T.C.
İstanbul Üniversitesi
Sosyal Bilimler Enstitüsü
Ekonometri Anabilim Dalı

Yüksek Lisans Tezi

Veri Madenciliğinde Kullanılan Sınıflandırma Yöntemleri ve Bir Uygulama

Mine Çelik
2501060294

Tez Danışmanı
Doç. Dr. Enis Sınıksaran

İstanbul 2009

T.C
İSTANBUL ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ

TEZ ONAYI

Enstitümüz **EKONOMETRİ** Anabilim Dalında **2501060294** numaralı **MİNE ÇELİK'İN** hazırladığı **"VERİ MADENCİLİĞİNDE KULLANILAN SINIFLANDIRMA YÖNTEMLERİ VE BİR UYGULAMA"** konulu **YÜKSEK LİSANS/** ~~**DOKTORA TEZİ**~~ ile ilgili **TEZ SAVUNMA SINAVI,** Lisansüstü Öğretim Yönetmeliği'nin 15.Maddesi uyarınca **11/11/2009 ÇARŞAMBA** günü **Saat: 10:00'da** yapılmış, sorulan sorulara alınan cevaplar sonunda adayın tezinin .....Kabul.....'ne* **OYBİRLİĞİ /**~~**OYÇOKLUĞUYLA**~~ karar verilmiştir.

JÜRİ ÜYESİ	KANAATİ(*)	İMZA
PROF.DR.KARUN NEMLİOĞLU	Kabul	
DOÇ.DR.HAKAN ONGAN	Kabul	
DOÇ.DR.ENİS SINIKSARAN	Kabul	
YRD.DOÇ.DR.AYLİN AKTÜKÜN	Kabul	
YRD.DOÇ.DR.MÜRÜVVET PAMUK	Kabul	

Besim Ömer Paşa Caddesi Beyazıt-34452-İstanbul
Tel:0212 440 00 00 –14221 – Fax: 0212 440 03 40

Veri Madenciliğinde Kullanılan Sınıflandırma Yöntemleri ve Bir Uygulama

Mine Çelik

ÖZ

Günümüzde, gelişen teknoloji ile birlikte elde tutulan veri miktarı artmış, saklanan ham veriyi bilgiye dönüştürmek geçmişe göre daha da önem kazanmıştır. Verinin bilgiye dönüşümü, karı arttırarak rekabetçi ortama daha kolay uyum sağlamaktan veri sahibini anlamaya kadar birçok avantaj sağlayabilmektedir. Veri madenciliği ham veriyi bilgiye dönüştürmede istatistiksel yöntemleri ve makine öğrenme algoritmalarını kullanan bir araçtır. Bu çalışmada veri madenciliğinde sınıflandırma yöntemleri incelenmiş, bir yardım derneğinden alınan veriler üzerinde, yardım edilme kararını etkileyen faktörleri inceleyen ve yardım kararı alınmasını model kurarak otomatize etme amacı güden bir tahmin modeli geliştirilmeye çalışılmıştır.

ABSTRACT

Nowadays, with the development of technology, the amount of data which is kept has increased and transforming the data into knowledge has become more important than the past. Transforming the data into the knowledge has a lot of advantages from accomodating oneself to competitive atmosphere by increasing the profit to understand the owner of the data. Data Mining is an instrument that uses statistical methods and machine learning algorithms to transform data into knowledge. In this paper, the classification methods of Data Mining are investigated and a forecasting model which analyses the factors affect giving assistancy and aims to automatize the decision of aiding using data taken from a charity house is developed.

ÖNSÖZ

Veri madenciliği, ilk tanıştığım günden bu yana bana, günlük yaşamı teori ile birleştirmekte daha esnek olduğum, hayattaki problemleri tanımlayarak, çözümleri pratiğe dökebilmeye daha yakın durduğum hissini veren bir alan. Bu anlamda beni özgürleştirdiğini, her öğrendiğim parçasında da beni yeniden heyecanlandırdığını düşünüyorum. Bu çalışmanın oluşması ise bu konuda bitmeyen öğrenme isteğimden kaynaklanmaktadır.

Çalışmada veri madenciliğinin ne olduğu anlatılmıştır. Çalışmanın amacı sınıflandırma yöntemlerinin incelenmesidir. Bununla birlikte veri madenciliğinde kullanılan diğer yöntemler de kısaca açıklanmıştır. Çalışma üç bölümden oluşmaktadır.

Birinci bölümde veri madenciliğinin tanımı yapılmış, neye hizmet ettiği ve süreçte karşılaşılabilecek kavramlar açıklanmış ve pratikte hangi alanlarda uygulandığı ve ne gibi uygulamalar yapıldığı anlatılmıştır. Veri madenciliği algoritmalarını kullanan ve bu alana yönelik tasarlanmış paket programlara da değinilmiştir.

İkinci bölümde veri madenciliği sürecinde kullanılan yöntemler tahminleyici ve tanımlayıcı olarak ayrılmış ve çalışma şekilleri ve özellikleri kısaca açıklanmıştır.

Üçüncü bölümde Deniz Feneri Derneği'nden alınan veriler kullanılarak, dernek tarafından ailelere yapılan yardımlar ile ilgili karar verici bir model geliştirilmesi amaçlanmıştır. Bu sebeple Yapay Sinir Ağı ve Karar Ağacı algoritmaları ile birlikte Lojistik Regresyon denenmiş ve sonuçları açıklanmıştır.

Çalışmam süresince bana yapıcı tutumu ile destek olan danışmanım Doç. Dr. Enis Sınıksaran'a ve her zaman her konuda yardımını esirgemeyen arkadaşım Elçin Timur Çakmak'a teşekkür ederim.

İÇİNDEKİLER

TABLO LİSTESİ

ŞEKİL LİSTESİ

GİRİŞ

Bilgisayar teknolojilerindeki gelişmeler ve bilgisayar donanımının ucuzlaması, büyük boyutlu verilerin depolanabilmesine olanak tanımıştır. Büyük veri tabanlarında saklanan bu verilerin kullanımı ile veri tabanlarında bilgi keşfi kavramı ortaya çıkmıştır. Veri madenciliği, istatistiksel yöntemler ile çeşitli bilgisayar algoritmalarını kullanarak veri tabanlarındaki veriden, bu anlamlı ve işe yarar bilginin çıkarımını ifade eden süreçtir. Bu sürecin ve kullanılan sınıflayıcı yöntem ve algoritmaların anlaşılması, çalışmanın ana konusunu oluşturmaktadır.

Çalışmanın ilk bölümünde, veri ambarları, veri tabanları, model ve öğrenme çeşitleri gibi temel kavramlar ile birlikte, veri hazırlama, modelin kurulması, değerlendirilmesi gibi veri madenciliği süreçleri açıklanmıştır. Ayrıca bankacılık-finans, perakende ve telekomünikasyon sektörlerindeki pratik uygulamaları ve bu uygulamalarla firmaların neler elde ettiği de anlatılmıştır. Genel olarak kullanılan paket programlar ve bu programların kullandığı algoritmalar da bu bölümde yer almaktadır.

İkinci bölüm daha kapsamlı olarak, veri madenciliğinin tanımlayıcı ve tahminleyici yöntemlerinin anlatımını içermektedir. Bu yöntemler, farklı amacı ve çıktısı olan modeller kurmak için kullanılmaktadır. Bu bağlamda, tahmin edici yöntemlerden karar ağaçları, yapay sinir ağları, regresyon yöntemleri, genetik algoritmalar ve karar destek makinaları, tanımlayıcı yöntemlerden kümeleme analizleri, birliktelik ve sıralı örüntü analizi, uç değer analizi açıklanmıştır. Regresyon, doğrusal regresyon ve lojistik regresyonu içermektedir. Karar ağaçlarının çalışma şekli, ağaçlardaki ayırma ve durma kriterleri anlatılmış, Chaid ve C&R tree algoritmaları da kısaca ele alınmıştır. Yapay sinir ağlarında ise yapay sinir ağının mimarisi ve sinir ağında öğrenme süreci ile ilgilenilmiştir. Karar destek makinaları, zaman serileri ve genetik algoritmalar tanımlar halinde verilmiştir. Kümelemenin tanımı yapılmış, kümelemede kullanılan algoritmalar gruplandırılmış, birliktelik ve sıralı örüntü analizi ve uç değer analizlerinin de mantığı kısaca verilmiştir.

Uygulama bölümünde Deniz Feneri Derneği'nin operasyonel veri tabanından Ağustos – 2008 tarihinde alınan veriler kullanılarak uygulama yapılmıştır. Veriler, Adana, Ankara, Samsun, İstanbul, İzmir ve Erzurum illerinden yardım talep eden ailelerin bilgilerini içermektedir. Bu bilgiler kullanılarak yardım kararını tahminleyen modeller geliştirilmiş ve karşılaştırmaları yapılmıştır. Modelin kurulmasında amaç, yardım kararının verilmesinde etkili olan değişkenleri belirlemek ve sonrasında yapılabilecek ek bir çalışma ile ilgilenilen modelin, sisteme yeni gelen aile bilgisi için kullanılabilir hale gelmesine yardımcı olmaktır. Bu bağlamda lojistik regresyon, yapay sinir ağları, karar ağaçları denenmiş ve sonuçları verilmiş, aralarında kıyaslama yapılmıştır.

1. VERİ MADENCİLİĞİNE GENEL BAKIŞ

Veri madenciliği, disiplinler arası bir alandır ve ham veriden kullanışlı olabilecek bilgiyi çıkarmak için gerekli yöntemler ile ilgilenmektedir. Veri madenciliğinde kullanılan yöntemlerin birçoğu iki ayrı araştırma dalı olan istatistik ve makine öğrenme olarak bilinmektedir.

Makine öğrenmenin geliştirilme amacı, veri türetme sürecine yardımcı olmak ve analistlere gözlenen verilerden gözlenmeyen olayları genelleyebilmelerine izin verebilen bir yapı oluşturabilmekti. İlk makine öğrenme modelini 1962 yılında Rosenblatt sunmuştur. Arkasından 1980'li yılların ikinci yarısında yapay sinir ağları geliştirilmiştir. Aynı dönemde bazı araştırmacılar karar ağacı teorisi ile ilgilenerek onları, sınıflandırma problemlerinde kullanılabilecek düzeye getirmişlerdir. İstatistiğin her dönemde modelleme için bir araç olduğu düşünülürse, 1980'li yılların ardından, bilgisayar teknolojilerinin de gelişmesi ile bilgisayarlı yöntemlerin istatistiksel analiz için önemi giderek artmıştır. 1990'lı yıllarda istatistikçiler makine öğrenme yöntemlerine de ilgi göstermişler, böylelikle metodolojinin gelişiminde büyük bir adım atılmıştır.[1]

1.1 Veri Madenciliğinin Tanımı

Temelleri klasik istatistiğe dayanan veri madenciliği, 1980'li yıllardan itibaren bilgisayarların da gelişmesi ile birlikte yapay zekâ ve makine öğrenme tekniklerini de içine katarak büyümüş, herhangi bir karar verme sürecine girdi hazırlayarak kullanımı ile sorunları daha anlaşılabilir hale getiren bir disiplin haline gelmiştir.

"Veri madenciliği, büyük ve karmaşık veri kümelerindeki ilişki ve örüntülerin açığa çıkarıldığı bir bilgi keşif sürecidir. Bu, belirli çıkarımları elde etmek için yapılan veri

[1] Paolo Guidici, **Applied Data Mining Statistical Methods for Business and Industry**, West Sussex, Willey 2003, s. 2

tutma gibi düşünülmemelidir.” Zira veri madenciliği, verinin ham haline bakarak birliktelikler ve kurallar çıkaran, iyi tanımlanmış algoritmalar kullanır.[2]

1.2. Genel Kavramlar

Karar Destek Sistemleri: Karar Destek Sistemleri, değişik kaynaklardan topladığı bilgileri düzenleyerek, kararı modelleyerek, bilgileri analiz ederek ve değerlendirme sonuçlarını sunarak karar vericiye seçim sırasında destek veren bilgisayar tabanlı sistemlerdir. “Bir karar verici için verilen kararın doğruluğu, onun yeteneklerine, deneyimine ve bilgi birikimine olduğu kadar sahip olduğu veri kümesinin yeterliliğine de bağlıdır. Diğer bir değişle kararın başarısında, verilerin doğru depolanması, doğru sınıflanması, doğru ayıklanıp işlenmesi ve doğru yorumlanması çok önemli bir rol oynar.” [3] Bu sebepten, veri madenciliği, Karar Destek Sistemleri için etkili bir araç olabilir.

Veri tabanları: Elde edilen verilerin tutulduğu alanlardır. “Bir veri tabanı sistemi, birbiri ile ilişkili verilerin birikimini içeren, veriye erişimi sağlayarak veriyi yönetmeye yardımcı olan yazılım programları kümesidir.”[4] Veri tabanları kullanım amaçlarına göre farklı isimler alırlar. Örnek olarak ilişkisel veri tabanları, işlemsel veritabanı, zaman serisi veritabanı verilebilir.

İlişkisel veritabanları, her biri farklı isimler alan tablolardan oluşur. Her tabloda her bir kaydın özelliklerinin değerlerini tutan alanlar ve her kayda ait bir tekil anahtar bulunur. Bir üniversitenin veritabanını ilişkisel veri tabanına örnek olarak verebiliriz. Zira her bir kişi için ayırt edici bir öğrenci numarası, hangi yılda kayıt yaptırdığı, hangi bölümde okuduğu gibi alanlar ile öğrenciye ait bilgiler saklanır. Buradan çeşitli sorgular ile hangi bölümde kaç öğrencinin okuduğu, geçtiğimiz yıl kaç kişinin belli bir bölüme kayıt yaptırdığı gibi soruların cevapları bulunabilir.

---

[2] Jing Luan, Terrence Willet, “Data Mining & Knowledge Management: A System Analysis for Establishing a Tiered Knowledge Management Model”, (Çevrimiçi), http://www.eric.ed.gov/ERICWebPortal/custom/portlets/recordDetails/, 04.Mayıs.2009

[3] Kaan Yaralıoğlu, **Uygulamada Karar Destek Yöntemleri,** İzmir, İlkem Ofset, 2004, s. 165

[4] Jiawei Han, Micheline Kamber, **Data Mining Concepts & Techniques,** San Francisco, Morgan Kauffmann Publishers 2006, s. 10

İşlemsel veritabanında her bir kaydın bir işlem olduğu varsayılır. Bir marketin veri tabanını düşünecek olursak, her an bir satış yapıldığını ve her bir satışın işlemsel veri tabanında bir kayıt olarak göründüğü varsayılabilir. Bu veritabanından, bugün, ilgilenilen üründen kaç tane satıldığı sorusunun cevabına ulaşılabilir.

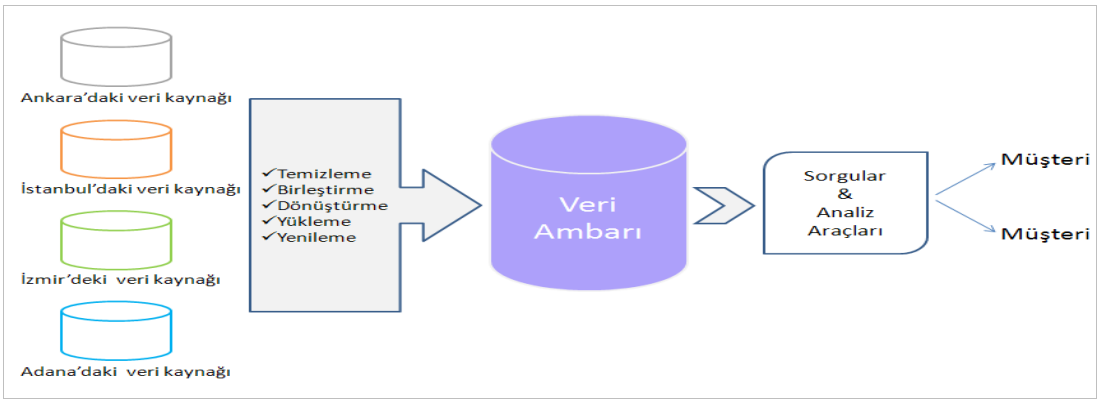
Zaman serisi veritabanı düzenli zaman aralıkları ile elde edilmiş (yıllık, haftalık, günlük) verilerin tutulduğu alanlardır. Örnek olarak borsa verilerinin, stok kontrolleri sonucu alınan verilerin, sıcaklık ölçümlerinden elde edilen verilerin depolanması gösterilebilir.

Veri Ambarları: "Veri ambarları, tüm operasyonel işlemlerin en alt düzeydeki verilerine kadar inebilen, etkili analiz yapılabilmesi için özel olarak modellenen ve tarihsel derinliği olan veri depolama sistematiği olarak tanımlanabilir."[5]

Günlük işlemler sonucu, farklı kaynaklardan toplanan veriler, temizleme dönüştürme, birleştirme gibi işlemlerden geçirilerek, daha önce inşa edilmiş veri ambarının yapısına uygun hale getirilerek veri ambarına aktarılır. Veri ambarları, üzerinde, verilerin yüklenmesi ve erişimi dışında herhangi bir işlem yapılmasına izin vermez. Veri ambarları belirli aralıklar ile güncellenirler.

Mimari açıdan veri ambarları üç farklı şekilde olabilir. İlki, işletmelerin farklı kaynaklardan (işletmenin kendi işlemsel veritabanı sistemleri ve dış kaynaklar dâhil olmak üzere) aldıkları tüm verilerin tutulduğu "işletme ambarları", ikincisi veri üzerinde çalışma yaparak karar alan kişiler için belirli kurallara göre oluşturulmuş "veri pazarları" , sonuncusu ise işlemsel veri tabanlarının görsel hali olan " görsel ambarlar" 'dır.

[5] Yaralıoğlu, **a.g.e.**, s. 165



Şekil 1.1 Bir Veri Ambarının Tipik Görünümü.[6]

OLTP (Çevrimiçi İşlem Süreçleri) : Organizasyonda satın alma, kaydetme, muhasebe, bankacılık gibi günlük işlemlerin yapıldığı işlemsel veritabanı sistemleridir. Detaylı bilgi içerir ve ayrıntılı görüntüye sahiptirler. Veriye erişim sağlanabilir, üzerinde oynama yapılmasına izin verir. Saklanan kayıt sayısı sınırlıdır.

OLAP (Çevrimiçi Analitik Süreçler) : Veri analizi ve karar verme için alt yapıyı oluşturan veri ambarı sistemleridir. İşlemsel veritabanı sistemlerinin aksine, bilgisel süreçler ile ilgilidir. Özet bilgi içerir ve çok boyutludurlar. Büyük boyutta kayıtlar saklanır.

Yukarıdaki kavramlar, Karar Destek Sistemleri'ne girdi sağlayan veri madenciliğinin daha kolay yapılabilmesi ve bunun için gerekli veri depolama, aktarma ve analize hazır hale getirme ile ilgilenir. Bununla birlikte, veri madenciliği analizleri ile ilgili olabilecek bazı kavramlar ise aşağıdaki gibidir.

Veri madenciliğinde kullanılan modeller, tahmin edici ve tanımlayıcı olmak üzere iki başlık altında incelenmektedir.

Tahmin Edici Modeller : Tahmin edici modellerde, sonuçları bilinen verilerden hareket edilerek bir model geliştirilmesi ve kurulan bu modelden yararlanılarak

[6] Han, Kamber, **a.g.e.**, s. 12

sonuçları bilinmeyen veri kümeleri için sonuç tahmin edilmesi amaçlanmaktadır.[7] Örneğin pasifleşmiş ve pasifleşmemiş müşterilerin özelliklerinden oluşan bir veri kümesine sahip isek, bağımlı değişkenimiz müşterilerimizin pasifleşme durumu, bağımsız değişkenlerimiz ise bu müşterilerin daha önce gösterdikleri özellikler olacak, kurulacak model ile sisteme katılan her bir müşteri için firmayı terk edip etmeyeceği tahmin edilebilecektir.

Tanımlayıcı Modeller : Tanımlayıcı modellerde, veri kümesinde bulunan gizli örüntülerin tanımlanması amaçlanmaktadır. Harcama miktarı ve geliş sıklığı düşük olup, A tipi kampanyalara geri dönüş yapma oranı yüksek olan kişiler ile harcama miktarı yüksek olup kampanyalara geri dönüşü çok düşük olan kişilerin satın aldıkları ürünlerin benzerlik göstermesinin belirlenmesi tanımlayıcı modellere örnek olabilir.

Denetimli Öğrenme : "Örnekten öğrenme olarak da isimlendirilen denetimli öğrenimde, analizi yapan kişiler tarafından 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 belirli kural cümleleri ile ifade edilmesidir."[8] Bu amaçla, bütün veri kümesinden bir öğrenme kümesi ayrılır ve model bu küme üzerinden kurulur. Ayrılan test kümesi ile de doğruluğu araştırılır. Modelin doğruluğu yeterli görülüp kullanılmak istendiği takdirde yeni gelen örneklere model uygulanır ve o örneklerin hangi sınıfa ait olduğunu kullanılan model belirler. Denetimli öğrenme sürecinin işlediği tekniklere karar ağaçlarını örnek verebiliriz.

Denetimsiz Öğrenme : Denetimsiz öğrenmede sınıflar önceden belirli olmayıp, veri kümesindeki verilerin özelliklerine göre sınıfların oluşturulması söz konusu olmaktadır. Denetimsiz öğrenme sürecinin işlediği tekniklere kümeleme tekniklerini örnek verebiliriz.

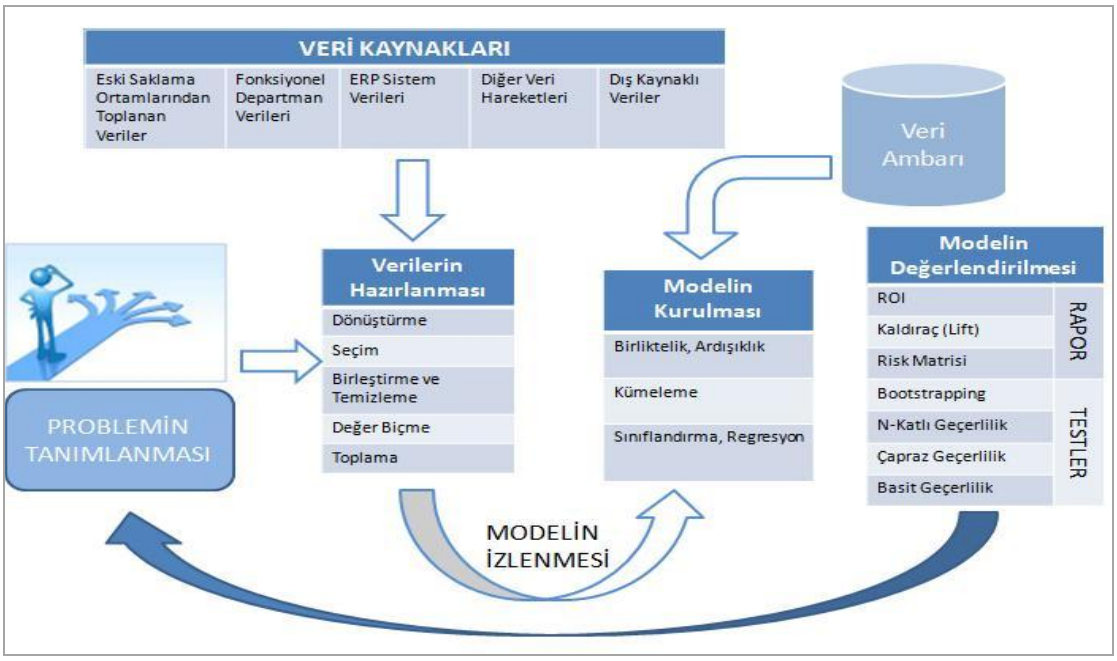
---

[7] Haldun Akpınar , "Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği", **İ.Ü. İşletme Fakültesi Dergisi**, Sayı:1 2000, (Çevrimiçi) http://www.isletme.istanbul.edu.tr/surekli_yayinlar/dergiler/nisan2000/1.htm , 10.Ocak.2009, s. 1-22

[8] Yaralıoğlu, **a.g.e.**, s. 175

1.3. Veri Madenciliği Süreci

Veri madenciliği süreci dört aşama ile tanımlanabilir. İlk aşamada problem tanımlanarak veri kaynakları değerlendirilir. İkinci aşamada veriler kullanıma uygun hale getirilmek için hazırlanır. Arkasından model kurulur ve nihai aşamada model değerlendirilerek kullanıma hazır hale getirilir.



Şekil 1.2 Veri Madenciliği Süreci.[9]

1.3.1. Problemin Tanımlanması

Amaç, işletme problemine verileri kullanarak çözüm getirmek olduğundan, ilk olarak ihtiyaç duyulan şey tam olarak tanımlanmalıdır. Bu problem, işletmenin ayrılmakta olan müşterisinin belirli özelliklerini tanımlayarak ona uygun davranmak olabildiği gibi, kendi kaynaklarını optimum kullanabilmek için yapacağı bir planlamada gelecek dönemdeki harcamalarını tahmin etmek şeklinde de olabilir.

[9] Akpınar, **a.g.e.**, s. 1-22

"Bu adımda ihtiyaç duyulan şeyin tanımlanması için cevaplanması gereken sorular neyin otomatize edilmeye değer olduğu ve neyin insan içeren süreçlere bırakılması gerektiği, amacın ne olduğu ve hangi performans kriterlerinin daha önemli olduğu, sürecin sonucunda elde edilecek çıktının keşif, sınıflandırma, özetleme gibi şeyler için kullanılıp kullanılmayacağı olabilir."[10]

Problemin tanımlanması durumunda ihtiyaç duyulan iş modelinin kalıbı da belirlenmiş olur.

1.3.2. Verilerin Hazırlanması

Modelin kurulması için gerekli bilgilerin hazırlandığı aşamadır. Öncelikle toplam, maksimum, minimum değer gibi dağılım ölçüleri; aritmetik ortalama, ağırlıklı ortalama gibi cebirsel ölçüler veya serpilme,dağılma diyagramı gibi grafiksel öğeler kullanılarak verilerin durumu hakkında bilgi edinilir. Verilerde eksik, hatalı, gürültülü bilgi olup olmadığı bu şekilde kontrol edilmiş olur. Eksik değerlerde kaydı dikkate almama, global sabit ile eksik değerleri doldurma, eksik değere o değişkenin ortalama değerini verme, gürültülü değerlerde regresyon ile belirli fonksiyonel kalıba sokma gibi yöntemler ile verilerdeki sıkıntı giderilebilir.

Farklı kaynaklardan gelen, aynı değişkene ait verilerin tiplerinde, alan isimlerinde uyuşmazlık olması halinde gerekli değişikliklere gidilerek tüm verileri bir arada tutabilecek yapı oluşturulmalıdır.

Bazı modellerin gereksinimlerini göz önünde bulundurmak açısından farklı dönüşümlere gitmek de veri hazırlanırken dikkate alınması gereken hususlardan olabilir. Örneğin bazı değişkenlerdeki değerler çok yüksek ise, bu değerleri normalize ederek, uzaklıklar ile çalışan kümeleme algoritmalarının öğrenme fazını hızlandırarak modelin oluşturulma aşaması için kolaylık sağlanmalıdır.

[10] S. Sumathi, S.N. Sivanandam, **Introduction to Data Mining and its Applications,** New york, Springer 2006, s. 189

Değişken sayısının çok yüksek olduğu, hangi değişkenlerin öneminin daha yüksek olduğuna karar verilemediği durumlarda faktör analizi, temel bileşenler analizi gibi yöntemler kullanılarak boyut indirgemeleri yapılmalıdır. Zira bu indirgemeler modele girecek değişken sayısını azaltarak modeli gereksiz bilgilerden ayıklar ve daha sağlıklı bir sonucun çıkmasına zemin hazırlarlar.

Gerektiğinde kategorik değişkenlerde kategori aralıklarını genişleterek kategori sayısını azaltma veya sürekli bir değişkeni kategorik hale getirmek de verinin hazırlanmasında dikkat edilmesi gereken unsurlardandır. Çok kategorili değişkenler duruma göre modelin çalışma süresini ve sürecin performansını olumsuz etkileyebilmektedir.

1.3.3. Modelin Kurulması

Modelin kurulması aşamasında birçok model denenerek veriyi en iyi temsil eden model seçilir. Verileri temsil eden en iyi modeli bulabilmek için çok sayıda model kurulmalı, en iyi sonucu alana kadar denemeye devam edilmelidir.

Modelin kuruluşu, amacımızın ne olduğuna, problemimizi ne şekilde çözmek istediğimize ve sonucun ne kadar işimize yarar olacağına göre değişebilir. Örneğin görmek istediğimiz gelecek dönemdeki tahmini ciromuz ise, sürekli bir değişkeni tahmin edeceğimiz doğrusal regresyon modelini; müşterilerimizin pasifleşme eğiliminde olup olmadıkları ise kategorik bir değişkeni tahmin edeceğimiz sınıflandırma modelleri olan karar ağaçlarını, yapay sinir ağını veya kategorik değişkenin olasılığını tahmin edeceğimiz lojistik regresyon modelini, hangi ürünlerimizin diğerlerine oranla daha çok beraber alındığı ise birliktelik analizi, beraber alınan bu ürünlerin hangi sırayla alındığı, nedensellikleri ise sıralı örüntü algoritmaları kullanılabilir. Ayrıca müşterilerimizin sahip oldukları alışveriş özelliklerine göre (gelme sıklıkları, uğradıkları mağazalar, satın aldıkları ürünler vb.) belirli gruplara ayırmak için kümeleme algoritmaları kullanılabilir.

Model kurulurken denetimli veya denetimsiz öğrenmeye göre farklı aşamalar uygulanmaktadır. Örneğin sınıflandırma algoritmaları kullanılırken tüm veri kümesi öğrenme ve test kümesi olarak ayrılmalı; modelin verilerden öğrenerek oluşturulması öğrenme kümesi, doğruluğunun kontrolü ise test kümesi ile gerçekleştirilmelidir.

Kurulan modellerde birbiri ile ilişkili olan veya anlamsız olan değişkenlerin elenmesine dikkat edilmelidir. Amaç bilgi çıkarımı olduğundan ve birbiri ile ilişkili olan değişkenler bize ekstra bilgi vermediğinden, diğerine göre daha anlamlı olan değişkeni modele katmak faydamıza olacaktır.

1.3.4. Modelin Değerlendirilmesi

Kurulan modellerin karşılaştırılarak veri kümesini en iyi temsil eden modelin seçildiği aşamadır.

Karşılaştırma için, sınıflayıcının tahmin ettiği sınıfların oranını belirten doğruluk oranı kullanılır. Sınıflayıcının doğruluk oranının görece yüksek olması, diğer modellere göre veri kümesini daha iyi ifade ettiğini gösterebilir. Doğruluğun testi için kullanılan geçerlilik yöntemleri basit geçerlilik yöntemi, çapraz geçerlilik yöntemi, n-katlı geçerlilik yöntemi olarak sıralanabilir.

Basit geçerlilik yönteminde verilerin bir kısmı test verisi olarak ayrılır, kalan kısım üzerinde modelin öğrenimi gerçekleştirildikten sonra ayrılan kısım üzerinde test işlemi yapılır. "Bir sınıflama modelinde yanlış olarak sınıflanan olay sayısının, tüm olay sayısına bölünmesi ile hata oranı, doğru olarak sınıflanan olay sayısının tüm olay sayısına bölünmesi ile doğruluk oranı hesaplanır."[11] Çapraz geçerlilik yöntemi daha az sayıda veri kümesine sahip olunduğu durumlarda kullanılabilir. Bu yöntemde veri kümesi rastgele seçilerek iki eşit gruba ayrılır, gruplar sırayla öğrenme ve test kümesi yapılarak elde edilen doğruluk oranlarının ortalaması kullanılır. N-katlı geçerlilik yöntemi de çapraz geçerlilik yöntemi gibi küçük veri kümeleri için

[11] Yaralıoğlu, **a.g.e.**, s.175

kullanılmaktadır. Veri kümesi birden fazla gruba ayrılır, bir tanesi test diğerleri öğrenim için kullanılır. Test kümesi değiştirilerek doğruluk oranı hesaplanır ve elde edilen oranların ortalaması kullanılır.

Risk matrisi geçerlilik yöntemlerini görselleştirmek için kullanılabilen bir araç olabilir. Yeni çıkan bir ürünü piyasaya sürmeden önce belli sayıda kişi ile görüşülerek ürünün tutup tutmayacağı konusunda bir araştırma yapıldığını ve ürün hakkındaki fikirleri iyi ya da kötü olarak sınıflandırmak istediğimizi düşünelim. Sonuçta karşılaştıracağımız sınıflandırma algoritmalarının doğruluğunu aşağıdaki şekilde görselleştirebiliriz.

	TAHMİN EDİLEN DEĞER		
GERÇEK DEĞER	İYİ	100	20
	KÖTÜ	50	40

Tablo 1.1. Risk Matrisi

1.4. Veri Madenciliğinin Kullanım Alanları

Günümüzde veri madenciliğinin, finanstan telekomünikasyona kadar çok geniş bir kullanım alanı bulunmaktadır. Bunlardan bazıları aşağıdaki gibidir.

1.4.1. Bankacılık – Finans

Bankacılık sektöründe veri madenciliği yoğunlukla kredi sahtekârlıkları tespiti, kredi risklerini değerlendirme, karlılık analizi, trend analizi ve müşteri yönetimi içindirektpazarlama kampanyalarında kullanılmaktadır. "Finansal pazarlarda ise portföy yönetimi, varlık fiyatlarının ve hatta finansal krizlerin tahminlenmesi gibi durumlarda karşımıza çıkar."[12]

[12] Mehmed Kantardzic, **Data Mining Concepts Models Methods and Algorithms,** NJ, Wiley-Interscience 2003, s. 344

Kredi geri ödemelerinin tahminleri ve müşteri kredilerinin analizi bir banka için önemli bir konudur. Tahminleme yöntemleri ile kişinin hangi olasılıkla temerrüde düşüp düşmeyeceği veya kredi talep edenlerin kredi verilmeye uygun olup olmadığı araştırılmakta, bunun sonucunda müşteriye özel stratejiler belirlenebilmektedir.[13] Müşteri yönetimi için müşteriler belirli özelliklerine göre kümelenerek ve her bir grup için ayrı öneri oluşturularak pazarlama yapılabilmektedir.

Ülkemizde, hemen her banka ve finans kuruluşu yukarıdaki analizleri başarı ile uygulamaktadır.

1.4.2. Perakendecilik

Son tüketiciyedirektulaşan perakende sektörü için veri madenciliği güçlü bir araçtır. Sektördeki firmalar, müşteri yönetimi için veri toplamayı plastik kartlar aracılığı ile yapabilmekte, müşterilerin her türlü bilgisini ve alışverişlerini, veri ambarı altyapılarında saklayarak, kişiye özel, hedef kitleli kampanyalar tasarlayabilmekte; bunun için ise bankalarda olduğu gibi kümeleme yöntemleri kullanarak kişilerin özelliklerini anlama ve buna göre müşterileri belirlenen yaşam tarzlarına atama veya değerine göre segmentlere ayırma, mağazalara uğrama sıklıklarına göre skorlama, sınıflandırma algoritmaları, regresyon gibi yöntemler ile müşteri ömrünü belirleme, pasifleşme eğilimi olanları tahminleme, müşterilerin geri dönüşlerinin belirlenmesi veya beraber alınan ürünleri yakalamak için birliktelik tespiti gibi çok çeşitli analizler kullanılabilmektedir.

Bu çalışmalar ile firmalar; en değerli ve bana en çok kazandıran müşterilerim kim, hangi ürünleri hangi raf düzeni ile satmalıyım ki ciromu yükseltebilirim, en çok tercih edilen ürünüm/hizmetim nedir, müşterilerimin yaşam tarzları nedir, ne gibi kampanyalardan hoşlanırlar ve geri dönüş yaparlar, daha ne kadar süre bana kazandırmaya devam edecekler gibi soruların yanıtlarını bulabilmekte, buna göre bütçe planlamasından hedef belirlemesine kadar birçok fayda sağlayabilmektedir.

[13] Han, Kamber, **a.g.e.**, s. 650

Belirtilen analizlerde ana amaç, müşteriyi ve onun tüketim alışkanlıklarını anlamak ve ona yönelik önerilerde bulunmaktır. Sonuçta müşteriden alınan veri, onu memnun etmek ve elde tutmak için yapılan kampanyalara dönüşmekte, firmalar için ise bilgi çıkarımı ile karlılığı yükseltme aracı olarak kullanılabilmektedir.

1.4.3. Telekomünikasyon

"Telekomünikasyon sektörü zaman içerisinde hizmet içeriğini farklılaştırarak sadece yerel ve uzun mesafeli telefon hizmeti sunmaktan çıkmış, fax, internet erişimi yolu ile veri transferi, cep telefonu ve bunun gibi diğer veri trafiklerinin alt yapısını sağlayan bir sektör haline dönüşmüştür." Bundan başka, telekomünikasyon sektörünün bazı ülkelerde yeniden düzenlenmesi, yeni bilgisayarların ve iletişim teknolojilerinin gelişmesi ile birlikte sektör daha da hızlı bir şekilde büyümekte ve rekabetçi bir hale gelmektedir. [14] Bu noktada veri madenciliği iş içeriğini anlamak, iletişim desenlerini tanımlayabilmek, sahtekârlıkları yakalayabilmek, veri kaynaklarını daha iyi kullanabilmek ve hizmet kalitesini arttırabilmek açısından önem kazanmıştır.

Telekom firmaları altyapılarında bulundurdukları arama süresi, bulunulan yer, arama zamanı, arama tipi gibi boyutlar ile birliktelik ve sıralı örüntü analizleri yaparak, kişilerin sonrasında oluşturacakları iletişim desenlerini tahmin edebilirler. Bunun dışında müşterilerinin pasifleşme eğiliminde olup olmadıklarını çeşitli sınıflandırma yöntemleri ile araştırıp pasifleşmeden yakalama şansı elde edebildikleri gibi kişileri özelliklerine göre kümeleyerek belirlenen segmentlere özel müşteri yönetimi kampanyaları gerçekleştirebilirler.

En yoğun olarak kullanılan sektörler olan bankacılık, finans, telekomünikasyon ve perakendecilikten başka veri madenciliği, astronomi, biyoloji, sigortacılık, tıp, mühendislik ve birçok başka dalda da uygulanmaktadır. Astronomide gökcisimlerini

[14] Han, Kamber, **a.g.e.**, s. 652

sınıflandırma, biyolojide gen yapılarını ayrıntılı tanımlama, tıp alanında ise kanserli hücrelerin anlaşılıp sınıflandırılması bu uygulamalara örnek olarak verilebilir.

1.5. Veri Madenciliğinde Kullanılan Yazılımlar

Piyasada birçok veri madenciliği yazılımı ve yeni algoritmalar üreten danışmanlık firmaları bulunmaktadır. Bu yazılımlardan en çok tercih edilenler Enterprise Miner ile SAS ve Clementine çözümü ile SPSS 'tir. Bununla birlikte Intelligent Miner, Viscovery, Unica, Angoss Knowledge Seeker da kullanılan çözümler arasındadır.

SPSS, 1998 yılından bu güne veri tabanlarında bilgi keşfi için analitik çözümler sunmaktadır. SPSS'in veri madenciliği çözümü olan Clementine, metodoloji olarak CRISP DM'i (Cross Industry Standart Processing for Data Mining) kullanmaktadır. Metodoloji; iş analizi, verinin anlaşılması, verinin hazırlanması, modelleme, değerlendirme ve uygulama adımlarını, içerdiği algoritmalar ile kullanıcının yarattığı bir akış içinde bir arada sunan bir yapıyı ifade etmektedir. Clementine, karar ağaçları, yapay sinir ağları, birliktelik, regresyon, zaman serileri analizleri için gerekli olabilecek tüm algoritmaları içerir. Açık ve anlaşılır ara yüze sahiptir.

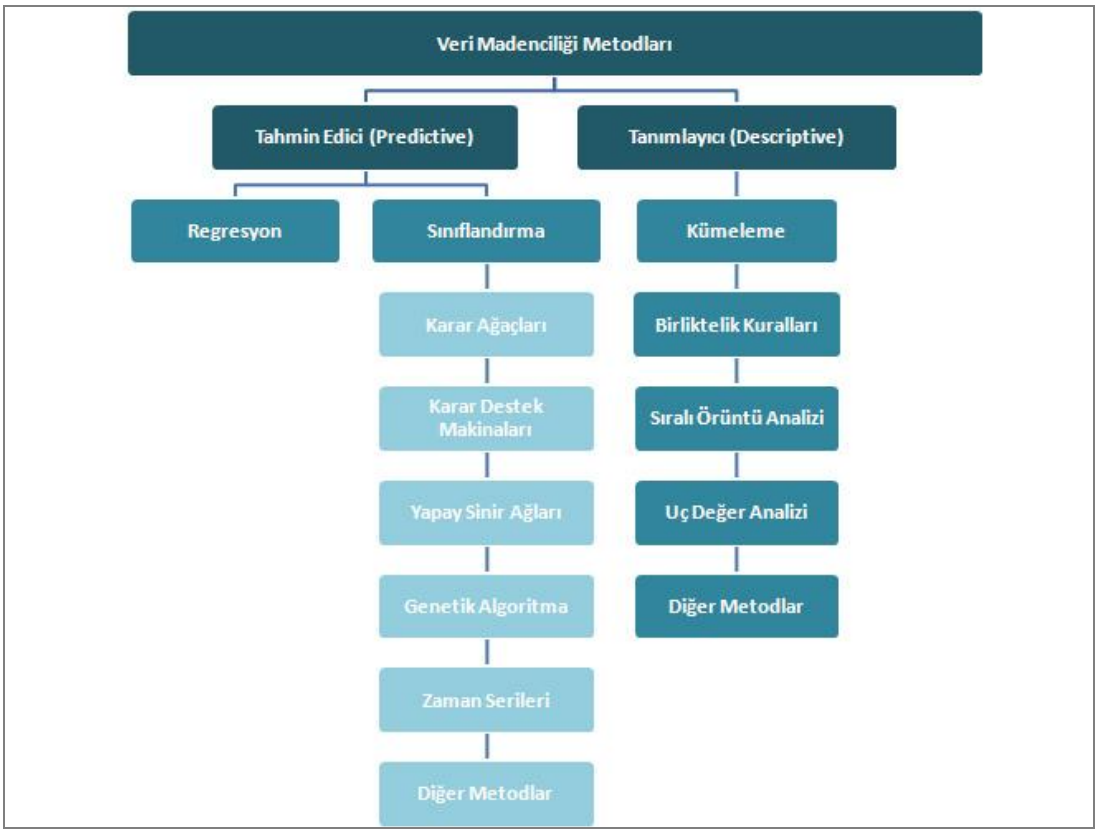
SAS'ın veri madenciliği çözümü olan Enterprise Miner ilk olarak 1997 yılında piyasaya sunulmuştur. Ara yüzü Clementine gibi açıktır ve karar ağaçlarından zaman serilerine kadar gerekli olabilecek tüm algoritmaları içerir.

IBM tarafından geliştirilen Intelligent Miner yazılımı da Clementine ve Enterprise Miner gibi yukarıda belirtilen algoritmaları içermekte, birçok analiz ve modelleme ihtiyacını karşılayabilmektedir. Angoss Knowledge Seeker, Viscovery SOMine gibi yazılımlar ise tahminleyici analizler üzerine yoğunlaşmış, karar ağaçları ile yoğun çalışanlara yönelik bir araçtır.

2. VERİ MADENCİLİĞİNDE KULLANILAN YÖNTEMLER

Veri madenciliği yöntemleri, tahmin edici ve tanımlayıcı olmak üzere iki ana başlık altında toplanmaktadır.

"Tahmin edici modellerde, sonuçları bilinen verilerden hareket edilerek bir model geliştirilmesi ve kurulan bu modelden yararlanılarak sonuçları bilinmeyen veri kümeleri için sonuç değerlerin tahmin edilmesi amaçlanmaktadır."[1] Tanımlayıcı modellerde ise veri kümesindeki örüntülerin bulunması amaçlanmaktadır.



Şekil 2.1. Veri Madenciliği Yöntemleri[2]

[1] Halil Kaya, Kemal Köymen, "Veri Madenciliği Kavramı ve Uygulama Alanları", **Doğu Anadolu Bölgesi Araştırma ve Uygulama Dergisi,** Şubat 2008 (Çevrimiçi), http://web.firat.edu.tr/daum/default.asp?id=79, 13.Ocak.2009

[2] Kaya, Köymen**, a.y.**

2.1. Regresyon

"Regresyon çözümlemesi, bir bağımlı değişkenin başka bağımsız değişkenlere olan bağımlılığını, bağımlı değişkenin ana kütle ortalama değerini, bağımsız değişkenin yinelenen örneklerdeki bilinen ya da değişmeyen değerleri cinsinden tahmin etme ve/veya kestirme amacı ile inceler."[3]

2.1.1. Doğrusal Regresyon

Doğrusal regresyon modeli, iki ya da daha fazla değişken arasındaki doğrusal ilişkiyi açıklar. Açıklanan değişkene bağımlı değişken, açıklayıcı değişkenlere ise bağımsız değişken adı verilir. Örneğin, gelir düzeyi ve eğitim düzeyi arasındaki ilişkiyi, öğrencilerin devamsızlık yaptığı günler ile başarıları arasındaki ilişkiyi açıklamak için regresyon modeli kullanılabilir.

Doğrusal regresyon modelinin matematiksel kalıbı tek bağımsız değişken için basitçe aşağıdaki gibi gösterilebilir.

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad i = (l, n) \tag{2.1}$$

Burada Y_i bağımlı değişkeni, X bağımsız değişkeni, β_0 sabit değeri, β_1 ise X değişkeninin Y_i değişkenini açıklama derecesini ifade etmektedir. ε_i ise, spesifikasyon hatasından veri girişinde yapılan hatalara kadar her türlü sıkıntıyı kapsayan hata terimini ifade etmektedir. Doğrusal regresyon modelinin günlük hayata uygulanabilirliği diğerlerine göre daha zor olabilmektedir. Zira modelleme için kullanılan ve en popüler yöntem olan En Küçük Kareler yöntemi ile modelin kurulabilmesi için belirli varsayımları sağlaması gerekmektedir. Bu varsayımlar aşağıdaki gibidir: [4]

[3] Damodar N. Gujarati, **Temel Ekonometri,** İstanbul, Literatür Yayınları, 2001, s. 16
[4] Şahin Akkaya, Vedat Pazarlıoğlu, **Ekonometri 1**, İzmir, Anadolu Matbaacılık, 2000, s. 93

Hata Terimi ε_i için:

- Ortalaması sıfıra eşit stokastik bir değişkendir.
- Normal dağılmaktadır.
- Hata teriminin değerleri arasında ilişki yoktur.
- Varyansı her X değeri için eşittir.

Bağımsız değişken X için:

- Hata terimi ile ilişkili olmayıp, stokastik değildir.
- Tekrar eden örnek değerlerine göre sabittir.
- Varyansı sonlu pozitif bir sayı olmalıdır.

Birden fazla bağımsız değişken olması durumunda modelin matematiksel kalıbı aşağıdaki gibi gösterilebilir.

$$Y_i = \beta_0 + \beta_{1i} X_{1i} + ... + \beta_{ki} X_{ki} + \varepsilon_i \qquad i = (l, n) \tag{2.2}$$

$$Y_i = \sum \beta_k X_{ki} + \varepsilon_i \qquad i = (l, n) \tag{2.3}$$

2.1.2. Lojistik Regresyon

Doğrusal regresyon modeli bağımlı değişken olarak sürekli değişkenleri alırken, kategorik değişkenlerin tahmini için farklı yöntemler geliştirilmiştir. Lojistik regresyon, bağımlı değişkenin iki veya daha fazla kategori içerdiği, bağımsız değişkenlerin ise sürekli veya kategorik bir yapıya sahip olduğu durumlarda bağımsız değişkenler ile bağımlı değişken arasındaki ilişkiyi araştırır.

Lojistik regresyon analizi iki farklı türde olup, türü bağımlı değişkenin kategorisi belirlemektedir. Bağımlı değişken iki kategoriye sahip ise ikili lojistik regresyon, ikiden fazla kategoriye sahip ise çoklu lojistik regresyon adı altında incelenebilir.

2.1.2.1. İkili Lojistik Regresyon

Basit doğrusal regresyon modelinde bağımlı değişken Y_i süreklidir ve bağımsız değişkenler de $-\infty$ ile $+\infty$ arasında değerler alırlar. Bağımlı değişken kategorik bir değişken olduğunda ve kesikli değerler aldığında bu kural bozulmaktadır.

$P(Y_i=1)$, i'inci gözlemin 1 değerini alma olasılığı olmak üzere beklenen değer aşağıdaki şekilde olmaktadır.

$$E(Y_i)= 1 \times P(Y_i=1) + 0 \times P(Y_i=0) = P(Y_i=1) \quad (2.4)$$

Kısaltacak olursak aşağıdaki regresyon denklemini elde ederiz.

$$E(Y_i)= P(Y_i=1) = \sum_{k=0}^{p} \beta_k X_{ik} \quad (2.5)$$

Sol tarafı 0-1 arasında olasılık değerleri alan bu denklem doğrusal olasılık modeli olarak adlandırılır. Bağımlı değişken kısıtlı değerler alırken, bağımsız değişkenlerin sınırsız değerler alması durumunda eşitlik sağlanamaz ve olasılık değeri $-\infty$ ile $+\infty$ arasında dönüşüme uğratılır. Yapılan dönüşümlerden en bilinenleri lojit ve probit dönüşümlerdir. Bu yöntemler birbirlerine yakın sonuçlar vermektedirler.

Lojit dönüşümde doğrusal olasılık modeli aşağıdaki dönüşümlere maruz kalarak bağımlı değişken $-\infty$ ile $+\infty$ arasına getirilir.

$$E(Y_i)=\log\left(\frac{P_i}{1-P_i}\right)= \sum_{k=0}^{p} \beta_k X_{ik} \quad (2.6)$$

$$P_i = \exp \frac{\sum_{k=0}^{p} \beta_k X_{ik}}{(1+\exp(\sum_{k=0}^{p} \beta_k X_{ik}))} \quad (2.7)$$

Adımları ile aşağıdaki nihai model aşağıdaki gibi gösterilebilir.

$$\text{Log}\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 * X_1 + ... + \beta_i * X_i \quad (2.8)$$

Buna göre;

P: İstenilen durumun gerçekleşme olasılığı

β_0: Sabit değer

$\beta_i : i = (1,n)$ olmak üzere her bir bağımsız değişkenin katsayısı

$X_i : i = (1,n)$ olmak üzere bağımsız değişkenleri ifade eder.

İkili lojistik regresyonun varsayımları aşağıdaki gibidir: [5]

- $Y_i \in (0,1)$
- $P(Y_i = 1 / X_i) = P_i$
- $Y_1,...,Y_n$ değerleri istatistiksel olarak bağımsızdır.
- Bağımsız değişkenler arasında ilişki yoktur.

"Modelin sonuç değişkeninin sınırlarını genişletmek için uygulanan lojit dönüşümün bazı özellikleri şöyle sıralanabilir:[6]

- P arttıkça lojit (P) de artar.
- P , 0 ile 1 arasında iken lojit (P) reel sayılar doğrusu üzerinde değerler alabilir.
- $P > 0.5$ olduğunda lojit $(P) < 0$ ve $P < 0.5$ olduğunda lojit $(P) > 0$ olur."

Lojistik model yorumlanırken, bağımsız değişkendeki katsayı değişiminin bağımlı değişkenin olma olasılığı üzerindeki etkisi şeklinde yorumlama yapılabilir. Örneğin bir şirketin müşterilerinin pasifleşme eğilimi araştırılıyor ise kullanılan lojistik modelin bağımlı değişkeni pasifleşecek ya da pasifleşmeyecek olarak iki kategoriye sahiptir ve pasifleşme olasılığının ciro bağımsız değişkenindeki katsayı değişimi kadar artacağı veya azalacağı ifade edilebilir.

Lojistik modelde katsayı tahminleri için kullanılan çözüm yöntemlerinden ikisi en çok olabilirlik yöntemi ve yeniden ağırlıklandırılmış iteratif en küçük kareler

[5] Hüseyin Tatlıdil, **Uygulamalı Çok Değişkenli İstatistiksel Analiz**, Ankara, Ziraat Matbaacılık 2002, s. 292

[6] Hüdaverdi Bircan, Yalçın Karagöz , "Lojistik Regresyon Analizi: Tıp Verileri Üzerinde bir Uygulama", **Kocaeli Üniversitesi Sosyal Bilimler Enstitüsü Dergisi**, 2004 (Çevrimiçi), http://iibf.erciyes.edu.tr:90/petas, 10.Haziran.2009

yöntemidir. “En çok olabilirlik yöntemi denklemin tahmin edilen tarafını yani kategorilerin gözlenme olasılığını maksimum yapacak değeri bulma ile ilgilenirken, yeniden ağırlıklandırılmış iteratif en küçük kareler yöntemi her bir değişkeni bir katsayı ile ağırlıklandırarak verilere en küçük kareler yönteminin uygulanmasından ibarettir.”[7]

2.1.2.2. Çoklu Lojistik Regresyon

Bağımlı değişkenin ikiden fazla kategori içerdiği lojistik regresyon modelleridir. Genelde iki grup lojistik modellerin çoklu grup durumunda da kullanılabilmesi mümkündür. Örnek olarak bağımlı değişken 0,1,2 gibi 3 kategoriye sahip olsun. Bu durumda iki tane farklı iki grup lojistik lojistik model söz konusudur. 0 kategorisi baz alındığında, 2 nolu kategoriyi 1 nolu kategori ile karşılaştıran fonksiyonlar aşağıdaki gibidir.[8]

$$g_1(X) = \text{Log}\left(\frac{P(Y=1/X)}{P(Y=0/X)}\right) = \beta_{10} + \beta_{11}X_1 + ... + \beta_{1p}X_p \quad (2.9)$$

$$g_2(X) = \text{Log}\left(\frac{P(Y=2/X)}{P(Y=0/X)}\right) = \beta_{20} + \beta_{21}X_1 + ... + \beta_{2p}X_p \quad (2.10)$$

Bu fonksiyonlardan hareketle üç kategori için koşullu olasılıklar k = 0,1,2 için aşağıdaki gibi olmaktadır.

$$P_k(X) = \frac{\exp(g_k(X))}{\sum_{t=0}^{2}\exp(g_t(X))} \quad (2.11)$$

Lojistik model yorumlanırken, bağımsız değişkendeki katsayı değişiminin bağımlı değişkenin olma olasılığı üzerindeki etkisi şeklinde yorumlanabilir. Buna göre 0 kategorisi sabit iken, 1 kategorisinin gerçekleşme olasılığı, 2 kategorisinin

[7] Şahin Akkaya, Vedat Pazarlıoğlu, **Ekonometri 2**, 1998, s. 89-90
[8] Tatlıdil, **a.g.e.**, s. 304

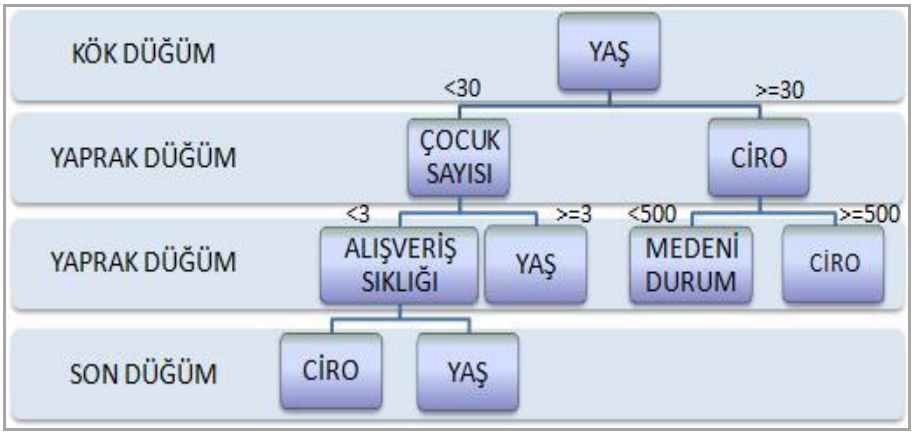
gerçekleşme olasılığına göre yüzde X kadar daha fazladır veya azdır yorumu yapılabilir.

2.2. Karar Ağaçları

"Karar ağaçları, tek bağımlı değişken ve çok sayıda bağımsız değişkene sahip olmaları açısından regresyon modellerine benzerler. Bununla birlikte, ek olarak, veriden regresyon modellerine alternatif olabilecek farklı ve kullanışlı örüntüler keşfederler." [9] Karar ağaçları, bağımlı değişkenin kategorik olduğu durumlarda lojistik regresyona alternatif oluşturabilecek bir yöntemdir.

Kolayca kural cümleciklerine çevrilebilir olmaları, sürekli ya da kesikli veriler ile çalışabilmeleri, eksik veya hatalı veriler ile tahminleme yapabiliyor olmaları karar ağaçlarının avantajlarındandır. Ayrıca parametrik olmayan yöntemler arasındadır. Bu, karar ağaçlarının uzay dağılımı veya sınıflayıcı yapısı ile ilgili varsayımlara uymak zorunda olmadığı anlamına gelir. Bununla birlikte, eksik veya hatalı verilere duyarsız olması ve yaprak düğümlerde mükerrerlik içermesi de dezavantajı olabilmektedir.[10]

Bir karar ağacı basitçe aşağıdaki gibi gösterilebilir.



Şekil 2.2. Karar Ağacı Örneği

[9] Louis Anthony Cox, "Data Mining and Causal Modelling of Customer Behaviours", **Telecommunication Systems**, Volume 21, 2002, s. 356

[10] Oded Maimon, Lior Rokach, **Data Mining and Knowledge Discovery Handbook**, Ramat-Aviv, Springer 2005, s. 183-184

Karar ağacının araştırdığımız sınıfı sağlayan başlangıç düğümüne kök düğüm, ara safhalardaki düğümlere yaprak düğüm, ağacın bittiği düğümlere ise son düğüm denir. Her bir düğümdeki gözlem sayısı düğümün büyüklüğünü ifade ederken, ağaçtaki dallanma sayısı ağacın derinliğini gösterir. Yukarıdaki örnekte ağacın derinliği 4'tür. Yukarıdaki ağacın bir ürünü alan kişilerin özelliklerini araştırdığını düşünelim. Bu durumda ağacı kısaca şöyle yorumlayabiliriz. Bu kişiler için en önemli özellik yaş olarak gösterilmiş, dallanma bu değişkenden başlamıştır. Bir sonraki düğümde ise 30 yaş altı grup için çocuk sayısının önemli bir gösterge olduğu, 30 yaş üstü için ise firmada yaptığı cironun ayırıcı bir özellik olduğunu görebiliriz. Aynı yorumlar diğer düğümler için de geçerli olmaktadır.

Karar ağacının her düğümünde değişkenler test edilir. "Karar ağacı algoritması, ağacın kökünde hangi değişken ile test edilmesi gerektiği sorusu ile başlayarak yukarıdan aşağıya doğru ağacı oluşturur. Bu işlemi her örnek değişken, eğitim örneklerinin sınıflandırmasına karar vermek için istatistiksel test kullanılarak değerlendirilir. En iyi değişken seçilir ve ağacın kök düğümünde test için kullanılır." Her bir düğüm için oluşturulacak dalların sayısı, test sonucunda kabul edilmiş olan değişkenin alabileceği değer sayısına göre farklılaşmaktadır.[11]

2.2.1. Karar Ağaçları'nda Ayırma Kriterleri

Ağacı devam ettirecek olan değişken seçilirken belirli kriterler gözetilmektedir. Bu kriterler sonucu her bir değişkenin aldığı değerlere göre seçim yapılmakta, ağaç dallandırılmaktadır.

Farklı ayırma kriterleri birbirlerinden farklı gibi görünmelerine rağmen performansları birbirine çok yakın olabilir. Bu durumun sebebi ayırmada kullanılabilecek değişkenlerin performanslarının birbirine yakın olmasıdır. Farklı ölçüler, farklı değişkenlerin seçilmesini sağlamasına rağmen, tüm bu ölçüler aynı

[11] Baha Vural Kök, Necati Kuloğlu, **"**Sollama Esnasında Taşıt ve Yol ile İlgili Faktörlerin Karar Ağacı Yöntemi ile İrdelenmesi", **Erciyes Üniversitesi Fen Bilimleri Enstitüsü Dergisi**, 2005 (Çevrimiçi), http://perweb.firat.edu.tr/personel/yayinlar/fua_522/522_20056.pdf, 10.Mayıs.2009

bakış açısını yakalamaya çalıştıklarından, sonuçta olacak modeller de benzer davranmaya eğilimli olabilirler.[12]

Kriterler tek değişkenli ve çok değişkenli ayırma kriterleri olarak iki grupta sınıflandırılabilir.

Tek Değişkenli Ayırma Kriterleri: Karar ağaçlarında bağımlı değişkenin kategorik olması durumunda karar ağaçlarında dallanma için bilgi kazanımı, gini, χ^2 yöntemleri kullanılırken, bağımlı değişken sürekli olduğu durumlarda ise F testi kullanılır.

Bilgi Kazanımı : Bilgi kazanımı, entropi ilkesine dayanır. Entropi bir sistemin düzensizliğini ifade eden kavramdır. Bu yöntem ayırma yöntemi olarak seçildiğinde, algoritma, entropiyi azaltan çözümler üretir. Zira sistemde düzensizliğin azalması ile elde edilen bilgi kazanımı artmaktadır.

Karar ağacındaki herhangi bir düğüm için, en fazla bilgi kazandırabilecek değişken, ayırıcı değişken olarak seçilir. Bu değişken, kayıtları sınıflandırmak için ihtiyaç duyulan bilgiyi minimize eder ve basit bir ağacın bulunma olasılığını arttırır. Bu da, minimum rassallık ve safsızlığı yansıtır. Zira beklenen bilginin küçük olması, ayrımların sağlığının büyük olması demektir.

X veri kümesindeki herhangi bir kaydı sınıflandırmak için beklenen bilgi 2.12'de gösterilmiştir.

$$\text{Bilgi } (X) = -\sum_{i=1}^{m} p_i \log_2(p_i) \qquad (2.12)$$

Burada p_i, X veri kümesindeki herhangi bir kaydın, bağımlı değişkenin kategorilerinden biri olan L sınıfına ait olma olasılığıdır. Örneğin bir ürünün alınıp alınmamasını araştırıyor isek, yukarıdaki değer, ürünü alan kişi sayısına ve almayan kişi sayısına göre toplam olarak hesaplanır.

[12] Michael J.A. Berry, Gordon S. Linoff, **Data Mining Techniques for Marketing, Sales and Customer Relationship Management**, 2004, s. 176

X veri kümesindeki bir kaydı sınıflandırırken her bir değişken için ihtiyaç duyulan bilgi,

$$\text{Bilgi}_k(X) = -\sum_{j=1}^{v} \frac{|X_j|}{|X|} * \text{Bilgi}\ (X) \tag{2.13}$$

gibidir. Bu değer, k değişkenindeki her bir kategori için ayrı ayrı hesaplanarak toplanır. $|X_j|/|X|$ oranı, k değişkeninin herhangi bir j kategorisi için, bağımlı değişkendeki sınıfların ağırlığını ifade eder. Örneğin, medeni durum değişkeninin yukarıda bahsi geçen ürünün alınıp alınmaması üzerindeki etkisini araştırıyor isek, k medeni durum olacak, $|X_j|/|X|$ sayısı da evliler ve bekârlar için ayrı ayrı hesaplanarak toplanacaktır.

Bilgi kazanımı, yukarıda açıklanan iki değerin farkı ile ifade edilmektedir.

$$\text{Kazanım}(X) = \text{Bilgi}_k(X) - \text{Bilgi}\ (X) \tag{2.14}$$

Değişkenlerin, o ağaca o noktada dallandırma yapıldığında kazandıracakları bilgi hesaplanarak hangi değişkene göre ilerleneceğine karar verilir. Zira en fazla bilgi veren, kazanımı en yüksek olan değişken ile ilerlenir.

Gini İndeksi : Gini indeksi, iki parçalı sonuçlar üretmek için kullanılan bir kriterdir. Yaklaşımı bilgi kazanımına benzer. X veri kümesindeki herhangi bir kaydı sınıflandırmak için gini indeksi aşağıdaki şekilde ifade edilir.

$$Gini(X) = 1 - \sum_{i=1}^{m} p_i^2 \tag{2.15}$$

p_i , X veri kümesindeki bir kaydın, L kümesine ait olma olasılığını gösterir. Yukarıdaki örnekle devam edecek olursak, bir ürünün alınma ve alınmama olasılıklarının karesini hesaplayarak bunların toplamlarını 1'den çıkarınca $\text{Gini}(X)$'e ulaşmış oluruz.

X veri kümesindeki bir kaydı sınıflandırırken her bir değişkenin gini indeksi ise aşağıdaki gibidir.

$$Gini_A(X) = \frac{|X_1|}{|X|} Gini(X_1) + \frac{|X_2|}{|X|} Gini(X_2) \quad (2.16)$$

Burada $|X_1|/|X|$ değeri, o kategori için o sonucun gerçekleşme oranıdır ki, örnek üzerinden düşündüğümüzde, bir ürünün alınıp alınmaması üzerinde A ile ifade edilen renk değişkeninin, 1 ile ifade edilen kırmızı kategorisinden toplam veri kümesinde kaç adet satın alındığını gösterir.

Değişken seçimi kararı için kullanılan gini indeksi 2.15 ve 2.16'daki iki ifadenin farkını içerir.

$$\Delta Gini(A) = Gini(X) - Gini_A(X) \quad (2.17)$$

Gini indeksi, sürekli değişkenlere de kategorik değişken gibi davranır. Olası tüm kesim noktalarından seçimler yaparak çalışır.

Tüm değişkenlerin gini değeri hesaplandıktan sonra, minimum indekse sahip değişken seçilerek karar ağacı dallandırılır. Minimum indeks değeri, maksimum safsızlık anlamına gelir.

χ^2 ve F Testleri: Karar ağaçları, χ^2 ve F testlerinin anlamlılığını kriter olarak kullanarak, bir potansiyel ayırıcı değişkenin tüm değerlerini değerlendirir. "Bağımlı değişkene göre istatistiksel olarak homojen olarak değerlendirilebilecek tüm değerleri birleştirir ve diğer tüm değerleri heterojen olarak değerlendirir. Ardından karar ağacındaki ilk dalın formuna göre en iyi ayırıcı değişkenin seçilmesiyle, her bir düğümün seçilen değişkenin homojen değerlerinin bir grubunu oluşturmasını sağlar."[13]

Çok Değişkenli Ayırma Kriterleri: **"**Çok değişkenli ayırma kriterlerinin birçoğu, bağımsız değişkenlerin kombinasyonlarına dayalı olarak oluşturulmaktadır. Burada optimal değişkeni bulma problemi, tek değişkenli parçalama kriterlerine göre daha

[13] Ayşe Oğuzlar, Selim Tüzüntürk, **"**Borsada İşlem Gören Şirketlerin Finansal Göstergelerinin Analizi", (Çevrimiçi), http://iletisim.atauni.edu.tr/eisemp/html/tammetinler/267.pdf, 18.Mayıs.2009

zordur. En uygun parçalayan değişkeni bulmak için, sonuca deneme yanılma yolu ile giden, doğrusal programlama, doğrusal diskriminant analizi gibi yöntemler kullanılmaktadır."[14]

2.2.2. Karar Ağaçları'nda Durma Kriterleri

Karar ağacının büyüklüğü modelin kalitesi için en önemli olan özelliklerden biridir. Çok küçük ağaçlar veri kümesini iyi tanımlayamayabilirler. Çok büyük ve çok fazla dallanmış, her dalında ufak miktarda veri barındıran ağaçların da temsil yeteneği düşük olabilir. Bu sebepten ağacın derinliği karar verilmesi gereken konulardandır.

Karar ağacı dallanmayı, belirtilen durma kriterlerinden biri ile karşılaşana kadar sürdürür. Ortak olarak kullanılan durma kriterlerinden bazıları şunlardır: [15]

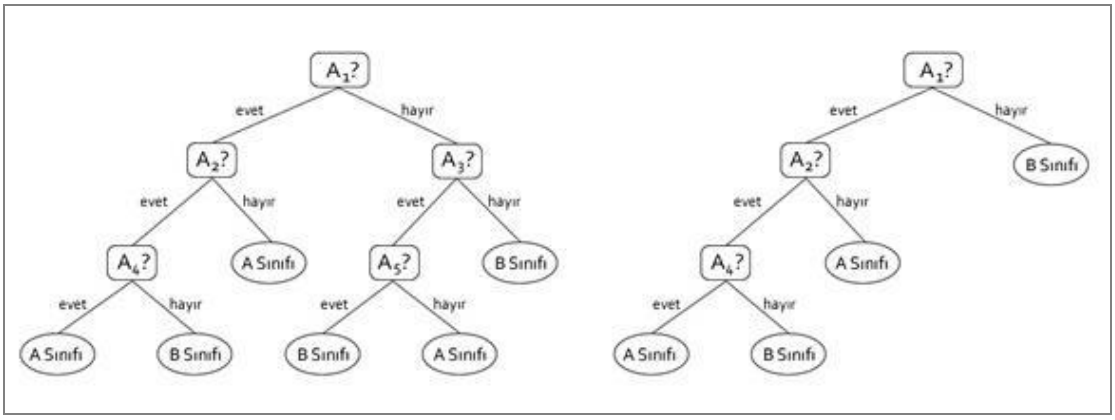
- Veri kümesindeki tüm örnekler tek bir sınıfa ait olduğunda veya belirli bir sayının altına indiği zaman.
- Ağacı oluşturan kişi tarafından belirlenen maksimum ağaç derinliğine ulaşıldığı zaman.
- Son düğümdeki örneklerin sayısı bir önceki düğümdeki minimum örnek sayısından küçük olduğu zaman.
- Kayıtların ayırma için sorgulanabilecek herhangi bir özelliği kalmadığı zaman.
- Ayırma kriteri, belirlenen eşik değerden daha büyük olduğu zaman.

2.2.3. Karar Ağaçları'nda Budama

Karar ağaçları oluşturulurken, çok fazla dallanmış, bazı dallarda sapmalı değerler anormallik yaratmış ve karar ağacı aşırı öğrenme gerçekleştirmiş olabilir. Bu durumda ağacın budanmasında fayda bulunmaktadır.

[14] Kolluru Venkata, Sreerama Murthy, "On Growing Better Decision Trees from Data", (Çevrimiçi), http://www.cbcb.umd.edu/~salzberg/docs/murthy_thesis/thesis.html, 13.Ocak.2009
[15] Maimon, Rokach, **A.g.e.** s. 174

Budama ile bazı istatistiksel ölçüler kullanarak fazla güvenilir olmayan dalları ayıklanır. Budanmış ağaçlar daha küçük ve daha az karmaşık olmaya eğilimlidirler ve böylelikle daha kolay yorumlanabilirler. Ayrıca eğitim kümesinden bağımsız olan test kümesini sınıflandırmada budanmamış ağaçlara göre daha iyi ve hızlıdırlar. Aşağıda budanmış ve budanmamış ağaç örneği görülmektedir.



Şekil 2.3. Bir Karar Ağacının Budanmamış ve Budanmış Versiyonları.[16]

Budama, ön budama ve son budama olarak sınıflandırılabilir. Ön budamada, ağacın dallandırılma aşamasında, ayırma için kullanılan istatistiki kriterler, gini indeksi veya kazanım oranı için belirli eşik değerler konularak ağacın o düğümden sonra büyümemesi esas alınır. Son budamada ise, bütün karar ağacı oluşturularak, son hali üzerinden küçültme işlemi gerçekleştirilir. Budama yöntemlerinden ikisi aşağıdaki gibidir.

Maliyet Karmaşıklığı Yöntemi : Bir ön budama yöntemidir. Bu yöntem, maliyet karmaşıklığını, ağaçtaki dal sayısının ve ağacın hata oranının bir fonksiyonu olarak kabul eder.

Buna göre ağacın her bir düğümü için maliyet karmaşıklığını hesaplar. Söz konusu düğüm budandığı vakit, daha düşük bir maliyet karşımıza çıkacak ise, o düğüm budanarak ağaç oluşturulur. Tüm bu hesaplamalar, eğitilen veri kümesinden bağımsız bir küme olan budama kümesi ile yapılır.

[16] Han, Kamber, **a.g.e.**, s. 305

Kötümser Budama Yöntemi : Maliyet karmaşıklığına benzer bir yöntemdir; fakat budama kümesi ayrı değildir.

Karar ağaçları, sonrasında kural cümleleri çıkarımı yapılabilmesi, bağımlı ve bağımsız değişkenler arasında doğrusallık olması gibi belirli varsayımlara bağlı kalmamaları ve yorumlanmaya diğer yöntemlerden daha müsait olmaları sebebi ile diğer veri madenciliği yöntemlerine göre avantajlıdır.

2.2.4. Bazı Karar Ağacı Algoritmaları

Bilinen en popüler karar ağacı algoritmaları C&RT, CHAID 'tir.

Her iki algoritma da sürekli ve kategorik bağımlı ve bağımsız değişkenler ile çalışabilir. Bu iki algoritma arasındaki en büyük fark, CHAID'in çoklu, C&RT 'nin ise ikili dallanma yapmaları ve dallanma için kullandıkları değişken belirleme şeklidir. CHAID, dallanma yaparken χ^2 ve F testi gibi istatistiksel ölçüler kullanırken, C&RT, χ^2 dışında, safsızlık ölçütü olan Gini indeksini de kullanabilir. Bununla birlikte, genel olarak ulaşılan sonuçlar iki ağaç için de birbirine yakın olabilir.

2.2.4.1. CHAID (Ki – Kare Otomatik İlişki Tespiti)

"Chaid algoritması, kategorik değişkenler için gözlenen sıklık değerlerinin χ^2 analizini yaparak bu değişkenlerin ne kadar iyi cevap verdiklerine karar verir. Chaid, popülasyondaki istatistiksel önemi olan grupları keşfetmek için kullanılır."[17]

Chaid algoritması, dallanan değişken ile bağımlı değişken arasındaki bağımlılığı test eder. Bağımlı değişken ile, ele alınan bağımsız değişken arasındaki ilişkiyi araştıran bu testin sonucu iki değişken arasında bağımlılığı ifade ediyorsa ağacın büyümesine,

[17] Rob Mattison, **Data Warehousing and Data Mining for Telecommunications,** Norwood, Artech House, 1997, s. 254

bağımsızlığı ifade ediyor ise ağacın durmasına sebep olur. Bu, beklenen bir durumdur. Çünkü amaç bağımlı değişkenin dallanarak açıklanmasıdır ve bağımsız olmaları o bağımsız değişkenin, bağımlı değişkeni açıklamadığını ifade etmektedir. Test sonucu olasılık değeri en küçük olan yani önem değeri en yüksek olan değişken dallanma için seçilir.

Chaid algoritması, kategorik bağımsız değişkenler ile çalışmayı tercih ettiğinden, modele giren bağımsız değişkenleri, sürekli olmaları halinde bölerek kategorik hale getirir. Bağımsız değişken çok fazla kategoriye sahip ise, bu durumda kategori sayısını indirgeyerek ağacı basitleştirme yoluna gider.

Chaid algoritmasını temel alan exhaustive chaid algoritması ise sürekli değişkenlerin kategorilerinin birleştirilmesi ve test edilmesi aşamasında basit chaid'e göre daha dikkatli bir yaklaşım sergiler. Özellikle kategorilerin birleştirilmesi işlemi, her bir değişken için iki kategori kalana kadar devam eder. Değişken seçimi chaid gibi olmasına karşın ayırma ve test etme aşamaları daha titiz olduğundan, çok fazla sürekli değişkene sahip büyük veri kümelerinde modelin geliştirilmesi uzun sürer.[18]

2.2.4.2. C&RT (Sınıflandırma ve Regresyon Ağacı)

C&RT algoritmaları, bağımlı değişkenin kategorik olduğu durumlarda sınıflandırma, sürekli olduğu durumlarda tahminleme modeli kuran bir karar ağacı algoritmasıdır. C&RT algoritmaları için birincil amaç, mümkün olan en iyi doğruluğu olan modeli kurabilmektir. En iyi doğruluk ise minimum maliyetli tahminler yapılmasını içerir. Minimum maliyetli tahminler yapılması, en düşük yanlış tahmin oranına yani yanlış sınıflandırılan verinin az olmasına sahip olunması demektir.

C&RT, dallanması sürecinde, her bir adımda tahminin doğruluğuna en fazla katkısı olan ayrımı yaparak ilerler. Ayrım ölçütü olarak Gini indeksi veya χ^2 gibi ölçütler kullanır. Dallanma, bütün durumlar en iyi şekilde sınıflandırılıncaya ya da tahmin

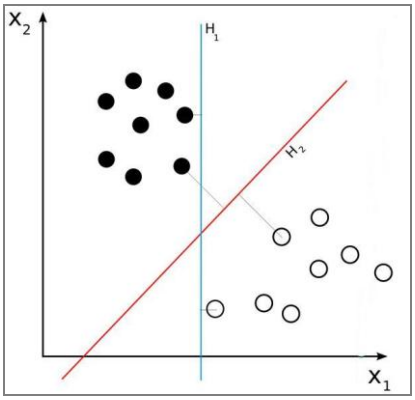
[18] http://www.statsoft.com/TEXTBOOK/stchaid.html#index

edilinceye kadar sürer. Bununla birlikte bazen ağacın yapısı orijinal veriden daha karmaşık bir yapıya bürünecek şekilde büyür ve bu, yeni gözlemleri tahmin etmek ya da sınıflandırmak için kullanışlı olmayabilir. Bu durumda C&RT, belirlenen bir ağaç derinliğine göre veya belirtilen diğer kriterlere göre büyümeyi durdurur.

Modelin doğruluğu, bağımlı değişkenin kategorik olması durumunda doğru tahmin edilen kayıtların oranı, sürekli olması durumunda ise ortalama hata kareler ile ölçülür.

2.3. Karar Destek Makineleri

Karar destek makineleri, doğrusal ve doğrusal olmayan verilerin sınıflandırılması ile ilgilenen bir yöntemdir. Doğrusal olmayan bir haritalama yöntemi ile orijinal veriyi daha yüksek boyutlara taşır. Taşıdığı bu boyutta, verileri sınıflandırmak için ayrıcı olabilecek doğrusal ayırıcı düzlemler araştırır ve optimum düzlemi yakalamaya çalışır. "Uygun bir haritalama yöntemi ve yeterli derecede yüksek boyutta iki farklı sınıfa ait veriler daima ayırıcı bir düzlem tarafından ayrılırlar. Buna göre algoritma belirtilen düzlemi, destek vektörleri (eğitim kümesi verileri) ve bu vektörler tarafından tanımlanmış mesafelerle bulur."[19]



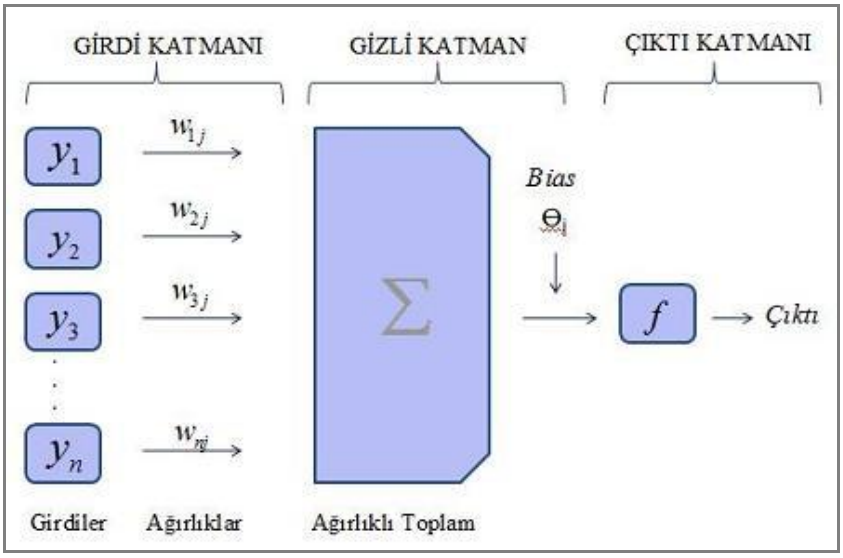
Şekil 2.4. Karar Destek Makineleri[20]

Şekilde ikili çıktı değişkenine sahip bir veri kümesi için oluşturulan farklı düzlemler görülmektedir.

[19] Han, Kamber, **a.g.e.**, s. 337
[20] http://en.wikipedia.org/wiki/Support_vector_machine

2.4. Yapay Sinir Ağları

"Yapay sinir ağları insan beyninin yapısından yola çıkarak tasarlanmış örüntü tanıma ve hata minimizasyonu üzerine kurulmuş bir yöntemdir." Bilgiyi içeriye alarak hafızasında tutan, her bir tecrübesinde yeni bir şey öğrenen ve veriler arasındaki ilişkiyi ortaya çıkaran bir yapıyı temsil etmektedir.[21] Yapay sinir ağları ile sinir sisteminin çalışma şekli örnek alınmış, nöronları içeren sinir hücreleri bir araya gelerek sinir ağını oluşturmuştur.



Şekil 2.5. Yapay Sinir Ağı [22]

Şekilde bir yapay sinir ağı görünmektedir. Basit bir sinir ağı girdi katmanı, gizli katman ve çıktı katmanından oluşmaktadır. Girdi katmanındaki her bir şekil bir değişkeni ifade etmektedir. Bu değişkenler biyolojik sinir ağındaki sinir hücrelerine karşılık gelir. Bu sinir hücreleri bir araya gelerek sinir ağını oluşturmuştur.

Sinir ağının işleme sürecinde öncelikle her bir değişken bir bağlantı ağırlığı ile çarpılır. Nöronlar giriş bilgilerini ağırlıklandırdıktan sonra toparlayarak doğrusal

[21] Olivia Parr Rud, **Data Mining Cookbook Modeling Data for Marketing, Risk and Customer Relationship Management**, New York, John Wiley, 2001, s. 16
[22] Han, Kamber, **a.g.e.**, s. 331

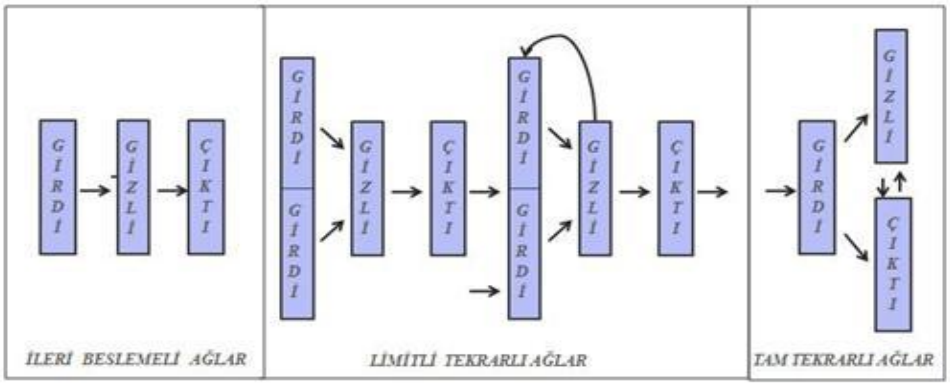
veya doğrusal olmayan bir fonksiyonda işlerler ve çıktı bilgisine dönüştürürler.[23] Bu bilgi, diğer nöronlar için girdi bilgisi olarak kullanılır. Bu işlemler her bir katmanda gereksiz bilgiler elenerek, diğer bütün katmanlarda da tekrarlanır ve sonuçta yapay sinir ağı modelini oluştururlar. Sinir ağlarının farklı yapılara sahip olmaları ve bu yapıların işleyişleri ağ mimarileri başlığı altında toplanabilir.

2.4.1. Ağ Mimarisi

Yapay sinir ağlarının üç farklı mimarisi bulunmaktadır. Bunlar ileri beslemeli ağlar, limitli tekrarlı ağlar ve tam tekrarlı ağlardır.
İleri beslemeli sinir ağlarında tüm işlemler tek bir akışla bitirilir. Öğrenme ve test süreçleri girdi katmanından başlar, gizli katmandan geçer ve çıktı katmanında son bulurlar. Bu süreç bir defa yaşanır. Girdi birimlerinin ilk andaki her bir değeri, o birim için aktivasyon değerini ifade eder. Çıktı değerleri aktivasyon değerleri ve bağlantı ağırlıklarına göre belirlenir. Aradaki süreçte değerler, genellikle sigmoid fonksiyonu olan bir aktivasyon fonksiyonu ile azalarak ya da artarak ilerler. Bu, sinir hücresine gelen sinyallerin şiddetlenmesi ya da hafiflemesi gibi düşünülebilir.

Limitli tekrarlı ağlarda girdilerin sırası önemli olabilir ve tüm önceki girdilerin değerleri tutularak bu değerler diğer katmandaki güncel değerler ile harmanlanır. Her an bir geri dönebilme söz konusu olduğundan geçmiş değerlerin tutulması, biyolojik ağlardaki gibi bir hafızaya sahip olunması söz konusudur. Aslında tamamen geri dönüş mekanizması tam tekrarlı ağlarda mümkün olabilmekte, limitli tekrarlı ağlarda bazı girdi kümelerinin değerlerine geri dönüş olabilmektedir. Bu durumda limitli tekrarlı ağların ileri beslemeli ve tam tekrarlı ağ mimarileri arasında bir geçiş olduğu düşünülebilir. Şekilde görüldüğü gibi, bazı girdilere geri dönüş ve girdilerin geçmiş bilgilerini kullanabilme söz konusu iken, bazıları için bu durum söz konusu değildir.

[23] Ayşe Yazıcı, v.d. , "Yapay Sinir Ağları'na Genel Bakış"**, Tıp Bilimleri Dergisi,** 2007, (Çevrimiçi), http://209.85.229.132/search?q=cache:nwaMpL4GkEJ:tipbilimleri.turkiyeklinikleri.com/download_pdf, 18.Mayıs.2009



Şekil 2.6. Yapay Sinir Ağı Mimarileri[24]

Tam tekrarlı ağlar ise bütün katmanlar arasında her türlü ileri ve geri harekete izin veren bir yapıya sahiptir. Aktivasyon değerleri birimlerin değerlerinin alt kümelerinden sınanarak ortaya çıkar ve sabit değildir. Her bir ileri geri harekette bu değerler değişmekte ve bu hareket, değerler sabitlenene kadar sürebilmektedir.

2.4.2. Yapay Sinir Ağı Öğrenme Süreci

Yapay sinir ağları hem denetimli hem de denetimsiz öğrenme için çeşitli yöntemler sunar. Denetimli öğrenmede amaç örnekler için daha önceden belirlenmiş çıktı değerlerinden yola çıkarak tahminsel bir modelleme geliştirmek iken, denetimsiz öğrenmede verileri özelliklerine göre gruplamaktır.

Bilinen yapay sinir ağı algoritmaları, mimarileri ve öğrenme şekilleri aşağıdaki gibidir.

MODEL	EĞİTİM ŞEKLİ	AĞ MİMARİSİ	BİRİNCİL FONKSİYONLARI
Geri Yayılım Algoritması	Denetimli	İleri Beslemeli	Sınıflandırma, Zaman Serileri
Tekrarlı Geri Yayılım Algoritması	Denetimli	Limitli Tekrarlı	Zaman Serileri
Radyal Tabanlı Fonksiyonlar	Denetimli	İleri Beslemeli	Sınıflandırma, Zaman Serileri

[24] Joseph P.Bigus, **Data Mining with Neural Networks**, USA, McGraw-Hill, 1996, s. 63,64

Uyarlamalı Rezonans Kuramı	Denetimsiz	Tam Tekrarlı	Kümeleme
Olasılıklı Sinir Ağları	Denetimli	İleri Beslemeli	Sınıflandırma
Kohonen Ağları	Denetimsiz	İleri Beslemeli	Kümeleme

Tablo 2.1. Yapay Sinir Ağı Algoritmaları[25]

Geri yayılım algoritması adını, hataları çıktı katmanından geriye doğru azaltmaya çalışmasından almaktadır. Denetimli öğrenme şekline sahip olup, sınıflandırma problemleri ile ilgilenmektedir. Geri yayılım algoritması, sinir ağının çıkış noktasındaki hata düzeyine göre bütün tabaka ağırlıklarını yeniden hesaplayarak çalışır. [26] Geri yayılım algoritmasında sinir ağlarının bütün katmanları bulunur ve birden fazla gizli katman olabilmesi olasıdır.

Kohonen ağları denetimsiz bir öğrenme metodu sunar. Tahmin edilmek istenen bir bağımlı değişken olmadığından bu ağlarda gerçek bir çıktı katmanı olduğu söylenemez. "Kohonen ağları bir girdi ve iki boyutlu kohonen tabakasından oluşmaktadır." [27] "Çok boyutlu girdi örüntülerinden daha düşük boyutlardaki çıktı kümeleri yaratan bir yapıya sahiptir. Bu kümeler, girdi verilerinin özellikleri arasında en sık gerçekleşen örüntülerdir." [28]

2.5. Genetik Algoritmalar

"Genetik algoritmalar, bir fonksiyonun optimizasyonu veya ardışık değerlerin tespitini içine alan birçok problem tipleri için çözüm arayan bir yöntemdir. Genetik algoritmalar, doğal seçilim ilkesine ve en iyinin korunumuna dayanırlar. Benzetim yoluyla bilgisayarlara uygulanan ve bilgisayar üzerinde oluşan bir evrim şeklidir. Genetik

[25] Bigus, **a.g.e.**, s. 77
[26] Evangelos Triantaphyllou, Giovanni Felici, **Data Mining and Knowledge Discovery Approaches Based on Rule Induction Techniques**, New York, Springer, 2006, s. 520
[27] Ayşe Oğuzlar, "Kümeleme Analizinde Yeni Bir Yaklaşım", **Atatürk Üniversitesi İ.İ.B.F Dergisi**, 2005, (Çevrimiçi), http://194.27.49.253/iibf/CV07.pdf, 20.Mayıs.2009
[28] David Taniar, **Research Trends in Data Mining Technologies and Applications**, 2007, s. 123

algoritmaların amacı, hem problemleri çözmek hem de evrimsel sistemleri modellemektir."[29]

"Genetik algoritmalar bir çözüm uzayındaki her noktayı, kromozom adı verilen ikili bit dizisi ile kodlar. Her noktanın bir uygunluk değeri vardır. Tek bir nokta yerine, genetik algoritmalar bir popülasyon olarak noktalar kümesini muhafaza eder. Her kuşakta, genetik algoritma, çaprazlama ve mutasyon gibi genetik operatörleri kullanarak yeni bir popülasyon oluşturur. Birkaç kuşak sonunda, popülasyon daha iyi uygunluk değerine sahip üyeleri içerir." Genetik algoritmalar, çözümlerin kodlanmasını, uygunlukların hesaplanmasını, çoğalma, çaprazlama ve mutasyon operatörlerinin uygulanmasını içerir.[30]

Genetik algoritmaların adımları aşağıdaki gibidir.[31]

- Tüm mümkün çözümler tanımlanır.
- Rastgele bir çözüm kümesi seçilir ve başlangıç populasyonu olarak değerlendirilir.
- Belirlenen çözümler için uygunluk fonksiyonu tanımlanır ve bu uygunluk fonksiyonlarına göre bireyler seçilir. Seçim işleminde uygun ve iyi olmayan bireyler elenir.
- Çaprazlama ve mutasyon yöntemleri ile yeni nesiller oluşturulur.
- Süreç belirlenen nesil sayısına ulaşıncaya kadar tekrarlanır.

2.6. Zaman Serileri

Gözlem sonuçlarının; dakika, saat, gün, hafta, ay, mevsim, yıl gibi herhangi bir zaman unsuru dikkate alınarak dizi haline getirilmesine zaman serisi denilmektedir.

---

[29] Arif Gülten, Şengül Doğan, "Genetik Algoritmalar Yönteminin Biyomedikal Verileri Üzerinde Uygulamaları", **Doğu Anadolu Bölgesi Araştırmaları Dergisi**, Ekim 2008, (Çevrimiçi), http://web.firat.edu.tr/daum/docs/71/03, 18.Mayıs.2009

[30] Gül Gökay Emel, Çağatan Taşkın, "Genetik Algoritmalar ve Uygulama Alanları", **Uludağ Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi**, 2002, s. 129-152 (Çevrimiçi), http://www.yapay-zeka.org/files/tez/genetik_algoritmalar_ve_uygulama_alanlari.pdf, 10.Mayıs.2009

[31] Lance D. Chambers, **Practical Handbook of Genetic Algorithms Complex Coding Systems Volume 3**, CRC, 1998, s. 31,32

Bu noktadan hareketle, serilerin geçmiş ve bu günkü değerleri kullanılarak gelecek dönem hakkında tahminler yapılmasının zaman serileri analizinin konusu olduğu söylenebilir.

Zaman serilerinde gözlem değerleri birbirlerine bağımlı olmaları özelliği kullanılarak ileriye dönük tahmin yapıldığından diğer serilerden bu noktada ayrılmaktadır. Düzensiz dalgalanmalardan meydana gelen zaman serisinin dalgalanmaları, serinin bileşenleri olan dört unsurdan kaynaklanır. Bu unsurlar trend, konjonktür dalgalanmaları, mevsimsel dalgalanmalar ve tesadüfi nedenler olup, unsurlar vasıtasıyla dalgalanmaların sebeplerinin araştırılması amaçlanmaktadır.

Zaman serilerinin en çok bilinen uygulaması finans kuruluşları tarafından finansal piyasalar ile ilgili tahminlerin yapıldığı durumlardır.

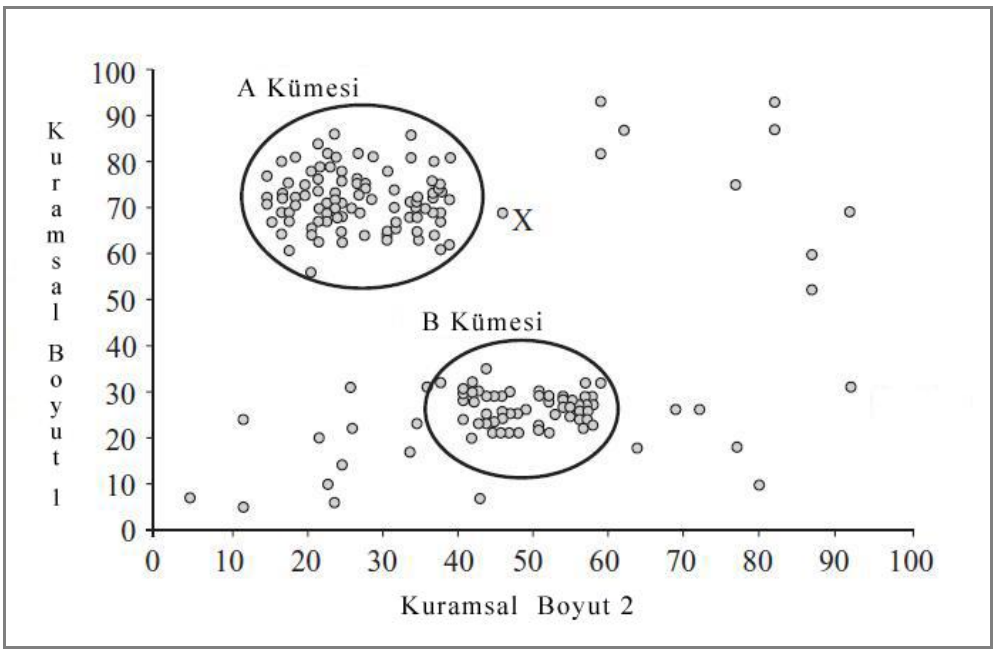
2.7. Kümeleme

Bir denetimsiz öğrenme metodu olan kümeleme, birbirine benzer verileri sınıflara ayırarak verileri özetleme veya fiziksel olarak gruplandırma sürecidir. Kümeleme yöntemi, büyük veri kümelerinin anlamlı şekilde parçalanarak alt kümelerine ayrıldığı ve benzer grupların bir arada ele alındığı bir süreçtir.[32] Kümeleme yönteminde bir çıktı değişkeni yoktur. Bu sebeple denetimsiz öğrenme metodu olarak bilinmektedir. Bu noktada kümeleme, veri kümelerinde verileri birbirinden ayıran başka bir yöntem olan diskriminant analizinden ayrılmaktadır. Zira kümelemede küme sayısı bilinmemekte ve analiz sonucunda veriden elde edilmektedir. Bununla birlikte kümelemede herhangi bir fonksiyon elde edilerek sonrasında diğer veriler için kullanılma durumu yoktur; çünkü ayırma işlemi tamamen o verilerin özellikleri kullanılarak yapılır.

"Küme, birbirlerine yakın bireylerin çok boyutlu uzayda oluşturdukları birlik olarak ifade edilebilir. Bu durumda küme kavramı, benzerlik ve uzaklık kavramlarını

[32] Lin Ohsuga, Liau Hu, **Foundations and Novel Approaches in Data Mining,** Warsaw, Springer, 2005, s. 121

çağrıştırmaktadır."[33] "Kümelemede verilerin normal dağılması gerektiği varsayımı olmakla birlikte, bu varsayım prensipte kalmakta, uzaklık değerlerinin normalliği yeterli görülmektedir."[34]



Şekil 2.7. Kümeleme [35]

Kümeleme süreci, her bir verinin farklı özelliklerini ifade eden değerlerinin, diğer bir veriden farklı ya da aynı olmasına göre işler. Farklılık ya da aynılık kümeleme algoritmaları için uzaklıkları ifade etmektedir. Kümeleme algoritmaları, öklid uzaklığı, minkowski uzaklığı, manhattan uzaklığı, canberra uzaklığı gibi farklı uzaklıkları kullanabilirler.

Bu uzaklıklarla çalışan bilinen kümeleme algoritmaları aşağıdaki gibidir.

1. Paylaştırma Yöntemleri : Veri kümesini, her bir kümede en azından bir veri olacak şekilde ve her bir veri sadece bir kümede olacak şekilde k kümeye

[33] Doğan İ. : "Kümeleme Analizi ile Seleksiyon", **Turkish Journal of Veterinary and Animal Sciences**, 2000, (Çevrimiçi), http://journals.tubitak.gov.tr/veterinary/issues/vet-02-26-1/vet-26-1-7-0007-1.pdf, 13.Ocak.2009
[34] Tatlıdil, **a.g.e.**, s. 329
[35] Glenn J. Myatt, **Making Sense of Data**, US, John Wiley & Sons Publication, 2007, s. 110

ayırır. Küçük ve orta büyüklükteki veri tabanlarında küre şekilli kümeleri bulmada iyidirler. Bu yöntemlerin en bilinen algoritmaları k-ortalamalar ve k-medoidler algoritmalarıdır.

2. Hiyerarşik Yöntemler : Önce gruplara ayırarak sonra kümeleme veya tüm veri kümesini aynı kümeye atayarak sonrasında ayırma gibi çeşitleri vardır. En bilinen algoritmalar BIRCH ve ROCK algoritmalarıdır.
3. Yoğunluk Tabanlı Yöntemler : Uzaklıklar ile çalışan yukarıdaki yöntemler genellikle küresel şekilleri bulmada iyidirler, yoğunluk tabanlı modeller bunların aksine rastgele şekilleri keşfetme özelliğine sahiptir. DBSCAN ve OPTICS tipik yoğunluk tabanlı algoritmalardır.
4. Izgara Tabanlı Yöntemler : Nesne uzayını, sonlu sayıda hücrenin bulunduğu bir uzaya indirgeyerek işlem yapar. STING algoritması bu yöntemlere örnek olarak verilebilir.
5. Model Tabanlı Yöntemler : Model tabanlı yöntemler her bir küme için bir model varsayımında bulunurlar ve veriye en uygun olan modeli bulmaya çalışırlar. EM ve COBWEB örnek olarak gösterilebilecek algoritmalardır.

Kümeleme yöntemleri, birçok sektörde birçok amaçla kullanılmaktadır. En sık rastlanan kullanımlarından biri pazarlama sektöründe müşterilerin segmentlere ayrılarak farklı segmentler için farklı pazarlama stratejilerinin geliştirildiği durumlardır.

2.8. Birliktelik Kuralları ve Sıralı Örüntü Analizi

Büyük veri kümelerinde beklenen birliktelikler dışında olan, farklı bilgi vererek çıkarım yapılabilmesini sağlayan birlikteliklerin tespit edilmesi birliktelik, ardışıklıkların (bir olayın gerçekleşmesinin arkasından ötekinin gerçekleşiyor olması) tespit edilmesi ise sıralı örüntü analizini ifade eder.

Birliktelik kuralları, bilgisayar bilimleri alanında geliştirilmiş ve genellikle Pazar sepet analizi, belirli müşteriler tarafından alınan ürünlerin birlikteliklerini ölçümlemek veya web sitelerine tıklayan kişilerin ardışık olarak ziyaret ettikleri

sayfaları görüntülemek gibi önemli analizlerde kullanılırlar. Amaç, belirli bir işlemler kümesinde tipik olarak birlikte olan kalemlerin altını çizmektir.[36]

"Veriler arasındaki ilişkileri açıklamak için eğer-sonra ifadeleri kullanılır ve *eğer* < bazı şartlar sağlanırsa > *sonra* < bazı niteliklerin değerlerini tahmin et > şeklinde belirtilebilir. Eğer bölümü ile ilişkili durumlar öncül, sonra bölümü ile ilişkili durumlar sonuç olarak adlandırılır. Öncül ve sonuç durumları X ve Y olarak ele alınırsa buradaki ilişki X → Y şeklinde sembolize edilebilir."[37]

Bu analizde önemli olan iki kavram destek ve güven kavramlarıdır. Destek, yukarıda belirtilen kural cümlesini sağlayan olayların birlikte gerçekleştiği durumların tüm durumlara oranı iken, güven, kural cümlesini sağlayan olayların birlikte gerçekleştiği durumların öncül bölümde gerçekleşen olaya oranıdır.[38] Örneğin, bir marketten satın alınan ürünlerin hangilerinin birlikte alındığını araştırdığımızı varsayalım. Kişi bir alışverişinde A,B,C ürünlerini, diğer bir alışverişinde K, L ve B ürünlerini, başka bir alışverişinde A,M ürünlerini son olarak da A,B ve P ürünlerini aldığını düşünelim. Toplam 4 defa alışveriş yapan bu kişi 2 defasında A ve B ürünlerini birlikte almıştır. Bu durumda A ve B ürünlerinin birlikte alındığı olay sayısı 2, tüm olay sayısı 4 olduğundan destek 2/4, A ve B'nin birlikte alındığı olay sayısı olan 2'nin öncül bölüm olarak A'nın bulunduğu olay sayısı 3'e bölümü olan güven de 2/3 olarak bulunacaktır. Bu durumda kişi, tüm işlemlerinin %50'sinde A ve B ürünün birlikte alırken, A ürünün aldığı işlemlerin %33'ünde de B ürünün aldığını söyleyebiliriz.

Sıralı örüntü analizinde gerçekleşen olaylarda ardışıklık olduğu ifade edilebilir. Örneğin bir web sitesinin arka arkaya ziyaret edilen sitelerin analiz edilmesi sonucu ortaya çıkabilecek farklı kurallar bütünü sıralı örüntü analizi olarak ifade edilebilir.

---

[36] Guidici, **a.g.e.**,, s. 121

[37] Gül Gökay Emel, Çağatan Taşkın, Arif Tok, "Pazarlama Stratejilerinin Oluşturulmasında bir Karar Destek Ağacı, Birliktelik Kuralı Madenciliği"**, Dokuz Eylül Üniversitesi Sosyal Bilimler Enstitüsü Dergisi**, 2005, (çevrimiçi), http://www.sbe.deu.edu.tr/adergi/2006/Cilt%207%20Sayi%203%202005/emel-taskin-tok.pdf , 18.Mayıs.2009

[38] Emel, Taşkın, Tok, **a.y.**

Birliktelik kuralları ve sıralı örüntü analizinin sınıflandırma algoritmalarından farkı, bir değişkenin iki yönlü olabilmesi, bazı durumlar için girdi, bazıları için çıktı değişken gibi tanımlanabilmesidir. En çok bilinen birliktelik algoritmaları Apriori, GRI ve CARMA algoritmalarıdır.

2.9. Uç Değer Analizi

Veri kümesinde diğer verilere göre çok farklı olan, şüphe uyandırarak diğerlerinden farklı değerlendirilmesi gereken gözlemlere uç değer denir. Uç değerler anormal davranışları tanımlarlar ve doğal veri değişkenliğinden sapmış verilerdir.

Uç değerler, yanlış girişi yapılmış bir gözlemi temsil edebildiği gibi, diğerlerinden çok farklı özellikleri olan bir gözlemi de temsil edebilir. Örneğin bir veri tabanında yaşı çok yüksek bir rakam olarak girilen bir kişi ile, çok yüksek bir harcamayı gerçekten yapan bir market müşterisi veya kredi kartı ile sahtekarlık yapan bir banka müşterisinin durumu da uç değerdir. Her üç durum da veriyi saptırmakta, modellemede bir takım yanlışlıklara sebep olmaktadır. Veri madenciliği çalışmasından önce uç değerler belirlenmeli ve uygun çözüm yöntemleri ile modele negatif etkisi önlenmelidir.

Veri kümesindeki uç değerlerin tespiti için kullanılan yöntemler, istatistiksel yöntemler, uzaklık tabanlı yöntemler, yoğunluk tabanlı yöntemler ve sapma tabanlı yöntemler olarak tanımlanabilir.

3. UYGULAMA

Bu bölümde, Deniz Feneri Derneği'nin operasyonel veri tabanından alınan veriler kullanılarak bir uygulama yapılmıştır.

Dernek, kendilerinden yardım talep eden kişi ve/veya ailelere yardımlar yapmakta, yardım yapılabilmesi için gerek sosyal incelemeler gerekse telefon görüşmeleri yaparak kişi ve/veya ailelerin sosyo – ekonomik düzeyleri, ihtiyaç duydukları yardım türü ve ne kadar süre bu yardıma ihtiyacı oldukları belirlenmektedir. Tüm bu inceleme ve görüşmeler sonucu gerekli belgeler toplanarak, yardım talep eden hakkında gerçekten yardıma ihtiyacı olup olmadığına, yardım yapılabilmeye uygunluğuna karar verilmektedir.

Uygulamadaki amaç, yardım talep eden kişiler hakkında, dernek tarafından toplanarak veri tabanına aktarılan bilgilerin ışığında, yardım kararı vermek için aracı olabilecek bir model geliştirmektir. Yapılan uygulamada 59.119 aile bilgisi kullanılmış olup, verilen sonuçlar bu ailelerin bilgilerinden çıkan sonuçlardır. Farklı örnek kümeleri farklı sonuçlar verebilmektedir. Bu sebeple çalışması yapılmayan aileler ile ilgili genel yargıya varılmamalıdır.

Yardım kararı için geliştirilen modeller, çıktı değişkeni kategorik olan tahminleme modelleridir. Lojistik Regresyon, Yapay Sinir Ağı ve Karar Ağacı (Chaid ve C&R Tree) algoritmaları aynı veriler üzerinde denenerek nihai modele karar verilmiştir. Veri hazırlanmasında ve modellemede SPSS firmasının veri madenciliği çözümü olan Clementine paket programı kullanılmıştır.

3.1. Verilerin Hazırlanması

Veriler, derneğin operasyonel veri tabanından Ağustos – 2008 tarihinde alınmıştır. Adana, Ankara, Samsun, İstanbul, İzmir ve Erzurum illerinden yardım talep eden

ailelerin bilgilerini içermektedir. Veriler hazırlanırken yapılan bazı kabullenmeler aşağıdaki gibidir.

- Veri tabanında tüm bilgiler, her biri bir aileyi ifade eden tekil id bazında tutulduğundan, modelleme öncesinde veriler bu tekil id bazında hazırlanmıştır. Her bir aile için farklı boyutları vurgulayan ve yardım kararını etkileyebilecek değişkenler, veri tabanında bulunan ana ve ara tablolardandirektolarak alınmış veya türetilmiştir.
- Verilerin alındığı anda, daha önce alınmış herhangi bir yardım kararı bulunmayan (yardım edilecek yada edilmeyecek şeklinde) aileler analize dahil edilmemiştir.
- Bir aileye en az 1 defa yardım kararı çıktı ise, aile yardım almış olarak işaretlenmiştir.
- Birden fazla sosyal inceleme yapılan aileler için, sosyal durumlarının en güncel hali son durumları olarak kabullenilmiş ve son incelemenin sonuçları girdi olarak alınmıştır.
- Gelir – gider durumları ile ilgili bilgiler kullanılırken, veri tabanının içeriği gereği kategorik ve sürekli halde alınmış olan bilgilerde tutarsızlık var ise elle düzeltmeler yapılmış, bunu yaparken kategorik değeri 0 yani belirtilen durum söz konusu değil şeklinde işaretleme yapılmış; fakat aynı bilginin tutarı 0'dan büyük ise, bu durumun söz konusu olduğu kabullenmesi yapılmış ve kategorik değer 1'e çekilmiştir. Örneğin, gelir kira değişkeni, o aile için 0 belirtilmiş; fakat gelir kira tutarı 0'dan büyük olarak yansıtılmış ise, o ailenin kira geliri vardır kabullenmesi yapılmıştır.

Yukarıdaki kabullenmeler sonucu yapılan gerekli düzeltmeler ile birlikte, tüm değişkenlerin tanımlayıcı istatistikleri incelenmiş, tek bir kategorinin çok fazla ağırlık gösterdiği, bilgi vermeyeceği düşünülen değişkenler elenmiştir. Uç değerler, o değişken için tüm verinin dağılımına bakılarak belirli bir noktaya çekilmiştir. Ayrıca boş değerler, o değişken içindeki diğer tüm kategorilerden daha yüksek miktarda ise

değişken çıkarılmış, aksi takdirde söz konusu durumu sağlamadığı şeklinde doldurulmuştur.

Tüm bu çalışmaların sonucunda, modellemede kullanılabilecek girdi değişkenleri ile "karar" olarak ifade edilen çıktı değişkeni hazırlanmıştır.

3.2. Modelin Kurulması

Bu aşamada verilere, lojistik regresyon, yapay sinir ağları, CHAID, C&R Tree modelleri uygulanmıştır. Çıktı değişkeni olarak *karar* alınmış, 20.000 aileyi içeren toplam veri kümesi içerisinde karar değişkeninin kategorileri eşit ağırlıklandırılmıştır. Sonrasında oluşturulan modeller yine 20.000 aileyi içeren bir veri kümesi ile test edilmiş, 19.119 aileyi içeren başka bir veri kümesi ile de doğrulanmıştır.

Veri hazırlığı sonucu oluşturulan değişkenler ve ifade ettikleri aşağıdaki gibidir. Bu değişkenlerin bir kısmı diğer değişkenler ile ilişkileri sebebi ile çıkarılmış, bir kısmı da anlamsız olduklarından sonradan modellere dahil edilmemişlerdir.

Karar: Karar değişkeni bağımlı değişkendir. Öncesinde yapılan başvurular neticesinde derneğin aldığı kararları ifade eder. Yardım edilme kararı alınmış ise 1, yardım edilmeme kararı alınmış ise 0 olarak işaretlenmiştir. Modeller, daha önce alınan bir kararı baz alarak her bir aile için bir tahmin kararı oluştururlar. Sonrasında kullanılacak olan modelin, kararı bağımsız değişkenlerin değerleri iledirektolarak tahminleyebiliyor olması amaçlanmaktadır.

Sürekli Tedavi: İki kategorili bir değişken olup, aile içinde sürekli tedavi gören bir kişinin olup olmadığını belirtmektedir.

Sigara Kullanan: Aile içerisinde sigara kullanan birinin olup olmadığını ifade etmektedir. İki kategorili bir değişkendir.

Cep Telefonu: Aile içerisinde herhangi bir kişide cep telefonu olup olmadığını ifade etmektedir. İki kategorili bir değişkendir.

Borç Yok: Ailenin borcu yok ise 1 var ise 0 şeklinde işaretlenmiş iki kategorili bir değişkendir.

Gelir Düzensiz Tutar: Ailenin, dernek tarafından araştırılan kira, maaş, tarım veya hayvancılıktan gelebilecek gelirlerinin dışında elde edilen ve belirli bir düzeni olmayan gelirleri ifade eder. Sürekli bir değişkendir.

Gider Fatura Tutar: Ailenin bir ayda yaptığı fatura harcamasının toplamını ifade eden sürekli bir değişkendir.

Gider Eğitim Tutar: Ailenin okuyan bireyleri için yapılan aylık toplam harcamayı belirten, sürekli bir değişkendir.

Toplam Kira: Ailenin verdiği aylık kira bedelini ifade eden sürekli bir değişkendir.

Oda Sayısı: Ailenin yaşadığı evde bulunan oda sayısını belirten kategorik bir değişkendir.

Görüşme Sayısı: Aile ile, ilk yardım talep edildiğinden itibaren yapılan direk, posta yolu ile veya telefonla yapılan iletişimin sayısını ifade eder. Bu görüşme, aileden birinin mektup yazarak durumunu ve yardım isteğini bildirmesi olabilmekle birlikte, yaşanılan ortamın görüldüğü sosyal incelemeler, komşularla, esnafla görüşülerek aile hakkında bilgilerin toplanabilmesi şeklinde de olabilmektedir.

Ailedeki Fert Sayısı: Ailedeki birey sayısını belirten kategorik bir değişkendir.

Ailede Okuyan Kişi Sayısı: Ailede herhangi bir öğrenim kurumuna kayıtlı kişi sayısını ifade eder.

Sosyal Güvencesi Olmayan Kişi Sayısı: Aile içerisinde herhangi bir sosyal güvencesi olmayan kişi sayısı ifade eder.

Gıda Yardımı Alınıyor mu: Ailenin, bağlı olduğu kaymakamlık, belediye gibi resmi kurumlardan veya herhangi bir vakıftan gıda yardımı alıp almadığını belirten iki kategorili bir değişkendir.

Ev Tipi: Ev tipi, ailelerin oturdukları evlerin apartman dairesi, müstakil ve gecekondu olup olmadığını veya bunların birleşimlerini ifade eden kategorik bir değişkendir. Ailenin ev tipine yönelik herhangi bir bildirim yapmaması halinde 0, müstakil ve gecekondu olarak bildirmesi halinde 1, gecekondu olarak bildirmesi halinde 2, hem apartman hem müstakil olarak bildirmesi halinde 3, sadece müstakil olarak bildirmesi halinde 4, sadece apartman olarak bildirmesi halinde ise 5 olarak işaretlenmiştir.

İkamet Türü: Ailelerin; ev sahibi, kiracı, bedelsiz oturma, geçici ikamet durumlarından herhangi birine veya bunların birleşimlerinden birine uyup uymadığını belirtir.

Isınma Tipi: Ailelerin; elektrikli ısıtıcı, soba, katalitik, kalorifer olarak belirlenen ısınma türlerinden herhangi birini veya bunların birleşimlerinden birini kullanıp kullanmadığını belirten kategorik bir değişkendir.

Ev Eşyası Var mı: Ailelerin evinde bulunan beyaz eşya harici, halı, çekyat, masa, koltuk, yatak vitrin gibi eşyalardan herhangi birinin bulunduğunu belirten iki kategorili bir değişkendir.

İlk olarak lojistik regresyon yöntemi uygulanmış, model uygulanmadan önce gerekli korelasyon analizi yapılarak birbiri ile ilişkisi yüksek değişkenler elenmiştir. Lojistik modelde çıkan anlamsız değişkenler de elenerek modelin son haline ulaşılmış, diğer modeller de bu değişkenler üzerinden çalıştırılmıştır.

3.2.1. Lojistik Regresyon Modeli

Modelin doğruluk oranları incelenmiş ve uygun değişkenlere sahip lojistik regresyon modeli belirlenmiştir. Belirlenen lojistik regresyon modeli, içerdiği değişkenler ve değişkenlerin tahmin edilen katsayıları aşağıdaki gibidir.

Variables in the Equation			
	B	S.E.	Sig.
SUREKLI_TEDAVI	.410	.057	.000
SIGARAKULLANAN	-.711	.046	.000
BORC_YOK	-.546	.061	.000
GELIR_DUZENSIZ_TUTAR	-.001	.000	.000
GIDER_FATURA_TUTAR	-.003	.001	.000
GIDER_EGITIM_TUTAR	.006	.001	.000
TOPLAM_KIRA	.008	.001	.000
GORUSME_SAYISI	.178	.004	.000
SOSYAL_GUV_YOK	.192	.008	.000
GIDA_YARDIMI_ALIYORMU2	.363	.055	.000
IKAMET_TURU	-.294	.008	.000

Tablo 3.1. Lojistik Regresyon Modeli/Bağımsız Değişkenler ve Modeldeki Katsayıları

Model incelendiğinde, modeli oluşturan değişkenler ve katsayıları istatistiksel olarak anlamlı ve mantıksal olarak uygun denilebilir. Modele göre, ailede sürekli tedavi gören birinin olması o aileye yardım edilmesi olasılığını arttırıyor görünmektedir. Aynı şekilde ailenin eğitim giderlerinin ve kira giderlerinin yüksek olması, ailede sosyal güvencesi olmayan kişi sayısının yüksek olması da aileye yardım edilmesi olasılığını arttırıyor gibi görünmektedir. Gıda yardımı alıyor mu değişkenine baktığımız zaman, başka bir resmi kurum tarafından gıda yardımı yapılıyor olması, hem ailenin başka kurumlar tarafından yardım ihtiyacının belirlenmiş olduğunu, hem de temel yaşam giderlerini dahi karşılayamıyor olduklarını belirtebileceğinden modele pozitif yönde etki ediyor olması akla yatkındır. Görüşme sayısının, derneğin aile ile kaç defa görüşme yaptığını ifade ettiğini düşünürsek, dernek, aileye ne kadar çok yakınlaşır ise ekonomik durumu hakkında o derece fikir sahibi olacağından, aile ile iletişim kurma sayısının pozitif yönde bir etkiye sahip olması mantığa uygun olabilir.

Ailede sigara kullanan birilerinin olması, sigara harcaması, aile içerisindeki bireylerin diğer harcamalarına ikame olarak yapıldığından yardım kararı alınmasında

negatif yönde etkisinin olması uygun görünmektedir. Zira ailede sigara içen biri var ise, aileye yardım edilme olasılığı azaldığını ifade etmektedir. Aynı şekilde ailenin borcunun olmaması, düzensiz gelirinin olması, fatura giderlerinin yüksek olması da aileye yardım edilme olasılığını azaltıyor gibi görünmektedir. İkamet türünde ise alt kategoriler ailenin geçici veya bedelsiz ikamet ettiğini belirttiğinden, ev sahibi veya kiracı olmasının yardım edilme olasılığı düşürmesini bekleyebiliriz. Değişkenin modelde negatif yönlü etkisi, bu beklentimizi doğruluyor görünmektedir. Modelin doğru tahmin etme oranlarını ifade eden Clementine çıktısı ise aşağıdaki gibidir.

SINIFLANDIRMA TABLOSU - EĞİTİM KÜMESİ				
		Tahmin edilen		
		KARAR		Doğruluk
Gözlenen		0	1	Oranı
KARAR	0	7.137	2.863	71,3
	1	1.697	8.303	83
Toplam Oran				77,2
SINIFLANDIRMA TABLOSU - TEST KÜMESİ				
		Tahmin edilen		
		KARAR		Doğruluk
Gözlenen		0	1	Oranı
KARAR	0	7.746	2.254	77,4
	1	2.178	7.822	78,2
Toplam Oran				77,8
SINIFLANDIRMA TABLOSU -DOĞRULAMA KÜMESİ				
		Tahmin edilen		
		KARAR		Doğruluk
Gözlenen		0	1	Oranı
KARAR	0	7.056	2.063	77,3
	1	1.852	8.148	81,4
Toplam Oran				76

Tablo 3.2. Lojistik Regresyon Modeli – Doğruluk oranları

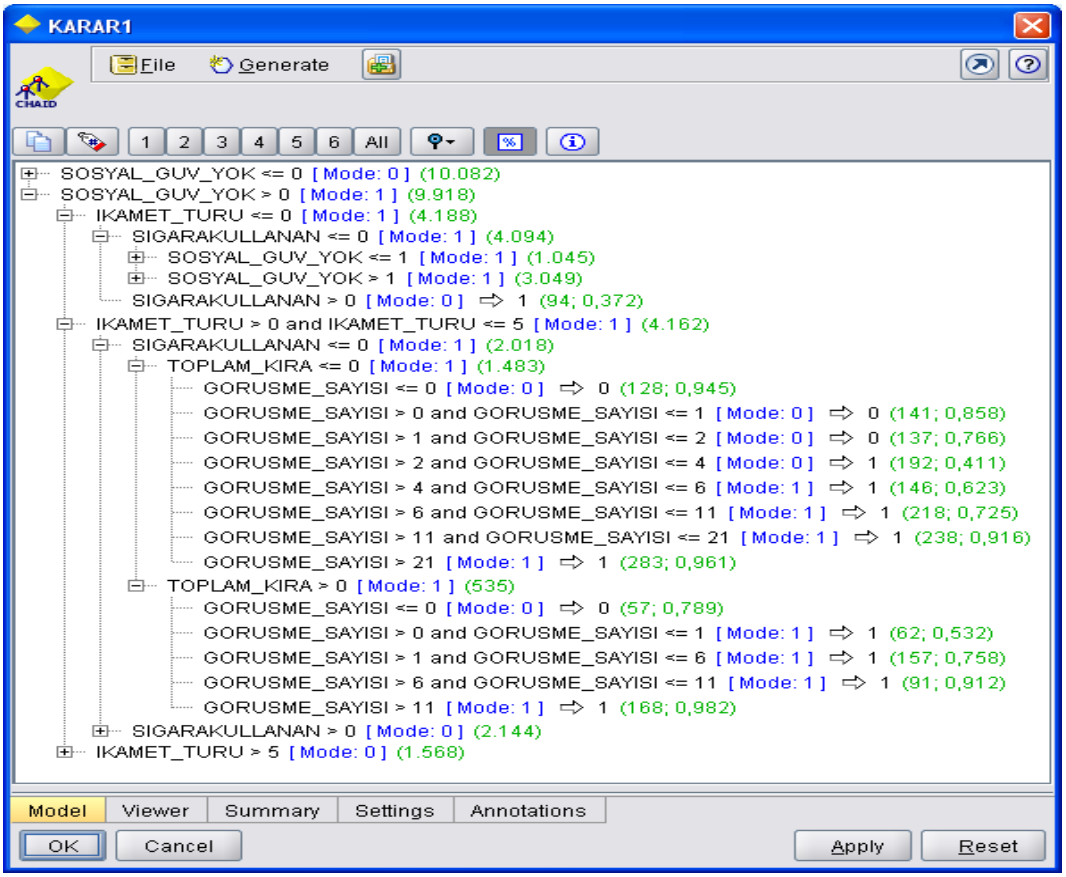
Göründüğü gibi model, daha önce yardım kararı alınmış aileleri %83 oranında, yardım kararı alınmamış aileleri ise %71,3 oranında tahmin etmektedir. Modelin toplam doğruluk oranı ise % 77'dir. Test ve doğrulama kümelerinde de durum benzer görünmektedir.

Yukarıda katsayıları ve doğru tahmin etme oranları görünen lojistik regresyon modeli aşağıdaki gibi ifade edilebilir.

$$Log\left(\frac{P(Karar)}{1-P(Karar)}\right) = 0{,}410*surekli\_tedavi - 0{,}711*sigara\_kullanan - 0{,}546*borc\_yok - 0{,}001*gelir\_duzensiz\_tutar + 0{,}003*gider\ fatura\_tutar + 0{,}006*gider\_egitim\_tutar + 0{,}178*görüşme\_sayisi + 0{,}192*sosyal\_guvenlik\_yok + 0{,}363*gida\_yardimi\_aliyormu - 0{,}294*ikamet\_turu$$

3.2.2. CHAID Modeli

Belirtilen değişkenler ile çalıştırılan chaid algoritması sonucunda elde edilen karar ağacının yardım kararı verilmesine ilişkin kurallarının bir bölümü aşağıdaki gibidir.



Şekil 3.1. Chaid Modeli

Karar ağacı incelendiğinde, yardım kararını etkileyen öncelikli değişkenin sosyal güvencesi olmayan kişi sayısı olduğu ve ağacın ilk dallanmasını bu değişken yardımı ile yaptığı görülebilir. Örneğin, yardım kararının alındığı dallar ile ilgilendiğimizi düşünelim. Bu durumda takip etmemiz gereken dal, yardım etme kararının yoğun olduğu, sosyal güvencesi olmayan kişi sayısının 1 ve daha fazla olduğu durumu belirten daldır. Hemen arkasından, bu düğümün dallanması için öncelikli değişken olan ikamet türü gelmiştir. Buna göre, ailenin bedelsiz oturuyor olması veya kiracı olması durumu, yardım edilme kararının, ev sahibi olarak ikamet eden ailelere oranla daha fazla alındığı durum olarak görünmektedir. Bu dallanmadan devam edildiğinde, ailede sigara kullanan birinin olmamasının da yardım edilmesinde etkili olduğu görülmekte; fakat sonrasında model yaptığı dallanmalarda karar değişkeninin tahmin değerini değişkenlere bakmaksızın yardım edilebilir olarak atıyor gibi görünmektedir.

Buna göre yardım edilen ailelerin bir kısmını yakalayan bir kurallar bütünü oluşturmak istersek bu; sosyal güvencesi olmayan kişi sayısı >=1 ve ikamet türü>0 ve ikamet türü<=5 ve sigara kullanan kişi sayısı =0 olanlar şeklinde olabilir.

Modelin doğruluğu, sezgisel olarak, değişkenlerin, yaptıkları dallanmalarla istenilen kitleyi ne oranda yakalamış olduğuna göre ölçülebildiği gibi, ağaçtaki kuralların gerçek hayata ne derece uygun olduğu ile de ölçülebilir.

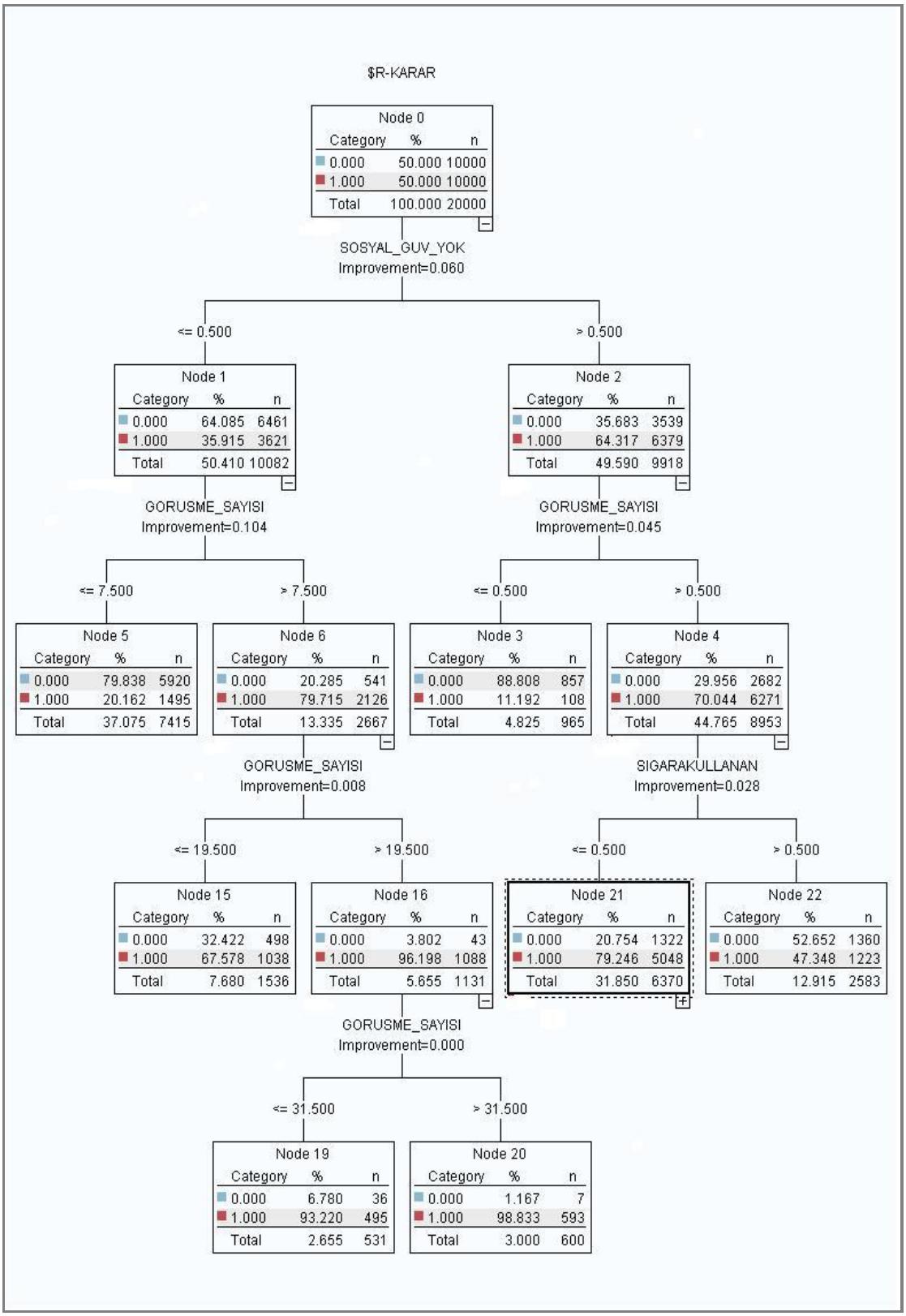
Modelin doğruluk oranları aşağıdaki gibidir. Oranlara bakıldığında, modelin tüm veriyi % 73 oranında doğru tahmin ettiğini, kararın tahmininde ise, bütün veriyi, yardım kararı verilecek şeklinde tahmin etme eğiliminde olduğunu açıkça görebiliriz. Karar yardım edilmeme şeklinde iken, kararın tahmin değerinin yardım edilmesi yönünde işaretlendiği 8.944 adet veri bulunmaktadır. Bu da %89 oranına tekabül etmektedir. Bununla birlikte, ağacın dallanması çok da anlamlı görünmemektedir. Zira sosyal güvencesi olmayan ve ikamet türü>0 ve ikamet türü<=5 görünen aileler için, ikamet türünün bütün kategorilerinde yardım edilmesi kararı daha fazla görünmekte, alt dallarda da bu durum devam etmektedir. Bu durumda modelin sosyal güvence ve ikamet türü dışındaki değişkenler ile fazla ilgilenmediği belirtilebilir.

SINIFLANDIRMA TABLOSU - EĞİTİM KÜMESİ				
		Tahmin edilen		
		KARAR		Doğruluk Oranı
Gözlenen		0	1	
KARAR	0	5.792	4.208	57,9
	1	1.056	8.944	89,4
Toplam Oran				73,6
SINIFLANDIRMA TABLOSU - TEST KÜMESİ				
		Tahmin edilen		
		KARAR		Doğruluk Oranı
Gözlenen		0	1	
KARAR	0	5.324	4.676	53,2
	1	1.546	8.454	84,5
Toplam Oran				68,8
SINIFLANDIRMA TABLOSU -DOĞRULAMA KÜMESİ				
		Tahmin edilen		
		KARAR		Doğruluk Oranı
Gözlenen		0	1	
KARAR	0	4.804	4.315	52,6
	1	1.318	8.682	86,8
Toplam Oran				70,5

Tablo 3.3 Chaid Modelinin Doğruluk Oranı

3.2.3. C&R Tree Modeli

C&R Tree modeli de aynı değişkenler için çalıştırılmıştır. Aşağıda, karar ağacı görülmektedir. Karar ağacına göre, birincil derecede değişken, sosyal güvencesi olmayan kişi sayısıdır. İkili dallanmalar yapan C&R Tree, en fazla yardım edilme kararını, sosyal güvencesi olmayan en az 1 bireye sahip olan, herhangi bir şekilde görüşme yapılmış ve sigara kullanan birinin olmadığı aileler için almış görünmektedir. Ağaç sol taraftan takip edildiğinde, aynı değişkenin farklı kategorileri için dallanmalar yapmış olduğu, modelleme yaparken ağacın budanması seçeneği kullanıldığı halde bu durumun gözlendiği göz önünde bulundurulursa, modelin gerçek hayata uygunluğunda da şüpheli olduğu düşünülebilir. Bununla birlikte, modelin eğitim, test ve doğrulama kümesi üzerindeki performanslarına bakıldığında bu modelin de tüm aileler için kararı, yardım edilmesi yönünde tahmin etme eğiliminin olduğu görünmektedir.



Şekil 3.2. C&R Tree

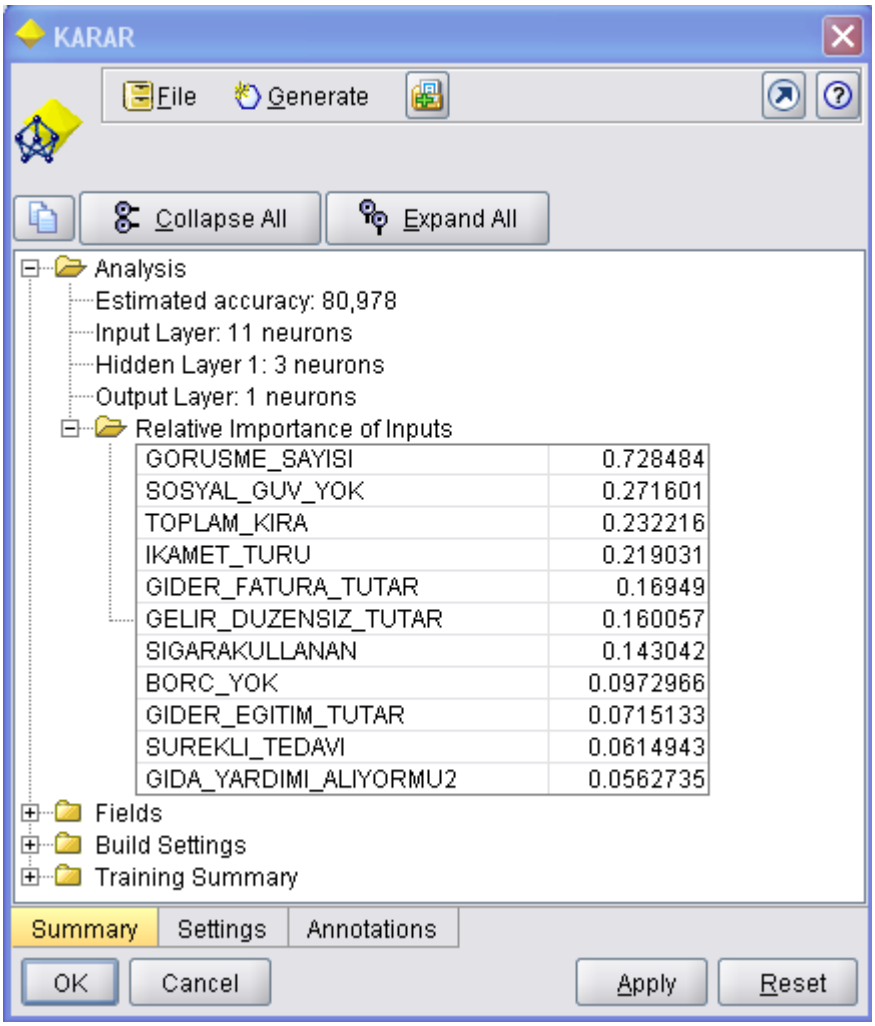
SINIFLANDIRMA TABLOSU - EĞİTİM KÜMESİ				
		Tahmin edilen		
		KARAR		Doğruluk Oranı
Gözlenen		0	1	
KARAR	0	6.301	3.699	63
	1	1.036	8.964	89,6
Toplam Oran				76,3
SINIFLANDIRMA TABLOSU - TEST KÜMESİ				
		Tahmin edilen		
		KARAR		Doğruluk Oranı
Gözlenen		0	1	
KARAR	0	6.743	3.257	67,4
	1	1.646	8.354	83,5
Toplam Oran				75,4
SINIFLANDIRMA TABLOSU -DOĞRULAMA KÜMESİ				
		Tahmin edilen		
		KARAR		Doğruluk Oranı
Gözlenen		0	1	
KARAR	0	6.164	2.955	67,5
	1	1.408	8.592	85,9
Toplam Oran				77,1

Tablo 3.4. C&R Tree Doğruluk Oranları

3.2.4. Yapay Sinir Ağı Modeli

Yapay sinir ağı modeli aynı örnek için oluşturulurken, clementine içerisindeki prune metodu kullanılmış, bir adet gizli katman ayarı yapılmıştır. Gizli katman için sinir sayısı elle ayarlanmış, 20 adet sinir belirlenmiştir. Model seçimi için, denemeler sonucu oluşturulan en iyi modelin çıktısı alınmış, bu modelle ilerlenmiştir.

Modeldeki değişkenlerin göreli önemi aşağıdaki gibidir. Buna göre sinir ağı modeli en yüksek ağırlığı görüşme sayısına vermiştir diyebiliriz.



Tablo 3.5. Yapay Sinir Ağı Çıktısı

Yapay sinir ağı modelinin doğruluk oranları aşağıda görünmektedir. Buna göre doğruluk oranlarının oldukça yüksek olduğu söylenebilir.

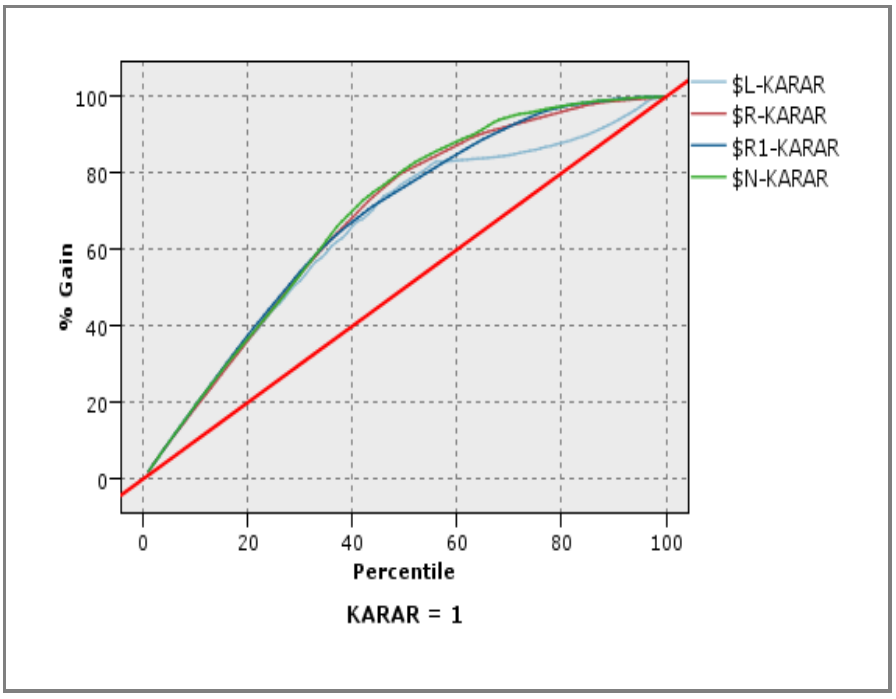
SINIFLANDIRMA TABLOSU - EĞİTİM KÜMESİ				
		Tahmin edilen		
		KARAR		Doğruluk Oranı
Gözlenen		0	1	
KARAR	0	8.176	1.824	81,7
	1	1.971	8.029	80,2
Toplam Oran				81
SINIFLANDIRMA TABLOSU - TEST KÜMESİ				
		Tahmin edilen		
Gözlenen		KARAR		Doğruluk

		0	1	Oranı
KARAR	0	8.600	1.400	86
	1	2.832	7.168	71,6
Toplam Oran				78,8
SINIFLANDIRMA TABLOSU -DOĞRULAMA KÜMESİ				
Gözlenen		Tahmin edilen		
		KARAR		Doğruluk Oranı
		0	1	
KARAR	0	7.924	1.195	86,8
	1	2.535	7.465	74,6
Toplam Oran				80,4

Tablo 3.6. Yapay Sinir Ağı Doğruluk Oranları

3.3. Model Karşılaştırması ve Seçimi

Modellerin doğruluk oranları yukarıda görünmektedir. Bütün modeller başlangıç anında aynı değişkenlerden oluşturulmuş, model eğitilirken farklı değişkenler önem kazanarak farklı kırılımlar vermişlerdir. Tüm modellerin karşılaştırmalı grafiği aşağıdaki gibidir.



Şekil 3.3. Modeller için Değerlendirme Grafiği

Grafikte n-karar sinir ağını, l-karar lojistik modeli, r-karar C&R Tree'yi ve r1-karar chaid modelini ifade etmektedir. Dikey eksen doğru tanımlanmış verilerin oranını, yatay eksen ise yanlış tanımlanmış verilerin oranını içermektedir. Bu durumda modelin kazanımının, köşegen olan kırmızı çizgiden uzak olması modelin görece uygun bir model olduğunu ifade edebilir. Yukarıda da incelendiği gibi, tüm modellerin kazanımı birbirine yakın görünmektedir.

Karar ağaçları anlamlı bölünmeler yapmaktan uzak ve bütün ailelere yardım edilsin anlamına gelen karar değerinin tahmin değerini 1 olarak atamaya meyilli gibi görünmektedirler. Sinir ağı ve lojistik regresyon modellerinin doğruluk oranları da yakın gibi görünmektedir. Son olarak iki modelin doğrulukları arasında anlamlı bir fark olup olmadığı test edilmiştir. Bunun için, bütün veriden ayrılan test verisinden 10 farklı ve rastgele küme oluşturulmuş, modeller tek tek denenerek hata oranları alınmıştır. Her iki model için oluşturulan 10 farklı hata oranına t testi uygulanmıştır. Sonuç aşağıdaki gibidir.

Paired Samples Test

		Paired Differences		t	Sig. (2-tailed)
		Mean	Std. Deviation		
Pair 1	LOJISTIK - SINIRAGI	,02000	,04214	1,501	,168

Tablo 3.7. Lojistik Regresyon ve Sinir Ağı'nın karşılaştırılması

H_0 = İki modelin doğruluğu arasında anlamlı bir fark bulunmamaktadır.
H_1 = İki modelin doğruluğu arasında anlamlı bir fark vardır.

Çıkan sonuca göre, hipotez red edilememiştir. Bu durumda, iki modelin doğruluğu arasında anlamlı bir fark olmadığı düşünülebilir. Bununla birlikte kullanılacak modelin, yukarıdaki sonuçlar yardımı ile karar verici trafından seçilmesi daha uygun görünmektedir.

SONUÇ

Veri madenciliği, veri tabanlarında depolanan verilerden, net olmayan ve ilk anda fark edilemeyen; ancak potansiyel olarak kullanışlı olabilecek bilgi ve örüntülerin çıkarılmasıdır. Bu da; regresyon, kümeleme gibi istatistik yöntemleri içerdiği gibi, çeşitli karmaşık bilgisayar algoritmalarını barındıran makine öğrenme tekniklerini de içermektedir.

Tez çalışmasında, veri madenciliği yöntemleri özetlenerek sınıflandırma yöntemleri üzerinde durulmuş ve eldeki veriler kullanılarak sınıflayıcı bir model geliştirilmiştir. Çalışmada Deniz Feneri Derneği'nden alınan 59.119 ailenin çeşitli bilgileri kullanılmıştır. Bu bilgiler çeşitli değişkenlerde değerlendirilmiş, aralarındaki ilişkilere bakılarak modellenmesi uygun görülen ve gerekli olan değişkenler saptanmıştır. Sonuçta bu değişkenler ile ailelere yardım kararı verilmesine ilişkin modeller geliştirilmiştir. Bunun için lojistik regresyon, yapay sinir ağları ve karar ağaçları denenmiş ve sonuçları incelenmiştir. Karar ağaçlarının doğruluk oranları sinir ağı ve lojistik regresyona göre daha düşük olduğu görülmüştür. Lojistik regresyon ile sinir ağının doğruluk oranları ise birbirine yakın görünmekte olup, yapılan test sonucu aralarındaki fark da istatistiki olarak anlamlı çıkmamıştır.

Çalışmada ulaşılan modeller yardım kararı verilirken ailede dikkat edilen özellikleri yansıtıyor gibi görünmektedir. Modellerin geliştirilmesinin esas amacı, sonrasında yardım için başvuran ailelerin aynı özelliklerine bakılarak yardım edilip edilemeyeceğine karar verilebilmesine yardımcı olabilecek bir yapı hazırlamaktır.

Sonuç olarak, geliştirilen modeller yüksek oranda doğru tahmin etme yeteneğine sahip gibi görünmektedirler. Bununla birlikte, modellerin pratikte uygulanabilirliğinin yüksek olabilmesi için, geliştirilen algoritmaların veri tabanına gömülerek otomatize edilmesi veya modellerin içerdiği değişkenlere ait bilgilerin girilebileceği ve otomatik karar ataması yapan bir ara yüz oluşturulması daha uygun olmaktadır.

KAYNAKÇA

Akkaya Ş., Pazarlıoğlu V. : **Ekonometri 1**, 4.baskı, İzmir, Anadolu Matbaacılık, 2000.

Akkaya Ş., Pazarlıoğlu V. : **Ekonometri 2**, 2.baskı, İzmir, Erkam Matbaacılık, 1998.

Akpınar H. : "Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği", **İ.Ü. İşletme Fakültesi Dergisi**, Sayı:1,Çevrimiçi http://www.isletme.istanbul.edu.tr/surekli_yayinlar/dergiler/nisan2000/1.htm,10.Ocak.2009, s. 1-22

Berry M.J.A., Linoff G.S. : **Data Mining Techniques for Marketing, Sales and Customer Relationship Management,** Second Edition, Wiley, April 2004.

Bigus J.P. : **Data Mining with Neural Networks,** USA, McGraw-Hill, 1996

Bircan H., Karagöz Y. : "Lojistik Regresyon Analizi: Tıp Verileri Üzerinde bir Uygulama", **Kocaeli Üniversitesi Sosyal Bilimler Enstitüsü Dergisi,** 2004 Çevrimiçi http://iibf.erciyes.edu.tr:90/petas, 10.Haziran.2009

Chambers L.D. : **Practical Handbook of Genetic Algorithms Comlex Coding Systems,** 1.st Edition, CRC, 1998.

Cox L.A. : "Data Mining and Causal Modelling of Customer Behaviours" **, Telecommunication Systems,** December 2002, Volume 21, pp. 349-381

Doğan İ. : "Kümeleme Analizi ile Seleksiyon", **Turkish Journal of Veterinary and Animal Sciences,** 2000, Çevrimiçi, http://journals.tubitak.gov.tr/veterinary/issues/vet-02-26-1/vet-26-1-7-0007-1.pdf, 13.Ocak.2009

Emel G.G.,Taşkın Ç.: "Genetik Algoritmalar ve Uygulama Alanları", **Uludağ Üniversitesi İktisadi ve İdari Bilimler**

Fakültesi Dergisi, 2002, Çevrimiçi, http://www.yapay-zeka.org/files/tez/genetik_algoritmalar_ve_uygulama_alanlari.pdf, 10.Mayıs.2009, s. 129-152

Emel G.G.,Taşkın Ç.,Tok A.: ” Pazarlama Stratejilerinin Oluşturulmasında bir Karar Destek Ağacı, Birliktelik Kuralı Madenciliği”, **Dokuz Eylül Üniversitesi Sosyal Bilimler Enstitüsü Dergisi**, 2005, Çevrimiçi, http://www.sbe.deu.edu.tr/adergi/2006/Cilt%207%20Sayi%203%202005/emel-taskin-tok.pdf, 18.Mayıs.2009

Guidici P. : **Applied Data Mining Statistical Methods for Business and Industry**, West Sussex, Wiley, 2003

Gujarati D.N. : **Temel Ekonometri**, 2.Baskı, İstanbul, Literatür Yayınları, Matris Matbaacılık, Mayıs 2001.

Gülten A, Doğan Ş. : “Genetik Algoritmalar Yönteminin Biyomedikal Verileri Üzerinde Uygulamaları”, **Doğu Anadolu Bölgesi Araştırmaları Dergisi** , Ekim-2008, Çevrimiçi, http://web.firat.edu.tr/daum/docs/71/03, 18.Mayıs.2009

Han J., Kamber M. : **Data Mining Concepts & Techniques**, San Francisco, USA, Morgan Kauffmann Publishers, 2006

Kantardzic M.: **Data Mining Concepts Models Methods and Algorithms**, Piscataway, NJ, Wiley-Interscience, 2003.

Kaya H., Köymen K. : “Veri Madenciliği Kavramı ve Uygulama alanları”, **Doğu Anadolu Bölgesi Araştırma ve Uygulama Dergisi,** Şubat 2008, Çevrimiçi, http://web.firat.edu.tr/daum/default.asp?id=79, 13.Ocak.2009

Kök V., Kuloğlu N. : “Sollama Esnasında Taşıt ve Yol ile İlgili Faktörlerin Karar Ağacı Yöntemi ile İrdelenmesi”, **Erciyes Üniversitesi Fen Bilimleri Enstitüsü Dergisi**, 2005, Çevrimiçi, http://perweb.firat.edu.tr/personel/yayinlar/fua_522/522_20056.pdf, 10.Mayıs.2009

Luan J.,Willet T. : "Data Mining & Knowledge Management: A System Analysis for Establishing a Tiered Knowledge Management Model", Çevrimiçi, http://www.eric.ed.gov/ERICWebPortal/custom/portlets/recordDetails/, 4.Mayıs.2009

Maimon O. Rokach L. : **Data Mining and Knowledge Discovery Handbook,** Ramat-Aviv ISRAEL, Springer, 2005.

Mattison R. : **Data Warehousing and Data Mining for Telecommunications**, Norwood, USA, Artech House, 1997

Myatt G. : **Making Sense of Data**, US, John Wiley & Sons Publication, 2007

Oğuzlar A., Tüzüntürk S. : "Borsada İşlem Gören Şirketlerin Finansal Göstergelerinin Analizi", Çevrimiçi, http://iletisim.atauni.edu.tr/eisemp/html/tammetinler/267.pdf, 18.Mayıs.2009

Oğuzlar A. : "Kümeleme Analizinde Yeni Bir Yaklaşım", **Atatürk Üniversitesi İ.İ.B.F Dergisi**, 2005, Çevrimiçi, http://194.27.49.253/iibf/CV07.pdf, 20.Mayıs.2009

Ohsuga L., Hu L. : **Foundations and Novel Approaches in Data Mining,** Warsaw, Poland, Springer, 2005.

Rud O.P. : **Data Mining Cookbook Modeling Data for Marketing, Risk and Customer Relationship Management**, New York, USA, John Wiley, 2001

Satman M.H. : "Ekonometrik Yöntem ve Sorunlara Genetik Algoritma Yaklaşımları ve İktisadi Uygulamalar, **İstanbul Üniversitesi Sosyal Bilimler Enstitüsü**, 2008

Sumathi S., Sivanandam S.N. : **Introduction to Data Mining and its Applications,** New York, USA, Springer, 2006

Taniar D.: **Research and Trends in Data Mining Technologies and Applications**, London, UK, Idea Group Publishing, 2007

Tatlıdil H.: **Uygulamalı Çok Değişkenli İstatistiksel Analiz,** Ankara, Ziraat Matbaacılık, Eylül 2002

Triantaphyllou E.,Felici G.: **Data Mining and Knowledge Discovery Approaches Based on Rule Induction Techniques,** New York, USA, Springer, 2006

Venkata K.,Murthy S. : "On Growing Better Decision Trees from Data", Çevrimiçi, http://www.cbcb.umd.edu/~salzberg/docs/murthy_thesis/thesis.html, 13.Ocak.2009

Yaralıoğlu K. : **Uygulamada Karar Destek Yöntemleri**, 1. Baskı, İzmir, İlkem Ofset, Mayıs 2004

http://en.wikipedia.org/wiki/Support_vector_machine

http://www.statsoft.com/TEXTBOOK/stchaid.html#index