)	% 7.5
Web Kullanım Madenciliği(8)	% 7.5
Sosyal Politikalar/Anket Analizi(8)	% 7.5
Tıp/Farmakoloji (8)	% 7.5
Güvenlik/Anti-terör (6)	% 5.6
Web içerik madenciliği (6)	% 5.6
Kamu/Askeri uygulamalar (4)	% 3.7
Seyahat (3)	% 2.8
Junk e-posta / Anti-spam tespiti (3)	% 2.8
Eğlence /Müzik (3)	% 2.8
Sosyal Ağlar(2)	% 1.9

Şekil 1.6. Veri madenciliği uygulama alanları (Kaynak:www.kdnuggets.com)

1.4 Veri Madenciliği Yöntemleri

Veri madenciliği yöntemleri genel olarak tanımlayıcı ve tahminleyici modeller olmak üzere iki ana başlık altında incelenebilir. Ancak bazı yöntemler hem tanımlayıcı hem de tahminleyici özelliğe sahip olabilmektedir.

Tahminleyici modellerde, sonuçları bilinen verilerden yola çıkarak bir model kurulması ve kurulan bu modelden sonuçları bilinmeyen verilerin sonuç değerlerinin tahmin edilmesi amaçlanmaktadır (Akpınar, 2000).

Tanımlayıcı modellerde, eldeki verilerden strateji geliştirme ve karar verme süreçlerinde kullanılabilecek bilgiler sağlanmaktadır. Tanımlayıcı modeller daha çok veriler arasındaki gizli kalmış ilişkiyi ortaya çıkarırlar (Silahtaroğlu,2008,s.30).

Tanımlayıcı modeller işlevine göre Kümeleme (Clustering) ve Birliktelik kuralları (Association Rules) olarak alt bölümlere ayrılmaktadır. Tahminleyici modeller ise Sınıflama (Classification) ve Regresyon (Regression) olarak iki alt başlıkta toplanır.

1.4.1 Kümeleme Modelleri

Kümele analizinde amaç, önceden sınıfları/grupları bilinmeyen elemanlardan birbirlerine çok benzeyen ancak özelliklerine göre diğerlerinden farklı olan elemanları bir araya getirmek yani sınıflamak/gruplamaktır.

Veri setini oluşturan örneklerin(elemanların) kaç değişik gruba ayrılacağı ve hangi sınıf/grupta yer alacağı örneklerin birbirine benzerliğine göre belirlenir. Belirlenen her bir sınıf/gruba küme ismi verilir.

Kümeleme modelleri genel olarak; hiyerarşik ve bölümlemeli (Partitioning) yöntemler olmak üzere iki ana başlıkta incelenir.

1.4.2 Birliktelik Kuralları ve Ardışık Zamanlı Örüntüler

Veritabanlarındaki çok büyük boyutlu veriler içerisinde kolaylıkla elde edilemeyen ilişkiler olabilir. Bu tip ilişkilerin elde edilmesi stratejik kararların alınmasına yardımcı olabilir. Ancak, bu ilişkilerin elde edilmesi kolay bir süreç değildir. Bu süreç veri madenciliğinde birliktelik kuralları olarak adlandırılmaktadır. Farklı olayların birbirleri ile ilişkili olduğunun ortaya çıkartılması, kullanılabilecek bilgiye dönüştüğünde oldukça önemli olmaktadır. Literatürde bu tür çalışmalara "pazar sepeti analizi" de denilmektedir. Pazar sepeti analizi ile müşterilerin alışveriş alışkanlıkları ortaya çıkarılır.

Ardışık zaman örüntülerine örnek olarak,"ilk üç taksitinden iki veya daha fazlasını geç ödemiş olan müşteriler %60 olasılıkla kanuni takibe gidiyor" cümlesi verilebilir (Alpaydın, 2000).

"X şirketinin hisse fiyatları ile Y şirketinin hisse fiyatları benzer hareket ediyor" örneğinde ise benzer zaman sıraları görülmektedir. Amaç zaman içindeki iki hareket serisi arasında bağıntı kurmaktır. Bunlar iki malın zaman içindeki satış miktarları olabilir. Örneğin dondurma satışları ile kola satışları arasında pozitif, dondurma satışları ile salep satışları arasında negatif bir bağıntı beklenebilir (Alpaydın, 2000).

1.4.3 Sınıflama ve Regresyon Modelleri

Eldeki verilerden hareket ederek geleceğin tahmin edilmesinde faydalanılan ve veri madenciliği teknikleri içerisinde en fazla kullanıma sahip olan sınıflama ve regresyon modelleridir. Resim ve örüntü tanıma, hastalık teşhisi ve dolandırıcılık tespiti konuları sınıflama tekniklerinin en yaygın kullanıldığı alanlardır.

Sınıflama ve Regresyon Modellerinden en çok kullanılan teknikler aşağıda listelenmiştir.

- Ağaç tabanlı yöntemler
    - Karar Ağaçları (Decision Trees)
    - Topluluk Yöntemler (Ensemble Learning)
        - Random Forests
        - Bagging
        - Boosting
- Yapay Sinir Ağları (Artificial Neural Networks),
- Naive-Bayes,
- Genetik Algoritmalar (Genetic Algorithms),
- K-En Yakın Komşu (K-Nearest Neighbor),
- Bellek Temelli Nedenleme (Memory Based Reasoning),
- Destek Vektör Makinesi (Support Vector Machines),
- Lojistik Regresyon (Logistic Regression),
- Bulanık Mantığa Dayalı Algoritmalar (Fuzzy Algoritms).
- Rough Sets

1.4.3.1 Ağaç Tabanlı Yöntemler

Ağaç tabanlı yöntemlerin temelini oluşturan karar ağaçları modellerinin ilk uygulamaları AID (Automatic Interaction Detector) algoritması ile yapılmıştır ve çeşitli algoritmalar ile sürdürülmüştür. Geliştirilen bu algoritmalar içerisinde CHAID (Chi-Squared Automatic Interaction Detector; G.V. Kass; 1980), CART (Classification and Regression Trees; Breiman, Friedman, Olshen ve Stone; 1984), ID3 (Quinlan; 1986), Exhaustive CHAID (Biggs, de Ville ve Suen; 1991), C4.5 (Quinlan; 1993), MARS (Multivariate Adaptive Regression Splines), QUEST (Quick, Unbiased, Efficient Statistical Tree; Loh ve Shih, 1997), C5.0 (Quinlan), SLIQ (Supervised Learning in Quest; Mehta, Agarwal ve Rissanen), SPRINT (Scalable Parallelizable Induction of Decision Trees; Shafer, Agrawal ve Mehta) başlıcalarıdır.

Burada sayılan algoritmalardan başka çok çeşitli ağaç tabanlı algoritmalar da geliştirilmiştir. Son yıllarda birden çok sınıflandırıcının bir araya getirilmesi ile

oluşan ve topluluk yöntemler ya da komiteler olarak adlandırılan algoritmalar önem kazanmıştır. Bu yöntemlerin, yapıyı tek bir ağaçla özetleyen CART, CHAID gibi yöntemlere göre tahminlerdeki hata payı daha düşük ve buna bağlı olarak kestirim gücü daha yüksektir. Bu anlamda en çok ön plana çıkan algoritmalar boosting, bagging ve Random Forests algoritmalarıdır.

www.kdnugets.com sitesinin en çok kullanılan veri madenciliği teknikleriyle ilgili 2007 yılında, web sitesi üzerinden analistlere yönelik olarak yaptığı anket sonuçları Şekil 1.7'de verilmiştir.

Son 12 ayda sıklıkla kullanıldığınız Veri madenciliği/analitik metotları nelerdir?(2007)	
Karar Ağaçları/Kuralları(127)	% 62.6
Regresyon (104)	%51.2
Kümeleme (102)	%50.2
İstatistik (descriptive) (94)	%46.3
Görselleştirme (66)	32.5
Birliktelik Kuralları (53)	%26.1
Sekans/Zaman Serileri Analizi (35)	%17.2
Sinir ağları (35)	%17.2
Support Vector Machine (32)	%15.8
Bayesian (32)	%15.8
Boosting (30)	%14.8
En yakın Komşuluk (26)	%12.8
Hybrid metotlar (24)	%11.8
Diğerleri (23)	%11.3
Genetik algoritmalar (23)	%11.3
Bagging (22)	%10.8

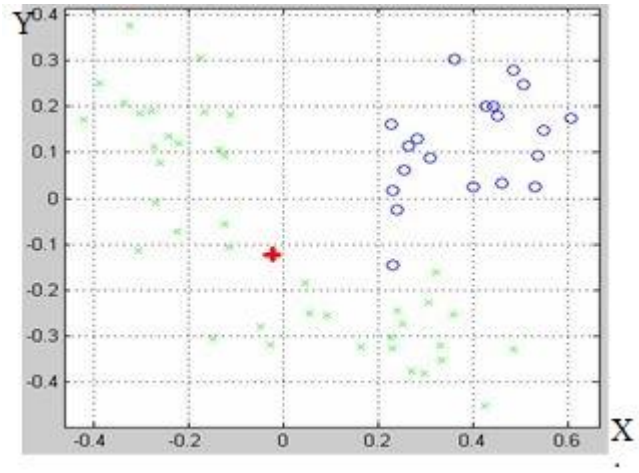
Şekil 1.7. En çok kullanılan veri madenciği tekniklerine ilişkin anket sonuçları

1.5. Veri Madenciliğinde Sınıflama ve Karar Ağaçları

Sınıflama denetimli öğrenme tekniklerindendir. Veri setinde bulunan her nesnenin bir dizi niteliği vardır ve bu niteliklerden biri de sınıf bilgisidir. Sınıflama genel anlamda, geçmişte toplanan verilerin hangi sınıfa ait olduğu bilindiğinde yeni gelen verinin hangi sınıfa ait olduğunu bulma işlemidir ve tahminsel bir modeldir. Hangi

sınıfa ait olduğu bilinen nesneler ile (öğrenme veri seti) bir model oluşturulur. Oluşturulan model öğrenme kümesinde yer almayan nesneler ile (test veri seti) denenerek başarısı ölçülür. Modelin doğru test edilebilmesi için; test veri seti, öğrenme veri setinden bağımsız olmalıdır. Buradaki öğrenme veri seti, tüm verilerin bulunduğu veri setinden, sınıf dağılımını bozmayacak şekilde rastgele kayıtların seçilmesi ile oluşturulmaktadır. Geriye kalan veriler ise test verisi olarak kullanılmaktadır. Öğrenme veri seti, genelde tüm verinin 2/3, test veri seti ise tüm verinin 1/3'ü büyüklüğündedir. Test verisi oluşturulan modelin doğruluk derecesi (accuracy), performans değerlendirme yöntemleriyle analiz edilir. Performans değerlendirme yöntemlerinden olan hata oranı en çok kullanılan performans ölçütüdür. Hata oranı, modele uygulandığında doğru olarak sınıflanmayan örneklerin sayısının, test veri setinde bulunan tüm örneklerin sayısına bölünmesi ile bulunmaktadır. Eğer modelin hata oranı kabul edilebilir bir değerde ise, model gelecek verilerin sınıflandırılmasında kullanılabilir. Aksi durumda ise, yeterli doğruluk sağlanana kadar model yeniden kurulur.

Sınıflamaya örnek olarak Şekil 1.8'de iki tahlil sonucu gösterilebilir. Şekilde tahlil sonuçları X ve Y eksenleriyle, hasta kişiler yuvarlaklarla, sağlam kişiler çarpılarla, hasta olup olmadığı merak edilen kişinin tahlil sonuçları artı ile gösterilmiştir (Amasyalı,M.F.,2008).



Şekil 1.8 Sınıflandırma problemi (Artı örneği hangi sınıftan)

Hangi sınıfta olduğu belirli olmayan nesnelerin(örneklerin) sınıfını belirlemek için sınıflama algoritmaları geliştirilmiştir. Bir kaydın önceden belirlenmiş bir gruba girebilmesi için sınıflama algoritması ile öğrenme verileri kullanılarak hangi sınıfların var olduğu ve bu sınıflara girmek için bir kaydın hangi özelliklere sahip olması gerektiği otomatik olarak keşfedilir. Test verileriyle de bu öğrenmenin testi yapılarak model kurulmuş olur.

1.5.1 Karar Ağaçları

Sınıflama modelleri içerisinde yer alan karar ağaçları yöntemleri tahmin edici ve tanımlayıcı özelliklere sahiptir. Karar ağaçları, yorumlanmalarının kolay olması, veri tabanı sistemleri ile kolayca entegre edilebilmeleri, güvenilirliklerinin daha iyi olması nedenlerinden dolayı sınıflama modelleri içerisinde en yaygın kullanıma sahip olan yöntemlerdir.

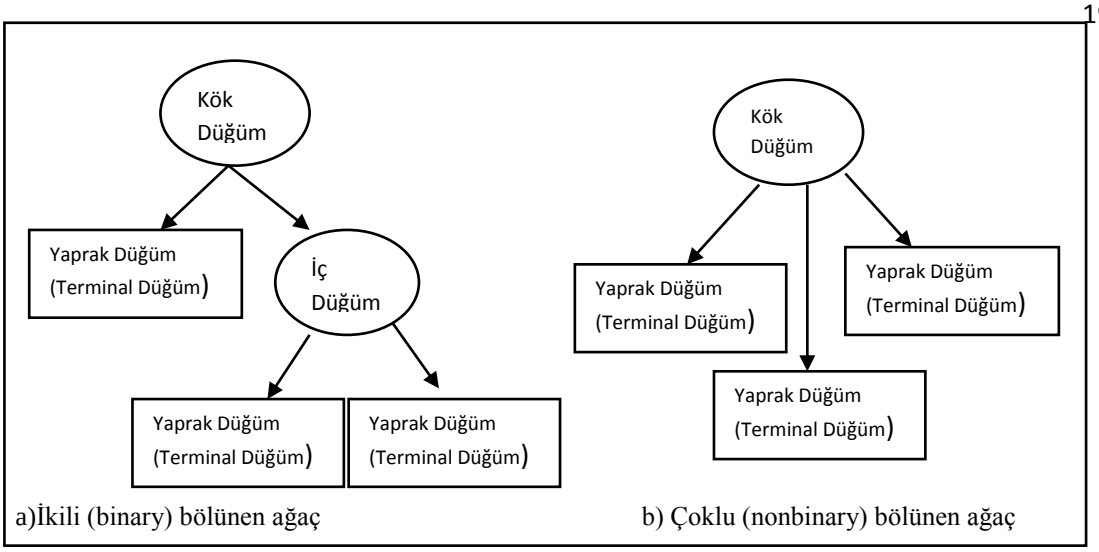
Karar ağaçları sınıflama, karar teorisi, kümeleme ve tahminsel fonksiyonlarda kullanılmaktadır. Öğrenme setindeki verileri var olan düzenlere göre gruplayıp kurallar çıkarmaktadır. Karar ağaçlarını oluşturacak verideki değişkenler (nitelik/özellik) kategorik veya sürekli olabilirler. Eğer değişkenler sürekli ise karar ağaçları genellikle regresyon ağaçları olarak adlandırılırlar. Eğer değişkenler kategorik ise buna sınıflama ağacı denilmektedir. Bu farklılığa rağmen karar ağaçları benzer biçimde kurulmaktadırlar. Karar ağaçları tıp alanında teşhis için, botanikte sınıflama için, felsefede karar teorisi için, ekonomide ise yatırım alternatiflerini belirlemek için sıklıkla kullanılır. Karar ağaçları kurulma biçimlerine göre birbirinden ayrılmaktadırlar. Bazı durumlarda yukarıdan aşağı doğru kurulurken bazı durumlarda soldan sağa doğru kurulabilirler (Omitaomu,O.,2006).

Karar ağaçları geçmiş veriye dayanarak yeni verilerin hangi sınıfa ait olduğuna, kurallar çıkartarak karar vermektedir. Karar ağacı, sorulan sorular ve alınan cevaplar doğrultusunda hareket eder ve sorulan sorulara alınan cevapları

birleştirerek kurallar oluşturur. Oluşan ağaç bir çok "eğer-ise"(if-then)'den oluşan kurallar bütünüdür de diyebiliriz. Soru sormaya verideki hangi değişkenden başlanacağına karar verildiğinde ilgili değişken ağacın kök düğümünü oluşturmuş olur. Kök düğümden başlayarak, cevabı veritabanında bulunan sorular sorulup alınan cevaplara göre yeni düğümler oluşturulmaktadır. Her düğüm kendinden sonra iki dala veya ikiden fazla dala ayrılmaktadır. Oluşan düğümden sonra yeni soru sorulamıyorsa dallanma bitmiştir ve bir sınıfı temsil eden yaprağa ulaşılmıştır. Şekil 1.9' da, kök düğüm, iç düğüm ve terminal düğüm (yaprak düğüm) gösterilmiştir.

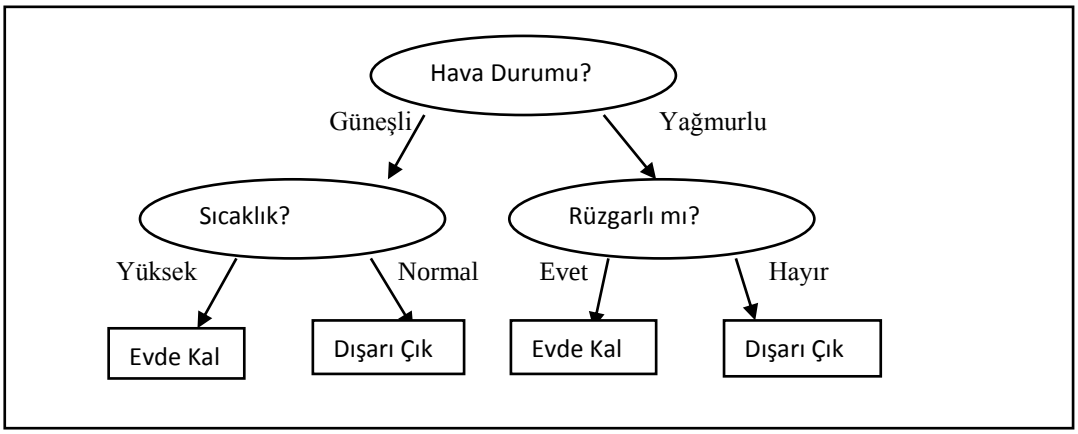
Karar ağaçları oluşturulurken kullanılan algoritmanın ne olduğu önemlidir. Çünkü kullanılan algoritmaya göre oluşturulan ağacın şekli değişebilir. Değişik ağaç yapıları da farklı sınıflandırma sonuçları verir. Kök düğümü oluşturan ilk düğümün farklı olması, en uçtaki yaprağa ulaşırken izlenecek yolu dolayısıyla sınıflamayı değiştirecektir. Gerek kök düğümün gerekse de sonraki her bir düğümün belirlenmesinde en önemli kriter, o noktadan dallara ayrıldığında veritabanının geri kalan kısmının benzer büyüklükte parçalara ayrılıp ayrılmadığıdır. Örneğin veritabanında bulunan cevap evet/hayır gibiyse iki eşit parçaya, evet/hayır/belki gibi üç değişkenliyse mümkün olduğunca üç eşit parçaya bölünmesi istenmektedir. Burada amaç en kısa yoldan istenilen yanıta veya sınıfa ulaşmaktır (Silahtaroğlu,G.,2008,S.47).

Karar ağacı oluşturulduktan sonra, bir test verisini sınıflandırmak oldukça kolaydır. Kök düğümden başlayarak kayda test koşulu uygulanır ve her sonuç için ona ait uygun dal takip edilir. Buradan ya yeni test koşulunun uygulanacağı başka bir iç düğüme, ya da bir yaprak düğüme ulaşılır. Böylece test verisinin hangi sınıfa ait olduğu hangi yaprakta sonlandığına göre belirlenmiş olur.



Şekil 1.9. Kök, iç, yaprak düğüm gösterimi ve bölünme türüne göre karar ağaçları

Örnek bir Karar ağacının gösterimi Şekil 1.10 da verilmiştir. Örnek, hava durumuna göre nasıl hareket edilmesini gerektiğine karar vermektedir.



Şekil 1.10 Örnek bir Karar Ağacı

Kök düğümden yaprak düğüme doğru giden herhangi bir yol bir karar kuralı oluşturmaktadır. Bir karar ağacında oluşan bütün kuralların toplamı da karar ağacının kendisini oluşturmaktadır. Şekil 1.10'da gösterilen karar ağacı için oluşturulan kurallar şu şekilde olacaktır.

- Eğer hava durumu güneşli ve sıcaklık yüksek ise, evde kal.
- Eğer hava durumu güneşli ve sıcaklık normal ise, dışarıya çık.
- Eğer hava durumu yağmurlu ve rüzgarlı ise, evde kal.
- Eğer hava durumu yağmurlu ise ve rüzgarlı değil ise, dışarıya çık.

Karar ağacı temelli analizlerin yaygın olarak kullanıldığı alanlar şunlardır:

- Belirli bir sınıfın olası üyesi olacak elemanların belirlenmesi,
- Çeşitli vakaların yüksek, orta, düşük risk grupları gibi çeşitli kategorilere ayrılması,
- Parametrik modellerin kurulmasında kullanılmak üzere çok sayıdaki değişkenden önemlilerinin seçilmesi,
- Gelecekteki olayların tahmin edilebilmesi için kurallar oluşturulması,
- Sadece belirli alt gruplara özgü olan iliksilerin tanımlanması,
- Kategorilerin birleştirilmesi ve sürekli değişkenlerin kesikli değişkenlere

dönüştürülmesidir (Akpınar, 2000).

1.5.1.1 Karar Ağaçları Oluşturma

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. 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 karar ağacı 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 (Tan ve ark., 2005).

Böl ve elde et (Hunt algoritması)

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

D_t , t düğümü ile bağlantılı öğrenim kayıtları kümesi ve $y=\{y_1,y_2,...,y_c\}$ 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 değişken test koşulu uygula
- Aynı şekilde oluşan her alt veri kümesine rekürsif olarak bu testi yaprak düğüme ulaşana kadar yinele.

1.5.1.2 Karar Ağacı Bölünme Kuralları

Yukarıdan-Aşağıya tekniği genellikle greedy yaklaşımla yapılır. Greedy her aşamada opsiyonel bazı kriterlere göre dallara ayırma yapılmasıdır. Sonuçta bu yaklaşım optimum olmayabilir. Buna rağmen greedy yaklaşım hesaplama yönünden oldukça etkin olduğundan popülerdir. Greedy yaklaşımla aşağıya doğru dallara ayırma yaparken karar vermemiz gereken üç önemli durum vardır. Bunlar, hangi değişken test koşulunun uygulanacağı, en iyi bölünmeyi sağlayacak hangi kriterlerin kullanılacağı ve bölünmenin ne zaman sonlandırılacağıdır. Hunt algoritması dallara ayırma için hangi kuralın uygulanacağını net olarak belirtmemiştir. Dallara ayırma için hangi değişkenin seçileceği bazı kriterlere göre belirlenmektedir. En optimum ağacı oluşturmak için dallanmada seçilecek değişken ve bu değişkenin seçilmesi için uygulanacak kriterler çok önemlidir. Birçok ağaç tabanlı algoritma entropi, bilgi kazancı, gini katsayısı gibi kriterler kullanmaktadır.

1.5.1.2.1 Entropi

Entropi, olayların olma olasılıklarıyla ilişkili olup belirsizliğin ölçülmesi için kullanılan bir ölçüttür. Entropi bilgi ile ilişkilidir ve belirsizlik arttıkça eldeki veriyi daha iyi tanımlamak için daha fazla bilgi gerekecektir. Entropi 0-1 arası değerler alır ve 1 değerine yaklaştıkça belirsizliğin arttığını gösterir.

ID3, C4.5, CART algoritmaları en iyi ayırıcı özelliğe sahip değişkeni bulmak için entropiden faydalanır. Entropiyi matematiksel olarak şöyle ifade edebiliriz;

Sınıf olasılık dağılımı P (p_1 , p_2 ,..., p_k) olan veri seti ***D*** olsun.

Bu durumda ***D***'nin entropisi Denklem 1.1'de verildiği gibidir.

$$E(D)= -\sum_{k=1}^{m} p_i \log_2(p_i) \qquad (1.1)$$

Pi; ***D*** veri setindeki ***i*** sınıfının olasılığıdır ve ***i*** sınıfına düşen örnek sayısının tüm veri setindeki toplam örnek sayısına bölünmesi ile elde edilir.

Karar ağaçları kurulurken amaç, veri setinin entropisini örneklerin hepsinin tek sınıf olarak ifade edildiği yaprak düğüm entropisi sıfır olana kadar düşürmeye çalışmaktır.

Çizelge 1.1'deki veri setinin(D) entropisini hesaplayalım.

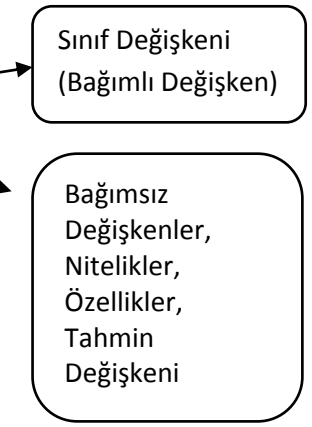
Çizelgede sınıf değişkeni "dolandırıcılık"tır. 3 kişi dolandıran ve 7 kişi dolandırmayan sınıfından olmak üzere toplam veri sayısı 10'dur. Sınıf dağılımı olasılıkları;

$P(p_1 , p_2)=\{\frac{3}{10}, \frac{7}{10}\}$'dur.

$E(D)= - [\frac{3}{10} \text{ x } \log_2(\frac{3}{10})+ \frac{7}{10} \text{ x } \log_2(\frac{7}{10})]= 0{,}8813$

Çizelge 1.1. Örnek veri seti (kredi dolandırıcılığı) ve değişkenlerin gösterilmesi.
(Veri kaynağı: Tan ve ark.,2005)

Kayıt No	Ev Sahipliği	Medeni Hal	Yıllık Gelir (Bin TL)	Dolandırıcılık
1	Evet	Bekar	125	Hayır
2	Hayır	Evli	100	Hayır
3	Hayır	Bekar	70	Hayır
4	Evet	Evli	120	Hayır
5	Hayır	Boşanmış	95	Evet
6	Hayır	Evli	60	Hayır
7	Evet	Boşanmış	220	Hayır
8	Hayır	Bekar	85	Evet
9	Hayır	Evli	75	Hayır
10	Hayır	Bekar	90	Evet



Tüm veri setinin entropisi 0,8813'dür. Entropinin 1 değerine yakın olması sınıf dağılımda yüksek bir değişkenlik olduğunu gösterir. Bu değerin 0'a yaklaşmış olması sınıf dağılımındaki değişkenliğin az olduğunu gösterir.

1.5.1.2.2 Bilgi Kazancı

Bilgi kazancı ID3, C4.5, CART algoritmalarında öğrenme 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. Bilgi kazancı şu şekilde bulunur:

Eğer veri seti ***D***, ***n*** tane alt bölüme ***X*** değişkeninden bölünecekse, X'e ait bilgi kazancı Denklem 1.2'deki gibi hesaplanır.

$$\text{Bilgi Kazancı}(D,X) = E(D) - \sum_{k=1}^{n} p(D_i)\, E(D_i) \qquad (1.2)$$

E(D), Veri setinin X üzerinden bölünmeden önceki entropisi;

$E(D_i)$, i alt bölümünün X üzerinden bölünme olduktan sonraki entropisi;

$p(D_i)$ i alt bölümünün X üzerinden bölünme olduktan sonraki olasılığı;

Bilgi kazancı hesaplanırken, öncelikle veri setinin alt bölümlere ayrılma olmadan önceki halinin entropisi bulunur, daha sonra her alt bölümün entropisi hesaplanır. Bu iki değer arasındaki farkın en yüksek olduğu değişken en iyi dallara ayırma kriteri olarak seçilir.

Örnek veri setindeki ilk düğüm olan kök düğümün, "*ev sahipliği, medeni hal veya yıllık gelir*" mi olacağına karar vermek için her değişkenin sağlayacağı bilgi kazançlarını bulmamız gerekir.

Bir üst bölümde örnek veri setinin entropisi 0,8813 olarak bulunmuştu.

Çizelge 1.1'deki veri setindeki "Ev Sahipliği" değişkeni için bilgi kazancını hesaplayalım;

Ev sahibi olan 3 kişi, ev sahibi olmayan 7 kişi vardır. Ev sahibi olan 3 kişiden 3 tanesi dolandırmayan sınıfındadır. Ev sahibi olmayan 7 kişiden 3 tanesi dolandıran, 4 tanesi dolandırmayan sınıfına aittir.

$E(\text{ev sahibi olmama}) = E(p_1,p_2) = E(\frac{3}{7}, \frac{4}{7}) = -[\frac{3}{7} \text{ x } \log_2(\frac{3}{7}) + \frac{4}{7} \text{ x } \log_2(\frac{4}{7})] = 0{,}985228$

$E(\text{ev sahibi olma}) = E(p_1,p_2) = E(\frac{0}{3}, \frac{3}{3}) = 0 + \frac{3}{3} \text{ x } \log_2(\frac{3}{3}) = 0$

$E(\text{ev sahipliği}) = \frac{3}{10} \text{ x } E(\text{ev sahibi olma}) + \frac{7}{10} \text{x } E(\text{ev sahibi olmama}) =$
$0 + \frac{7}{10} \text{x } 0{,}985228 = 0{,}68966$

Bilgi Kazancı(Ev Sahipliği) = 0,8813 - 0,68966 = 0,19164

Çizelge 1.1'deki veri setindeki "Medeni Hal" değişkeninin bilgi kazancını hesaplayalım;

$E(\text{bekar}) = E(p_1,p_2) = E(\frac{2}{4}, \frac{2}{4}) = -[\frac{2}{4} x \ \log_2(\frac{2}{4}) + \frac{2}{4} x \log_2(\frac{2}{4})] = 1$

$E(\text{evli}) = E(p_1,p_2) = E(0,1) = -[0 + 1 \ x \ \log_2(1)] = 0$

$E(\text{boşanmış}) = E(p_1,p_2) = E(\frac{1}{2}, \frac{1}{2}) = -[\frac{1}{2} x \log_2(\frac{1}{2}) + \frac{1}{2} \ x \ \log_2(\frac{1}{2})] = 1$

$E(\text{Medeni Hal}) = \frac{4}{10} \ x \ 1 + \frac{4}{10} x \ 0 + \frac{2}{10} x \ 1 = 0,6$

Bilgi Kazancı(Medeni Hal) = 0,8813 – 0,6 = 0,2813

Çizelge 1.1'deki veri setindeki "Yıllık Gelir" değişkeninin bilgi kazancını hesaplayalım;

Yıllık gelir değişkeni sayısal bir değişkendir. Oluşacak ağacın yanlara doğru genişlemesi için verilerin temizlenmesi ve dönüştürülmesi gerekmektedir yani yıllık gelir değişkeni gruplara ayrılmalıdır. Yılık gelirin 80.000 TL'den küçük ve 80.000 TL'den büyük şeklinde iki gruba bölündüğünü varsayalım. Bu durumda örnek çizelgemiz Çizelge1.2'deki gibi olacaktır.

Çizelge1.2. Kredi dolandırıcılığı çizelgesi için dönüştürülmüş veriler

No	Ev sahibi olma	Medeni Hal	Yıllık Gelir (Bin TL)	Dolandırma
1	Evet	Bekar	>80	Hayır
2	Hayır	Evli	>80	Hayır
3	Hayır	Bekar	<80	Hayır
4	Evet	Evli	>80	Hayır
5	Hayır	Boşanmış	>80	Evet
6	Hayır	Evli	<80	Hayır
7	Evet	Boşanmış	>80	Hayır
8	Hayır	Bekar	>80	Evet
9	Hayır	Evli	<80	Hayır
10	Hayır	Bekar	>80	Evet

$E(\text{yıllık gelir} \leq 80 \text{ bin TL}) = E(p_1,p_2) = E(0,1) = 0 \ x \log_2(0) + 1 \ x \log_2(1) = 0$

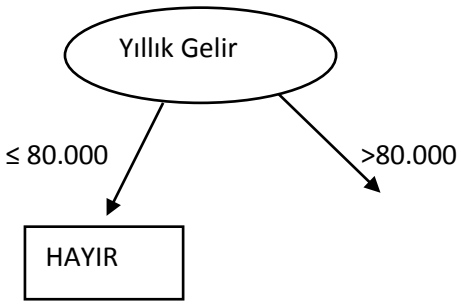
$E(\text{yıllık gelir} > 80 \text{ bin TL}) = E(p_1,p_2) = E(\frac{3}{7}, \frac{4}{7}) = -[\frac{3}{7} x \log_2(\frac{3}{7}) + \frac{4}{7} x \log_2(\frac{4}{7})]$

$= 0,9852$

$$E(\text{yıllık gelir}) = \frac{3}{10} \text{ x } 0 + \frac{7}{10} \text{ x } 0{,}9852 = 0{,}68964$$

Bilgi Kazancı(Yıllık Gelir) = 0,8813 - 0,4854 = 0,3959

Yukarıdaki hesaplamalardan bilgi kazancı en büyük olan değerin yıllık gelir olduğu görülmektedir. Bu nedenle bölünmeye yıllık gelirden başlanacaktır. Şekil 1.11'de oluşturulan kısmi karar ağacı ve dallanma gerçekleştikçe karar ağacında kullanılan veri setleri gösterilmiştir.



Yıllık Gelir ≤ 80.000 TL için yeni alt veri kümesi altkümesi

Ev Sahipliği	Medeni Hal	Yıllık Gelir (Bin TL)	Dolandırma
Hayır	Bekar	≤80	Hayır
Hayır	Evli	≤80	Hayır
Hayır	Evli	≤80	Hayır

Yıllık Gelir > 80.000 TL için yeni veri

Ev Sahipliği	Medeni Hal	Yıllık Gelir (Bin TL)	Dolandırma
Evet	Bekar	>80	Hayır
Hayır	Evli	>80	Hayır
Evet	Evli	>80	Hayır
Hayır	Boşanmış	>80	Evet
Evet	Boşanmış	>80	Hayır
Hayır	Bekar	>80	Evet
Hayır	Bekar	>80	Evet

Şekil 1.11. Kısmi karar ağacı ve dallanmaya göre alt veri kümeleri gösterimi

Yıllık Gelir ≤ 80.000 TL ise Dolandırma = HAYIR olarak bulunup dal sonlanmış ve yaprak düğüm oluşmuştur. Yıllık gelir > 80.000 TL'den büyük kayıtlar için tekrar yukarıda hesaplandığı gibi alt veri setlerine bölünmemiş mevcut veri setinin entropisi

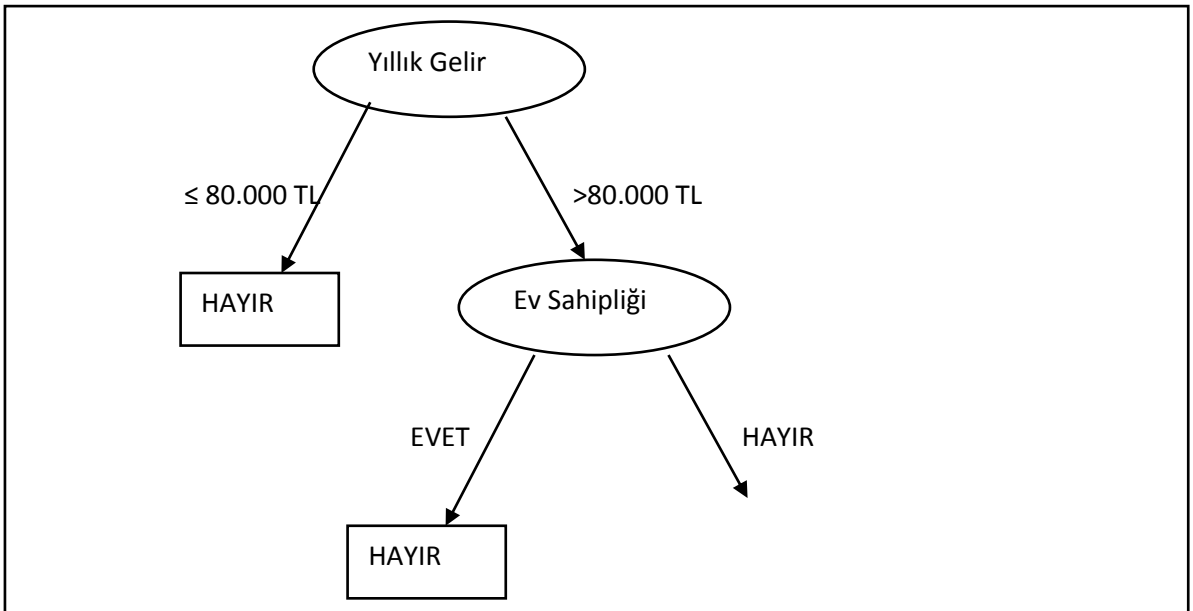
hesaplanacak, daha sonra mevcut veri seti için ev sahipliğinin ve medeni halin entropisi bulunarak her iki değişken için bilgi kazancı hesaplanacaktır. Bilgi kazancı yüksek olan değişken iç düğümü oluşturup, o değişkenden dallara ayrılma gerçekleştirilecektir.

Yıllık Gelir > 80.000 TL alt veri kümesi için ve alt veri kümesinin alt bölümleri için bilgi kazançları aşağıdaki gibi bulunmuştur.

Bilgi Kazancı(Ev Sahipliği)= 0,52164

Bilgi Kazancı(Medeni Hal)= 0,30596

Ev sahipliği değişkeninin bilgi kazancı daha büyük olduğu için bu aşamada bölünmeye ev sahipliği değişkeninden başlanılır. Bu durumda kısmi karar ağacı Şekil 1.12.'deki gibi olur.

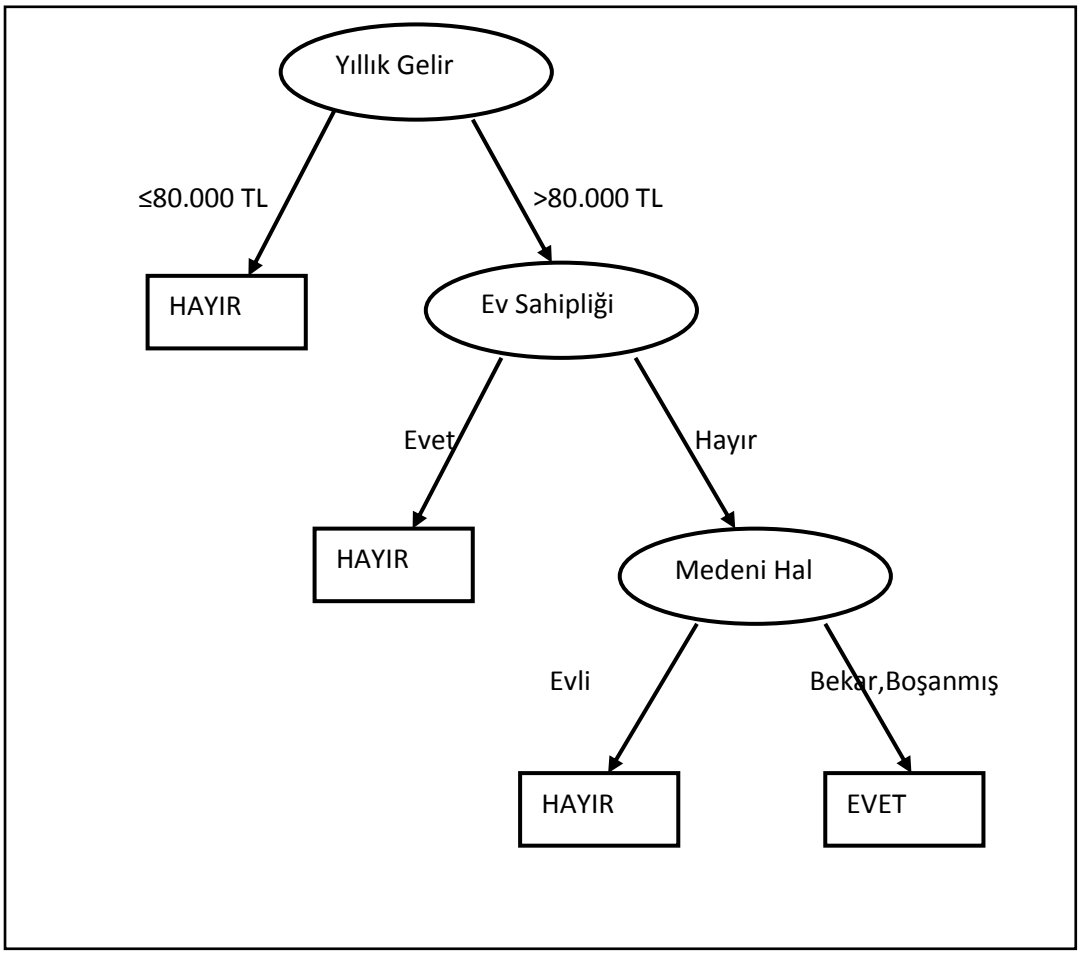


Ev Sahipliği	Medeni Hal	Yıllık Gelir (Bin TL)	Dolandırma
Evet	Bekar	>80	Hayır
Evet	Evli	>80	Hayır
Evet	Boşanmış	>80	Hayır

Ev Sahipliği	Medeni Hal	Yıllık Gelir (Bin TL)	Dolandırma
Hayır	Evli	>80	Hayır
Hayır	Boşanmış	>80	Evet
Hayır	Bekar	>80	Evet
Hayır	Bekar	>80	Evet

Şekil 1.12 Kısmi Karar ağacı ve dallanmaya göre alt veri kümeleri gösterimi

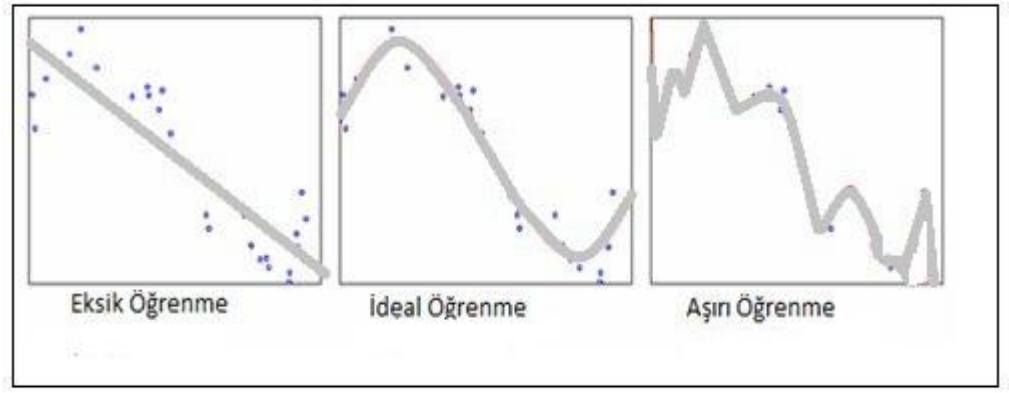
Karar ağacının bu aşamasında Ev sahipliği=EVET ise Dolandırma=HAYIR olarak bulunup dal sonlanmış ve yaprak düğüme ulaşılmıştır. Ev sahipliği=HAYIR olduğunda Medeni Hal değişkenine göre Dolandırma=EVET ve Dolandırma = HAYIR olan kayıtlar mevcuttur. Bu durumda Medeni hal değişkeninden aşağıya doğru bölünme yapılacaktır. Bu aşamadan sonra her sınıflara yerleşmemiş veri olmadığından Şekil 1.13'de görülen nihai karar ağacı oluşturulmuş olacaktır.



Şekil 1.13 Örnek veriler için oluşturulan nihai karar ağacı

1.5.1.2.3 Budama

Teorik olarak, öğrenme veri setindeki verileri kullanarak sıfır hata oranına sahip karar ağacı oluşturulabilir. Ancak oluşturulan tahminsel model için aşırı öğrenme (overfitting) söz konusu olabilir. Bu durumda oluşturulan model öğrenme veri seti için %100'e yakın doğru sınıflama sağlarken, yeni gelen veriler için doğruluk oranı çok düşük olabilir. Aşırı öğrenme, karar ağaçları çok fazla detayı karakterize ettiğinde veya öğrenme verisinde çok fazla gürültülü veri olduğunda ortaya çıkmaktadır. Şekil 1.14'de model oluşturulması sırasında karşılaşılan farklı öğrenme tipleri gösterilmiştir. Öğrenme veri setindeki verilerin önemli noktalarını dikkate almadan geliştirilen karar ağaçlarının doğruluk oranları düşük olacaktır. Bu gibi durumlarda Şekil 1.14'de görüldüğü gibi eksik öğrenme (underfit) söz konusu olur.



Şekil 1.14. Model oluşturulması sırasında karşılaşılan farklı öğrenme tipleri

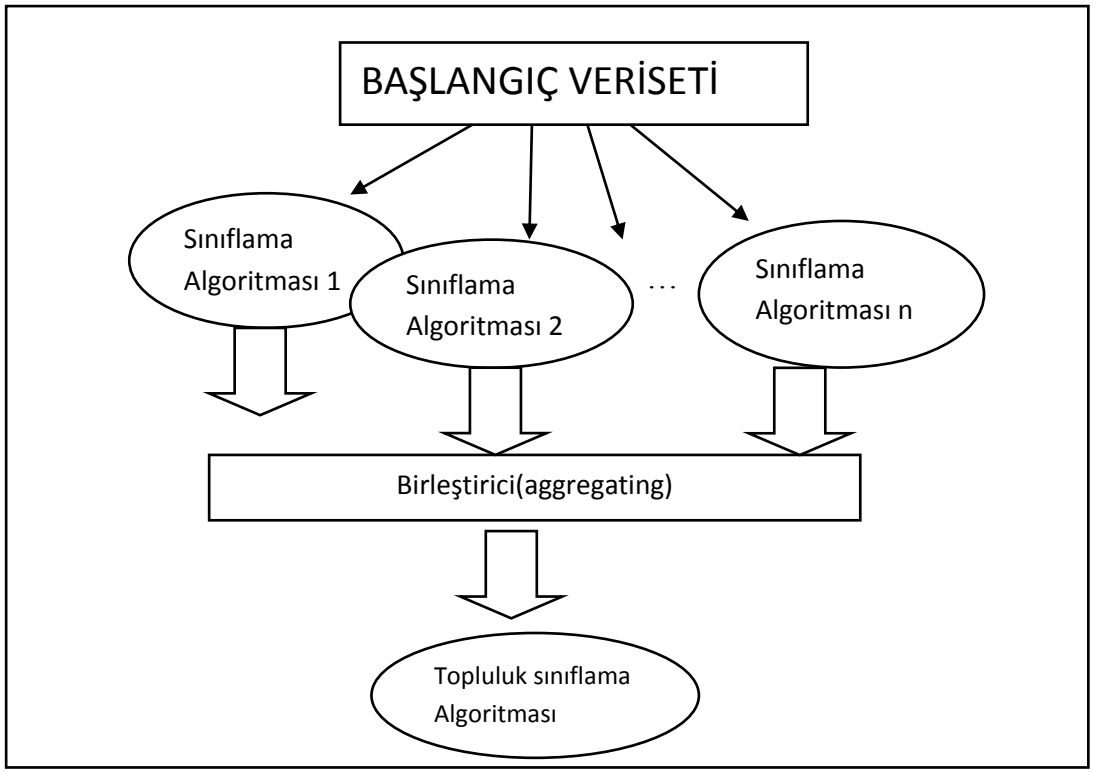
İdeal öğrenmede (ideal fit) ise verilerin önemli ve gerekli noktaları dikkate alınarak model geliştirilmektedir. İdeal öğrenme mevcut öğrenme veri seti için uygun bir doğruluk oranı sağlarken, sonradan gelecek veri için de yeterli doğruluk oranını korumaktadır. İdeal öğrenme Şekil 1.14'de gösterilmiştir.

Aşırı öğrenme durumunda, gürültü oluşturan veya yanlış bilgi veren aşırı değerler (outliers), oluşturulan modeli hatalı hale getirmektedirler. Bu durumda model tekrar gözden geçirilmelidir. İdeal öğrenmeyi sağlamak için, oluşturulan ağaç üzerinde yeni bir operasyon yapmak gerekmektedir. Bu operasyona ***budama (pruning)*** denilmektedir. Budama, önbudama (pre-pruning) veya sonradan budama (post-pruning) şeklinde yapılabilir. Çoğu algoritma genellikle sonradan budama yöntemini kullanmaktadır. Sonradan budama, veri setindeki tüm veriler kullanılarak

tüm yaprak düğümlere ulaşılan ağaç oluşturulduktan sonra belirlenen alt dalları ağaçtan çıkararak veya iki ayrı dalı birleştirerek uygulanmaktadır. Bu şekilde yapılan budama sonrası ağaçların boyutları küçülürken, modelin sınıflama hata oranı da azalmaktadır.

1.6 Ağaç Tabanlı Topluluk Yöntemler

Topluluk öğrenme yöntemlerinde(Ensemble Learning) birden çok sınıflayıcının ortaya koyduğu sonuçlar bir araya getirilerek, topluluk adına tek bir karar verilmektedir. Bu yöntemler, birbirinden farklı çok sayıda sınıflayıcının yaptığı sınıf tahminleri oylamaya tabi tutar ve oylama sonucunda en çok oyu alan sınıfı topluluğun (komitenin) sınıf tahmini olarak sunar. Topluluk öğrenme yöntemleri, çok sayıdaki sınıflayıcının kararını birleştirdiği için daha güvenilir tahminler ortaya koymaktadır. Şekil 1.15.'te topluluk öğrenme stratejisi gösterilmiştir.



Şekil1.15. Topluluk Öğrenme Stratejisi

Topluluk öğrenme yöntemleri temel veya tekil öğrenme algoritmalarının ortaya koyduğu tahminlerin doğruluk oranını arttırmaktadır ve bu sebepten dolayı tekil öğrenme yöntemlerine göre daha başarılıdırlar. Bu yöntemlerden Bagging (Breiman 1996) ve Boosting en çok üzerinde çalışılan ve bilinen topluluk öğrenme algoritmalarıdır. Breiman, 2001 yılında Bagging yöntemine göre daha fazla rastgelelik (randomization) sağlayan Random Forests yöntemini önermiştir. Karar ormanını oluşturan her bir karar ağacı, orijinal veri setinden rastgele ve yerine koyarak seçilen bootstrap örnekleme ile oluşturulmaktadır. Her bir karar ağacı, veri setini oluşturan tüm değişkenlerden rastgele seçilen az sayıdaki değişken kullanılarak oluşturulmaktadır.

Öğrenme yöntemlerinde öncelikle en az örnek sayısıyla çok fazla verinin sınıflandırmasını yapmak için uygun metot bulunması amaçlanmaktadır. En ideal öğrenici (sınıflayıcı) en az zamanda ve olabildiğince en az öğrenme veri seti kullanarak tahmin yapabilen öğrenicidir. Random Forests yöntemi bu açıdan bakıldığında en ideal sınıflayıcılar arasında yer almaktadır

1.6.1 Bagging Yöntemi

Bagging (Bootstrap Aggregating) bir topluluk yöntemi olup sınıflama ve regresyon modelleri için uygulanmaktadır. Aşırı öğrenmeye karşı güçlü olan bu yöntem sınıflamada doğru sınıflama oranını arttıran ve varyansı düşüren bir yöntemdir. Veri setinde kayıp verilerin olduğu durumlarda da sınıflamada oldukça başarılıdır.

Bagging yöntemi, bir çok sınıflama modeline uygulanabilmekle birlikte daha çok karar ağaçları için kullanılmaktadır. Bagging yöntemi veri setinden sınıf yapısını bozmayacak şekilde rastgele örnekler seçilerek (bootstrap) oluşturulan çok sayıdaki karar ağacının yaptığı sınıf tahminleri oylamaya tabi tutularak en çok oyu alan sınıfı nihai sınıf tahmini olarak belirleyen öğrenme yöntemidir. Bagging yönteminde art arda oluşturulan ağaçlar önceden oluşturulan ağaçlara bağımlı değildirler ve ağaçlar orijinal veri setinden bootstrap örnekleme yapılarak oluşturulmaktadır. Bootstrap yöntemi bölüm 1.6.3.4'de anlatılmıştır.

1.6.2 Boosting Yöntemi

Boosting, Kearns(1988) tarafından; "Çok sayıda zayıf öğrenicinin oluşturduğu sınıflayıcı grup bir araya gelerek güçlü bir öğrenici oluşturabilir mi?" sorusundan etkilenerek oluşturulmuş bir topluluk öğrenme yöntemidir. Zayıf öğrenici, doğru sınıflama ile çok az ilişkili bir sınıflayıcı iken güçlü öğrenici doğru sınıflama ile çok fazla ilişkili olan bir sınıflayıcıdır.

Boosting, belirli bir algoritmayla kısıtlı değildir, ancak çok sayıda boosting algoritması belirli bir dağılıma tabi olarak zayıf sınıflayıcıların sonuçlarını tekrarlı bir şekilde toplayıp en son güçlü sınıflayıcıyı oluşturmaktadır. Her yeni zayıf öğrenici bir sonuç ortaya koyduğunda veri yeniden ağırlıklandırılmaktadır. Burada toplama işlemi devam ederken, önceki oluşturulan zayıf öğrenicilerden yanlış sınıflama yapanlara daha fazla odaklanılarak yeni zayıf öğreniciler oluşturulmakta ve sonuç olarak güçlü bir sınıflayıcı ortaya konulmaktadır.

Boosting yöntemi de çok sayıda oluşturulan karar ağaçlarının sonuçlarını ağırlıklı oylamaya tabi tutarak son sınıf tahminini yapmaktadır. Boosting yönteminde sonradan oluşturulan ağaçlar önceden oluşturulan ağaçlara bağımlıdırlar. Daha önceki oluşturulan ağaçlardan yanlış tahminde bulunan tahmin edicilere daha fazla ağırlık verilerek ard arda yeni ağaçlar oluşturulmaktadır. Sonunda da final tahmin için ağırlıklı oylama yapılmaktadır. En çok bilinen boosting algoritması Yoav Freund ve Robert Schapire(1996) tarafından formüle edilen Adaboost'tur.

1.7 Random Forests (RF) Yöntemi

1.7.1. Tanımı ve Algoritması

Random Forests, yukarıda bahsedildiği gibi topluluk öğrenme yöntemidir. Bireysel olarak oluşturulan karar ağaçları bir araya gelerek karar ormanını oluşturmaktadır. Karar ormanındaki her karar ağacı, orijinal veri setinden bootstrap tekniği ile farklı

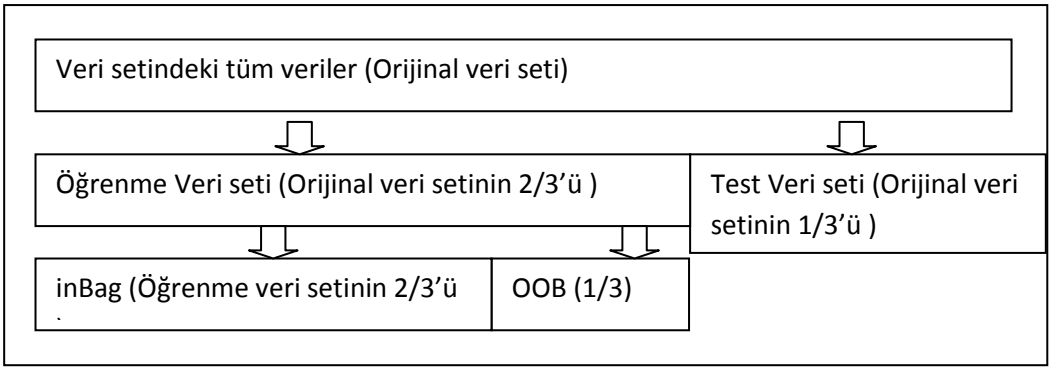
örneklemler seçilerek oluşturulmaktadır. Orman, ağaçların yapmış olduğu sınıf tahminleri bir araya getirerek, nihai sınıf tahminini yapmaktadır.

Random Forests yöntemi, Breiman (1996) tarafından önerilen Bagging tekniği ile Ho (1998) tarafından önerilen Random Subspace tekniğini birleştirmiştir. Bagging yönteminde karar ağaçları veri setinden bootstrap tekniği ile örneklem seçilerek birbirinden bağımsız olarak oluşturulmaktadır. Random Subspace yöntemi ise, karar ağacının her düğümünde en iyi dallara ayırıcı değişkenin, tüm değişkenler arasından rastgele (random) seçilen az sayıdaki değişken içinden seçilmesidir.

Random Forests yönteminde sonradan gelen veriye ait tahmin yapılmasının yanında, değişkenlerin önem derecesi de hesaplanmaktadır. Veri setinde çok sayıda değişken varsa, değişken önem derecesinin hesaplanması model indirgemesi açısından oldukça kullanışlıdır. Örneğin binlerce değişkenin bulunduğu veri setinde, Random Forests yöntemiyle elde edilen önem derecesine göre, kurulacak yeni modelde önem derecesi yüksek değişkenler kullanılarak daha etkin tahminlerin yapılması sağlanabilir.

Random Forests yönteminde, model kurulurken, modeli test etmek için ayrı bir test veri seti yoksa veya orijinal veri setinden test veri seti ayrılmamışsa, sınıf dağılımına bağlı kalınarak orijinal veri setinin 2/3'ü öğrenme veri seti (inBag), 1/3'ü ise test veri seti (Out-Of-Bag (OOB)) olarak ayrılmaktadır. Eğer ayrı bir test veri seti varsa veya orijinal veri setinden test veri seti ayrılmışsa, modelin kurulması için ayrılan öğrenme veri seti, kendi içinde 2/3 oranında öğrenme veri seti (inBag), 1/3'ü ise test veri seti(OOB) olarak ayrılmaktadır. Şekil 1.16'de karar ağacı oluşturmak için orijinal veri setinden "inBag" verisi ve "OOB" verisinin hangi oranlarda seçildiği gösterilmiştir. Karar ormanı kaç karar ağacı ile oluşturulacaksa o kadar sayıda bootstrap tekniği ile örneklem oluşturulur ve her örneklem için inBag ve OOB verisi ayrılır. Kurulan her karar ağacının, o ağaç için ayrılan OOB verisi ile testi yapılarak hata oranı tahmini yapılmaktadır. Tüm karar ağaçları için yapılan OOB hata oranlarının ortalaması alınarak, karar ormanının(modelin) OOB hata oranı kestirimi hesaplanmaktadır. OOB verisi ile yapılan teste modelin iç testi de denilmektedir. Modelin testi, ayrı bir test veri seti varsa veya orijinal veri setinden

test veri seti ayrılmışsa bu veri setleri ile de yapılabilir. Bu şekilde yapılan test ile ortaya çıkan hata oranı, OOB hata oranına yakın bir değer olmaktadır. Yapılan çalışmalar, yeteri kadar karar ağacı oluşturulursa OOB hata oranının sapmasız olarak doğru hesaplandığını ortaya koymaktadır.



Şekil.1.16 Random Forests yönteminde veri seçimi. (Kaynak**:** Pater, N., 2005)

Random Forests yönteminde her bir karar ağacı, bootstrap tekniği ile orijinal veri setinden seçilen örneklemle ve her düğümde tüm değişkenlerden belirlenen sayıda rastgele değişken seçilmesi ile oluşturulmaktadır. Rastgele seçilen değişkenlerden en iyi bölünme sağlayacak olan belirlenip, dallara ayrılmaya o değişkenden başlanılır. Bu yöntemde oluşturulan karar ağaçları üzerinde budama işlemi yapılmaz. Diğer taraftan RF yönteminin uygulanması da oldukça kolaydır. Uygulamada analistin belirlemesi gereken iki parametre vardır, bunlar oluşturulacak ağaç sayısı ve seçilecek değişken sayısıdır. Bununla birlikte, bu parametrelerin seçilecek değerlerin genellikle sonuca etkisi azdır.

RF yöntemiyle sınıflama ağaçları veya regresyon ağaçları kurulabilmekte ve kümeleme yapılabilmektedir.

Karar ormanını oluşturan karar ağaçları, bootstrap yöntemiyle karar ağacını oluşturacak veri seçildikten sonra, CART algoritması ile oluşturulmaktadır. CART algoritması veri setinin hangi değişkenden başlayarak dallara ayrılacağına bilgi kazancını kullanarak karar verir. Ayrıca değişkenin uygun sınıflama için cut-off değeri gini katsayısı ile belirlenir.

Ağaçlar kurulduktan sonra, test veri setindeki her biri farklı bir deneğe (kişiye) ait olan veriler satır satır önceden kurulmuş ağaçlar üzerinde yukarıdan aşağıya doğru yerleştirilir. Bu işlem veri setindeki tüm veriler için tekrar edilir. Her karar ağacı her bir deneği var olan sınıflardan birine yerleştirir. 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. Sınıflama ağaçları için bütün ağaçlardan ağırlıklı olarak en çok oyu almış olan sınıf nihai sınıf tahmini olarak belirlenir. Veri setindeki her biri farklı bir deneğe ait olan verilerin satır satır hangi sınıfa yerleştirildiğine nihai olarak orman karar vermiş olur. Regresyon ağaçları içinse yapılan oylamanın ortalaması alınarak nihai tahmin yapılır.

Sınıflama veya regresyon ağaçları için RF algoritması şu şekildedir.

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 ağacını şu şekilde oluştur;

    a. 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.

    b. Belirlenen tahmin değişkeni için en iyi dallanma kriterini gini indeksi ile hesapla ve hesaplanan değere göre veri setini her düğümde iki alt dala ayır

    c. Madde a ve b'deki 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 değişkenlerin(bağımsız değişkenlerin) 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

Öğrenme veri setinden hata oranını hesaplamak içinse ;

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

Andy (2002), yeteri kadar ağaç oluşturulursa OOB hata oranının oldukça doğru kestirildiğini, yeteri kadar ağaç oluşturulmadığı durumda ise OOB hata oranının olduğundan büyük kestirildiğini belirtmiştir.

1.7.2. Değişken Önem Derecesi

Değişken önem derecesinin hesaplanması oldukça zordur. Çünkü bir değişkenin önem derecesi diğer değişkenlerle olan ilişkisinin derecesinden de kaynaklanıyor olabilir. 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 ağaçlar için tekrar edilir. Oluşan d_i'lerin ortalaması alınarak m. değişken için kabaca değişkenin önemi belirlenmiş olur(d). orijinal veri setindeki tüm değişkenler için aynı işlem tekrar edilir ve tüm değişkenlerin kabaca değişken önem derecesi bulunmuş olur. 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 bölünerek, bulunan değer ilgili değişkenin önem derecesi skorunu oluşturmaktadır (Eşitlik 1.3).

$$\text{Önem Derecesi Skoru} = \frac{\overline{d}}{SEdi} \qquad (1.3)$$

Gini Yöntemi: Bu yöntemde, her m. değişkenden dallara ayırma gerçekleşmeden önce gini değeri hesaplanmaktadır. Daha sonra m. değişkenden alt dalla ayrıldığında bölünen veri için tekrar gini değeri hesaplanmaktadır. 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 fark bulunur ve tüm ağaçlar oluştuktan sonra aradaki farklar toplanır. Bulunan değer m. değişkenin gini önem derecesini verir. Bu işlemler tüm değişkenler için hesaplanır. Bu yöntemle elde edilen değişken gini önem derecesi; çoğu zaman OOB verisindeki m. değişkenin tüm değerleri yer değiştirerek bulunan değişken önem derecesine paralel bir değer olmaktadır.

Değişken önem derecesi ile sonucun oluşmasında fazla bir rolü olmayan değişkenlerin modelden çıkartılması sağlanarak model indirgemesi yapılabilir. RF

yönteminde binlerce değişken olsa da yöntem bu tür verileri de işleyebilecek özelliğe sahiptir. Binlerce değişkenden önemli değişkenler belirlenebilir. Ancak çok fazla değişkenin modele katılması sistem kaynaklarını zorlayabilir ve uygulama zaman alabilir.

1.7.3. Örnekler Arası Yakınlık (Proximity)

Oluşturulan ağaçların birbirinden farklılığı, ağacın her düğümünde rastgele değişken seçilmesi ve seçilen bu değişkenlerden en çok bilgi sağlayan değişkenin bölünme için kullanılması ile sağlanmaktadır. Burada sağlanan çeşitlilik çok önemlidir çünkü RF algoritmasının performansı oluşan ağaçlardan herhangi ikisinin arasındaki korelasyonun derecesi ile yakından ilişkilidir. Ağaçlar arasındaki korelasyon arttıkça, ormandaki bütün ağaçların toplam performansı azalmaktadır. Ağaçlar arasındaki korelasyonun azaltılması tüm değişkenler arasından rastgele değişkenler seçmek ve her düğümde seçilen değişkenleri kullanarak ağaç oluşturmak ile mümkündür. Bu da oluşan ağaçların daha güçlü olmasını ve sınıflayıcının performansının artmasını sağlayacaktır. Çok çeşitli ağaçlardan oluşan orman ile daha etkili kararların alınması sağlanmaktadır.

Proximity derecesi veri setindeki kayıtlı verilerin birbirleriyle ne düzeyde ilişkili olduğunu gösterdiğinden çok önemlidir. 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. Genellikle proximity matrisi NxN boyutlu olup, N ağaç oluşturulurken veri setindeki toplam denek sayısıdır. Örneğin; 3 ayrı denekten alınan veri, 5 ağaçta kullanılmış olsun. Verilerin beraber olarak kaç kez terminal düğümde sonlandığını gösteren proximity matrisi Çizelge 1.3a'daki gibi olabilir.

Çizelge1.3a NxN Proximity Matrisi

	Denek1	Denek2	Denek3
Denek1	5	3	1
Denek2	3	5	4
Denek3	1	4	5

Çizelge 1.3b Normalleştirilmiş Matris

	Denek1	Denek2	Denek3
Denek1	1	0,6	0,2
Denek2	0,6	1	0,8
Denek3	0,2	0,8	1

Çizelgedeki her değer, ormandaki ağaç sayısına bölünerek normalleştirilmiş proximity matris oluşturulur. Oluşturulan matris simetrik bir matristir. Çizelgede görüleceği üzere, Denek1 ve Denek2, 3 ağaçta birlikte bulunmuşken, Denek1 ve Denek3 sadece 1 ağaçta beraber bulunmuşlardır. Çizelge 1.3b'de Çizelge 1.3a'daki verilerin normalleştirilmiş hali gösterilmiştir.

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. Aşırı değerlerin, sınıflama neticesinde aynı sınıfta yer aldığı tahmin edilen diğer verilerle arasında proximity değeri oldukça düşüktür. Proximity değerleri kullanılarak veri setindeki eksik değerlerin tahmin edilmesi de sağlanmaktadır. Proximity matrisi ile birbirine yakın verileri bir araya toplayarak kümeleme de yapılabilmektedir.

1.7.4. Bootstrap Örnekleme

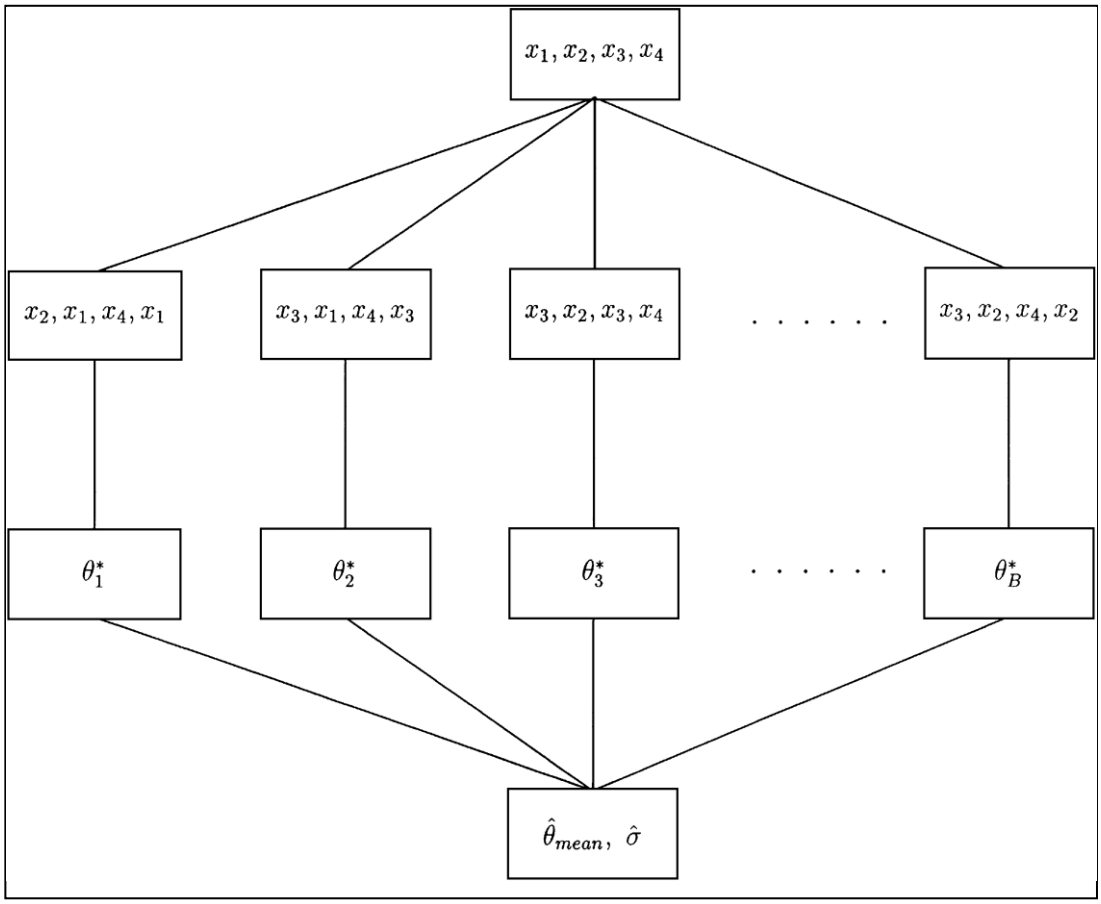
Mevcut veri setindeki verilerden her defasında yerine koyarak farklı örnekler seçip yeni bir veri seti oluşturmak mümkündür. Bu şekilde yeni veri setleri oluşturmaya bootstrap metodu denilmektedir. Bu yöntem 1979 yılında Bradley Efron tarafından önerilmiştir ve istatistiksel çıkarımlar için çok güçlü bilgisayar destekli kazanımlar sağlamıştır. Dağılımı bilinmeyen veriler için yaklaşık güven aralıklarının tahminini yapmak için sık kullanılan bir metot olmuştur (Clarke, B. ve ark, 2009).

Bootstrap metodu uygulanma kolaylığı ve yararlılığının yanı sıra başka avantajlara da sahiptir. Klasik istatistikte, incelenen değişkenlerin normal dağılış gösterdiği varsayımına dayalı olarak tahminler yapılırken, bootstrap metodunda ise veri setinden şansa bağlı örnekler alınarak istatistiksel tahminlemeler yapılmaktadır. Bu metot ile,

- Çok küçük veri setlerinde bile doğru yanıtlar alınabilmektedir.
- Büyük veri setlerinde ise klasik istatistik yöntemlerle elde edilen sonuçlarla paralel sonuçlar elde edilmektedir.
- Hemen hemen tüm istatistiksel analizler yapılabilmektedir.

Bootstrap metodu, çok karmaşık matematik formüllerin çözülmesinde hesaplama yükünü azaltması sebebiyle de avantajlıdır. Ayrıca, verilerin dağılışı hakkında herhangi bir varsayım gerektirmemekte ve herhangi bir istatistiğin varyansı hakkında bilgi verebilmektedir. Bu nedenle, diğer metotların kullanımının uygun olmadığı ya da bilinen varsayımların geçersiz olduğu durumlarda Bootstrap metodu tercih edilebilmektedir (Yakupoğlu,Ç.,Atıl,H.).

Bootstrap metodu ile örnek seçimini şu şekilde açıklayabiliriz. N adet gözlemden oluşan veri seti X= $(x_1, x_2, x_3, x_4, \ldots\ldots\ x_N)$ olsun. Bu veri setinden 1/N kadar olasılıkla şansa bağlı olarak bootstrap örnek veri seti X_i^* = $(x_1^*, x_2^*, x_3^*, x_4^*, \ldots\ldots, x_N^*)$ elde edilmektedir. Bu işlem ne kadar örneklem oluşturulmak isteniyorsa o kadar tekrar edilerek, istenilen kadar bootstrap veri seti oluşturulabilir. Bootstrap metodunun şematik gösterimi Şekil 1.17.'de verilmiştir.



Şekil 1.17. Bootstrap yönteminin şematik gösterimi (Kaynak:Sacchi, M.D., 1998)

1.7.5. Gini Katsayısı

Random Forests algoritmasının alt yapısını, CART algoritması ile oluşturulmuş çok sayıda karar ağacı oluşmaktadır. CART sınıflama ve regresyon tekniklerini birleştirmiştir. Ağacın aşırı büyümesini önlemek için ikili bölme (binary-split) yöntemi geliştirilmiştir. RF yönteminde, CART algoritmasından farklı olarak, karar ağaçlarına herhangi bir budama işlemi yapılmaz ve ağaçlar oluşturuldukları şekilde bırakılır. CART/RF algoritmasında veri setindeki hangi değişkenin en iyi dallara ayırma değişkeni olacağına karar verilirken entropi tabanlı bilgi kazancı kullanılır. Hangi değişkenden bölüneceği saptandıktan sonra ilgili değişkenin hangi değeri ile dallara ayrılacağı ***gini indeksi*** kullanılarak belirlenir.

Gini katsayısı eşitlik 1.4'teki gibi hesaplanır.

Eğer veri seti ***D***, ***n*** tane sınıftan örnekler içeriyorsa,

$$Gini(T)=1-\sum_{j=1}^{n}(p_j^2) \quad (1.4)$$

p_j, ***D*** veri setindeki ***j*** sınıfının relatif frekansıdır.

Eğer D veri setindeki örnek büyüklüğü sırayla N_1 ve N_2 olan, D_1 ve D_2 alt veri setlerine bölünmüşse; bölünmüş verinin gini indeksi ***n*** tane sınıftan örnekler içermektedir. Bu durumda gini indeksi eşitlik 1.5'teki gibi bulunur.

$$Gini_{bölünmüş}(D) = \frac{N1}{N}Gini(D_1)+\frac{N2}{N}Gini(D_2) \quad (1.5)$$

En küçük $Gini_{Bölünmüş}(D)$ değerine sahip olan değişken değeri bölünme değeri olarak seçilir.

1.7.6. Random Forest Modelinin Kurulması

Random Forests modelinin kurulmasında aşağıdaki aşamalar izlenir.

- orijinal veri setinden, öğrenme veri seti ayrıldıktan sonra, öğrenme veri setinden sınıf dağılımına sadık kalacak şekilde bootstrap yöntemiyle rastgele örneklemler seçilir.
- Oluşturulan örnek veri setinden kullanıcının belirlediği sayı kadar değişken ağaç yapısında kullanılmak üzere rastgele seçilir. Eğer orijinal veri setinde, sınıf değişkeni hariç, toplam değişken sayısı *M* ise, ağaç yapısında analistin belirlediği *R* tane değişken kullanılacaktır. Ağacın çok fazla büyümemesi ve aşırı öğrenme sorunu oluşmaması için ağaç yapısında bütün değişkenler kullanılmamaktadır. Burada $R<M$ olmak zorundadır.
- Karar ağacı oluşturulurken her düğümde, belirlenen R tane değişkenden dallara ayrılmaya bilgi kazancı en yüksek değişkenden başlanılır. Dallara ayrılacak değişken belirlendikten sonra her düğümden aşağıya doğru iki dal oluşturulur. Dalların hangi değere göre ayrılacağına gini indeksi kullanılarak karar verilir. Bu işlem her düğüm için yeni oluşturulacak dal kalmayıncaya kadar tekrar edilir.

- RF yönteminde, orijinal veri setinden öğrenme veri seti ayrıldıktan sonra, öğrenme veri setinin 2/3'ü ise karar ağaçlarını kurmak için kullanılır. Öğrenme veri setinin 1/3'ü ise modelin iç hata oranını hesaplamak için test veri seti (OOB) olarak ayrılır. Bu test modelin kendi kendini test etmek ve iç hata oranını (OOB hata oranı) hesaplamak için yapılmaktadır.
- OOB hata oranı, modelin düzgün kurulmasını ve oluşturulan her ağaca bir ağırlık değeri verilmesini sağlar. En düşük hata oranına sahip ağaca en yüksek, en yüksek hata oranına sahip olan ağaca en düşük ağırlık verilir. Bu şekilde kurulan tüm ağaçlara, ağacın hata oranının değerine göre göreceli olarak bir ağırlık verilir.
- Ormanın sınıflaması yapılırken, her ağaç oluşturduğu sınıflardan birine ağırlıklı olarak oy verir. Orman, veri setindeki her deneğe ait sınıf tahminini, tüm ağaçların yaptığı tahminlerin bir araya getirilmesi neticesinde, ağırlıklı olarak en çok oyu almış sınıfı seçerek yapar. Mesela, veri setindeki 1. sıradaki denek için oluşturulan bütün ağaçlarda düştüğü sınıf belirlenir. Daha sonra bu ağaçların ağırlıkları OOB hata oranı değerlerine göre hesaplanır ve her bir sınıf için ağaçların aldığı ağırlığa göre verdiği oyların toplamı ayrı ayrı belirlenir. Hangi sınıfın ağırlık toplamı daha büyük ise söz konusu denek o sınıfa atanır. Diğer denekler içinde benzer işlemler yapılarak RF yöntemine göre sınıflama yapılmış olur.
- RF yönteminde bireysel olarak oluşturulan ağaçlar üzerinde budama işlemi yapılmaz. Her ağaçta kullanılan veri ve değişkenler farklı olduğundan, RF yöntemi aşırı öğrenmeye ve gürültülü veriye karşı güçlüdür.

RF yöntemiyle oluşturulan ormandaki tekil bir ağacın elde edilme aşamalarını aşağıdaki örnek ile gösterebiliriz.

Sınıf değişkeninin firmanın finans durumu (0-kötü,1-iyi) olduğu toplam 5 örnekten oluşan örnek veri seti Çizelge 1.4a.'da verilmiştir. Sınıflamada Firma Türü, Ödeme Türü, Müşteri Sayısı ve Bulunduğu Şehir olmak üzere 4 değişken kullanılacaktır. Ağaç yapısında p toplam değişken sayısını göstermek üzere $p^{1/2} = 4^{1/2} = 2$ değişken kullanılacaktır. 4 değişkenden rastgele olarak Müşteri Sayısı ve Firma Türü seçilmiş ve N=5 genişliğinde bootstrap örneklem seçilerek Çizelge 1.4b. oluşturulmuştur.

Çizelge 1.4a. Firmaların finans durumunu gösteren örnek veri seti

Firma no	Değişkenler				Sınıf
	Firma Türü	Ödeme Türü	Müşteri Sayısı	Bulunduğu Şehir	
1	3	Nakit	31	Ankara	1
2	1	Kredi Kartı	30	İstanbul	0
3	2	Nakit	6	Bursa	0
4	4	Nakit	15	Bursa	1
5	4	Kredi Kartı	10	Ankara	0

Çizelge 1.4b. Bootstrap örnekleme ve rastgele değişken seçimi ile oluşturulmuş örnek veri seti

Firma no	Değişkenler		Sınıf
	Firma Türü(FT)	Müşteri Sayısı(MS)	
3	2	6	0
5	4	10	0
4	4	15	1
2	1	30	0
1	3	31	1

Örneğimizde firmaların finans durumu(0-kötü,1-iyi) rastgele seçilen Müşteri Sayısı ve Firma Türü değişkenleriyle tahmin edilmeye çalışılacaktır. Veriyi ilk olarak bölüm 1.5.1.2.2.'de hesaplama yöntemi verilen bilgi kazancı değerinin en yüksek olduğu değişkenden başlayarak dallara ayıralım. Örnek veri seti için en çok bilgi kazancı sağlayan değişken MS olarak bulunmuştur ve dallara ayrılmaya ilk MS değişkeninden başlanacaktır. MS için mümkün bölünme değerleri $6 \leq \boldsymbol{x} \leq 31$ aralığında değişmekte olup $\boldsymbol{x}$ bölünme değeridir. Veri setimize göre MS'in alabileceği değerler; MS ≤ 6, MS ≤ 10, MS ≤ 15, MS ≤ 30, MS ≤ 31 'dir. En ideal bölünme kriterini belirlemek için ***gini indeksi*** hesaplanacak ve en küçük gini indeksine sahip değer, ilk bölünme kriteri olacaktır. MS ≤ 6 kriteri için veri seti RF yöntemine göre parçalandığında Çizelge 1.5 deki değerler elde edilir.

Çizelge 1.5. MS≤ 6 değişkeni için sınıf değerleri

Değişken Kriteri	Sınıflara düşen Kayıt Sayısı		
	0 (kötü)	1 (iyi)	N=5
MS ≤ 6	1	0	n_1=1
MS > 6	2	2	n_2=4

Çizelge 1.5'te ki değerlere göre Eşitlik 1.4 kullanılarak Gini(MS ≤ 6) ve Gini(MS > 6) ve $Gini_{Bölünmüş}$ değerleri aşağıdaki gibi elde edilir.

$Gini\ (MS \leq 6) = 1-(1^2+0^2) = 0$

$Gini(MS > 6) = 1-(2/4^2+2/4^2) = 0,5$

Eşitlik 1.4 kullanılarak $Gini_{Bölünmüş}$ aşağıdaki gibi elde edilir.

$Gini_{Bölünmüş} = 1/5x0 + 4/5x\ 0,5 = 0,4$

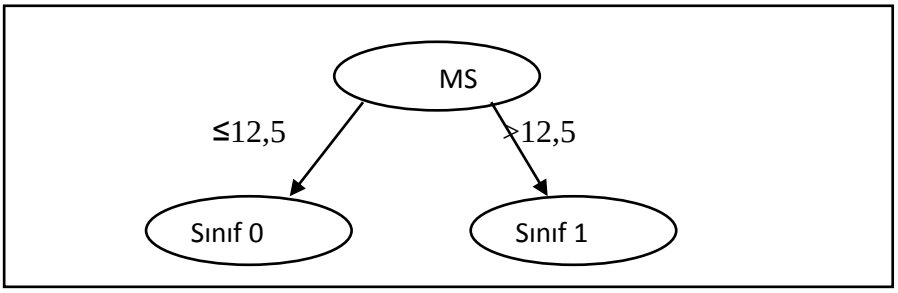
MS değişkenine ait diğer tüm $Gini_{Bölünmüş}$ değerleri Çizelge 1.6'de verilmiştir.

Çizelge 1.6 MS değişkeninin olası tüm değerleri için hesaplanan gini katsayıları

$Gini_{Bölünmüş}$	Hesaplanan Değer
$Gini_{Bölünmüş}$(MS ≤ 6)	0,4000
$Gini_{Bölünmüş}$(MS ≤ 10)	0,2671
$Gini_{Bölünmüş(}$ MS ≤ 15)	0,4671
$Gini_{Bölünmüş}$ (MS ≤ 30)	0,3000
$Gini_{Bölünmüş}$(MS ≤ 31)	0,4800

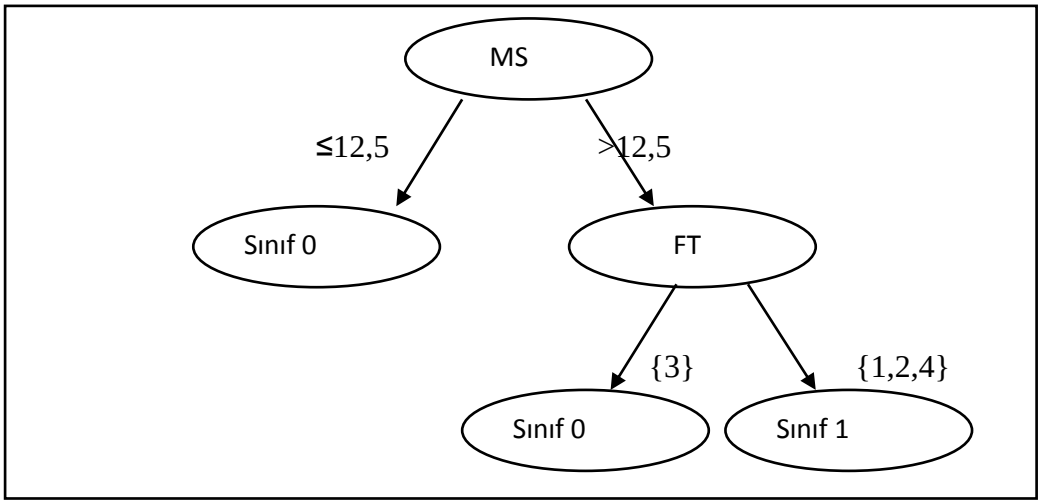
Random Forest yönteminde en küçük $Gini_{Bölünmüş}$ değerini veren kriter bölünmeyi sağlayacak kriter olarak seçilir. MS değişkeni sürekli olduğundan art arda gelen değerlerin orta noktası en iyi bölünme kriteri olarak belirlenir. Örneğimizde

MS ≤ 10 bölünme değeri en küçük $Gini_{Bölünmüş}$ değerine sahiptir. Bu durumda bölünme değeri MS=(10+15)/2 =12,5 olarak seçilir. Veri setini ilk bölecek değişken ve bölünme değeri belirlendikten sonra karar ağacı Şekil 1.18'de verildiği gibi oluşturulur.



Şekil 1.18: CART algoritması ile oluşturulan kısmi karar ağacı

Örneğimizdeki diğer değişken olan FT için de benzer işlemler uygulandığında nihai karar ağacı Şekil 1.19'deki gibi oluşturulur.



Şekil 1.19 CART/RF algoritması ile oluşturulan nihai karar ağacı

Şekil 1.19 da verilen karar ağacına göre karar kuralları aşağıdaki gibidir.

- Eğer MS ≤ 12,5 ise Sınıf değerimiz 0'dır.
- Eğer MS > 12,5 ve FT=3 ise Sınıf değerimiz 0'dır.
- Eğer MS > 12,5 ve FT={1,2,4} ise Sınıf değerimiz 1'dir.

Şekil 1.19 'da ki ağaç RF yönteminin bir adımı olan ve CART algoritması ile oluşturulan tekil bir ağaçtır. RF yönteminde, farklı bootstrap örneklem ve rastgele seçilmiş farklı değişken setleri ile çok sayıda ağaç oluşturulmaktadır. Örneğimizde ormanın 6 ağaçtan oluştuğunu ve Şekil 1.19'da verilen ağacın "Ağaç 1" olduğunu varsayalım. Örnek olarak, sınıfını tahmin etmemiz için {MS=25, FT=1} şeklinde yeni bir kayıt geldiğini varsayalım. "Ağaç 1" in {MS=25, FT=1} verisi için yapacağı tahmin "sınıf 1" olacaktır. Ormanı oluşturan 6 ağaç olduğunu varsayalım. Diğer 5 ağacında bu veri için belirlediği sınıflar Çizelge 1.7'de verildiği gibi olsun. OOB hata oranına göre ağaçlara verilen ağırlıkların da çizelgedeki gibi olduğunu varsayalım.

Çizelge 1.7 Ormandaki ağaçların belirlediği sınıflar ve bu ağaçların ağırlıkları

Ağaç	Sınıf	Ağırlık
Ağaç 1	1	7
Ağaç 2	1	6
Ağaç 3	0	4
Ağaç 4	1	4
Ağaç 5	0	5
Ağaç 6	0	2

Sınıf 1'in aldığı toplam oy: 7+6+4=17
Sınıf 0'ın aldığı toplam oy: 4+5+2=11

Ormanın kararını vermek için, ağaçların yapmış olduğu oylar ağırlıklı olarak toplandığında sınıf 1'in aldığı toplam oy 17 iken, sınıf 0'ın almış olduğu oy 11'dir. Örneğimizde Müşteri Sayısı 25 ve Firma Türü 1 olan bu firmanın finansal durumu "iyi" olarak tahmin edilmiştir.

1.7.7. 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 ormanın hata oranı (generalization error) ormandaki ağaç sayısı arttıkça belirli bir limite yakınsamaktadır. Ağaç yapılı sınıflayıcıların oluşturduğu ormanın hata oranı, bireysel oluşturulan ağaçların güçlü olmasına ve aralarındaki korelasyonun az olmasına bağlıdır. Bununla beraber her düğümde rastgele değişkenlerin seçilmesi, hata oranının Adaboost'a (Boosting algoritması) göre daha düşük olmasını sağlamakla beraber modelin gürültüye karşı da daha güçlü olmasını sağlamaktadır. İki sınıflı model için sınıflama matrisi Çizelge 1.8.'da verilmiştir.

Çizelge 1.8. Sınıflama Matrisi

	Gerçek Sınıf		
Modelin Sınıf Tahmini		Pozitif	Negatif
	Pozitif	Doğru Pozitif Sayısı (TP)	Yanlış Pozitif Sayısı (FP)
	Negatif	Yanlış Negatif Sayısı (FN)	Doğru Negatif Sayısı (TN)

Modeli oluşturan toplam örnek sayısı=N=TP+FP+FN+TN (1.6)

Modelin doğru sınıflama oranı (Accuracy) $=\frac{TP+TN}{N}$ (1.7)

Modelin sınıflama hata oranı $=\frac{FP+FN}{N}$ (1.8)

Eğer modeldeki sınıf sayısı ikiden fazla ise modelin performansı Eşitlik 1.9'da ki gibi, modelin sınıflama hata oranı ise Eşitlik 1.10'deki gibi hesaplanmaktadır.

Modelin doğru sınıflama oranı= $\frac{\sum_{i=1}^{n}(Doğru\ Sınıflama\ )i}{Toplam\ \ Örnek\ \ Sayısı}$ (1.9)

Modelin hatalı sınıflama oranı = 1- Modelin doğru sınıflama oranı. (1.10)

Denklemdeki *n,* modeldeki toplam sınıf sayısını, "doğru sınıflama" sayısı ise modelin sınıflara göre doğru tahmin sayısını ifade etmektedir.

Sınıflama yöntemlerinde modelin performansını değerlendirmek için kullanılan başka ölçütlerde vardır. Bu ölçütlerden başlıcaları; Doğru Pozitif Oranı (Duyarlılık), Doğru Negatif Oranı (Seçicilik), Seçicilik (Sensivity) ve Duyarlılık (Specifity) kullanılarak çizilen ROC (Reciever Operating Curve) eğrisi'dir. Çizelge 1.8'a göre bu değerlerin hesaplanması Eşitlik 1.11 ve Eşitlik 1.12'de gösterilmiştir.

Doğru Pozitif Oranı(Duyarlılık) = $\frac{TP}{(TP+FN)}$ (1.11)

Doğru negatif Oranı(Seçicilik) = $\frac{TN}{(TN+FP)}$ (1.12)

1.7.8. Hata Oranı Tahmini

En çok kullanılan ölçme değeri her algoritma için modelin hata oranını hesaplamak ve bu değere göre en iyi modeli seçmektir. Hata oranı, test veri setinin öğrenme veri setiyle oluşturulan model üzerinde uygulanmasıyla hesaplanmaktadır.

Modelin hata oranını hesaplamak için üç temel yöntem vardır.

1.7.8.1. Holdout Yöntemi

Holdout yöntemi hata oranını hesaplamak için kullanılan en kolay ve en temel tekniktir. Hata oranını hesaplamak için orijinal veri seti, öğrenme veri seti ve test veri seti olmak üzere önceden belirlenen oranlarda iki parçaya bölünmektedir. Genellikle veri setinin 2/3'ü öğrenme, 1/3'ü ise test veri seti olarak ayrılmaktadır. Öğrenme veri seti sınıflama algoritmasını oluşturmada kullanılırken, test veri seti oluşturulan

algoritmaya göre sınıflayıcının başarısını ölçmek için kullanılır. Holdout yönteminde test veri seti oluşturulan algoritmaya sadece bir kez uygulanır.

1.7.8.2. Tekrarlı Holdout Yöntemi

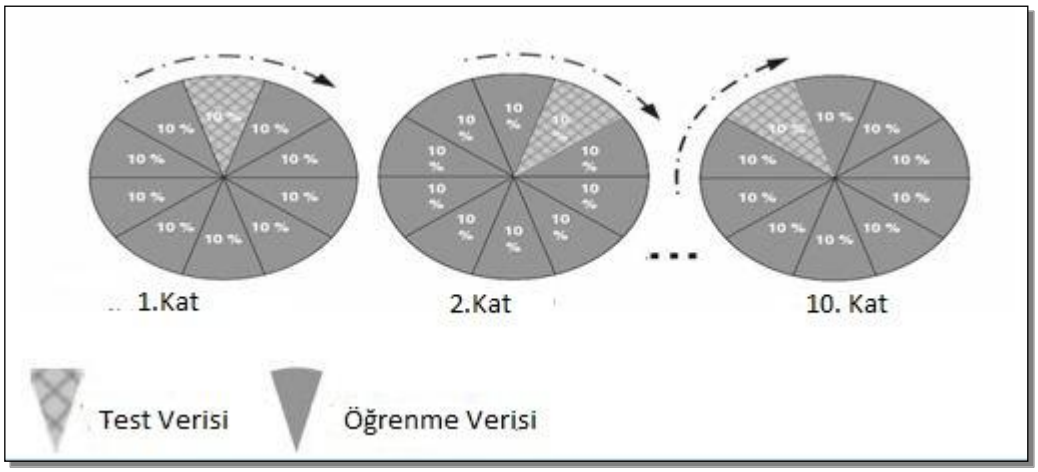
orijinal veri seti birbirinden farklı birçok test ve öğrenme veri setlerine bölünüp her test ve öğrenme veri seti için holdout tekniği uygulandığında daha gerçekçi sonuçlar elde edilir. Bu yöntem "tekrarlı holdout tekniği" olarak adlandırılmaktadır. Bu metotla her aşamada rastgele belirlenen oranda test ve öğrenme veri seti oluşturulup model için hata oranı hesaplanmaktadır. Her aşamada hesaplanan hata oranlarının ortalaması alınarak genel bir hata oranı elde edilmektedir.

1.7.8.3. Çapraz-Doğrulama Yöntemi

K-katlı (k-fold) çapraz doğrulama metodunda, ana veri seti örneklem genişlikleri eşit ***k*** sayıda alt veri setine bölünmektedir. Oluşturulan ***k*** adet alt veri setinden bir tanesi modeli test etmek için doğrulama verisi olarak seçilir, geri kalan ***k-1*** tane alt veri seti ise öğrenme veri seti olarak kullanılır. ***k*** tane alt veri setinden her defasında bir tanesi doğrulama verisi olarak kullanılarak ve her alt veri seti bir kez doğrulama verisi seçilerek bu işlem ***k*** defa tekrar edilir. *k*. adımdan sonra, hesaplanan k tane hata oranının ortalaması alınarak genel hata oranı bulunur. Çapraz Doğrulama tahmini (CVA) Eşitlik 1.13'de gösterilmiştir.

$$\text{CVA}=\frac{1}{k}\sum_{i=1}^{k} Ai \qquad (1.13)$$

Denklem'de A_i doğrulama metriği, k ise doğrulama sayısını göstermektedir. Veri madenciliğinde genelde 10 katlı çapraz doğrulama yöntemi tercih edilmektedir. Deney 10 defa tekrar edilerek, örnekleme ile test ve öğrenme veri seti olarak seçilen veriden kaynaklanan sapmalar en aza indirilmektedir. Şekil 1.20'de 10 katlı çapraz doğrulama yöntemi gösterilmiştir.



Şekil 1.20 10-katlı çapraz doğrulama yöntemi gösterimi. (Kaynak: Delen,D. Advanced Data Mining Techniques.S.142)

1.7.9. Random Forest Algoritmasının Üstün Yönleri ve Kısıtları

Random Forests algoritmasının diğer algoritmalara göre üstünlükleri olduğu gibi kısıtları da mevcuttur. Bu özellikler aşağıda verilmiştir.

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 verasetlerinde 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 verasetleri 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 matris oluşturarak tespit edilebilmektedir. Proximity, kümeleme yaparken, aşırı değerleri (outlier) tespit etmek için kullanılmasının yanında verinin görsel olarak gösterilmesin 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.

Kısıtları:

- Tek bir karar ağacında olduğu gibi ortaya çıkan sonuç ağaç yapısında görsel olarak görülemez. Model karmaşık olduğu için bir çok 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 vermemektedir.
- Random Forests yöntemi için geliştirilen programlar, oluşan her ağacı sistem belleğinde tuttuğundan için çok fazla bellek kullanmakta ve bu nedenle düşük bellekli bilgisayarlarda uygulaması zorlaşmaktadır.

2. GEREÇ VE YÖNTEM

2.1. Uygulama Verisi

Bu çalışmada Random Forests yöntemini uygulamak için kullanılan veriler, Gazi Üniversitesi Diş Hekimliği Fakültesi Periodontoloji bölümünden temin edilmiş olup yapılan ölçümler periodontal hastalık şikayeti ile gelen 175 kişiye aittir.

Hastalar sistemik ve dental muayene değerlerine göre üç ana gruba ayrılmıştır. Bu gruplar; hem sistemik olarak kalp hastalığı hem de dental olarak periodontal hastalığı olan bireyler (PER+AMI), sistemik olarak sağlıklı ancak dental olarak periodontal hastalığı olan bireyler (PER-AMI), hem sistemik hem dental olarak sağlıklı bireyler (Kontrol Grubu)'den oluşmuştur.

Bu çalışma kapsamında hastalardan sistemik, demografik, dental ve serolojik özelliklere ait ölçümler alınmıştır. Bu ölçümlere ait detaylı bilgiler Çizelge 2.1'de verilmiştir.

Çizelge 2.1. Hastalardan alınan ölçüm değişkenleri

SİSTEMİK ÖZELLİKLER		
Değişken Adı	**Değişken Tipi**	**Açıklama**
CRP , mg/l	Sürekli Nicel	Creaktif protein,Kalp hastalığında artar
CHOLMGDL , mg/dl	Sürekli Nicel	Total kolesterol
LDLMGDL , mg/dl	Sürekli Nicel	LDL kolesterol
HDLMGDL, mg/dl	Sürekli Nicel	HDL kolesterol
TGMGDL , mg/dl	Sürekli Nicel	Trigliserit
GLUCMGDL , mg/dl	Sürekli Nicel	Glukoz (açlık kan şekeri)
WBC	Sürekli Nicel	Beyaz kan hücresi
DIABETES	Kategorik 0=yok1=var	Diyabet varlığı
HYPERTENS	Kategorik 0=yok1=var	Hipertansiyon varlığı
HYPERLIP	Kategorik 0=yok 1=var	Yüksek lipidemi(yağ oranında artış)
DEMOGRAFİK ÖZELLİKLER (Hastalara doldurulan anket sonuçlarına göre belirlenmiştir)		
Değişken Adı	**Değişken Tipi**	**Açıklama**
AGE	Sürekli Nicel	Yaş
GENDER	Kategorik 1=erkek 2=kadın	Cinsiyet
SMOKING	Kategorik 1=hiç içmemiş 2=bırakmış	Sigara içme durumu

	3=sigara içen	
BMI , kg/m^2	Sürekli	Vücut kitle indeksi
SOSECST	Kategorik 1=düşük 2=orta 3=yüksek	Sosyoekonomik düzey
EDUCATION	Kategorik 1=ilköğretim 2=lise 3=üniversite	Eğitim durumu
DENTAL ÖZELLİKLER (Klinik olarak kaydedilmiştir)		
Değişken Adı	**Değişken Tipi**	**Açıklama**
NTEETH	Kesikli Nicel	Ağızda mevcut diş sayısı
NEXTRACT	Kesikli Nicel	Çekilmiş diş sayısı
NTEETHPDLESS4	Kesikli Nicel	Cep derinliği 4 mm'den az olan diş sayısı
NTEETHPD4TO5	Kesikli Nicel	Cep derinliği 4mm ve 5mm' ye eşit diş sayısı
NTEETHPDOVER5	Kesikli Nicel	Cep derinliği 5 mm'den fazla olan diş sayısı
NTEETHCAL4TO5	Kesikli Nicel	Klinik ataşman düzeyi 4 ve 5 mm olan diş sayısı
NTEETHCALOVER5	Kesikli Nicel	Klinik ataşman düzeyi 5mm'den fazla olan diş sayısı
NSİTESPD4TO5	Kesikli Nicel	Cep derinliği 4 ve 5 mm olan diş bölgesi sayısı
NSİTESPDOVER5	Kesikli Nicel	Cep derinliği 5 mm'den fazla olan diş bölgesi sayısı
NSİTESCAL4TO5	Kesikli Nicel	Klinik ataşman düzeyi 4 ve 5 mm olan diş bölgesi sayısı
NSİTESCALOVER5	Kesikli Nicel	Klinik ataşman düzeyi 5 mm'den fazla olan diş bölgesi sayısı
PLIMEAN	Kategorik 0=plak yok 1=gözle görülemeyen ve aletin ucuna gelen plak varlığı 2=ince film şeridi şeklinde plak varlığı 3=dişin 2/3'ünü kaplayan plak varlığı	Plak indeks , ağızda biriken ve periodontal hastalığa sebep olan bakteri ürünleri ve yiyecek artıklarının oluşturduğu bir yapıdır.
GIMEAN	Kategorik 0=sağlıklı dişeti 1=iltihap başlangıcı, kanama yok 2=ilerlemiş iltihap ve adacıklar şeklinde kanama 3=ilerlemiş iltihap ve spontan kanama	Gingival indeks, dişeti iltihabını sınıflayan bir indeks sistemidir
PDMEAN	Sürekli Nicel	Cep derinliği , Periodontal hastalığa bağlı olarak meydana gelen kemik yıkımını göstermektedir
CALMEAN	Sürekli Nicel	klinik ataşman düzeyi , kemik yıkımı ile beraber yumuşak dokulardaki kayıpları gösterir ve her bir hasta için ortalama değer verilir
BOPMEAN	Sürekli Nicel	Sondlamada kanama , hastalığın akut durumda olup olmadığını gösterir ve her bir hasta için ortalama değer verilir
SEROLOJİK ÖZELLİKLER (Hastalardan alınan serum örneklerinde belirlenmiştir.)		

Değişken Adı	Değişken Tipi	Açıklama
LIMULUSLPS	Sürekli Nicel	Serumda belirlenen lipopolisakkarit düzeyi
AAPAL1000RATIO	Sürekli Nicel	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 Nicel	Mikroorganizmanın (A.a) hastalık yapan yüzey antijenlerinden PAL a karşı antikor cevabınının serumda 1/2000 dilüsyonda ölçülmesi
AALPS1500RATIO	Sürekli Nicel.	Mikroorganizmanın (A.a) hastalık yapan yüzey antijenlerinden lipopolisakkarite karşı antikor cevabınının serumda 1/1500 dilüsyonda ölçülmesi
AALPS3000RATIO	Sürekli Nicel	Mikroorganizmanın (A.a) hastalık yapan yüzey antijenlerinden lipopolisakkarite karşı antikor cevabınının serumda 1/3000 dilüsyonda ölçülmesi
AAOMP1000RATIO	Sürekli Nicel	Mikroorganizmanın (A.a) hastalık yapan yüzey antijenlerinden outer membrane proteine (dış membran proteini) karşı antikor cevabınının serumda 1/1000 dilüsyonda ölçülmesi
AAOMP2000RATIO	Sürekli Nicel	Mikroorganizmanın (A.a) hastalık yapan yüzey antijenlerinden outer membrane protein (dış membran proteini) karşı antikor cevabınının serumda 1/2000 dilüsyonda ölçülmesi
PGLPS100RATIO	Sürekli Nicel	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 Nicel	Mikroorganizmanın (P.g) hastalık yapan yüzey antijenlerinden lipopolisakkarite karşı antikor cevabınının serumda 1/200 dilüsyonda ölçülmesi
PGOMP200RATIO	Sürekli Nicel	Mikroorganizmanın (P.g) hastalık yapan yüzey antijenlerinden outer membrane protein (dış membran proteini) karşı antikor cevabınının serumda 1/200 dilüsyonda ölçülmesi
PGOMP400RATIO	Sürekli Nicel	Hastalık yapan yüzey antijenlerinden outer membrane protein (dış membran proteini) karşı antikor cevabınının serumda 1/200 dilüsyonda ölçülmesi

Veri setini oluşturan değişkenlerden yola çıkarak RF algoritması ile bir model oluşturulacaktır. Bu modelin amacı veri setindeki değişkenlerden hangilerinin modele (sınıflamaya) katkı sağladıklarını bulmak, model kurulduktan sonra aşırı değer olan denekleri tespit etmek, sınıflama yapmak ve yapılan sınıflamanın performansını ölçmektir.

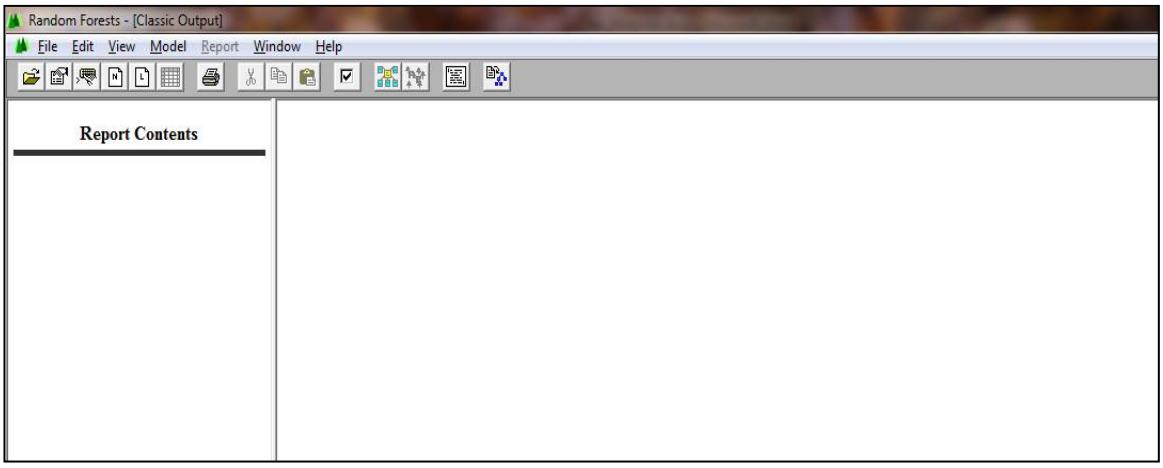
Sınıflama yöntemi sağlık bilimlerinde sıklıkla tanı koyma amaçlı kullanılmaktadır. Model oluşturulduktan sonra hangi değişkenlerin modele katkı sağladığı belirlenmiş olacağından, aynı tür bir hastalığın tespit edilmesi yani tanı konulması için tüm tetkik verilerine gerek duyulmaksızın sadece sınıflamaya katkısı olan verilerin elde edileceği tetkiklerin yapılması, hem sağlık alanında çalışan doktorlara zaman kazandıracak hem de ekonomik anlamda daha az maliyetli olacaktır. Sınıflamaya katkısı en çok olan değişkenlerin değerlerini içeren bir veri geldiğinde model, verinin oluşturulan gruplardan hangisine ait olduğunu belirli bir hata oranıyla verecektir.

2.2. Veri analizinde kullanılan program

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

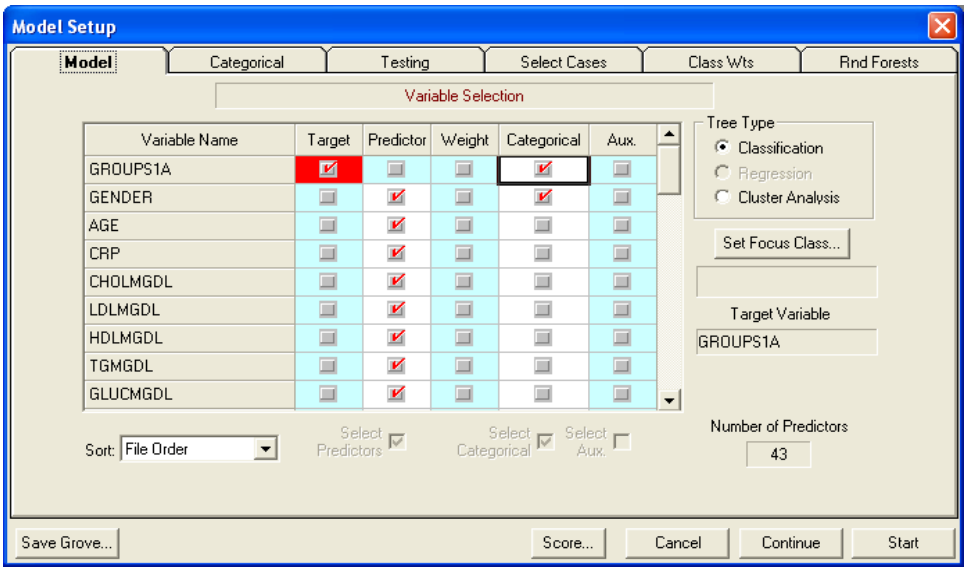
RandomForests Programının ana ekranları

RandomForests Programının ana menüsü Şekil 2.1'te görüldüğü gibidir. Program SPPS, SAS, Excel ve benzeri veri dosyalarını okuyabilmektedir. Veri dosyası açmak için menüden ***File-->Open*** seçeneğiyle sisteme aktarılabilir.



Şekil 2.1. RandomForests programının ana menüsü

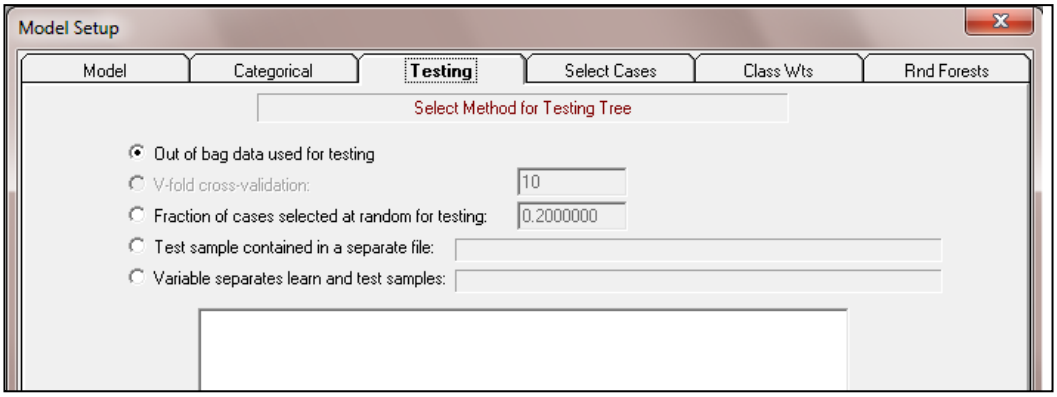
Veri dosyası sisteme aktarıldıktan sonra ***Model*** ekranı açılır. ***Model*** ekranından hangi değişkenin target (sınıf değişkeni) olduğu, hangi değişkenlerin predictor (tahmin değişkeni) olduğu seçilir. Şekil 2.2'de Model ekranı görülmektedir. Değişkenler içerisinde kategorik olanlar, "***Categorical***" butonu kullanılarak işaretlenmelidir. Amaç sınıflama yapmak ise "***Tree Type***" bölümünde ***Classification*** seçilmelidir.



Şekil 2.2. RandomForests programının değişken seçme ekranı

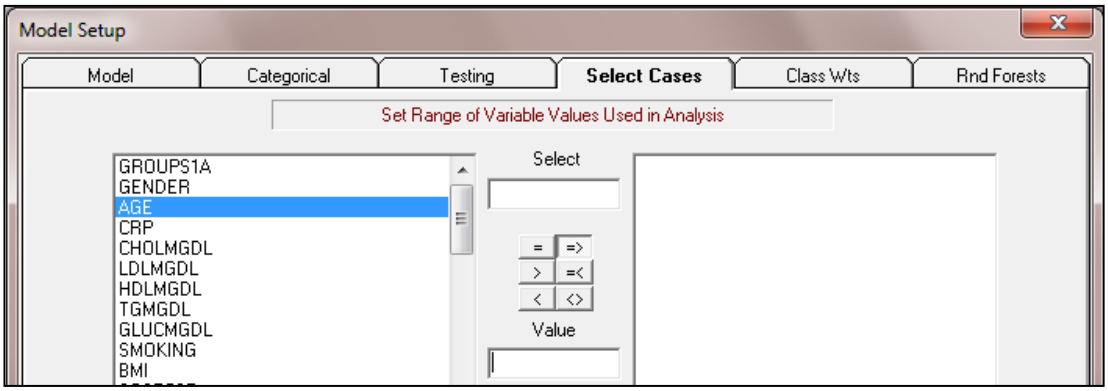
Modeli test etmek için yöntem seçimi "***Model setup***" ekranında ***Testing*** bölümünden yapılır (Şekil 2.3). Modeli test etmek için ayrı bir test veri seti olmadığı

durumda veya orijinal veri setinde denek sayısı az olduğu zaman "***Out of bag data used for testing"*** butonu seçilir. Bu yöntem tekrarlı çapraz geçerlilik (k-fold cross validation) yöntemine benzer. Orijinal veri setinden alınan bootstrap örneğindeki verilerin bir bölümü model geliştirmede bir bölümü test amacıyla kullanılır ve daha sonra tersi işlem yapılır. Böylece tüm denekler üzerinde hem model geliştirme hem de test işlemi yapılmış olur. Bu yöntemin tahminleri ile bağımsız test verisi kullanılarak yapılan tahminlerin uyumlu sonuçlar verdiği gösterilmiştir. Ancak orijinal veri setinde denek sayısı çok olduğu zaman orijinal veri setinin bir bölümü öğrenme veri seti amacıyla ayrılmalı bir bölümü ise sadece test verisi olarak kullanılmalıdır. Bu durumda "***Fraction cases selected at random for testing"*** butonu seçilir. Test veri seti ayrı bir dosyada saklandığında, "***Test sample contained in a separate file***" butonu, veri setinde test için kullanılacak veri ayrı bir değişken tanımlanarak belirlendiğinde ***"Variable separates learn and test samples"*** butonu işaretlenmelidir.



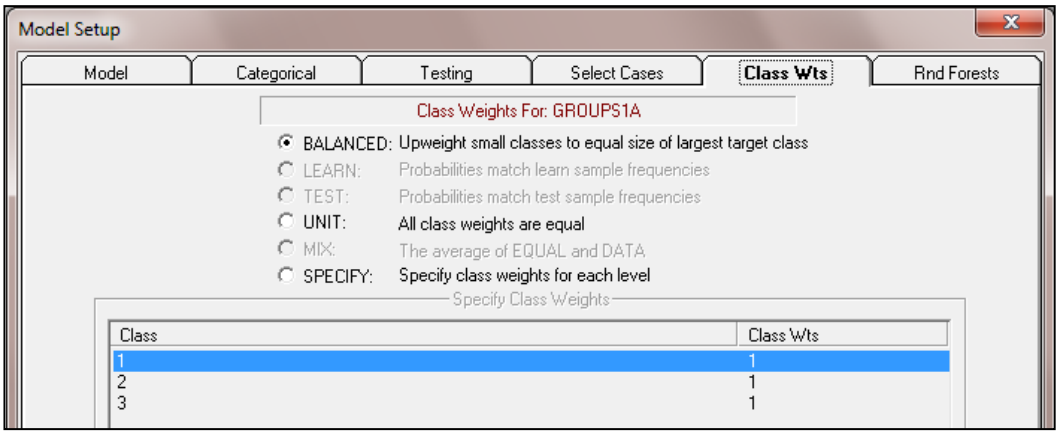
Şekil 2.3 RandomForests programının test ekranı

Değişkenler üzerinde herhangi bir dönüştürme işlemi yapmak veya bazı kısıtlar koymak istendiğinde ***"Model Setup"*** ekranındaki ***"Select Cases***" bölümü kullanılır.



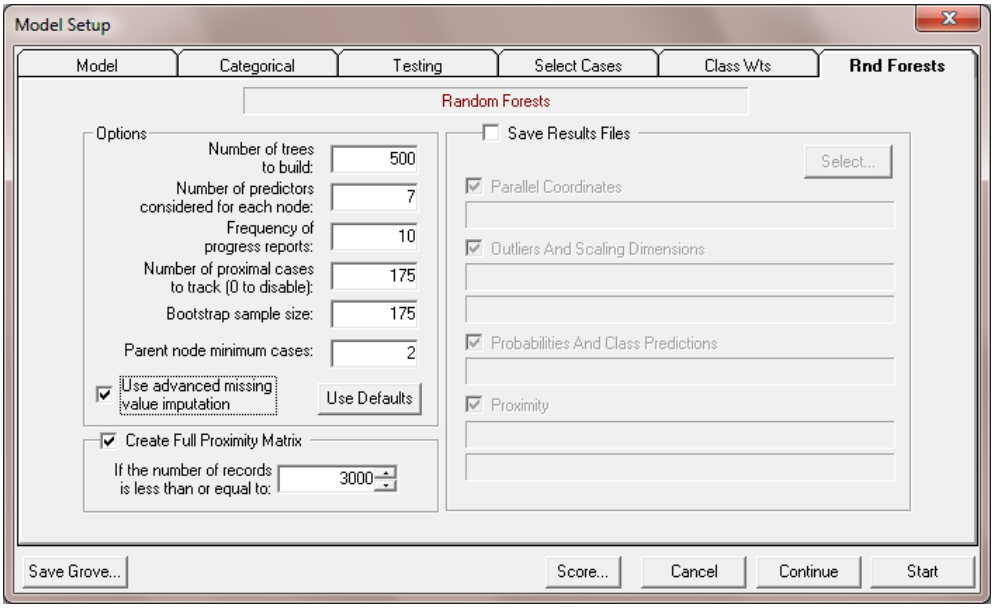
Şekil 2.4 RandomForests programının veri kısıtlama veya dönüştürme ekranı

Modeli kurarken sınıfların dağılımları, Şekil 2.5'deki "***Class Wts***" bölümünden yapılabilir. Sınıflama değişkeninde sınıflara düşen denek sayılarının belirlenmesinde "***Balanced***", "***Unit***" veya "***Specify***" seçeneklerinden biri işaretlenir. Sınıflara düşen denek sayıları çok farklı olduğunda "***Balanced"*** seçeneği işaretlenir ve otomatik olarak denek sayısı fazla olan sınıfa düşük ağırlık, denek sayısı az olan sınıfa yüksek ağırlık verilir. Böylece örnekteki sınıfların dağılımında denge sağlanmış olur. Sınıflara farklı farklı ağırlıklar verilmek istenilirse ***"Specify"*** seçeneği , sınıflara eşit ağırlık verilemek istenilirse "***Unit***" seçeneği işaretlenir.



Şekil 2.5 RandomForests programının sınıf ağırlıkları ekranı

"***Rnd Forests"*** ekranından , "***Oluşturulacak Ağaç Sayısı***", her düğümde rastgele seçilecek "***Değişken Sayısı***", "***Bootstrap Örnekleme Sayısı***" ve diğer parametreler girilerek "***Start"*** butonuna basıldığında Random Forests modeli kurulmaya başlanır.



Şekil 2.6 RandomForests programının parametre girme ekranı

Model kurulduktan sonra modeli kaydetmek için ***"Save Grove"*** butonu seçilir. Kurulan model ile yeni bir veri değerlendirilmek istendiğinde ana pencereden ***"Model"*** butonu menüsü ve "***Score Data"*** seçeneği seçilmelidir.

3. BULGULAR

Çalışmamızda 175 deneğin sistemik, demografik, dental ve serolojik özelliklerine ilişkin 43 değişkeninden veriler toplanmıştır. Bağımlı değişken, deneklerin bulunduğu "PER+AMI", "PER-AMI" ve "KONTROL" gruplarıdır. Gruplara düşen denek sayıları Çizelge 3.1'de verilmiştir.

Çizelge 3.1 Sınıflara düşen veri sayıları

Sınıf Değeri	Sınıflara Düşen Denek sayısı
PER+AMI	80
PER-AMI	80
Kontrol Grubu	15
Toplam	175

Veri setini oluşturan tüm değişkenlere ait tanımlayıcı istatistikler Çizelge 3.2'de verilmiştir.

Çizelge 3.2. Veri setindeki değişkenlerin özet istatistikleri

Değişken	N	Ortalama	Std.Sapma	Min	Max
AGE	175	51.240	6.699	35.0	73.0
CRP	160	11.685	21.128	0.0	126.0
CHOLMGDL	160	198.331	41.708	104.0	377.0
LDLMGDL	160	125.635	35.190	32.0	248.0
HDLMGDL	160	44.625	10.767	4.0	91.0
TGMGDL	160	147.031	83.203	28.0	562.0
GLUCMGDL	160	118.736	60.678	11.0	497.0
BMI	175	26.413	3.383	18.65	40.460
WBC	160	12.947	9.548	4.69	59.3
NTEETH	175	22.560	3.648	14.0	28.0
NEXTRACT	175	5.491	3.662	0.0	14.0
NTEETHPD	175	22.560	3.648	14.0	28.0
V21_A	160	12.669	3.170	5.0	20.0
V22_A	160	10.956	4.390	1.0	22.0
NTEETHCA	160	16.200	3.733	6.0	23.0
V24_A	160	13.369	3.936	2.0	25.0

NSITESPD	160	26.400	11.008	8.0	65.0
V26_A	160	16.538	6.945	2.0	44.0
NSITESCA	160	29.831	10.818	10.0	68.0
V28_A	160	20.388	6.841	6.0	46.0
PDMEAN	175	3.204	0.850	1.7	5.9
CALMEAN	175	4.168	1.206	1.7	8.1
BOPMEAN	175	0.367	0.216	0.1	1.0
LIMULUSL	148	0.596	0.624	0.061	4.638
AAPAL100	55	0.560	0.537	0.038	2.226
AAPAL200	55	0.544	0.577	0.050	2.628
AALPS150	175	1.865	1.169	0.048	5.495
AALPS300	175	2.216	2.129	0.031	11.923
AAOMP100	175	1.533	0.743	0.309	4.071
AAOMP200	175	1.761	1.026	0.394	6.796
PGLPS100	175	1.864	0.844	0.350	4.606
PGLPS200	175	1.844	0.937	0.386	5.076
PGOMP200	175	1.551	0.574	0.398	2.802
PGOMP400	175	1.691	0.824	0.378	4.664

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 yapılacaktır. Periodontoloji veri setindeki örnek sayısı çok fazla olmadığı için, bootstrap örneklemedeki örnek sayısı, orijinal veri setinin örnek sayısı olan 175 olarak seçilmiştir.

Sınıflara düşen denek sayıları birbirine yakın olmadığından sınıflara düşecek denek sayılarını dengelemek için gerekli ağırlıklar **"Class Wts"** menüsünden **"BALANCED**" seçeneği işaretlenmiş ve Çizelge 3.3'deki değerler elde edilmiştir.

Çizelge 3.3 Sınıf değişkenlerini dengelemek için uygulanacak ağırlıklar

Hedef sınıf	**Ağırlık**
PER+AMI	1
PER-AMI	1
Kontrol grubu	5,333

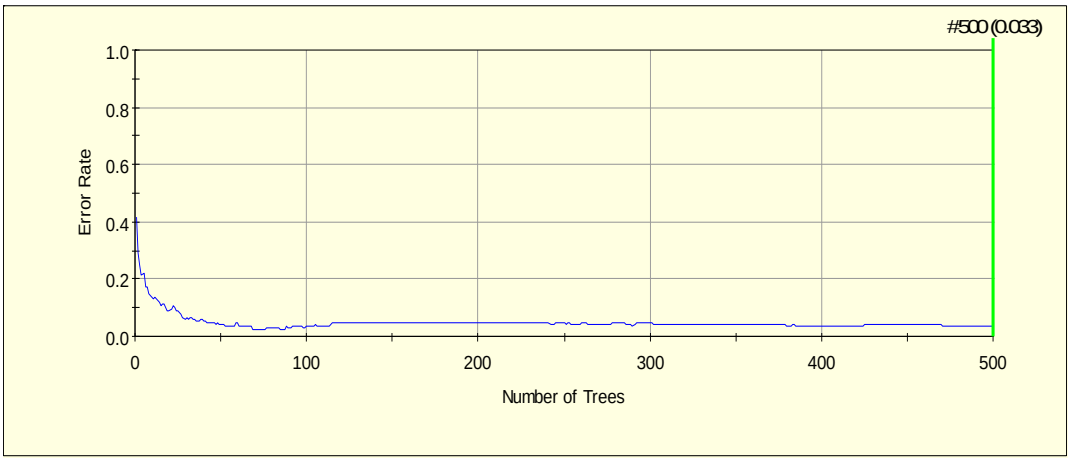
Her düğümde rastgele seçilecek değişken sayısı, toplam değişken sayısının karekökünün tam sayıya yuvarlandığı yöntem kullanılarak belirlenmiştir. Bu

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.

Ağaç sayılarına göre, ormanın hata oranı Şekil 3.1'te 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.



Şekil 3.1. 500 karar ağacından oluşan karar ormanın genel hata oranı

Periodontoloji veri setindeki sınıf değişkenleri PER+AMI, PER-AMI ve Kontrol grubu içinse hata oranları sırasıyla; 0.0625, 0.0375 ve 0.0 olarak hesaplanmıştır.

Yapılan sınıflama sonucunda, doğru ve yanlış sınıflandırılan denek sayıları, doğru ve yanlış sınıflama oranları Çizelge 3.4'de gösterilmiştir. Tabloda 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.

Çizelge 3.4. 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

$$\text{PER+AMI için hata oranı} = \frac{5}{80} = 0{,}0625 \tag{3.1}$$

$$\text{PER-AMI için hata oranı} = \frac{3}{80} = 0{,}0375 \tag{3.2}$$

$$\text{Kontrol grubu için hata oranı} = \frac{15}{15} = 0{,}0 \tag{3.3}$$

Random Forests algoritması sınıflardaki farklı denek sayılarını dengelemek için Çizelge 3.3'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 Çizelge 3.3'de verilen ağırlığa göre Eşitlik 3.4'te ki gibi ağırlıklı toplam denek sayısına bölmektedir.

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

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'da ki gibi bulunacaktı. Bulunan bu değer modelin kurulmasındaki varsayımları dikkate almadığından modelin hata oranını doğru yansıtmayabilir.

$$\text{Sınıflama hata oranı} = \frac{8}{175} = 0{,}457 \tag{3.5}$$

Bu çalışmada yapılan tüm değerlendirmeler ve karşılaştırmalar, Eşitlik 3.4'te verilen hata oranı hesaplama yöntemine göre yapılmıştır.

Yapılan sınıflama sonucunda, Çizelge 3.5'deki sınıflama çizelgesi elde edilmiştir. Modelin doğru sınıflama oranı, çizelgedeki doğru sınıflanan deneklerin sayısının toplam denek sayısına bölünmesi ile elde edilmiştir.

Çizelge 3.5 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			

Ormanın doğru sınıflama oranı= $\frac{75+77+15}{175}$ =0,9543

PER+AMI doğru sınıflama oranı = $\frac{75}{80}$ = 0,9375

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

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

RF yöntemi, değişken önem derecesini iki farklı yöntemle hesaplamaktadır. Bu yöntemler, önem derecesi hesaplanan değişkenin tüm değerlerinin yerlerinin değiştirilmesi ile hesaplanan standart yöntem ve gini önem derecesi ile hesaplanan gini yöntemidir. Random Forests programı yardımıyla, değişkenlerin standart yöntemle ve gini yöntemiyle önem derecesi hesaplanmıştır. Modele katkısı olan değişkenler yüksek önem derecesi skoru alırken, katkısı az olan değişkenler daha düşük önem derecesi skorları almaktadır. Çizelge 3.6a'da değişkenlerin gini yöntemiyle hesaplanan önem derecelerine ait skorlar gösterilmiştir. Çizelge 3.6b'de

ise değişkenlerin standart yöntemle hesaplanan önem derecelerine ait skorlar gösterilmiştir.

Çizelge 3.6a 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		

Çizelge 3.6a'ye göre CALMEAN, PDMEAN, WBC değişkenleri sınıflara ayırmada en çok bilgi sağlayan ilk 3 değişken olarak görülmektedir.

Çizelge 3.6b 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		

Çizelge 3.6b'ye 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. Çizelge 3.6a'da ve Çizelge 3.6b'de önem derecesine göre en önemli bulunan ilk 10 değişken ile yeniden model kurulmuş ve bu modellere ilişkin hata oranları Çizelge 3.7'de verilmiştir.

Çizelge 3.7 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

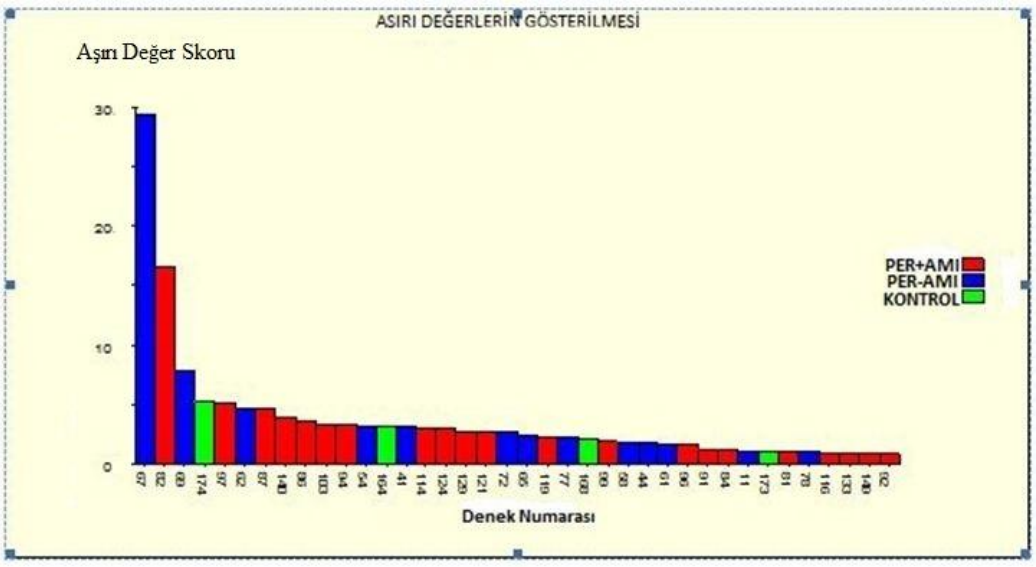
Çizelge 3.7'ye göre gini yöntemi kullanılarak en önemli ilk 10 değişken modele alındığında 700 ağaç oluşturmak en iyi sonuç vermektedir. Standart yöntemle belirlenen en önemli ilk 10 değişken modele alındığında 900 ağaç daha iyi bir sonuç vermektedir. Standart yöntemle belirlenen en önemli ilk 10 değişkenle 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 periodontoloji veri seti için değişken önem derecesinin standart yöntemle hesaplanması önerilebilir.

Çizelge 3.8'de farklı değişken sayılarına göre 500 ve 1000 ağaçlık ormanların hata oranları gösterilmiştir. Çizelgede görüldüğü gibi hata oranları bir çok değişken sayısı için aynı kalmaktadır. Bu sonuca göre değişken sayısının model üzerinde önemli bir etkisinin olmadığı söylenebilir.

Çizelge 3.8 Değişken sayılarına göre ormanın hata oranları

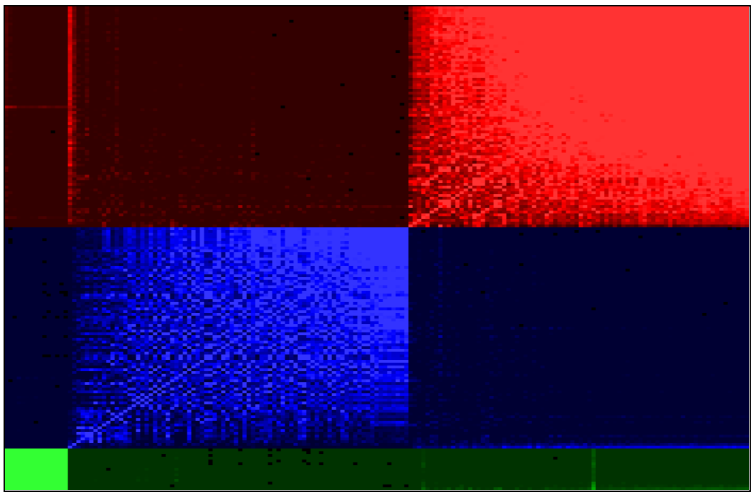
Ağaç Sayısı	Değişken Sayısı	Hata Oranı	Ağaç Sayısı	Değişken Sayısı	Hata Oranı
500	3	0,033	1000	3	0,033
500	7	0,033	1000	7	0,042
500	14	0,046	1000	14	0,042
500	21	0,033	1000	21	0,033
500	28	0,033	1000	28	0,038
500	35	0,033	1000	35	0,033

RF yöntemi ile model kurulduktan sonra veri setindeki hangi değerlerin aşırı değer (outlier) oldukları benzerlik veya farklılık ölçüleri yardımıyla hesaplanır. Her bir bireyin ait olduğu gruptaki diğer deneklerden ortalama uzaklığı hesaplanır. Daha sonra bu uzaklık değerleri ortanca değer ve interquartile range kullanılarak standardize edilmiş uzaklık değerleri elde edilir. 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 3.2'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 3.3'e 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 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.



Şekil 3.2. Aşırı değer olan deneklerin gösterilmesi.

RF yönteminin önemli diğer özelliği ise kümeleme yapabilmesidir. RF yönteminde proximity matrisi ile deneklerin birbirlerine olan yakınlığı ölçülebilmektedir ve ölçüm değerleri birbirine yakın olan denekler bir gruba toplanarak kümele yapılabilmektedir. Şekil 3.3'te RF yöntemine göre yapılan hesaplamalar sonucu proximity yoğunluk haritası gösterilmiştir. Grafikte kırmızı, mavi, yeşil olan renklerle 3 ayrı küme yani üç ayrı sınıf olduğu anlaşılmaktadır



Şekil 3.3. Proximity yoğunluk haritası (Proximity heat map)

Random Forests programı ile yapılan sınıflama başarısı sonuçları yukarıda verilmiştir. Random Forests programı veri setindeki deneklerin hangi sınıfa düştüğünü tek tek göstermektedir. Ayrıca ağaçların yapmış olduğu oyların neticesine göre ilgili deneğin sınıflara hangi olasılıkla düşebileceğini hesaplamaktadır. Periodontoloji veri setindeki deneklerin bir bölümünün hangi sınıfa hangi olasılıkla düşmüş olabileceği Çizelgede 3.9.'de verilmiştir.

Çizelge 3.9. Veri setindeki deneklerin sınıf tahminleri

Denek No	Yapılan Tahmin	PER+AMI sınıfında olma olasılığı	PER-AMI sınıfında olma olasılığı	Kontrol Grubu Olma Olasılığı	Gerçek Sınıf
1	PER-AMI	0,09	0,91	0,00	PER-AMI
2	PER-AMI	0,10	0,82	0,08	PER-AMI
3	PER-AMI	0,08	0,86	0,06	PER-AMI
4	PER-AMI	0,09	0,82	0,09	PER-AMI
5	PER-AMI	0,07	0,92	0,01	PER-AMI
6	PER-AMI	0,13	0,70	0,17	PER-AMI
7	PER-AMI	0,06	0,82	0,11	PER-AMI
8	PER-AMI	0,10	0,86	0,04	PER-AMI
9	PER-AMI	0,13	0,79	0,08	PER-AMI
10	PER-AMI	0,07	0,88	0,05	PER-AMI
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
167	KONTROL	0,13	0,06	0,81	KONTROL
168	KONTROL	0,16	0,05	0,80	KONTROL
169	KONTROL	0,13	0,06	0,81	KONTROL
170	KONTROL	0,11	0,07	0,82	KONTROL
171	KONTROL	0,14	0,05	0,81	KONTROL
172	KONTROL	0,13	0,06	0,81	KONTROL
173	KONTROL	0,10	0,08	0,82	KONTROL
174	KONTROL	0,14	0,07	0,79	KONTROL
175	KONTROL	0,12	0,06	0,82	KONTROL

3.1 RF Yöntemi ve Bagging yönteminin karşılaştırılması

Bagging, bootstrap yöntemiyle seçilen örneklemlerle oluşturulan çok sayıda karar ağacının yapmış olduğu tahminlerin toplayarak nihai sınıf tahmini yapan bir yöntemdir. orijinal veri setindeki tüm değişkenleri kurulan ağaç yapısında 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 Çizelge 3.10'da görüldüğü gibi hesaplanmaktadır. Bagging yöntemi uygulandığında modelin hata oranı 0,054 olarak bulunmaktadır.

Çizelge 3.10 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

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.

3.2 RF Yöntemi ve CART yönteminin karşılaştırılması

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.

Periodontoloji veri seti için CART yöntemi ile sınıflama yapılmış ve modelin hata oranı Çizelge 3.11'de gösterildiği gibi 0,0875 olarak bulunmuştur.

Çizelge 3.11. 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

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.

4.TARTIŞMA

Bu tez çalışmasında, 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 (Lempitsky,V.ve ark. ,2009) Çalışmamızda örnek veri setine Random Forests 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 Random Forests 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(2001), RF yönteminin, diskriminant analizi, destek vektör makineleri ve yapay sinir ağları gibi sınıflandırıcılar dahil birçok diğer sınıflandırıcılara göre de daha iyi sonuçlar verdiğini bunun yanında yöntemin aşırı öğrenmeye karşı da oldukça güçlü olduğunu belirtmiştir. Son yıllarda RF yöntemi araştırmacılar tarafından, çok farklı veri setleri kullanılarak çeşitli yöntemlerle karşılaştırılmıştır. Clark ve ark.(2009) 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. Wu ve ark.(2003) 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 Random Forest yöntemlerinin sınıflama başarılarını karşılaştırmış ve Random Forests yönteminin diğerlerinden daha başarılı sonuçlar verdiği gözlemlemişlerdir.

Xu ve ark.(2009), terörist profili çıkarmak ve veri setlerindeki bilgilerden yola çıkarak terörist olabilecek kişileri tespit etmek için Random Forests yöntemiyle sınıflama yapmışlar ve bulunan sonucu standart tek bir karar ağacı ve fuzzy kümeleme yöntemi ile elde edilen sonuçlarla karşılaştırmışlardır. Bu çalışmada Random Forests yönteminin diğer iki yöntemden daha başarılı olduğu sonucuna varılmıştır. Yoonhee ve ark.(2007), 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. Statnikov ve ark.(2008), kanser hastalığına sebep olan en önemli genlerin tespit edilmesi için yaptıkları mikroarray çalışmasında Random Forests ve destek vektör makineleri yöntemlerinin sonuçlarını karşılaştırmışlar ve destek vektör makinelerinin Random Forests yönteminden daha iyi performans gösterdiğini ortaya koymuşlardır. Ruiz-Gazen ve ark.(2007), sınıf dağılımları dengesiz olan meteoroloji veri seti için lojistik regresyon ve Random Forests 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 Random Forests yöntemine göre daha hızlı olduğunu ve yorumlanmasının daha kolay olduğu için meteorolojistler için tercih edilebileceğini söylemişlerdir. O'Leary ve ark.(2009), lösemi hastalığı ile ilgili veri setleri için bayesyen sınıflama ağaçları(BCART) ile Random Forests yöntemini karşılaştırmışlardır. BCART yönteminin üç ayrı veri seti içinde hastalığa neden olan önemli genlerin tespitinde ve sınıflama tahmininde Random Forests yönteminden daha başarılı olduğunu ortaya koymuşlardır.

RF yöntemi bireysel oluşturulan ağaçların sonuçlarını toplayan bir topluluk yöntemdir. Uygulama bölümündeki 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. Değişken sayısı 3 veya 7 olarak belirlendiğinde ormanın hata oranı % 3,33 olduğu gözlemlenmiştir. Değişken sayısı 14 seçildiğinde hata oranının % 4,6 olduğu gözlemlenmiştir. 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 olarak, gini yöntemiyle hesaplanan önem derecesine göre ise % 5,8 olarak bulunmuştur. Tüm değişkenlerin modele girdiği durumda ise hata oranı %3,33 olarak bulunmuştur. 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, aşırı değerle ait ölçümler klinik olarak incelenebilir. Yapılan inceleme sonucunda aşırı değerlerin veri setinden çıkartılmasına veya kalmasına karar verilebilir. Eğer çıkartılmasına karar verilirse, model yeniden kurulmalı ve analizler yeniden yapılmalıdır. Uygulama veri setinde 67., 82. ve 69. denekler aşırı değer olarak bulunmuştur.

5. 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 metodu, bir topluluk yöntem olmasına rağmen, topluluk yöntemlerden farklı olarak modele ayrı bir katman olan rastgelelik 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 değişken önem derecesi hesaplayarak, tanı koymada hangi değişkenlerin en önemli rol oynadıklarını tespit ederek, daha az sayıda tetkik yapılmasını önerebilir. RF yöntemi özelikle çok sayıda değişkenin (genler) olduğu microarray çalışmalarında ve DNA veri seti gibi binlerce gen arasından önemli olanları tespit etmek için yapılan çalışmalarda kullanılabilir (Archer, K.,J., Kimes, R.,V.,2008).

Tez çalışmasında 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.

ÖZET

Veri Madenciliğine Genel Bakış ve Random Forests Yönteminin İncelenmesi: Sağlık Alanında Bir Uygulama

Karar vericilere, eldeki verilerden yola çıkarak doğru ve etkin kararlar almasına yardımcı olmak amacıyla veri madenciği yapılmaktadır. 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 tez çalışmasında öncelikle veri madenciliği yöntemleri genel olarak tanıtılmış, veri madenciliğinde önemli yer tutan ve sınıflama modellerinden olan karar ağaçları anlatılmıştır. Ayrıca ağaç tabanlı yöntemlerden olan Random Forests (RF) yöntemi incelenmiş ve periodontoloji bilim dalından elde edilen bir veri seti üzerinde uygulaması yapılmıştır.

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. Bu özelliği sebebiyle RF yöntemi oldukça başarılı sonuçlar vermektedir.

RF yöntemiyle % 95,4 oranında başarılı bir sınıflama yapılmıştır. Oluşturulan karar ormanının hata oranı ise % 3,33 olarak bulunmuştur. Aynı veri seti için Bagging ve CART yöntemi ile de sınıflama yapılmıştır. Bagging yöntemi ile hata oranı % 5,4 , CART yöntemi ile % 8,75 olarak bulunmuştur.

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

SUMMARY

An Overview of Data Mining Techniques and Analysis of Random Forests Method: An Application On Medical Field

Data Mining is processed in order to help policy makers for giving valid and efficient decisions using the available data on the subject. In general, data mining has descriptive and predictive perspectives. In medicine, especially its predictive aspects are used.

Within this thesis study, data mining techniques are introduced briefly. Further, decision trees, which has an important place in data mining, are explained. Also, tree-based data mining method Random Forests (RF) is analyzed and applied on periodontology data set.

In RF method, decision trees which form decision forest are created with different data sets. These data sets are bootstrapped samples from original data set. 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. Using these properties RF gives fairly good results.

Using RF method, 95,4 % of successful classification rate is achieved. Decision Forest's error rate was found 3,33 % . Classification was made by Bagging method and CART method for the same data set and the error rates were found 5,4 % and 8,75 % respectively.

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 ones from large amount of DNA data set which has thousands of predictors(genes).

Key Words: Data Mining, Classification, Random Forests, Decision Trees, Decision Forest

KAYNAKLAR

AKPINAR, H. (2000). Veri tabanlarında bilgi keşfi ve veri madenciliği, *İ.Ü İşletme Fakültesi dergisi*, c.**29**, s.1,2000

ALPAYDIN, E. (2004). Introduction To Machine Learning. The MIT Press s.:7.

ALPAYDIN, E. Zeki Veri Madenciligi: Ham veriden altın bilgiye ulaşma yöntemleri , Erişim: [www.cmpe.boun.edu.tr/~ethem/files/papers/veri-maden_2k-notlar.doc]. Erişim Tarihi:20.08.2009.

AMASYALI,M.F. (2008). Yeni Makine Öğrenmesi Teknikleri ve İlaç Tasarımına Uygulaması, yayınlanmamış doktora tezi.

ANDY, L., MATTHEW W. (2002). Classification and Regression by RandomForest, *R News* Vol. **2/3**,s.18-22.

ARCHER, KELLIE J., KIMES, RYAN V. (2008). Empirical characterization of random forest variable importance measures, *Computational Statistics & Data Analysis*, Volume **52**, Issue 4, 10 Ocak 2008, s.:2249-2260

BREIMAN, L. (2001). Random Forests. *Kluwer Academic Publishers*., **45**,s.:1-30.

BREIMAN, L. (2004) Manual on setting up, using, and understanding random forests. Erişim: [http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm].Son Güncelleme: 3 Mart 2004. Erişim Tarihi:02.10.2009.

CLARKE, B., FOKOUE, E., ZHANG, H.,H. (2009) Principles and Theory for Data Mining and Machine Learning, Springer Series In Statistics.

DE VEAUX, R., D. (2009). Data Mining –Five Lessons Learned in the Pit, Erişim: [http://www.williams.edu/Mathematics/rdeveaux/lessons.pdf]. Erişim Tarihi:15.10.2009.

HAN, J. (2004). Data Mining, Concepts and Techniques, San Francisco, CA, Morgan Kaufmann Publishers, s.:7.

KDNUGGETS.COM. (2008). Veri madenciliği uygulama alanları, Erişim: [http://www.kdnuggets.com/polls/2008/data-mining-applications.htm]. Erişim Tarihi:28.09.2009.

KDNUGGETS.COM. (2007). En çok kullanılan veri madenciliği metotları, Erişim: [Http://www.kdnuggets.com/polls/2007/data_mining_methods.htm]. Erişim Tarihi:28.09.2009.

LAROSE,D.T. (2004). Discovering Knowledge In Data. An Introduction To Data Mining, A John Wiley & Sons, Inc. Publication.

LEMPITSKY,V., VERHOEK, M., NOBLE,A., BLAKE. (2009). A. Random forest classification for automatic delineation of myocardium in real-time 3d echocardiography. Erişim: [http://research.microsoft.com/pubs/81125/fimh.pdf]. Erişim Tarihi:02.10.2009

NATHAN P. (2005). Enhancing Random Forest Implementation in Weka, Machine Learning Conference Paper for ECE591Q.

O'LEARY,R.A., FRANCİS R.W., CARTER K.W., FİRTH M.J., KEES U.R.,

KLERK N.H.,(2009), A comparison of Bayesian classification tress and random forest to identify classifiers for childhood leukaemia.18th World IMACS/MODSIM Congress,Australia,13-17 July 2009

OMITAOMU,O. (2006). Lecture Notes In Data Mining, World Scientific Publishing

Co. Pte. Ltd., s.:39.

RUIZ-GAZEN A., VILLA N.,(2007) Storms Prediction: Logistic Regression Vs Random Forest For Unbalanced Data, *Case Studies in Business, Industry and Government Statistics* **1, 2** *(2007) s:91-101*

SACCHI, M.D. (1998). A bootstrap procedure for high-resolution velocity analysis. Geophysics, vol:**63**, no.5.

SİLAHTAROĞLU,G. (2008). Kavram ve Algoritmalarıyla Temel Veri Madenciliği,Papatya Yayıncılık Eğitim.s.29-60.

STATNIKOV A., WANG L.,F ALIFERIS C.,(2008), A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification, *BMC Bioinformatics* 2008, Vol.**9**:3

TAN,P., STEINBACH, M., KUMAR,V. (2005). Introduction To Data Mining, Addison Wesley,s.:152.

WU B., ABBOTT T., FISHMAN D., MCMURRAY W., MOR G., STONE K.,
WARD D., WILLIAMS K., ZHAO H.,(2003), Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data., *Bioinformatics.* 2003 Sep 1;**19(13),s:**1636-43.

XU J, CHEN J,LI B.,(2009), Random forest for relational classification with application to terrorist profiling, *Granular Computing, GRC apos;2009.*

YAKUPOĞLU,Ç.,ATIL,H. Bootstrap metodu ve uygulanışı üzerine bir çalışma, Erişim: [http://dondurma.ksu.edu.tr/sempkonf/tbt4s_Cd/tbt4s_ozet/print/atil.doc]. Erişim Tarihi:10.10.2009

YOONHEE K., HO K.,(2007), Application of Random Forests to Association Studies Using Mitochondrial Single Nucleotide Polymorphisms, *Genomics & Informatics* Vol. **5(4)** 168-173, December 2007

ZHOU,Z. (2003). Three Perspectives Of Data Mining, Elsevier Science Publishers Ltd. Essex, UK, s.:139-146.

ÖZGEÇMİŞ

- **Bireysel Bilgiler**

Adı ve Soyadı: : Muhammet AKMAN
Doğum Yeri ve Tarihi : Kdz. Ereğli,12.03.1976
Uyruğu : T.C.
Medeni Durumu : Evli
Askerlik Durumu : Yaptı
E-posta : muhakman@yahoo.com, makman@ssm.gov.tr
Telefon :506 4757625

- **Eğitim**

Ankara Üniversitesi / Biyoistatistik / Yüksek Lisans (2010)
Orta Doğu Teknik Üniversitesi / İstatistik / Lisans (1999)
Kdz.Ereğli Anadolu Lisesi (1993)
Yabancı Dil: İngilizce- KPDS:84

- **Ünvanlar**

Bilgi Yönetimi Uzmanı
Çözümleyici
Programcı

- **Mesleki Deneyim**

Savunma Sanayii Müsteşarlığı- Bilgi Yönetimi uzmanı (2007-
Gümrük Müsteşarlığı- Çözümleyici (2001-2007)
İçişleri Bakanlığı- Programcı (2000-2001)

- **Bilimsel İlgi Alanları**

İstatistik, Biyoistatistik,Veri Madenciliği,Bilgisayar Bilimleri

- **Bilimsel Etkinlikler**

1. Genç Y., Öztuna D., Akman M. "Kümelenmiş Verilerde Oranların Karşılaştırılması: İteratif Olmayan Yöntemleri İçeren Programın Tanıtılması", 5.İstatistik Kongresi, Antalya, 2007

2. Genç, Y., Ateş, C., Öztuna, D., Gültekin, S., Tüccar, E., Akman, M. "Kümelenmiş Verilerde İşlem Karakteristiği Eğrisi (İKE) Altında Kalan Alanın Tahmini", 10. Ulusal Biyoistatistik Kongresi, Sivas, 2007