ANKARA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

YÜKSEK LİSANS TEZİ

CHAID VE CART ANALİZ YÖNTEMLERİNİN UYGULAMALI KARŞILAŞTIRILMASI

Yasin İNAĞ

BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

ANKARA
2015

TEZ ONAY SAYFASI

Yasin İNAĞ tarafından hazırlanan "**CHAID ve CART Analiz Yöntemlerinin Uygulamalı Karşılaştırılması**" adlı tez çalışması 14/01/2015 tarihinde aşağıdaki jüri tarafından oy birliği ile Ankara Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı'nda **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Danışman : Yrd. Doç. Dr. Recep ERYİĞİT

Jüri Üyeleri:

Başkan: Doç. Dr. Süleyman TOSUN
Hacettepe Üniversitesi / Bilgisayar Mühendisliği Anabilim Dalı

Üye : Yrd. Doç. Dr. Semra GÜNDÜÇ
Ankara Üniversitesi /Bilgisayar Mühendisliği Anabilim Dalı

Üye : Yrd. Doç. Dr. Recep ERYİĞİT
Ankara Üniversite / Bilgisayar Mühendisliği Anabilim Dalı

Yukarıdaki sonucu onaylarım.

Prof. Dr. İbrahim DEMİR
Enstitü Müdürü

ETİK

Ankara Üniversitesi Fen Bilimleri Enstitüsü tez yazım kurallarına uygun olarak hazırladığım bu tez içindeki bütün bilgilerin doğru ve tam olduğunu, bilgilerin üretilmesi aşamasında bilimsel etiğe uygun davrandığımı, yararlandığım bütün kaynakları atıf yaparak belirttiğimi beyan ederim.

14/01/2015

Yasin İNAĞ

ÖZET

Yüksek Lisans Tezi

CHAID VE CART ANALİZ YÖNTEMLERİNİN UYGULAMALI KARŞILAŞTIRILMASI

Yasin İNAĞ

Ankara Üniversitesi
Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Danışman: Yrd. Doç. Dr. Recep ERYİĞİT

Veri madenciliği, bilgi keşfedilmesini sağlayan bir tekniktir. Anlamsız büyük verilerden anlamlı veriler elde etmek için, birçok alanda yaygın olarak kullanılmaktadır. Veri madenciliğinde farklı modeller kullanılmaktadır. Sınıflama ve Regresyon ağaçları (Classification and Regression Tress, CART) parametrik olmayan bir metottur. Herhangi bir objenin sınıf üyeliğini bir veya daha fazla bağımsız değişken kullanarak tahmin etmeye yarayan ağaç modelidir. Otomatik ki-kare etkileşim tekniği (Chi-squared Automatic Interaction Detector, Chaid) sınıflandırılmış bağımlı değişkenlerin tanımlanması ve analizi için kullanılır.

Akademik erteleme eğilimi, bireylerin yerine getirmesi gereken sorumlulukları belirlenen süre içerisinde yapmama veya geciktirmesidir. Üniversite öğrencilerinin akademik erteleme eğilimini etkileyen nedenler Cart ve Chaid teknikleri analiz edilip, kullanılan yöntemler karşılaştırılmıştır. Kullanılan yöntemlerle elde edilen ağaç yapıları belirlenip farklılıklar nedenleri ile sunulmuştur.

Ocak 2015, 69 sayfa

Anahtar Kelimeler: Sınıflama ve Regresyon Ağacı (CART), Otomatik ki-kare etkileşimi (CHAID), Akademik erteleme eğilimi, Gini, Twoing

ABSTRACT

Master Thesis

CHAID AND CART AS A PRACTICAL COMPARISON OF ANALYSIS METHODS

Yasin İNAĞ

Ankara University
Graduate School of Natural and Applied Sciences
Department of Computer Engineering

Supervisor: Asst. Prof. Dr. Recep ERYİĞİT

Data mining engineering is a technique that provides to explore information. It is commonly used in several areas in order to derive specific and meaningful data from large and meaningless data. Different models are used in data mining engineering. Classification and Regression Trees (CART) is a nonparametric method. It is a tree model which estimates class membership of any object by using of one or more independent variable. Chi-squared Automatic Interaction Detector (CHAID) is used for the aim of identifying and analysing classified dependent variables.

Academic postponement tendency is individual's attitudes in order to not fulfil the responsibilities or postpone the responsibilities within the time specified. The reasons are affecting undergraduates academic postponement tendencies have been analysed by using of Cart and Chaid techniques and the methods were used have been compared. Derived tree structures by used methods have been determined and the differences provided with the reasons.

January 2015, 69 pages

Key Words: Classification and Regression Trees (CART), Chi-squared Automatic Interaction Detector (CHAID), Academic Postponement Tendency, Gini, Twoing

TEŞEKKÜR

Bu çalışma süresince tüm bilgilerini benimle paylaşmaktan kaçınmayan, her türlü konuda desteğini benden esirgemeyen ve tezimde büyük emeği olan, aynı zamanda kişilik olarak ta bana çok şey katan danışman hocam, Sayın Yrd. Doç. Dr. Recep ERYİĞİT´e (Ankara Üniversitesi Bilgisayar Mühendisliği Anabilim Dalı Öğretim Üyesi), araştırmalarımın her aşamasında bilgi, öneri ve yardımlarını esirgemeyerek akademik ortamda olduğu kadar beşeri ilişkilerde de engin fikirleriyle yetişme ve gelişmeme katkıda bulunan Doç. Dr. Murat KAYRİ'ye (Muş Alparslan Üniversitesi Bilgisayar Yazılımı Anabilim Dalı Öğretim Üyesi), Çalışmalarıma katkılarından dolayı Araş. Gör. Görkem CEYHAN ve Araş. Gör. Fuat ELKONCA'ya, çalışmam boyunca maddi manevi desteklerini esirgemeyen çok kıymetli aileme sonsuz minnet ve teşekkürlerimi sunarım.

Yasin İNAĞ
Ankara, Ocak 2015

İÇİNDEKİLER

ŞEKİLLER DİZİNİ

ÇİZELGELER DİZİNİ

1. GİRİŞ

Dünya üzerinde teknolojinin gelişmesi ile birlikte, gelişen pazar piyasası müşteri ve veri odaklı çalışmalar veri kirliliği sebebiyle hızla zorlaşmaktadır. Verilerin toplanması, güvenli depolama ve analiz gibi sorunlar ortaya çıkmıştır. Her geçen gün olağanüstü bir hızla artmaya devam eden verilerin çok küçük bir kısmı kullanılabilmektedir. Verilerin boyutlarının çok büyük oluşu, herhangi bir araç kullanmaksızın verilerin etkin bir biçimde analiz edilmesini ve karar destek aşamasında kullanılmasını imkansız hale getirmiştir (Westpal ve Blaxton 1998).

Toplanan verilerin iyi kullanılmaması veri tabanlarında insanlar için sorun haline gelmektedir. Toplanan veri miktarı artıkça verilerin karmaşıklığı da artmaktadır. Bu sebeple verileri çözümleme tekniklerine olan ihtiyaç artmaktadır. Çözümleme teknikleri veri madenciliği ve veri madenciliği bilgi keşfi kavramlarıyla tanımlanmıştır. Eğer veri tabanlarında bilgi keşfi süreci başarılı ise; keşfedilen bilgi, organizasyonların karar verme sürecini geliştirmek amacıyla kullanılabilir (Kaya vd. 2005).

Bir firmanın önceki dönemlere ilişkin satış bilgileri, "hangi müşteri grubunun hangi ürünü alabileceği" şeklinde bir bilgiyi içerebilir. Bu bilgi firmanın satışlarının artmasına sebep olabilir (Freitas 2002).

Veri madenciliği, veri tabanları bilgi keşfi süreci içerisinde, verilerin anlamlı bir bütün halini alması için değerlendirme ve uygun modelin belirlenmesi hususunda önemli bir kısmını oluşturmaktadır. Çeşitli veri tabanları ve kaynaklardan verinin toplanması yöntemi ile başlayan bilgi keşfi süreci, verilerin değerlendirilmesi için uygun hale getirilmesi ile devam etmektedir. Ancak veri ambarına sahip olan kuruluşlarda, gerekli verilerin veri deposu olarak adlandırılan daha küçük veri ambarlarına aktarılması ile doğrudan veri madenciliği işlemlerine başlanabilmesi de mümkündür (Akpınar 2000).

Akademik erteleme eğilimi, öğrencilerin üniversite yaşamları boyunca farklı sebeplerden dolayı yerine getirmesi gereken sorumlulukları, belirlenen zaman içinde yapamama veya sürenin sonlarına doğru yapması olarak tanımlanmaktadır.

Veri madenciliğinde, analiz için veri tipine göre farklı algoritmalar geliştirilmiştir. Kategorik sürekli verilerde Cart ve Chaid analizleri kullanılabilir. Tezin çalışma planı şu şekildedir. Tezin ikinci bölümünde veri madenciliği konusu açıklanmış büyük verilerden anlamlı veri elde etmek için yapılması gereken adımlar, veri madenciliğinin kullanım alanları açıklanmıştır. Tezin üçüncü bölümünde kullanılan yöntemler ve karar ağaçları açıklanmıştır. Dördüncü bölümde Cart ve Chaid algoritmalarının çalışma prensipleri, ayırma yöntemleri ve sonlandırma durumları ile doğruluk analizlerinin algoritması açıklanmıştır. Tezin beşinci bölümünde akademik erteleme eğilimi Cart ve Chaid algoritmaları ile incelenmiş ve bu iki algoritmanın performansları karşılaştırılmış ve sonuçlar yorumlanmıştır.

2. VERİ MADENCİLİĞİ VE BİLGİ KEŞFİ

2.1 Veri Madenciliği Kavramı

Veri madenciliği ile ilgili farklı tanımlar yapılmıştır. Bu tanımlardan bazılarına aşağıda yer verilmiştir.

Veriden anlamlı ilişkiler ve örüntüler çıkarma sürecine, "veri madenciliği", "bilgi çıkarımı", "bilgi keşfi", "veri arkeolojisi" ve "veri şablon işleme" gibi isimler verilmektedir. İlk olarak 1989 yılında bir atölye çalışmasında, veri işleme sürecinde bilginin son ürün olduğunu vurgulamak için "veri tabanlarında bilgi keşfi" tanımlaması kullanılmıştır (Piatetsky ve Shapiro 1989).

Veri madenciliği; büyük miktardaki verinin, bu veriden anlamlı örüntü ve kurallar çıkarılması amacıyla analiz edilmesidir (Berry ve Linoff 2000). Veri madenciliği kümeleme, tahmin, tanımlama ve görselleştirme, sınıflandırma ve birliktelik kuralları gibi tekniklerden yararlanmaktadır.

Veri tabanlarında zengin bilgiye sahip olan pek çok organizasyon, bu bilgiyi yönetmenin çok zor olması sebebiyle, analizi gerçekleştirecek bilgisayar ve benzeri teknolojik ürünler kullanmaktadır. Bu ürünler kullanılarak veriler içerisinden anlamlı bilgilerin çıkarılması, veri madenciliği olarak tanımlanmıştır (Adrians ve Zantinge 1997).

Büyük miktarda veri içinden, gerçekleşebilecek durumlarla ilgili tahmin yapmamızı sağlayacak birlikteliklerin ve modellerin uygun bilgisayar programları kullanılarak aranmasıdır. Veri analizi yapılarak, bir insanın yaşam şartlarına göre davranışları belirlenebilir, bir mal için bir sonraki ayın satış tahminleri yapılabilir, müşteriler satın aldıkları mallara bağlı olarak gruplandırılabilir, yeni bir ürün için potansiyel müşteriler belirlenebilir, müşterilerin zaman içindeki hareketleri incelenerek market sahiplerine ürün güncelleme konusu belirlenebilir. Binlerce verinin olabileceği düşünülürse bu

analizin gözle ve elle yapılamayacağı, bilgisayar ortamında yapılmasının gerektiği ortaya çıkar ve veri madenciliği bu noktada devreye girer (Alpaydın 2004).

Büyük miktarda ve oldukça hızlı toplanan verilerin, çeşitli çözümlemeler sonucunda anlamlı bilgilere dönüştürülmesi noktasında devreye giren bir süreçtir (Özmen 2004). Veri madenciliği tanımları incelendiğinde; dünyamızda çok fazla miktarda çoğunluğu çözümlenmemiş anlamsız verinin depolandığı ve bu verilerden anlamlı bilgiler elde edilmesidir.

Veri madenciliği, verilerinizdeki gizli desenleri öngörebilen teknikler kullanarak ortaya çıkarır. Verilerinizde bulunan bu gizli desenler karar verme süreçlerinde çok kritik rollere sahiptir, çünkü iş süreçleriniz ile ilgili birçok akışı ve bilgiyi ortaya çıkartır. Veri madenciliği ile müşterilerinizle olan etkileşimlerinizdeki karlılığınızı arttırabilir, hilekarlık durumlarını tespit edebilir ve risk yönetiminizi geliştirebilirsiniz. Veri madenciliği, ortaya çıkan bilgilerle doğru zamanda daha iyi kararlar verebilmektir (http://www.spss.com.tr/Veri.html 2014).

Veri madenciliği, "büyük miktarda veri içinde gömülü olan bilgilerin çıkarılması ve bu bilgilerin kuruluşa stratejik karar desteği sağlama amaçlı kullanılması" olarak tanımlanmıştır. Müşteri ile ilgili detaylı ve yüksek miktardaki verinin, onlara sunulan hizmetlerin geri dönüşünü sağlamak ve yapılan işin değerini arttırmak amacıyla bilgiye dönüştürülmesinin, ham verideki bilinmeyen ilişkilerin ortaya çıkarılmasının, güçlü analitik gereçleri bir arada sunan veri madenciliği çözümleri ile mümkün olduğu belirtilmektedir.

Veri yığınları içinden kuruluş yöneticileri için en gerekli olan verilerin seçilmesi, düzenlenmesi ve modellenmesi süreçlerini içermektedir. Veri madenciliği, karar vericilerin kullanabileceği yeni bilgiler oluşturabilmek için yapay zeka bir sonraki adımı otomatik olarak belirleyebilecek teknoloji içeren yöntemler kullanmaktadır (http://www.sas.com/technologies/analytics/datamining/index.html 2013).

2.2 Veri Madenciliği Süreci

Veri madenciliği çok fazla miktarda veri içerisindeki bilgi keşfidir. Veri madenciliği belirli bir süreç sonucu elde edilen anlamlı bilgilerdir, süreçteki adımlar birbirini takip ettiğinden dikkatle izlenmesi gerekmektedir çünkü bir aşamanın çıktısı diğer aşamanın girdisi olacaktır.

Veri madenciliği süreci düzgün işlemediği takdirde istenilen verilere ulaşmak çok daha zor olacaktır. Bu sebeple veri madenciliği için standart bir süreç söz konusudur.

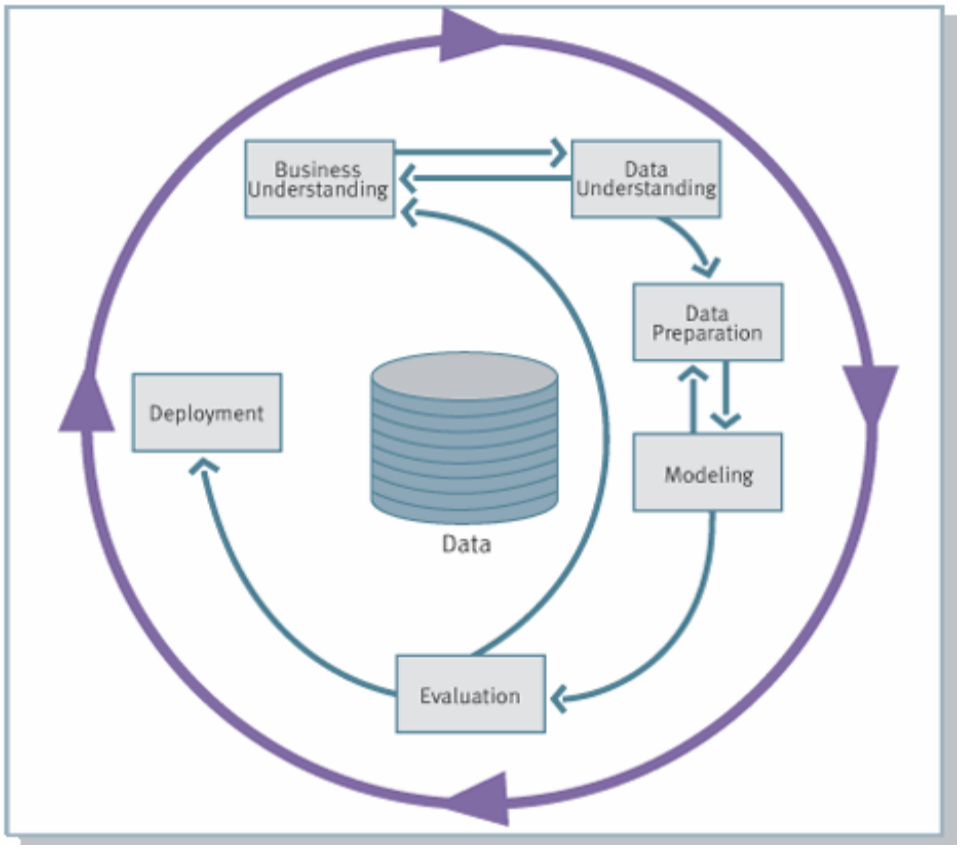
Bu standart süreç bir konsorsiyum tarafından belirlenmiştir. The Cross- Industry Standard Process for Data Mining (CRISP-DM) konsorsiyumu, 1996 yılının sonlarına doğru veri madenciliğinin yeni olduğu bir dönem de üç firma tarafından kurulmuştur.

Bu üç firmanın ilki olan Daimler Chrysler (önceden Daimler-Benz) birçok endüstriyel ve ticari organizasyona, veri madenciliği tekniklerini uygulama konusunda öncü olmuştur.

SPSS firması 1990 yılından beri veri madenciliği üzerine çeşitli hizmetler sağlamış ve ilk ticari veri madenciliği çalışma platformu olan Clementine'i 1994 yılında harekete geçirmiştir. NCR, müşterilerine değer katma işini sağlayabilmek ve alıcılarının ihtiyaçlarına hizmet edebilmek için birçok veri madenciliği danışmanlığı ve teknoloji uzmanlığı takımları kurmuştur.

Bu gelişmelerden bir yıl sonra, sözcüklerin baş harfleri "Cross-Industry Standard Process for Data Mining" açılımında olan CRISP-DM konsorsiyumu oluşturulmuş, Avrupa Komisyonundan fon elde edilmiş ve başlangıç fikirleri oluşturulmaya başlanmıştır.

CRISP-DM süreci 6 aşamadan oluşmaktadır. Şekil 2.1'de bu süreç görülmektedir (Helberg 2002).



Şekil 2.1 Veri madenciliği süreci

2.2.1 Problemin tanımlanması

Problemin tanımlanması bütün bilim dallarında çözümü noktasındaki en temel ve en önemli adımdır. Veri madenciliğinde de başarılı bir sonuç elde etmenin en temel şartıdır. Yapılacak olan araştırmanın ve istenilen veri analizinin elde edilmesinde, veri madenciliğinin ne amaçla yapılacağını belirlemektir. Elde edilecek olan sonuçların başarı yüzdelerinin nasıl ölçüleceği tanımlanmalıdır. Çizelge 2.1'de veri madenciliğinin analiz yapmada nasıl yardımcı olduğuna dair bazı sorulara yer verilmiştir (Giovinazzo 2002).

Veri madenciliğinde problemin tanımlanması aşamasında öncelik elde edilmek istenen veri seti analizidir. Bu sebeple elde edilmek istenen sonuca ulaşacak veri analizi problemin başlangıcı olacaktır, veri madenciliğini gerçekleştirecek kişi, veri madenciliği

uygulamasının sonucunu etkileyen faktörleri belirlemelidir. Analizin doğru sonuç vermesi için veri analiz projesinin dikkatle planlamalı ve özel, gerçekleştirilebilir, ölçülebilir bir sonucun olmasına bağlıdır (Giovinazzo 2002).

Çizelge 2.1 Veri madenciliğinin cevap bulabileceği bazı sorular

Müşteri Davranışı	Firmamızdan sıklıkla ürün alan müşterilerimizin özellikleri nelerdir? Karlı müşterilerimizin sıklıkla aldıkları ürünler nelerdir?
Müşterinin Demografik Özellikleri	Karlı müşterilerimizin cinsiyetleri nedir? Karlı müşterilerimizin yaşadıkları yerler neresidir? Karlı müşterilerimizin daha az alışveriş yaptığı bir dönem var mıdır?
Zamana Bağlılık	Karlı müşterilerimiz hangi sıklıkla alışveriş yapmaktadır?

Market sahibi bir kişinin müşteri memnuniyetini ve bağımlılığını artırmak için "çapraz satış tekniklerini kullanmayı" planlaması, müşteri kitlesinin dikkatini çekecek bir kampanya ile karı artırma, maliyeti düşürmek, yeni ürün geliştirme veya pazar payını artırmayı istek gibi hedefler belirlenebilir (Moss 2003).

2.2.2 Verinin hazırlanması

Veri madenciliği, eldeki birçok veriden anlamlı sonuçlar elde eden bir çalışma prensibi olduğundan doğru ve yeterli veri elde etmek çok önemlidir. Elde edilecek veriler ile kurulacak model de çıkacak sorunlarda yeniden veri hazırlama kısmına geri dönülecektir. Verinin hazırlanması; verinin toplanması, verinin birleştirilmesi ve temizlenmesi ile verinin dönüştürülmesi aşamalarından oluşmaktadır (Moss 2003).

2.2.2.1 Verinin toplanması

Doğru ve güvenilir veri toplama veri madenciliğinde belirlenecek ve kullanılacak veri ile doğru sonuç elde etmede çok önemlidir. Bu aşamada elde var olan veya toplanması gereken verilerin büyük bir titizlikle sonuca hizmet edecek bir şekilde güvenilir kaynaklardan elde edilmelidir (Giovinazzo 2002).

Veri elde etme aşamasında gösterilecek özen analiz yapacak olan kişinin doğru sonuçlar elde etmesinde en önemli faktördür. Bu aşamadaki doğru adımlar ilerleyen süreçte problemlerle karşılaşma olasılığını minimuma indirir.

2.2.2.2 Verilerin birleştirilmesi ve temizlenmesi

Veri madenciliğinde kullanılacak verilerin farklı kaynaklardan toplanması veri uyumsuzluğuna sebep olacaktır. Farklı zamanlarda elde edilmesi, farklı birimlerin kullanılması, farklı kodlamaların kullanılması, farklı ölçü birimlerinin kullanılması, farklı sosyal yapılardan edilmesi gibi birçok durum veri uyuşmazlığına sebep olabilir. Bu uyuşmazlıklardan kaynaklı elde edilen verilerin birleştirilmesi ve temizlenmesi gerekmektedir (Moss 2003).

Aşağıdaki çizelge 2.2'de bir veri tabanında kişilerin cinsiyetlerine göre kodlamaları görülmektedir. Analizi uygulayan kişi bu kodlamayı bir çok farklı şekilde gerçekleştirebilir.

Çizelge 2.2 Veri tabanlarında olası kodlama biçimleri

ERKEK	**KADIN**
0	1
1	0
1	2
"ERKEK"	"KADIN"
"erkek"	"kadın"
"E"	"K"

Tabloda da görüldüğü gibi elde edilmek istenen veri farklı kodlamalarla elde edilmiştir. Analizi yapacak kişi bu verileri birleştirmek zorundadır.

Ayrıca veride bulunan eksik veya kayıp bilgilerde gözden geçirilmelidir. Örneğin veri tabanındaki kişilerin yaşları biliniyorken bazı kişiler için yanlış bilgiler olabileceği gibi hiçbir veri de olmayabilir. Bu durumdaki eksiklikler "kayıp veriler" olarak tanımlanmalıdır. Aynı şekilde bazı kayıtlar aşırı uç değerler veya yanlış girilmiş değerler olabilir. Bu tür verilere gürültü denilmektedir (Jiawei ve Micheline 2001).

Doğru ve güvenilir bir analiz elde etmek için elde edilen verilerin eksik kısımları zamanla tamamlanmalıdır. Bir diğer yöntem ise verilerin tahmin yolu ile tamamlanmasıdır. Veri girme aşamasında verilerin yanlış girilmesi mümkündür, bu verilerin göz ardı edilmesi analiz sonucunu etkileyebilmektedir. Bu neden ile uç değerler gözden geçirilip veri tabanından temizlenmelidir. Tüm bu düzenlemeler gerçekleştirildikten sonra veriler aynı ortamda aynı kodlama sistemi ile toplanmalıdır (Özmen 2003).

Veri tabanında ki veriler aynı sonucu veriyorsa bu değişkenlerden biri fazladır. Örneğin bir veri listesinde kişilerin yaşları ve doğum yılları mevcutsa bunlardan biri fazladır. Bu verilerden biri veri tabanından kaldırılmalı veya birleştirilmelidir (Jiawei ve Micheline 2001).

2.2.2.3 Verinin dönüştürülmesi

Dönüştürme aşamasında, kullanılacak model ile bağlantılı olarak bazı veri gruplarının uygun şekilde tanımlanması gerekmektedir. Bu uygunluk mevcut değil ise uygun şekle dönüştürülmelidir. Örneğin 0/1 olarak tanımlanmış bir verinin, modele uygun olarak Erkek/Kadın biçimine dönüştürülmesi sonuç açısından daha faydalı olacaktır. Eğer bir problemde karar ağacı modeli kullanılacaksa gelir düzeyinin Düşük/Orta/Yüksek seklinde tanımlanması, yapay sinir ağları modeli kullanılacaksa kişinin borç bilgisi için Evet/Hayır şeklinde ifadelerin kullanılması daha anlamlı olacaktır (Akpınar 2000).

2.2.3 Modelin kurulması

Veri madenciliğinde istenilen sonucu elde etmekte doğru güvenilir veri toplamayla birlikte bu verileri anlamlı kılacak analizi yapmak için doğru modelin kullanılması gerekmektedir. İyi kurgulanmış veri modeli, analiz sonucunun kalitesini de etkileyecektir. İyi bir veri madenciliği uygulayıcısı, analiz sonucunda hangi veri örüntülerinin bulunabileceğini tahmin edebilir. Eğer model doğru kurulmazsa ise, veri setinde bulunabilecek veri ve veri ilişkilerinin doğru sonuç verilmesi mümkün olmaz. Önemli örüntüler tespit edilemez. Dolayısı ile modelden başarılı sonuç elde etme olasılığı düşük olur (Berry ve Linoff 1998).

2.2.4 Modelin kullanılması

Veri madenciliği süresince izlenilmesi gereken yöntemlerin sonuncusudur. Yapılacak olan analize ve istenilen sonuçlara ulaşmamızı sağlayacak adımların yapılmasına denir. Bu bir yöntem olabileceği gibi başka bir modelin alt parçası da olabilir. Veri madenciliğinde kurulan model canlıdır yani sürekli bir gelişim ve değişim içindedir. Bu sebeple modelin kullanılması da yaşayan bir süreçtir. Kullanılan model süreç içinde değişebilmektedir (Akpınar 2000).

2.3 Veri Madenciliğinde Karşılaşılabilecek Önemli Sorunlar

Veri madenciliği sistematik çalışılması gereken bir ver yönetim sistemidir. Veriyi elde tutma zorlukları olası karşılaşılabilecek sorunlar doğurmaktadır. Verinin büyük, dinamik olması veriyi saklama ve anlamlı kılma noktasında sorunlar ortaya çıkarabilmektedir. Karşılaşılabilecek sorunlar, saklanan verinin yeterliliği ve konuyla ilgili olup olmaması ile ilgilidir (Baykal 2003).

2.3.1 Veri tabanlarının boyutları

Veri madenciliği büyük veri tabanları kullanıldığından kaynaklı anlamlı sonuçlar çıkarma konusunda sorunlar oluşur. Veri madenciliği yöntemlerinden herhangi birinin

ham veri ile başarıya ulaşma olasılığı yoktur. Veri madenciliği yöntemleri, bu şekilde elde edilen sonuçların tüm veri tabanını temsil edeceği var sayılarak örnekleme yöntemi kullanılır.

Örnekleme yöntemleri veri tabanı küçültme işlemi olarak ta adlandırılır. Veri alanında örnekleme ile küçültme işlemi genellikle rasgele bazı veriler seçilerek oluşturulur. Özellik alanına göre veri tabanı küçültme işlemi ise bütün veri kaynaklarının belirli özellikleri seçilerek oluşturulur. Yine birçok özellik varsa bunlarda rasgele seçilir.

2.3.2 Dinamik veri yapısı

Veri tabanları dinamik yapılardır. Her zaman veri ekleme, çıkarma, güncelleme işlemlerine tabidir. Bir diğer değişle veri tabanları sürekli gelişim halinde olduğundan dolayı canlı olarak ta tanımlanabilir. Bu dinamik yapı doğru veriye ulaşma noktasında problemler ortaya çıkarmaktadır. Veri tabanı üzerinde yapılacak olan küçültme ve ya örnekleme çalışmalarında sürekli gelişir olması sorun oluşturabilmektedir.

2.3.3 Eksik veya kesin olmayan veri

Veri madenciliğinde bir veri tabanı, basit öğrenme veya anlamlı sonuçlar elde etmek için farklı yöntemler ile tasarlanmaktadır. Veri tabanlarında toplanan veriler eksik olabilir. Bu eksiklikler hatalara yol açmaktadır. Örneğin; hasta veri tabanında kan hücre bilgileri eksik olan hasta ile ilgili bazı kan hastalıkları teşhisleri konulamaz. Sorunu çözmek için bu tür veriler geliştirilir.

2.3.4 Gürültü

Nitelik veya sınıflandırma bilgilerinde oluşan hatalar “gürültü” olarak adlandırılır. Veri tabanlarındaki bazı kayıtlarda olması beklenen değerlerden çok uzak veriler, aşırı uç değerdeki veriler veya yanlış girilmiş veriler olabilmektedir. Bu tür verilere gürültü denilmektedir. Örneğin; yaş verisi sürekli olarak tanımlanmış ise veri tabanına rakamlardan farklı bir karakter girilmesi, ortalama insan ömrü yetmişlerdeyken çok

daha yüksek bir yaş girilmesi veri tabanlarında gürültü olarak tanımlanmaktadır. Herhangi bir veri tabanında gürültüsü olmayan veri elde etmek oldukça zordur. Fakat tanımlanmış kuralar ile toplam doğruluğa ulaşılabilir. Veri setinden gürültünün elenmesi istenir (Jiawe ve micheline 2001).

2.3.5 Eksik değerler

Veri tabanlarında veri toplama işlemi süresince veriler eksik doldurulabilir. Elde edilmek istenen veri bilinmiyor veya yanlışlıkla girilmemiş olabilir. Veri madenciliğinde her veri için özellik sayısı gerektiğinden, eksik değerler sorun yaratabilmektedir. Eksik değerler yüzünden karşılaşılacak sorunları çözmek için verilere tanımlanmış özeliklerin varsayılan bir değer ile tanımlanması gerekmektedir. Eksik olan veriler varsayılan olarak atanması ile problem çözüme kavuşturulur. Bir diğer çözüm yöntemi ise eksik veriye sahip olan kayıtların tamamı veri tabanından silinir (Baykal 2003).

2.4 Veri Madenciliği ve İlişkide Olduğu Disiplinler

Veri madenciliği, veri tabanı sistemleri, istatistik, yapay sinir ağları, makine öğrenimi, veri görselleştirme gibi konular ile ilişkilidir (Berry ve Linoff 2000).

2.4.1 Veri tabanı yönetim sistemleri

Veri tabanı yönetim sistemleri sanal ortamda çok büyük boyutlu verilerin düzenlenmesi ve yönetilmesi için kullanılan sistemdir. Verilerin tekrar tekrar kullanılmasını önlemek için düzenli bir yapıya çevirme amaçlı kullanılan bu sistem veri madenciliğine büyük katkı sağlamaktadır. Günümüzde birçok kuruluş veri tabanı yönetim sistemlerinden faydalanmaktadır. Düzenli ve sistematik veri kullanımı için bankalar, şirketler, genel müdürlükler, bakanlıklar özel veya kamuya ait kuruluşlar bu sistemden faydalanmaktadır (Yarımağan 2000).

2.4.2 İstatistik

İstatistik verinin analizi sonucu yapılan değerlendirmelerdir. Veri madenciliği de verinin anlamlı bir bilgi topluluğu olduğu düşünüldüğünde bu iki disiplin ilişkide olmakla birlikte aynı amaca hizmet eden iki ayrı kuram gibidir. Bu iki disiplini birbirinden ayıran en belirgin özellik, veri madenciliğinin daha geniş kapsamla, ve daha farklı disiplinlerle ilişkide olmasıdır. İstatistik ve veri madenciliğinde model kuramı önemlidir. Ayrıca veri madenciliği, istatistik çalışmalarında nümerik olmayan verilerle çalışma ve ana kütle ile çalışma sorunlarını ortadan kaldır (Hand 1999).

2.4.3 Makine öğrenimi

Yapay zeka olarak ta adlandırılan bu disiplin bilgisayarın veya makinelerin öğrenme yeteneği kazanması esasına dayanır. Makinelere girdi olarak verilen verilerin hangi çıktıyı vermesi gerektiği olaylar bazında işlenir. Öğrenme bilgi toplama, bilgi de veri madenciliği anlamına gelir. Makinalar öğretilmiş olaylara veya bilgilere benzer durumlara rastladığı zaman karar verir ve problem çözülür. Makine öğrenimi daha geniş olarak "Bilgisayarın bir olay ile ilgili bilgileri ve tecrübeleri öğrenerek, gelecekte oluşacak benzeri olaylar hakkında kararlar verebilmesi ve problemlere çözümler üretebilmesidir" şeklinde tanımlanabilir (Öztemel 2003).

2.4.4 Veri görselleştirme

Veri madenciliği büyük verileri anlamlı kılma amaçlı kullanılmaktadır. Anlamlı veri kolay anlaşılır ve kullanıma uygun olmalıdır. Görüntüleme teknikleri verideki anlaşılırlığı, teknik dağılımları, kümeleri ve sınır dışı öğeleri daha anlaşılır ve çekici kılmaktadır. Bunun için bilgisayar grafikleri, veri dağılım haritaları, eğriler, üç boyutlu şekiller ve yüzeyler kullanılır.

Veri madenciliğinde analizin daha anlaşılır olması için seçilecek olan görselleştirme metodu, kullanılacak olan veriye ve modele göre değişebilmektedir (Hand 1999).

2.4.5 Diğer disiplinler

Veri madenciliği yukarda açıklanan disiplinler dışında Veri ambarları, yapay sinir ağları ve genetik algoritmalar ile de ilişkilidir.

2.5 Veri Madenciliği Uygulama Alanları

Veri madenciliği verinin anlamlı bir bütünlüğü tanımlaması amaçlanmıştır. Günümüzde hemen hemen her sektörde veri ve bu verileri anlamlandırma çalışmaları yapılmaktadır. Bütün işletmeler sahip oldukları müşterilerin davranışlarını tahmin etmek istemektedir. Veri madenciliği bu amaçla kullanılacak bir yöntemdir. Bu nedenle veri madenciliği pazarlama, perakendecilik, bankacılık ve finans, ulaşım, tıp ve daha birçok alanda kullanılmaktadır.

2.5.1 Pazarlama yönetimi

Günümüzde serbest piyasa, müşteri odaklı şirketler, serbest rekabet ortamında, zaman ve müşteri verilerinin analizi karar destek amaçlı bilgi haline dönüşmüştür. Bu bilgiler pazarlama sektöründe müşteri profilinin isteğine ve şirketlerin gelecek ile ilgili profillerini belirlemesinde önemlidir. Doğru strateji ve bu stratejilerin hayata geçirilmesi, doğru zamanda ve bilgilere uygun yapılması işletmelerin varlığını sürdürebilmeleri açısından çok önemlidir.

Pazar analizlerinde kullanılacak verinin kaynağı; kredi kartı işlemleri, indirim kuponları, müşteri şikayet aramaları, yaşam stili çalışmaları, müşteri kayıtları, finans kayıtları olabilir. Bu kapsamda yapılan çalışmalar ilgi alanı, gelir düzeyi, harcama alışkanlıkları gibi özellikler açısından benzer niteliklerdeki müşteriler için bir model belirlenmesi ve hedef pazarın saptanması; müşterinin fiyat artışı ile değişen satın alma alışkanlıklarının belirlenmesi; çapraz satış analizleri ile ürün satışları arasındaki birlikteliklerin ve ilişkilerin belirlenmesi ve bu bilgilere dayanılarak ürün satış tahminleri yapılması; müşteri profili belirleme çalışmaları kapsamında hangi özelliklerdeki müşterilerin hangi ürünleri satın aldıklarının belirlenmesi, müşteri

ihtiyaçlarının belirlenmesi kapsamında farklı müşteri tipleri için en iyi ürünlerin neler olduğunun belirlenmesi ve yeni müşterileri çekmede hangi faktörlerin etkili olacağının tahmini şeklinde özetlenebilir (Berry ve Linoff 1998).

Veri madenciliği ile pazarlama sektöründe kazancın artırılması ile birlikte maliyetlerin azaltılması da mümkündür. Eldeki verilere göre müşteri alım alışkanlıkları tespit edilerek şirketin büyüme yönü veya alışkanlıklara uygun reyon planlaması ile müşteri memnuniyeti ve daha çok satış sağlanarak kazanç elde edilebilir. Telefonla veya posta yolu ile kampanyalara cevap verme olasılığı çok düşük olan müşteriler elenerek, bilgi toplama maliyetlerinde tasarruf sağlanabilir (Berry ve Linoff 2000).

2.5.2 Perakende sektörü

Perakende sektörü, satış bilgileri, müşteri profili, ürün nakil bilgileri, servis kayıtları gibi pek çok veriyi işlediğinden veri madenciliği tekniklerini kullanmaktadır. Ayrıca günümüzde çoğalan elektronik ticaret uygulamaları ile perakende sanayi veri madenciliği için zengin bir kaynak sunmaktadır.

Perakende sektöründe veri madenciliği sepet analizi ve satış tahminleri uygulamaları için kullanılmaktadır. Sepet analizi müşterinin bir ürün aldığında o ürünle birlikte en çok hangi ürünleri tercih ettiği belirlenerek mağaza raf, reyon planlaması yapılır ve promosyon stratejileri belirlenir (Moss 2003).

Perakende sektöründe satış tahminleri stok kontrolünde kullanılır. Stok analizi yapılarak gerekli ürün piyasaya sürülür. Eğer bir müşteri bugün alışveriş yaparsa bir sonraki alışveriş yapacağı zaman tespit etmek için kullanılır.

2.5.3 Sahteciliğin tespiti

Şirketler sahtekarlığa yönelik eğilimleri veri madenciliği teknikleri ile izleyebilir; geçmişe ait verileri kullanarak bunlara ait bir model kurabilir. Geliştirilen bu model

kullanılarak sahteciliğe meyilli olanlar tespit edilebilir ve bu eğilimlere benzer davranan müşteriler takibe alınarak sahtekarlık oranı düşürülebilir.

Sahtecilikleri belirlemenin önemli olduğu alanlar sigortacılık sektörü, finans sektöründe kredi kartı servisleri, perakendecilik sektörü ve telekomünikasyon sektörüdür. Sahtecilik şu alanlarda görülebilir (Dinkla 2004).

- Kredi kartı dolandırıcılığı
- İnternet işlemleri, e- nakit dolandırıcılığı (e-nakit internet ortamında müşteri ve üye işyeri arasındaki alışverişlerde ödeme kolaylığı sağlayan sanal bir ön ödeme sistemidir. Kullanıcılara özel bilgi girme zorunluluğu getirmeden güvenli bir şekilde alışveriş imkanı sağlamaktadır.)
- Sigorta dolandırıcılığı
- Kara para aklama
- Bilgisayar ve bilgisayar ağlarına izinsiz girme
- İletişim sektöründeki dolandırıcılıklar
- Abonelik dolandırıcılığı

2.5.4 Bankacılık ve finans sektörü

Veri madenciliği finans sektöründe işletme riskinin azaltılması, risk derecelendirme tahmini, maliyetlerin düşürülmesi, karlılık analizleri, ATM'lere gün içinde dağıtılacak para miktarının tespiti, doğru ve etkin kredi kararı verebilme, müşteri ve çalışan memnuniyetinin arttırılması için kullanılmaktadır (Şimşek 2006).

Bankacılık alanında “Churn Analizi” ile bir uygulama gerçekleştirmişlerdir. Bir bankanın bugünkü müşterilerinin gelecek dönemde, rakip firmalara geçip geçmeyeceği tahmin edilmeye çalışılmıştır. Churn analizi haricinde veri madenciliği, bankaların kredi başvurularını doğru bir şekilde değerlendirmelerine olanak vermektedir. Bireysel kredi başvurularında yaş, cinsiyet, meslek, iş tanımı, gelir ve giderler baz alınarak kişinin borcunu ödeyemez duruma düşmesi olasılığının tespiti yapılır (Chye ve gery 2002).

Veri madenciliği teknikleri farklı müşteri gruplarına uygun finansal ürünlerin veya poliçe türünün tespitinde, yüksek getirili yatırımlarla ilgilenebilecek müşterilerin profillerinin analiz edilerek etkin çapraz satış fırsatı yaratmak ve yatırım süreleri dolanların sistemde kalmasını sağlayacak pazarlama planları hazırlamak ve pazarlama maliyetlerini düşürmek amacıyla da kullanılır (Şimşek 2006).

2.5.5 Sağlık

Veri madenciliği yöntemleri kullanılarak sağlık alanında erken teşhis koyma, olası belirti ve semptomlardan gerekli incelemeleri yapmaya teşvik etmede kullanılmaktadır.

Kronik hastalıkların yaşanma sıklığı, kişi yaş ve önceki hastalıkları gibi verilerle tespit edilerek erken önlem alınabilir. Laboratuvar ortamında hatalı sonuç veya suiistimal olaylarının tespitinde, klinik karar destek ünitelerinin geliştirilmesinde, hasta odaklı sağlık hizmetleri sunumu ile kalitenin geliştirilmesinde kullanılmaktadır (Koyuncugil ve Özgülbaş 2009).

Hastanelerin, belirli zaman dilimlerinde oluşabilecek hasta yoğunluklarının tespitinde kullanılmaktadır. Hastaneyi tercih edenlerin yaş ortalamaları, genel sağlık durumlarının tespiti ile mevsimsel gerçekleşen hastalıkların oluşma yoğunluğuna göre hastane işletme yöntemleri belirlenmektedir. Veri madenciliğinin sağladığı bu olanak ile hastaların mağdur olmaması için gerekli önlemler alınabilmektedir (Irmak ve diğerleri 2012).

2.5.6 Diğer alanlar

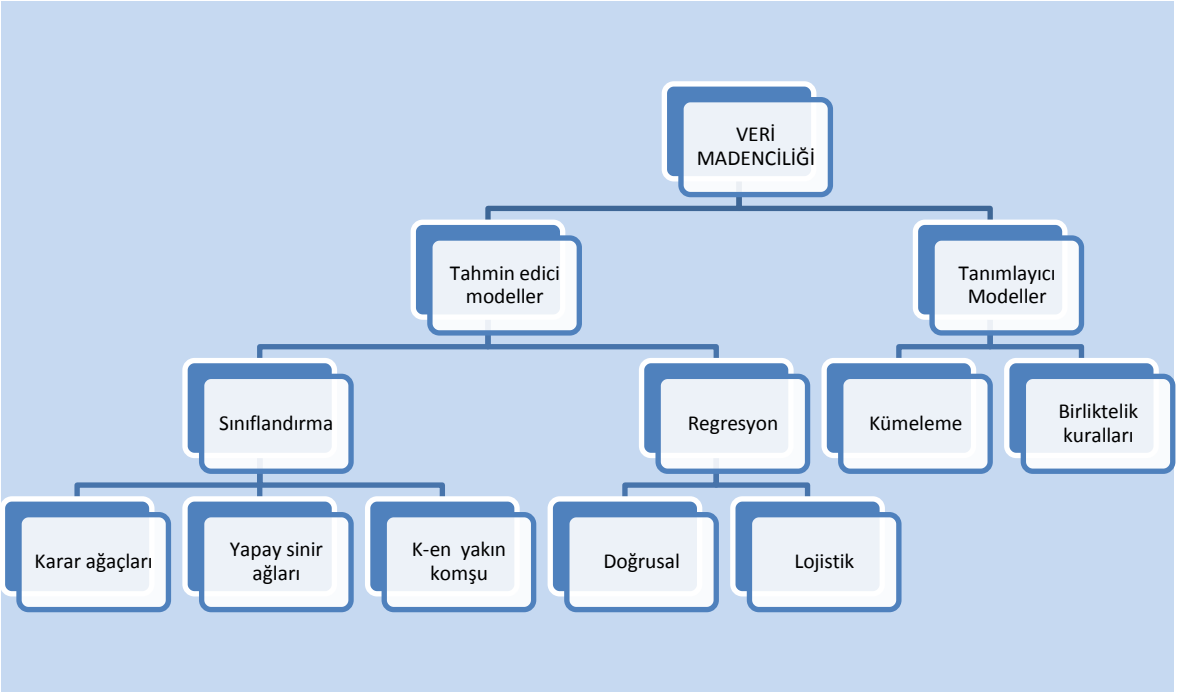
Veri madenciliği, büyük miktarda verilerden anlamlı örüntüler çıkarma amaçlı kullanıldığından günümüzde birçok alanda kullanılmaktadır. Müşteri ilişkileri yöntemleri, eğitim alanında ölçme değerlendirme, hayvanların gelişim ve faydalanma yöntemleri gibi birçok alanda kullanılmaktadır.

Ülkemizde resmi olarak Türkiye istatistik enstitüsü tarafından toplanan veriler ile enflasyonun tespiti, büyüme oranları gibi birçok alanda da kullanılmaktadır.

3. VERİ MADENCİLİĞİNDE KULLANILAN MODELLER

Veri madenciliği ve programlama araçları, ulaşılmak istenen bilgileri çıkarmaya yardımcı olur. Bu bilgilere ulaşmak için birçok farklı veri madenciliği modeli mevcuttur. Mevcut veri ve ulaşılmak istenen sonuca göre model belirlenir. Belirlenen model tahmin edici, tanımlayıcı veya hem tahmin edici hem tanımlayıcı olarak belirlenebilir (Bounsaythip ve Runsala 2001).

Veri Madenciliğinde kullanılan modeller, temel olarak şekil 3.1'de görüldüğü üzere tahmin edici ve tanımlayıcı olmak üzere iki ana başlık altında incelenmektedir.



Şekil 3.1 Veri madenciliğinde kullanılan modeller

3.1 Tanımlayıcı Modeller

Tanımlayıcı modellerde eldeki veriler ile karar verme aşamasında verilerdeki örüntülerin tanımlanmasını sağlamaktır. Veri setlerine dayandırılan sonuçlar için benzer koşularda farklı durumların örüntülerini ve ilişkilerini çözümlemede kullanılır.

3.1.1 Kümeleme analizi

Kümeleme analizinde amaç benzer özellikleri gruplamaktır. Kümeleme analizi çok değişkenli istatistik yöntemlerinde biridir. Benzer özelliklerde veriler küme içinde değişken durumda olan verileri gözlemede kolaylık sağlar. Dolayısıyla elde edilen kümelerdeki veriler homojendir (Hair vd. 1998).

Kümeleme analizinde birincil öncelik, dağınık bir halde bulunan benzer özelliklere sahip verileri bir araya getirip işlenebilir bir halde sınıflandırmaktır (Koç 2005).

Kümeleme analizi sınıflandırma yöntemiyle birçok noktada benzerlik göstermektedir. Bu iki yöntem arasındaki en temel fark, benzer özellikteki verilerin daha önceden belirlenmemiş olmasıdır. Kümeleme analizinde genellikle yapay sinir ağları ve istatistiksel yöntemlerden yararlanılmaktadır (Akpınar 2000).

Kümeleme analizinde, küme üyelerinin birbirine çok benzemesi ile elde edilen kümelerin özeliklerinin farklı olması esastır. Veri deposundaki verilerin farklı kümelere bölünmesi ve işlenebilir bir halde işlenmesi analizin temel dayanak noktasıdır. Oluşturulmuş kümeler ulaşılmak istenen veri sonuçları için güçlü birer temsilci olacaktır.

3.1.2 İlişki analizi

İlişki analizi birliktelik kuralları ve ardışık zamanlı örüntüler gibi alt kategorilerde yaygın olarak kullanılmaktadır.

3.1.2.1 Birliktelik kuralları

Veri madenciliği araştırma yöntemlerinde büyük veri setleri içinde aralarında ilişki bulunan nesneleri belirlemede kullanılır. Birliktelik kuralları ticaret, mühendislik, fen ve sağlık sektörlerinin de bulunduğu birçok alanda yaygın olarak kullanılmaktadır.

Birliktelik kuralları belirli ölçütlere göre bir araya getirilen veri setlerinden oluşur. Dağınık anlamsız bir şekilde duran verilerin, belirli ölçütlerle birliktelik kuralarına uygun bir araya getirilmesi ile veri tabanı bilgi keşfi süreci başlatılmış olur. Büyük ve dağınık veri tabanlarında başlangıçta fark edilmeyen ilişkilerin oluşturulabilmesi ve sürecin yönetimi noktasında yardımcı olur.

Günümüzde online alışveriş, süpermarket, alışveriş merkezleri gibi aynı ortamda farklı ürünleri temin edebildiğimiz tüketim ortamları yaygınlaşmasından kaynaklı birliktelik kuralları ile oluşturulan veri madenciliği yöntemi önem kazanmıştır. Literatürde Pazar sepeti olarak adlandırılan çözümleme yöntemi ile müşteri ile ilgili analizler yapılmaktadır. Müşterinin bir sonraki tüketim malzemesine göre yapılan analiz ile hem müşteri hem satıcı odaklı anlamlı bir veri seti oluşturulur.

Müşterinin aldığı ürünle birlikte veya sonrasında hangi ürünü alacağı tahmin edilerek satış stratejisi belirlenir (Yen ve Lee 2006). Ayrıca birbirleri ile ilişkileri olan farklı alanlar arasındaki ilişkilerin belirlenmesi ile anlamlı veri birliktelikleri oluşturulmasında da birliktelik yöntemi kullanılır.

Veri madenciliği yöntemlerinin en iyi örneklerinden biri olan birliktelik kuralları anlamsız veri setlerinin potansiyel ilişkisini tanımlama da kullanılır. Birliktelik kuralları oluşturulması kararlı kurallar ile güvenilirliğe önem vererek oluşturulur (Tuğ 2006).

Müşteri tüketim analizi ile oluşturulmuş Pazar sepeti analizi için, veri seti alındı-alınmadı bilgisi ile oluşturulmuş birliktelik kuraları, tüketim malzemeleri arasındaki ilişkiyi, destek ve güven kriterlerine göre hesaplar.

Destek:

P(X ve Y) = X ve Y Mallarını Satın Almış Müşteri Sayısı / Toplam Müşteri Sayısı

Güven: P(X/Y) = P(X ve Y) / P(Y)

X ve Y Mallarını Satın Almış Müşteri Sayısı / Y Malını Satın Almış Müşteri Sayısı

Destek kriteri, iki ürünü birlikte alan müşteri oranın belirlenmesi ile elde edilir. Güven kriteri Y ürünün almış olan kişilerin X malını da almış olma olasılığı hesaplanması ile elde edilir. Elde edilen sonuçlar ne kadar yüksek olursa bu iki ürünün birliktelik kuralları ile yapılan analizi sonucu o kadar başarılı olacaktır. Yapılan destek ve güven analizlerinin sonucuna göre raf düzenlemesi, reklam, satış kampanyaları belirlenir (Alpaydın 2004).

3.1.2.2 Ardışık zamanlı örüntüler

Tanımlayıcı modeller içinde yer alan ardışık zamanlı örüntüler olayların sıralı takibi ile elde edilen analiz yöntemidir. Gerçekleşen bir olaydan sonra gerçekleşme olasılığı oluşan başka bir olayın meydana gelme olasılığına dayanan veri madenciliği yöntemidir. Ardışık zamanlı örüntüler birliktelik kuralları yöntemi ile tümleşik analiz edilerek elde edilen sonuçlar, güvenilirlik kriterinin daha yüksek elde edilmesini de sağlar. Örneğin; profesyonel bir fotoğraf makinası alıp bu alanda yeni başlayan müşteri birkaç ay sonra yeni bir lens alma ihtiyacı duyacaktır. Mağaza yetkilileri fotoğraf makinası için yaptığı kampanyadan birkaç sonra lens ürünleri içinde bir kampanya düzenleyerek satış oranına katkı sağlayabilir.

X ameliyatı yapıldığında 15 gün içinde %45 olasılıkla Y enfeksiyonu oluşacaktır (Akpınar 2000).

3.2 Tahmin Edici Modeller

Veri madenciliğinde kullanılan modellerden bir diğeri olan tahmin edici modellerde sonuçları bilinen verilerden hareket edilerek, yeni bir model oluşturulması ve bu modelden yararlanarak yeni veri kümeleri için sonuç tahmin edilmesi amaçlanmaktadır. Örneğin bankalar müşterilerine kredi verip verme noktasında karar kılmak için, daha önce kredi vermiş olduğu müşteri bilgilerinden yola çıkarak tahminde bulunabilir. Banka müşterilerin tüm bilgilerine sahip olduğundan kredi geri ödeme noktasında,

hangi kriterlerin daha önemli olduğunu tespit edip bu tespitler sonucunda yeni kredi vermede müşteri ile ilgili tahminde bulunabilir.

3.2.1 Sınıflandırma

Sınıflandırma yönteminin amacı önceden bilinmeyen verilerin sınıfını tahmin etmektir. Yeni bir verinin belirlenmiş sınıflardan hangisine ait olup olmadığı belirlenir. Belirlenmiş sınıflar, veri tabanından alınan verinin sınıflandırılması için model uygulamada kullanılır (Bergero 2002).

Sınıflandırma belirli bir hastalığın semptomları ile hastalığın belirlenmesi için, bir ürünün özellikleri ile müşteri özelliklerinin eşleşmesi gibi sınıflandırma yöntemlerinde kullanılır. Örneğin, genç bayanlar küçük araba satın alır; yaşlı ve zengin erkekler, büyük ve lüks araba satın alır.

Şirketler, tanıtım politikalarını belirlerken sınıflandırma yöntemi ile tahmin edilen bu durumun sonucu olarak, genç bayanların okuduğu dergi, magazin veya gittikleri mekanlarda küçük model araçların tanıtımlarını sağlayarak gerçekleştirir (Alpaydın 2004).

Sınıflandırma veri madenciliğinde en çok kullanılan yöntemlerden birisidir. Görüntü tanıma, hastalık tanıları, kalite kontrol ve pazarlama konuları gibi veri madenciliğin yaygın kullanıldığı alanlarda sınıflandırma yöntemi sıklıkla kullanılır. Sınıflandırma tahmin edici bir model olup eldeki veri setine ve sınıflara göre yeni verilerin veya olayların sonraki aşaması tahmin edilir.

3.2.2 Regresyon analizi

Veri madenciliğinde tahmine dayalı kullanılan regresyon modelinde, bağımlı bir değişkenin, bir veya birden fazla bağımsız değişken ile arasındaki ilişkiyi tahmin etmeyi

amaçlar. Tahmin edilen ilişki matematiksel bir fonksiyon haline çevrilerek uygulanır. Uygulama sonucunda bağımlı değişkenin değeri tahmin edilir (Orhunbilge 1996).

Regresyon ve sınıflandırma modelleri arasındaki en temel fark, bağımlı değişkenin süreklilik gösteren bir değere sahip olmasıdır. Bağımlı değişken sürekli ise kullanılacak model regresyon, değil ise sınıflandırma modeli olacaktır.

Sınıflama ve regresyon modellerinde kullanılan başlıca teknikler;

- Karar ağaçları
- Yapay sinir ağları
- Lojistik regresyon
- Genetik algoritmalar
- K-en yakın komşu
- Bellek temelli nedenleme
- Naïve-Bayes,
- Bulanık Küme Yaklaşımı

4. KARAR AĞAÇLARI

Bir veya birden fazla bağımsız değişkeni farklı olasılık ve ikili homojen dağılımlarla, kategorik veya sürekli değişkenler ile bağımlı değişkeni tahmin etmeye, değişkendeki değişimi belirlemeye yarayan ve şekil yönünden ters ağaç görüntüsü oluşturan modellemeye ağaç modelleri denir (Marian vd. 2002).

Verilerin görsel olarak modellenmesi, bağımsız değişkenler ile bağımlı değişkenler arasındaki ilişkilerin belirlenmesi ve bu ilişkiden yola çıkarak tahmin yürütülmesi için karar ağaçları kullanılır.

Ağaç modelleri, bağımsız değişkene ait temel sorulardan aldığı cevaplar ile oluşturduğu ağaç dalları ile bağımlı değişkeni hangi bağımsız değişkenlerin etkilediğini belirler (Orekeci-Temel 2004).

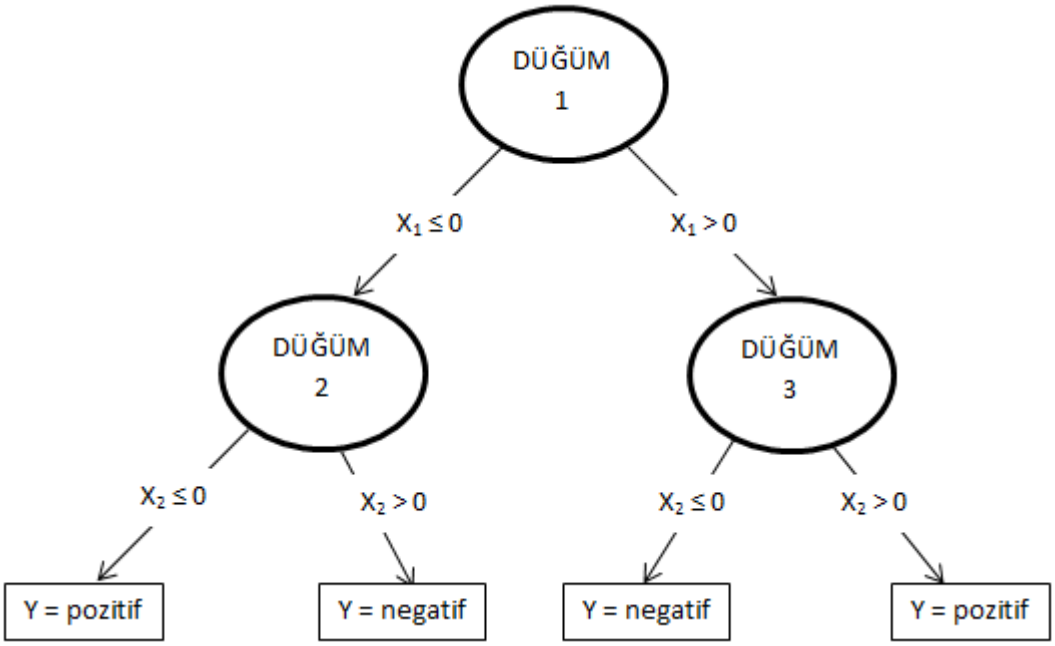
4.1 Ağaç Modeli

Ağaç modelinde, X_1 ve X_2 ; [-1;+1] aralığında değişen düzgün olasılık dağılımdan rastgele seçilen n_1 ve n_2 örneklerin içerdiği bağımsız değişkenler olduğunda, bağımlı değişken Y ise bağımsız değişken değerlerine çarpma kuralı uygulanarak elde edilen sonuç değişkeni ise; çarpma kurallarına göre;

$X_1 * X_2 \geq 0$ ise Y pozitiftir.

$X_1 * X_2 < 0$ ise Y negatiftir.

Bağımlı değişken pozitif ve negatif diye iki seviyeye sahiptir. Ölçme seviyelerine bakılmaksızın, şekil 4.1 çarpma kuralının ağaç yapısını göstermektedir.



Şekil 4.1 Basit bir ağaç yapısı

Ağaç modellerinde karar verme noktalarına düğüm denir. Şekil 4.1'deki ağaç modelinde başlangıç düğüm Düğüm 1'dir (Speybroeck vd. 2003). Başlangıç düğümü bağımsız değişkenlerin tüm etki olasılıklarını içerdiği için en karmaşık düğümdür. Bu düğüme aile düğümü ya da kök düğümü de denir. Kök düğümü iki alt çocuk düğümüne ayrılır. Şekil 4.1 deki örnek ağaç modelinde çocuk düğümler düğüm 2 ve düğüm 3'tür. Başka bir değişle her aile düğümü çocuk düğümlerin bir araya gelmesi ile meydana gelir. Kök düğümü oluşturan yeni düğümler belirlenen özeliğin tüm koşularını tanımlamak zorundadır. Kök düğüm iki alt gruba bölündüğünde mutlak şart tüm olasılıkları barındırmak zorunda olmasıdır. Her çocuk düğümü bir sonraki adımda aile düğümü olur. Bu ifade matematiksel olarak aşağıdaki gibi ifade edilir.

Düğüm 1 = Düğüm 2 U Düğüm 3

Kök düğümünden her bölünmede oluşan çocuk düğümleri kök düğümüne göre daha homojendir. Örnek şekilde çocuk düğümlerinde henüz karar verme gerçekleşmemiştir ve dolayısıyla çocuk düğümler saf değildir. Amaç çocuk düğümlerini belirlenmiş ayırma kriterlerine göre ikili bölümlere ayırarak karar noktalarını yani terminal

düğümlerini elde etmektir. Terminal düğümlerde sapma sıfır olarak kabul edilir. Terminal düğümler ağaç yapısındaki en homojen düğümler olduğu söylenebilir. Terminal düğümlerden sonra bölünme gerçekleşmez (Bremner ve Tablin 2002).

Ağaç modellemede temel amaç; Kök düğümden başlayarak ayırmalara başlayarak homojen alt gruplara ayırarak karar noktalarına ulaşıp bağımlı değişkeni tanımlamaktır. Düğüm 'deki gözlemler sıfıra eşit veya sıfırdan küçük iseler düğüm 2'ye aksi takdirde düğüm 3'e atanırlar. Aynı şekilde terminal düğümlere ulaşana dek gözlemler bağımsız değişkenlere göre diğer düğümlere atanırlar. Şekil 4.1'den de anlaşılacağı üzere bağımlı değişkene ait grupları sınıflandırma birden fazla bağımsız değişken kullanılır. Böylece, bağımsız değişkenler arasındaki etkileşim ve bağımsız değişkenlerin bağımlı değişken üzerindeki etkisi ağaç modelleriyle tanımlanabilmektedir (Lewis 2000).

4.2 Ağaç Modellerinin Amaçları

Verilerin etkileşimi belirlenmesi ve modelin elde edilmesinden sonra tahminler yapılabilmesi için istatistiksel analizlere başvurulur. Ağaç modelleri, veriler arasındaki etkileşimi görsel olarak sunar ve istatiksel analizlerde kolaylık sağlar. Ağaç modelleri, çalışma yürütücüsüne problemi ayrı ayrı her bir bağımsız değişkenin etkisini incelemesinde kolaylık sağlar ve karmaşık problemlerin olası alt kümelerinin aşamalı olarak değerlendirilmesine olanak sağlar (Orekeci-Temel 2004).

Ağaç modellerinin sınıflama ve tahmin olmak üzere iki temel amacı vardır. Sınıflama, veri setini sistematik, anlaşılır ve tahmin edilebilir bir şekilde modeller. Tahmin ise gözlenmeyen verileri sınıflandırma sonuçlarına göre güvenli bir meydana gelme olasılığından yola çıkarak yorumlamaktır. Sınıflandırma ve tahmin etme aşamalarında kolaylık sağlayan ağaç modelleri kullanılmaktadır (Orekeci-Temel 2004).

5. SINIFLAMA VE REGRESYON AĞAÇLARI

5.1 Tanımı

Sınıflama ve regresyon ağaçları (CART) kategorik ya da sürekli bağımlı değişkenlerin analiz ve tahmin etmek üzere geliştirilmiş parametrik olmayan istatistiksel bir metottur. Bağımlı değişkenin sürekli olup olmamasına göre CART sınıflama ve regresyon diye iki farklı isimde adlandırılır. Bağımlı değişken kategorik ise sınıflama, sürekli ise regresyon ağacı ismini alır (Fu 2000). CART bağımlı değişkeni etkileyen bağımsız değişkenlerin, aralarındaki etkileşime göre, ikili alt gruplara ayrılması ile elde edilen ağaç modelidir. Tekrarlı ikili alt gruplara ayırma, karar noktalarına ulaşıncaya kadar devam eder (Chipman vd. 2000).

CART kök düğümden başlayan, ikili alt gruplara ayırma ile terminal düğümler elde edilene dek devam eden mekanik bir öğrenme metodudur. İkili alt gruplara ayırma işlemi ile her adımda daha homojen alt gruplar elde edilir (Bevilacqua vd. 2003).

5.1.1 Kullanım alanları

Sağlık bilimlerinde CART tekniği son yıllarda büyük bir gelişme göstermiştir. CART yaygın olarak tıp biliminde tanı koymada, botanikte sınıflamada ve karar teorisinde kullanılmıştır. CART Frydman, Altman ve Kao tarafından 1985 yılında risk altındaki firmaların sınıflandırılmasında, Marais, Patell ve Wolfson tarafından ticari borçların sınıflandırılmasında kullanılmıştır.

Uluslar Arası Gıda Politikası Araştırma Enstitüsü (IFPRI) tarafından hane halkı kıtlık seviyesi belirlenmesinde 1985 yılında CART kullanılmıştır. Ekolojik verilerin değerlendirilmesinde 1992 yılında Staub,1993 yılında Baker ve 1999 yılında Rejwan tarafından kullanılmıştır (De'ath ve Fabricius 2000).

5.1.2 CART' ın avantajları

İstatiksel analizlerde CART'ın kullanılmasının temel nedenleri şunlardır.

- CART parametrik olmayan bir modeldir.

- Değişkenlerin türünün sürekli, kategorik veya bunların karışımı olması konusunda bir varsayım yoktur.

- Bağımlı ve bağımsız değişkenlerin dağılımı ile ilgili bir varsayım gerektirmediğinden değişkenlere herhangi bir dönüşüm uygulanmasına gerek yoktur (Orekeci-Temel 2004).

- Bağımlı değişkenin bağımsız değişkenler ile olan etkileşimi ikili ağaç modelleri ile tanımlandığından yorumlaması daha kolaydır.

- Tanımlanan bağımlı değişken için etkileşim içinde olan bütün bağımsız değişkenleri ve onların tüm kombinasyonlarını modelde belirlemesi ile mümkün olan en doğru sınıflandırmayı gerçekleştirir. Değişkenlerin tüm etkileşim olasılıklarına bakıldığı için daha esnek ve daha geniş bir bakış açısı ile elde edilir. Böylece bağımlı değişkeni etkileyen bağımsız değişkenlerin önem sırası da belirlenmiş olur.

- Büyük ve karmaşık veri setlerinde dahi doğru tahmin yapılabilir.

- Bağımlı ve bağımsız tüm değişkenler için kayıp veya eksik değerler ile aşırı uç değerlerden etkilenmeyen bir metottur.

- Metot sıralamasında analizciye düzeltme olanağı tanır (Lewis 2000).

- Belirlenen ayırma kriterlerine göre bağımsız değişken alt gruplara bölünebilir.

- Bağımlı değişkeni etkileyen bağımsız değişkenlerin belirlenmesinde ve aralarındaki etkileşimi belirlemede kullanılır(Lewis 2000).

5.2 Sınıflama Ağacının Oluşturulması

Sınıflama ağacının oluşturulmasında temel amaç, bağımsız değişkenlerin değerlerine göre bağımlı değişkenin alabileceği değerleri tahmin etmektir. Bu amaç doğrultusunda, N adet deney ünitesine ait verileri içeren bir başlangıç seti kullanılır. Bu veri setine Learning Sample denir ve terminolojide L ile gösterilir.

$X = \{x_1, x_2, \ldots\ldots\ x_n, \ldots\ldots\ x_N\}$

$J = \{j_1, j_2, \ldots\ldots\ldots\ j_n, \ldots\ldots\ldots\ j_N\}$

$L = \{(x_1, j_1), (x_2, j_2), \ldots\ldots\ldots\ldots, (x_n, j_n), \ldots\ldots\ldots, (x_N, j_N)\}$

$x_n \in X\ ; j_n \in (1,2,3, \ldots, C)\ ;\ n = 1,2,3 \ldots\ldots N'dir$

Burada;

L : Başlangıç veri seti(bağımsız değişkenlere ait bilgilerin tümünü içeren alan vektörüdür)

J : Deney ünitelerinin ait olduğu kategoriler vektörü

X_n: n. birey yada objeye ait bağımsız değişkenin değeri

J_n: n. birey yada objenin ait olduğu kategori

C: Deney ünitelerinin ait olduğu kategori sayısı

N: Toplam deney ünitesi sayısı

Ölçüm vektöründe yer alan değişkenler rakamsal olarak ölçülebilen sayılardan oluşuyorsa o vektöre sürekli değişkenler vektörü, sayımla belirtilen değerlerden oluşuyor ise kategorik değişken vektörü denir.

Sürekli değişkene kişinin ağırlığını, boy uzunluğunu, yaşını kategorik değişkene ise kişinin cinsiyetini, ebeveynlerin hayatta olma durumunu örnek olarak verebiliriz.

Sınıflama ağaçları, kök düğümden başlayarak ve her düğümde o düğüme ait deney ünitelerine sorulan evet/hayır cevaplı soruların sonucunda şekillenen bir ağaç modeli oluşturur. Her aşamada sorulan bu sorulara ayıraç denir. Her ayıraç t düğümünü iki alt gruba ayırır. Bu işleme ayırma işlemi denir ve terminolojide $\delta(t)$ ile gösterilir. Bağımsız değişkenin sürekli veya kategorik oluşuna göre sınıflama ağaçlarında ayırmalar ikiye ayrı şekilde tanımlanır.

Eğer X sürekli bir değişken ise ayıraç "s gerçek bir sayı olmalıdır ve düğümü alt gruplara ayırmak için gerekli olan " xn ≤ s midir? " sorusudur. Yanıt evet ise sol düğüme, hayır ise sağ düğüme atanır. Bağımsız değişkenin k adet farklı değeri var ise en fazla k-1 ayraç ve dolayısıyla k-1 adet farklı ayırma mümkündür.

Eğer X kategorik bir değişken ise ayraç" A C X olmak üzere xn ∈ A mıdır ?" sorusudur. Soruya alınan cevap evet ise sol düğüme, hayır ise sağ düğüme atanır. Kategorik değişken k adet farklı kategorik değişken içeriyorsa değişken en fazla 2^{k-1}adet alt kümesi vardır ve dolayısıyla 2^{k-1} - 1 adet farklı ayırma mümkündür(De'ath ve Fabricius 2000).

Birden fazla bağımsız değişkenin mevcut olduğu durumlarda yukarıda tanımlanan işlemler kullanılabilir. Ancak kullanılan ayraç bütün bağımsız değişkenler ve kombinasyonları için tek tek uygulanmalıdır. Mümkün olan tüm olası ayırmalar belirlenmeli ve mevcut durumlar için kullanılmalıdır.

Sınıflama ağaçlarında, herhangi bir t düğümde kullanılabilecek tüm ayıraçlar belirlendikten sonra, mümkün olan ayırmalar için ayırmanın uygunluk derecesi hesaplanır. Bu hesaplamada ayırma fonksiyonu kullanılır. Matematiksel olarak ayırma fonksiyonu aşağıdaki gibi tanımlanır;

$$\Delta\big(\delta,(t)\big) = i(t) - P_L\, j(t_L) - P_R j(t_R)$$

Burada;

P_L : t. düğümden sol çocuk düğümüne atanan deney ünitelerinin oranı.

P_R : t. düğümden sağ çocuk düğümüne atanan deney ünitelerinin oranı.

$i(t)$: t düğümünün safsızlık ölçüsü

$i(t_L)$: Sol çocuk düğümünün safsızlık ölçüsü

$i(t_R)$: Sağ çocuk düğümünün safsızlık ölçüsü(Breiman vd. 1993)

Ayırma fonksiyonu ile olası bütün ayırmaların uygunluk derecesi hesaplandıktan sonra, en iyi uygunluk derecesine sahip ayırma ile t düğümü alt gruplara ayrılır.

Düğümlerin heterojenlik değeri safsızlık ölçüsü olarak adlandırılır ve bu değer safsızlık fonksiyonu ile belirlenir. Bu değer sıfır hesaplanması düğümün tamamen homojen olduğu anlamına gelmektedir.

Sınıf numaraları j = 1,2……,k olan kategorik bir değişken için, t düğümünün safsızlık ölçüsü matematiksel olarak aşağıdaki gibi ifade edilir;

$$i(t) = \varphi\{p(1|t), p(2|t), \dots\dots\dots, p(k|t)\}.$$

Burada;

$i(t)$: t düğümünün safsızlık ölçüsü

φ : Safsızlık(impurity) fonksiyonu

$p(j|t)$: t.düğümde, bağımlı değişken j sınıfına atanan deney ünitelerinin oranıdır.

Safsızlık ölçüsü belirlenmeye çalışılan düğümde her kategori eşit orana sahip olursa fonksiyon maksimum değerine ulaşır. Tek bir kategoriyi kapsar ise minimum değerine ulaşır.

Safsızlık ölçüsünü belirlemede birçok yöntem kullanılmaktadır. Safsızlık ölçüleri literatürde en iyi ayırma kriterleri olarak adlandırılır. Bu yöntemler arasında en yaygın kullanılan ayırma kriterleri Gini diversity index(Gini) ve Twoing kuralı'dır.

5.2.1 Ayırma kriterleri

5.2.1.1 Twoing kuralı

Twoing kuralı Gini'den oldukça farklıdır. Twoing düğüme ait verilerin %50'sini içeren ve farklı sınıfları birleştirmeye çalışır. Düğümlerin içerdiği değerler göz önüne alınarak iki ayrı gruba ayrılır. Bunlara aday bölünme denilir. Her hangi bir t düğümünde kümeler t_{Sol}, $t_{Sağ}$diye iki ayrı alt gruba ayrılır. Aday bölünmelerin her biri için Psol ve $P(j|t_{Sol})$ olasılıkları hesaplanır. $P(j|t_{Sol})$ ifadesi bir j kategorisinin sol tarafta olma olasılığını hesaplama için kullanılır. Olasılık hesaplamasını matematiksel olarak şöyle ifade edebiliriz;

$$P_{Sol} = \frac{t_{Sol}\text{' daki herbir nitelik değerinin ilgili nitelik sütunundaki tekrar sayısı}}{\text{Eğitim kümesindeki kayıtların sayısı}}$$

$$P(j|t_{Sol}) = \frac{t_{Sol}'\text{daki kayıtların j sınıfları sayısı}}{t_{Sol}'\text{daki herbir nitelik değerinin ilgili nitelik sütunundaki tekrar sayısı}}$$

Aynı işlemler sağ tarafta olma olasılığı için aşağıdaki gibi hesaplanır;

$$P_{Sağ} = \frac{t_{Sağ}\text{' daki herbir nitelik değerinin ilgili nitelik sütunundaki tekrar sayısı}}{\text{Eğitim kümesindeki kayıtların sayısı}}$$

$$P(j|t_{Sağ}) = \frac{t_{Sağ}'\text{daki kayıtların j sınıfları sayısı}}{t_{Sağ}'\text{daki herbir nitelik değerinin ilgili nitelik sütunundaki tekrar sayısı}}$$

$\Phi(s|t)$, t düğümündeki s aday bölümlerinin uygunluk ölçüsü olsun. Uygunluk ölçüsü aşağıdaki gibi hesaplanır:

$$\Phi(s|t) = 2P_{Sol}P_{Sağ}\sum_{j=1}^{n}|P(j|t_{sol}) - P(j|t_{Sağ})|$$

Yapılan işlemler sonucu hesaplanan uygunluk ölçüsü değerleri içerisinden maksimum değere sahip olan seçilir ve bu değerin ilgili olduğu aday bölünme satırı dallanmanın yapılacağı düğümü belirtir.

5.2.1.2 Gini algoritması

Sınıflama ağaçlarında her hangi bir düğüm için, mümkün olan en iyi ayırma yöntemini belirlemede kullanılan ayırma kriterlerinde Gini en yaygın kullanılan yöntemlerin başında gelir. Gini bağımlı değişkeni tanımlamak için kullanılan bağımsız değişkenler için sorulan sorulara verilen cevaplar ile oluşturulan ağaç yapısında bölütlemede en geniş kategoriyi ayırarak işlem yapar. Gini ayırma kuralının amacı ayırma fonksiyonunu maksimum yapacak seviyede bir sınıfı diğerlerinden ayırmaktır. Gini'nin ayırma kriterlerine göre hangi sınıf ayrılırsa ayrılsın sol çocuk düğümünün safsızlık ölçütü sıfırdır. Ancak ayırma fonksiyonunun maksimum olabilmesi için sağ çocuk düğümünün

safsızlık ölçüsü de düşük olmalıdır. Tek bir sınıfı diğerlerinden tamamen homojen bir şekilde ayırabilecek bir ayıraç bulmak normal şartlarda zor yada imkansızdır. Gini safsızlık ölçüsü matematiksel olarak aşağıdaki gibi ifade edilir;

Her bir nitelikle ilgili sol ve sağ taraftaki bölümler için Gini_{Sol} ve $\text{Gini}_{\text{Sağ}}$değerleri hesaplanır.

$$\text{Gini}_{\text{Sol}} = 1 - \sum_{i=1}^{k} \left(\frac{L_i}{|T_{Sol}|}\right)^2$$

$$\text{Gini}_{\text{Sağ}} = 1 - \sum_{i=1}^{k} \left(\frac{R_i}{|T_{Sağ}|}\right)^2$$

Bu bağlantılarda yer alan ifadelerin anlamları şu şekildedir:

k: Sınıfların sayısı

T: Bir düğümdeki örnekler

$|T_{Sol}|$: Sol taraftaki örneklerin sayısı

$T_{Sağ}|$: Sağ taraftaki örneklerin sayısı

L_i: Sol taraftaki i kategorisindeki örneklerin sayısı

R_i: Sağ taraftaki i kategorisindeki örneklerin sayısı

Her j niteliği için, n eğitim kümesindeki satır sayısı olmak üzere aşağıdaki bağıntının değeri hesaplanır:

$$\text{Gini}_{\text{j}} = \frac{1}{n}(|T_{Sol}|\text{Gini}_{\text{Sol}} + |T_{Sağ}|\text{Gini}_{\text{Sağ}})$$

Her j niteliği için hesaplanan Gini_{j} değerleri arasından en küçük olan seçilir ve bölünme bu nitelik üzerinden gerçekleştirilir.

5.2.1.3 Gini ve Twoing ayırma kurallarının karşılaştırılması

Mümkün olan en iyi ayırmayı elde etmek için, ayırma kuralı belirlerken bazı durumlar a dikkat edilmelidir. İki seviyeli kategorik bağımlı değişkenin hata oranı 0.50'den az olduğu tahmin ediliyorsa Gini, 0.80'den az olduğu tahmin ediliyorsa Twoing ayırma kuralları tercih edilmelidir. Bağımlı değişkenin seviyesi dörtten daha büyük olduğu durumlarda twoing ayırma kuralı tercih edilmelidir (Özkan 2013).

5.2.2 Risk matrisi

Bir sınıflama modelinde yanlış olarak sınıflanan olay sayısının, tüm olay sayısına bölünmesi ile hata oranı, doğru sınıflandırılan olay sayısının tüm olay sayısına bölünmesi ile de doğruluk oranı tespit edilir.

Doğruluk oranı = 1 - hata oranı

Hata oranlarına karar vermek için risk matrisi kullanılır. Risk matrisi yapılması gereken düzenlemelerin önceliğini belirlemede kullanılır. Verilerin sınıflandırılması sonucunda ortaya çıkan her hangi bir düğüme atanacak olan en uygun sınıf aşağıdaki gibi belirlenir.

$C(j/i)$: i sınıfını j sınıfı gibi sınıflamanın maliyeti (risk matrisi katsayıları)

π_i : i sınıfının önceki olasılığı

N_i : Learning sample'da i sınıfında bulunan deney üniteleri sayısı

$N_i^{(t)}$: t düğümünde i sınıfında bulunan deney üniteleri sayısı olmak üzere;

$\frac{C(j/i)\pi_i N_i^{(t)}}{C(i/j)\pi_j N_j^{(t)}} > \frac{N_i}{N_j}$ eşitsizliği j'nin bütün değerleri için sağlıyorsa t düğümüne en yakın olan i sınıfı atanır.

Yapılan hesaplamalar sonucunda, birden fazla sınıfı bu eşitsizliği sağlayabilir veya hiçbir sınıf bu eşitsizliği sağlamayabilir. Bu durumda en uygun sınıfı belirlemek için çoğulculuk kuralı veya minimum risk kuralı uygulanır.

5.2.3 Çoğulculuk kuralı

Verilerin sınıflandırılması ile tespit edilen hatalı sınıflandırma maliyetini bütün düğümler için eşit sayarak, düğüm içerisinde en büyük orana sahip olan sınıfı en uygun sınıf olarak atar.

5.2.4 Minimum risk kuralı

Minimum risk kuralı hatalı sınıflandırılan düğümler içerinde maliyeti en düşük olan sınıfın seçilmesidir. Örnek olarak, bir problemde iki sınıf (S1,S2) olsun;

$r_1^{(t)}$: Sınıf 1'in t düğümüne atanma maliyeti

$r_2^{(t)}$: Sınıf 2'in t düğümüne atanma maliyeti

π_1 :Sınıf 1'in ön olasılığı

π_2 :Sınıf 2'in ön olasılığı

$r_1^{(t)} = \pi_1 C\left(\frac{2}{1}\right)$ $r_2^{(t)} = \pi_2 C(1/2)$ olarak tanımlanmış olsun.

Eğer; $r_1^{(t)} < r_2^{(t)}$ ise düğüm t, sınıf 1'e aksi halde sınıf 2'ye atanır.

C(2/1) = C(1/2) ise çoğulculuk kuralına başvurulur. Ön olasılıkları en yüksek olan sınıf 1 bu düğüme atanır.

5.2.5 Doğruluk tahmini

Daha önce belirtildiği gibi, Bir sınıflama modelinde yanlış olarak sınıflanan olay sayısının, tüm olay sayısına bölünmesi ile hata oranı, doğru sınıflandırılan olay sayısının tüm olay sayısına bölünmesi ile de doğruluk oranı tespit edilir. Bağımlı değişkenin kategorik olduğu durumlarda doğruluk tahmini üç farklı yöntemle yapılabilir.

5.2.5.1 Revizyon ve yeniden yerine koyma tahmini

Learning sample'ın tamamı ağaca uygulanarak yeniden sınıflama yapılır ve sonucunda hatalı sınıflandırılmış deney ünitelerinin oranı hesaplanır.

$R(T) = \frac{1}{N}\sum_{i=1}^{N} X(d(x_n) \neq J_n)$ Burada :

X(.) = indikatör (gösterge) fonksiyondur. Bağımlı değişkenin tahmin edilen sınıf üyeliği gerçekte ait olduğu sınıf üyeliğine eşit ise "1" değerini değilse "0" değerini alır.

d(x) = Deney ünitesinin tahmin edilen bağımlı değişken sınıfıdır.

J_n = Deney ünitesinin gerçekte ait olduğu bağımlı değişken sınıfıdır.

N = Toplam deney ünitesi sayısı

5.2.5.2 Test sample tahmini

Test sample tahmin yöntemi, veri setinin yaklaşık üçte birlik kısmını test sample, üçte ikilik kısmını da learning sample olarak iki ayrı gruba ayırır. Learning sample olarak ayrılmış veri seti ile sınıflandırma yapılır ancak, hata oranı test sample ile hesaplanır. Kullanılan yöntem veri setinin üçte birlik kısmını test sample olarak ayırdığı için büyük veri setlerinde kullanılması gerekir.

$$R(T_{ts}) = \frac{1}{N_2}\sum_{i=1}^{N} X(d(x_n) \neq J_n)$$

N_2 = Test sample'daki toplam deney sayısı

5.2.5.3 Çapraz geçerlilik testi

Sınırlı veri olduğu durumlarda kullanılan bu yöntemde, mevcut deney üniteleri rastgele iki eşit gruba(a, b) ayrılır. İki aşamalı yürütülen bu testte ilk olarak a grubu learning samble b grubu test sample olarak, ikinci aşamada ab grubu learning sample a grubu test sample olarak ele alınır. Bu iki aşamada da hata oranı hesaplanır ve elde edilen sonuçların aritmetik ortalaması alınarak ağacın hata oranı hesaplanır.

Mevcut deney üniteleri v eşit gruba ayrılması durumunda, çapraz geçerlilik testi v katlı uygulanmış olur. Ağacın hata oranı v tane hata oranının aritmetik ortalaması sonucu elde edilir. Çapraz geçerlilik testi matematiksel olarak aşağıdaki gibi ifade edilir.

$$R(T_{ts}^{cv}) = \frac{1}{N_v}\sum_{i=1}^{N} X(d^v(x_n) \neq J_n)$$

$$N_v = \frac{N}{v}$$

5.3 Regresyon Ağaçları

Bağımlı değişkenin sürekli verilerden oluştuğu durumlarda CART regresyon ağacı üretir. Regresyon ağacının amacı, bağımlı değişken ile bağımsız değişken arasındaki yapısal ilişkiyi ortaya koymak ve bağımsız değişkenlere ait ölçüm vektöründen bağımlı değişkenin değerini tahmin etmektir.

Regresyon ağacının oluşturulması sınıflama ağacının oluşturulması ile benzerdir. Ancak ayırma kuralları, uygun kriterin belirlenmesi ve doğruluğunun belirlenmesi farklıdır. Regresyon ağaçları terminal düğümlerde ortalama ve standart sapma değerlerine sahiptir.

5.3.1 Ayırma kuralları

Regresyon ağaçlarında amaç iki çocuk düğümdeki homojenliği maksimum hale getirmektir. Homojenlik en uygun bölme kriteri belirlenerek düğümdeki heterojenlik giderilmesiyle elde edilir. Bu şekilde düğümler arası farklılık en üst seviyeye ulaşır.

Regresyon ağaçlarında Least Squares (LS), Least Absolute Deviation (LAD) ve Clark&Pregibon (CP) olmak üzere üç ayırma kuralı vardır. Ayırma kurallarında amaç heterojenliği minimize etmektir. Heterojenlik ölçüsü i(t) ile tanımlanır.

LS ve LAD kuralları arasındaki temel fark, LS kuralında heterojenlik ölçüsü i(t) düğümdeki ortalama etrafında bağımlı değişkenin kareleri toplamı iken, LAD medyan etrafında bağımlı değişkenin kareleri toplamıdır.

5.3.1.1 Least squared kuralı

$$\boldsymbol{i(t) = \sum_{i=1}^{N} (Y(i) - \overline{Y}(t))^2}$$

Burada;

i(t) = t. düğümdeki heterojenlik

Y(i) = t. düğümdeki bağımlı değişkenin değeri

$\bar{Y}$(t) = t. düğümdeki bağımlı değişkenin ortalama değerini göstermektedir.

5.3.1.2 Clark & Pregibon kuralı

Bu kurala göre sapma düğümdeki bütün gözlemlerin sapmalarının toplamıdır. Amaç hata kareler toplamını (RSS) minimize etmektir.

RRS = $\sum_{i \in L}(y_i - \bar{y}_L)^2 + (y_i - \bar{y}_R)^2$

Burada;

y_i = Sol düğümdeki bağımlı değişkenin değerini

$\bar{y}_L$ = Sol düğümdeki bağımlı değişkenin ortalama değeri

$\bar{y}_R$ =Sağ düğümdeki bağımlı değişkenin ortalama değerini göstermektir.

5.3.1.3 En iyi ayırma kriterlerinin tespiti

$\Phi(t) = i(t) - i(t_R) - i(t_L)$

$i(t_R)$ = Sağ çocuk düğümdeki ortalama etrafındaki kareler toplamı

$i(t_{L)=}$ Sağ çocuk düğümdeki ortalama etrafındaki kareler toplamı

En iyi ayırmada amaç $\Phi(t)$'yi maksimize etmektir. Yani $i(t_R) + i(t_L)$ 'yi minimize etmektir.

5.3.2 Regresyon ağaçlarında doğruluk tahmini

Regresyon ağaçlarında doğruluk tahminlerinin işleyişi sınıflama ağaçlarında olduğu gibidir. Sadece formüller aşağıdaki gibi değişir.

5.3.2.1 Revizyon tahmini veya yeniden yerine koyma tahmini

$$R(d) = \frac{1}{N}\sum_{i=1}^{N}(Y_i - d(x_i))$$

Burada;

R(d) = Hata oranıdır.

$Y_i =$ Sürekli bağımlı değişkenin gerçek değeridir.

$d(x_i) =$ Bağımlı değişkenin tahmin edilen değeridir.

5.3.2.2 Test sample estimate

$$R_{(\boldsymbol{d})}^{ts} = \frac{1}{N_2}\sum_{(x,y)\in L}^{N}(Y_i - d(x_i))^2$$

5.3.2.3 V-Fold cross validation

$$\boldsymbol{R_{(d)}^{cv} = \frac{1}{N}\sum_{y}\sum_{x,y_n}(Y_i - d_{(x_n)}^{y})^2}$$

6. CHAID ANALİZİ

6.1 CHAID Analizinin Genel Yapısı

CHAID(chi-squared Automatic Interaction Detection) otomatik ki-kare etkileşim tekniği sınıflandırılmış bağımlı değişkenlerin tanımlanması ve analizi için kullanılır. Veri madenciliğinde sıkça kullanılan bu analiz yöntemin amacı, veri setini bağımlı değişkenleri analizde kullanılan bağımsız değişkenleri daha homojen alt kategorilere bölmektir. Analizin güvenli ve doğru sonuçlar vermesi veri kümesinin homojen alt kategorilere bölünmesine bağlıdır. Problemin çözümü için yapılan analizlerin güvenirliği için veri setinin büyüklüğü önemlidir. Büyük veri setlerini bağımlı değişkeni destekleyecek ve açıklayacak diğer değişkenleri ve bu değişkenlerle ilgili verileri homojen bir şekilde ortaya koymak CHAID analizinin temel mantığıdır. CHAID bağımlı değişken için gerekli homojen alt grupları belirler. Oluşturulan alt gruplarla sağlanmak istenen amaç tahmin edici veri kümeleri elde etmektir. Bu amaçla oluşturulan alt gruplar ilerleyen adımlarda bağımlı değişkenine bağlı analizleri tahmin etmek için kullanılır. Regresyon problemlerinde de kullanılan CHAID analizi karar ağaçları oluşturulmasında da kullanılan bir analiz yöntemidir. CHAID analizinde bir çok çapraz tablo kullanılması ve istatiksel önem oranının hesaplanmasından kaynaklı ki-kare ismi de kullanılır (Hoare 2004).

CHAID analizi; Bağımlı değişkeni en iyi bir şekilde açıklamak için kullanılır. Bağımlı değişkenleri açıklamak için kullanılan veriler, sınıflayıcı, sıralayıcı ve aralık ölçekli olabilir. Veri analizinde kayıp verileri bir alt kategori olarak alıp p değeri hesaplamada kullanır. Bağımlı değişkeni açıklamak için oluşturulan alt kategorileri bağımsız olarak yeniden tanımlayıp kategorize eder. Homojen alt gruplara ayırma işlemindeki her adımda her bölünme işlemi yeniden bağımsız olarak gerçekleştirir (Pehlivan 2006).

CHAID analizi, büyük veri setlerinde değişkenleri kategorize eder. Belirlenen özelikler bakımından bir birini yakın olan kategorileri birleştirir, istenilen değişkenlere göre bölerek veri setini analiz edilebilir durumda özetlemiş olur. Her bağımsız değişken için oluşturulmuş alt kategoriler birleştirilir ve bağımlı değişkene göre kontenjans tabloları

oluşturulur. Bonferroni p değerleri ile x2 değerleri hesaplanır. Bağımsız değişkenler için hesaplanan p değerleri kıyaslanarak en küçük değere sahip olan bağımsız değişkene göre alt kategorilere ayrılır (Pehlivan 2006).

CHAID analizi, Bağımlı değişkeni açıklayıp analiz etmek için homojen alt gruplara ayırır. Bağımlı değişkenin tahmin edilebilmesi için belirli özelliklere göre ayrılan bu homojen alt gruplar bağımsız olarak yeniden kategorileştirilir. Belirli bir süreç ve yöntem ile birleştirme işlemine tabi tutulur. Birleştirme işlemi tüm alt gruplar birbirleriyle incelenip sonuçlanana kadar devam eder. Bu işlemin sonlandırılması Bonferroni dizisindeki p değerine göre belirlenir (Doğan ve Özdemir 2003).

Bonferroni yaklaşımında, belirlenmiş her bir alt kategori için ortalama vektörü hesaplanır. Hesaplanan vektörlerin, alt grupların hesaplanmış olan genel ortalama vektörlerinden farkları belirlenir. Alt kategorilerin vektörleri ile genel ortalama vektör farkının sıfır olup olmadığı kontrol edilir. Genel ortalama vektörü $\overline{\times}$ ve her grubun i.değişkene göre ortalama vektörleri aşağıda belirtildiği gibi gösterilir.

$$\overline{\times}=\begin{bmatrix}\overline{\times}_{1.}\\ \overline{\times}_{2.}\\ .\\ .\\ \overline{\times}_{p.}\end{bmatrix} \quad \overline{\times}_{.1}=\begin{bmatrix}\overline{\times}_{11}\\ \overline{\times}_{21}\\ .\\ .\\ \overline{\times}_{p1}\end{bmatrix} \quad \overline{\times}_{.2}=\begin{bmatrix}\overline{\times}_{12}\\ \overline{\times}_{22}\\ .\\ .\\ \overline{\times}_{p2}\end{bmatrix} \quad \ldots. \quad \overline{\times}_{.g}=\begin{bmatrix}\overline{\times}_{1g}\\ \overline{\times}_{2g}\\ .\\ .\\ \overline{\times}_{pg}\end{bmatrix}$$

Her grubun ortalama vektörünün, genel ortalama vektöründen farkları değişkenlere göre aşağıdakilere göre belirlenir.

$$\boldsymbol{d_1} = \overline{\times}_{.1} - \overline{\times} = \begin{bmatrix}\overline{\times}_{11}\\ \overline{\times}_{21}\\ .\\ .\\ \overline{\times}_{p1}\end{bmatrix} - \begin{bmatrix}\overline{\times}_{1.}\\ \overline{\times}_{2.}\\ .\\ .\\ \overline{\times}_{p.}\end{bmatrix} \ldots . \boldsymbol{d_g} = \begin{bmatrix}\overline{\times}_{1g}\\ \overline{\times}_{2g}\\ .\\ .\\ \overline{\times}_{pg}\end{bmatrix} - \begin{bmatrix}\overline{\times}_{1.}\\ \overline{\times}_{2.}\\ .\\ .\\ \overline{\times}_{p.}\end{bmatrix}$$

k. grup ile j. grup i. değişken ortalamaları arasındaki ortalama farklarına ait 1-α güven aralığı,

$$(d_{ki}-d_{ji}) = (\bar{x}_{ki}-\bar{x}_{ji}) \mp t\left(\frac{\alpha}{pg(g-1)}, (N-g)\right)\sqrt{\left(\frac{1}{n_k}+\frac{1}{w_j}\right)\frac{w_{ii}}{N-g}}$$

eşitliği ile hesaplanmaktadır. Burada, N=n1+ n2 + n3+…+ng, p değişken sayısı, g grup sayısı ve $\mathbf{w_{ii}}$ ise W matrisinin köşegen elemanlarını ifade etmektedir. W matrisi gruplar için değişimi göstermekte olup,

$$\boldsymbol{W} = \sum_{i=1}^{g}\sum_{J=1}^{n_i}(\bar{x}_{ij}-\bar{x}_{i})(\bar{x}_{ij}-\bar{x}_{i})'$$

denklemi ile hesaplanmaktadır. Burada, g grup sayısını, ni ise i. gruptaki birim sayısını ifade eder. Her bir değişken için gruplar ikişerli olarak dikkate alınır. Yapılan hesaplamalar ile elde edilen i. değişken için aralığın sıfır değerini içerip içermediği kontrol edilir. Değerin sıfırdan farklı olması, belirlenen alt kategoriler ile istatiksel olarak net bir farkın olup olmadığı anlaşılır. Sıfırdan farklı ise elde edilen değer kategoriler arasındaki ilişki için farklılıkların olduğu yorumu yapılabilir aksi durumda farklılık söz konusu olmaz (Özdamar 2004).

6.2 CHAID Analizinin Algoritması

Chaid Analizi'nin genel algoritmasını şu şekilde izah etmiştir. Bağımlı değişkenin kategori sayısı d ≥ 2 olsun. Analiz edilecek olan belirli bir açıklayıcı değişken c ≥ 2 sayıda kategoriye sahip olsun. Burada amaç, c x d kotenjans tablosunu açıklayıcı değişkenindeki uygun kategorileri birleştirme yolu ile en anlamlı j x d tablosuna indirgemektir. Kavramsal olarak ilk hesaplanacak değer Tj(i) istatistiğidir. Tj(i) j x d tablosunu oluşturmadaki i. metot için, $\chi 2$ istatistiğidir (j:2,3,4,….c; i'nin değişim aralığı açıklayıcı değişkenin tipine bağlıdır). Tj(*)=maxTj(i) ise en iyi j x d tablo için , $\chi 2$ istatistiği elde edilmiş olur. Yani, en önemli Tj(i) seçilir. Monotonik ya da dichotomous serbest açıklayıcı değişkenin varlığında Tj(i) Fisher metoduna göre

bulunabilir. Bu dinamik program c2 hesaplarına dayanır. $d \geq 3$ ve açıklayıcı değişken sıralı kategorilere sahip değilse Fisher metodundan yararlanılamaz (Kass,1980).

Algoritma 3 adımda gerçekleştirilir; birleştirme, dağıtma ve dondurma

6.2.1 Birleştirme

1.Adım: Her bir açıklayıcı değişken için sırasıyla, açıklayıcı değişkenin kategorileri ile bağımlı değişkenin kategorilerinin çapraz tablosu bulunur ve adım 2 ile 3 uygulanır.

2. Adım: Sadece açıklayıcı değişkenin tipi tarafından belirlenen uygun çiftler göz önüne alınarak, 2 x d alt tablosunda anlamlılığı düşük olan açıklayıcı değişken kategori çiftleri bulunur. Eğer önem derecesi kritik bir değere ulaşmıyorsa, bu iki kategori birleştirilir. Bu birleşim tek bir kategori olarak ele alınır ve bu adım tekrarlanır. Bu işlem açıklayıcı değişkenin kendi içindeki birleşmeleri anlamsız oluncaya kadar devam eder.

3.Adım: Açıklayıcı değişkenin tipi tarafından oluşturulan ve orijinal kategorilerin 3 veya daha fazlasının birleştirilmesi ile meydana gelen; her bir bileşik kategori için, birleşmenin tekrar ayrılabileceği en önemli ikili bölünme bulunur. Eğer önem derecesi kritik değerin üzerindeyse, bölünme gerçekleştirilir ve 2. adıma dönülür.

6.2.2 Dağıtma

4.Adım: Optimal bir şekilde birleştirilmiş olan, her bir açıklayıcı değişken için önem derecesi hesaplanarak, en büyük önem derecesine sahip olan, diğerlerinden ayrılır. Eğer bu önem derecesi, verilen kriterlerin değerlerinden büyük ise, veri kümesi seçilen açıklayıcı değişkenin birleştirilmiş kategorilerine göre alt gruplarına bölünür.

6.2.3 Durdurma

5.Adım: Verinin analiz edilememiş her bir grubu için, 1. adıma dönülür.

6.3 Açıklayıcı Değişkenlerin Belirlenmesinin Önemi

Algoritmanın 4. Adımında, önceki adımlarla birleştirilmiş olan tabloların dağıtıma işlemi sonucu elde edilen kontenjans tablolarının güvenilirliği test edilmelidir. Testler sonucunda eldeki kontenjans tablolarında daha fazla dağıtma işlemi gerçekleştirilemiyorsa χ^2 testi kullanılabilir. Bu test bağımsız değişkenin kategori sayısına bağlıdır. Sonuçlardan emin olunmadığı sürece veya olasılık tablosu indirgenmemiş ise Bonferroni sonuçlarının kullanılması tercih edilir (Pehlivan,2006).

CHAID analizinde kullanılan üç tip açıklayıcı değişken bulunmaktadır. Bu değişkenlerin Bonferroni çarpanlarının formülleri aşağıdaki gibidir.

6.3.1 Monoton açıklayıcı değişkenler

Monoton açıklayıcı değişkenler, kategorileri sıra belirten bir ölçek üzerinde olan açıklayıcı değişkenlerdir. Bu da bitişik kategorilerin sadece birbiri ile gruplandırılabileceği anlamına gelir. Bonferroni çarpanı, kolayca elde edilen binomial çarpandır.

$$B_{\text{monotonik}} = \binom{c-1}{r-1}$$

Yas değişkeni ile eğitim düzeyi açıklayıcı değişkenler olsun, eğitim düzeyi için, bütün kategori düzeyleri ilkokuldan başlayarak sırası ile ortaokul, lise ve üniversite seklinde sıralanarak devam ediyor ise monoton açıklayıcı değişkene örnek gösterilebilirler.

6.3.2 Serbest açıklayıcı değişkenler

Kategorileri tamamen nominal olan açıklayıcı değişkenlere serbest açıklayıcı değişkenler denir. Bu da kategorilerin her şekilde gruplandırılabileceği anlamına gelir. Aşağıdaki gibi formüle edilir.

$$B_{\text{serbest}} = \sum_{i=0}^{r-1} (-1)^{i} \frac{(r-1)^{c}}{i!(r-i)!}$$

Meslek grupları değişkeni ya da bireyin kimlerle birlikte yaşadığını gösteren açıklayıcı değişkenler bu tip açıklayıcı değişkenlere örnek olabilir.

6.3.3 Hareketli açıklayıcı değişkenler

Hareketli açıklayıcı değişkenler, kategorileri sıralayıcı ölçek ile ölçülmüş olup hem sıraya uymayan hem de sıralayıcı ölçekteki pozisyonun bilinmediği tek bir kategorinin olduğu açıklayıcı değişkenlerdir. Kayıp ya da bilinmeyen bir kategori olması durumunda, bu durum ortaya çıkar. Kayıp bilgi Preece tekniği gibi bazı tekniklerle yeniden kodlanabilir. Preece tekniği, kayıp bilginin yerine muhtemel bir sayı ya da değer atama yoluyla kayıp bilgiyi yeniden kodlar.

Hareketli kategori dışındakiler aynı monoton açıklayıcı değişkende olduğu gibi gruplandırılır. Hareketli kategorilerde gruplama yapılırken, hareketli kategori tek basına bırakılabilir veya başka bir kategori ya da kategori grubuyla birleştirilebilir. Bonferroni çarpanı, monoton durumdaki çarpan genişletilerek, basit bir şekilde elde edilir.

$$B_{\text{hareketli}} = \binom{c-2}{r-2} + \binom{c-2}{r-1} = \frac{r-1+r(c-r)}{c-1} B_{monotonik}$$

Kategorileri, sıra belirten bir ölçek üzerinde olan eğitim düzeyi monoton açıklayıcı değişkeni için, bir eğitim düzeyine ait, örneğin ortaokul mezunu olanlar ile ilgili veriler kayıp ise bu açıklayıcı değişken ise bu açıklayıcı değişken artık hareketli açıklayıcı değişken grubuna girer (Pehlivan 2006).

7. AKADEMİK ERTELEME EĞİLİMİ

Bireyler yaşamlarının her alanında çeşitli sorumluluklar almaktadır. İnsanların yaşamlarını sürdürebilmeleri için toplumun ve insani değerlerin belirlediği sorumluluklar çerçevesinde yaşamaktadır (Çelikkale ve Akbay 2013). Üniversite okuyan öğrencilerinde sorumlulukları artmaktadır. Üniversite eğitimine başlayan bireylerin, akademik sorumlulukları önem arz etmektedir. Hayatının geriye kalan kısmını şekillendirecek olan eğitim süreci sonunda, meslek edinme, toplum içindeki mevkiinin ve hayat standartlarının belirleneceği öğrenciye akademik alanda sorumluluklar yüklemektedir. Bireylerin zamanı doğru kullanamaması sonucu oluşan erteleme eğilimi en önemli sorunlardandır. Bireylerin zaman yönetiminden kaynaklı erteleme eğilimine girmeleri olumsuz sonuçlar doğurmaktadır. Erteleme eğilimi, bireylerin bir görevi yerine getirmeyi, sorumluluk almayı ya da karar vermeyi ilerleyen zamana bırakmasıdır (Kachga vd. 2001). Erteleme eğilimin bireylerin farklı düşünsel ve toplumsal etkilerden kaynaklanmaktadır. En temel nedenlerden biri, bireylerin zaman yönetimini doğru gerçekleştirememesidir (Mccown vd. 1987). Bireylerin sorumluluk duygusunun yetersiz olması, konuya yoğunlaşmada güçlük çekmesi erteleme eğiliminin nedenlerindendir (Balkıs vd. 2006). Kişilerin alması gereken sorumluluklarda sürekli başarısız olma korkusu, kendince geliştirmiş olduğu ancak gerçekçi olmayan performans ve zaman yönetimi stratejileri ve mükemmeliyetçi yaklaşımlar erteleme eğilimini kişiler üzerinde gerçekleştirmektedir (Yaakub 2000). Yapılan araştırmalarda erteleme eğilimi; sorumluluk, başarı güdüsü, bireyin kendine olan güveni, gelecek beklentisi ile ters orantılı; sorumluluktan kaçma, başarısızlık endişesi ve mükemmeliyetçilik ile doğru orantılı bir şekilde gerçekleşmektedir (Özer ve Altun 2011).

Akademik erteleme eğilimi ödevlerin, sınavlara hazırlanmanın, belirlenmiş bir zamanda teslim edilmesi gereken ödevlerin son zamana bırakılması veya yapılmaya çalışılmasıdır (Solomon ve Rothblum 1984). Öğrencilerin yerine getirmesi gereken sorumlulukları yüksek kaygı duyuncaya kadar yerine getirmemesidir. Akademik erteleme eğilimi; bireylerin sürekli veya nadiren görevleri geciktirmesi veya kaygı yaşaması olarak iki aşamalı bir durum gibi incelenmektedir (Rothblum vd. 1986). Anlaşılacağı üzere

erteleme eğilimi kaygı ve stresle paralellik göstermektedir. Erteleme eğilimi kaygıyı tetiklediği gibi, kaygı ve stres altındaki bireylerde de erteleme eğilimi de gerçekleşmektedir. Etkili olmayan öğrenme yöntemleri, not ortalamasının düşük olması, sıkılma, sorumlulukları yerine getirmede zorlanma, plansız çalışma, gerçekçi olmayan mazeretler, kaygı, özgüven eksikliği, kişisel bilgi yetersizliği gibi durumlardan kaynaklı akademik erteleme eğilimi gerçekleşmektedir (Haycock vd. 1998). Motivasyon bireylerin akademik erteleme eğilimini etkilemektedir. Kişisel ve çevresel motivasyonu yüksek olan bireylerde akademik erteleme eğilimin daha düşük olduğu belirlenmiştir (Brownlow ve Reasigner 2000). Yetkinlik inancı akademik erteleme eğilimini azaltan faktörlerden biridir. Yetkinlik inancı bireyin yeteneklerine olan inancı doğrultusunda, farklı şartlar veya ortamlarda yeteneklerini, becerilerini ortaya koyabilmesidir (Bandura 1995). Yerine getirilmesi gereken sorumluluklar noktasında ki bireysel istek ve özverili çabası olarak ta tanımlanmıştır (Bandura 1997). Bireylerin inançları motivasyonlarını sağlamada en önemli etmenlerdendir bu sebeple inançla erteleme eğilimi arasında büyük bir ilişki vardır. Yetkinlik inancı yüksek olan bireylerin erteleme eğilimi daha düşüktür.

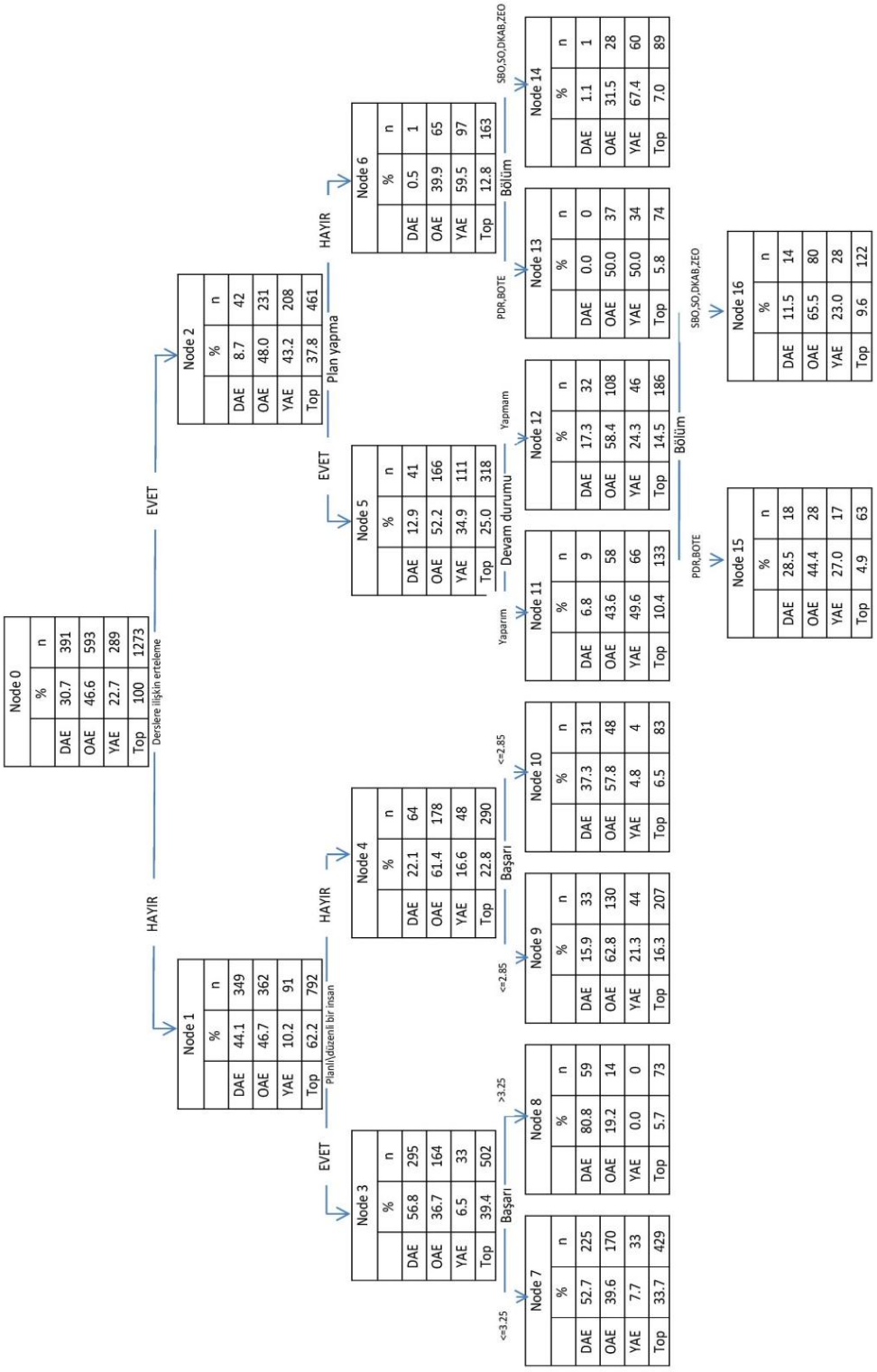
Örneklemdeki bireylere uygulanan ölçeğe ait maddeler puanlanmış ve ölçek maddelerine ilişkin puanlar girilmiştir. Örneklemdeki bireylerin akademik erteleme eğilimlerinin belirlenmesi ve sonucu etkileyen değişkenlerin sınıflandırılması CHAID ve CART Analizi ile yapılmıştır. Örneklemde yer alan öğrencilerin aynı evrenden gelmeme olasılığı üzerine, İki Aşamalı Kümeleme Analizi ile çalışma örneklemi homojen alt sınıflara ayrılmıştır. Toplam puana ait İki Aşamalı Kümeleme Analizi sonuçları çizelge 7.1'de gösterilmiştir.

Çizelge 7.1 Toplam puana ilişkin İki Aşamalı Kümeleme Analizi sonucu

Kümeleme	N	$\overline{X}$	%	SS
Düşük Düzeyde Akademik Erteleme	391	21,94	30,7	3,85
Orta Düzeyde Akademik Erteleme	593	34,53	46,6	4,08
Yüksek Düzeyde Akademik Erteleme	289	47,21	22,7	4,31

Çizelge 7.1'deki veriler ele alındığında, İkinci kümede, madde toplam puan ortalamaları 34.53 ± 4.08 değerinde olan grup yer almaktadır ve bu grupta 593 (% 34.53) birey bulunmaktadır. Elde edilen bu küme eşik değer olarak ele alınacak olup, ölçek puanı eşik değerin üstünde olan bireyler yüksek düzeyde akademik erteleme eğilimine sahip, eşik değerin altında puan alan bireylerin ise düşük düzeyde akademik erteleme eğilimine sahip oldukları söylenebilecektir. Çizelge 7.1'de görüldüğü gibi, düşük düzeyde akademik erteleme eğilimine sahip bireylerin olduğu 1. kümede madde toplam puan ortalamaları 21.94 ± 3.85 değerinde olan grup yer almaktadır ve bu grupta 391 (% 30.7) birey bulunmaktadır. Üçüncü kümede ise yüksek düzeyde akademik erteleme eğilimine sahip olan ve madde toplam puan ortalamaları 47.21 ± 4.31 değerinde olan grup yer almaktadır ve bu grupta 289 (% 22.7) birey bulunmaktadır.

7.1 Akademik Erteleme Eğiliminin CART Yöntemiyle İncelenmesi



Şekil 7.1 Üniversite öğrencilerinin Akademik Erteleme eğilimleri üzerinde etkili olan yor dayıcılara ilişkin ağaç yapısı (Gini algoritmasına göre)

Şekil 7.1'deki ağaç yapısı incelendiğinde, araştırmaya katılan öğrencilerin % 30.7'si düşük düzeyde akademik erteleme eğilimi, % 46.6'sı orta düzeyde akademik erteleme eğilimi ve % 22.7'si yüksek düzeyde akademik erteleme eğilimi göstermektedir. Araştırmaya katılan öğrencilerin akademik erteleme eğilimleri üzerinde başat etkiye sahip olan değişken, öğrencilerin derslere ilişkin erteleme durumları olmuştur. Derslere ilişkin erteleme eğilimi gösteren öğrencilerin % 8.7' si düşük, % 48.0'ı orta ve % 43.2'si yüksek düzeyde akademik erteleme eğilimi gösterirken, derslere ilişkin erteleme eğilimi göstermeyen öğrencilerin % 44.1' i düşük, % 45.7'si orta ve % 10.2'si yüksek düzeyde akademik erteleme eğilimi göstermektedir.

Derslere ilişkin erteleme eğilimi gösteren öğrencilerin, akademik erteleme eğilimleri üzerinde etki düzeyi en yüksek olan değişken öğrencilerin ''plan yapma'' durumlarıdır. Elde edilen sonuçlara göre, araştırmaya katılan öğrencilerden plan yapanların % 12.9'u düşük, % 52.2'si orta ve % 34.9'u yüksek düzeyde akademik erteleme eğilimi gösterirken, plan yapmayanların %0.6'sı düşük, %39.9'u orta ve %59.5'i yüksek düzeyde akademik erteleme eğilimi göstermektedir.

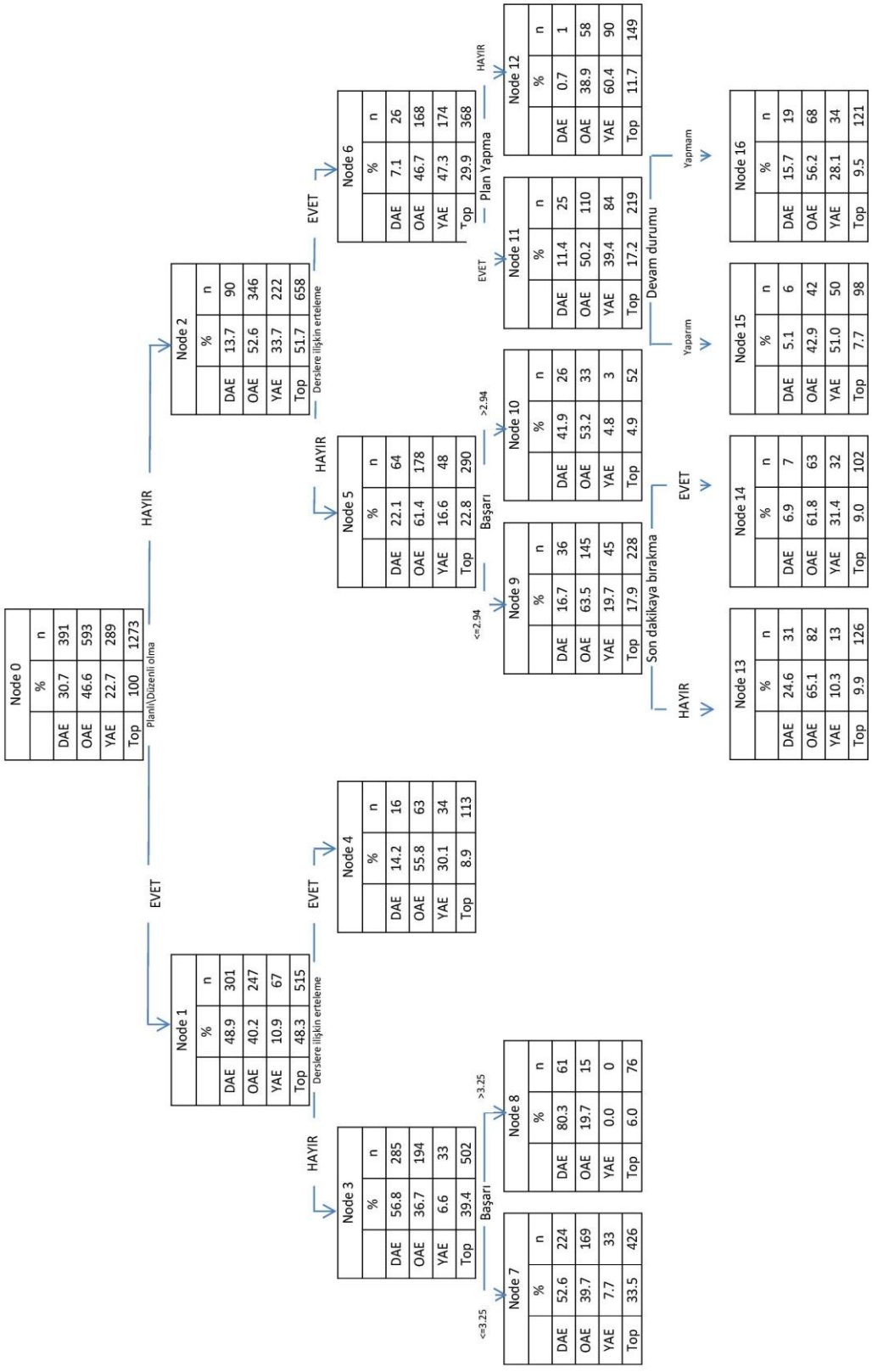
Plan yapmayan öğrencilerin akademik erteleme eğilimleri üzerinde etkisi en yüksek olan değişken, öğrencilerin öğrenim gördükleri bölümleridir. Elde edilen sonuçlara göre; plan yapmayıp PDR veya BOTE bölümlerinde öğrenim gören öğrencilerin % 50'si orta, % 50'si yüksek düzeyde akademik erteleme eğilimi gösterirken, plan yapmayıp SBO, SO, DKAB veya ZEO bölümlerinde öğrenim gören öğrencilerin % 1.1'i düşük, % 31.5'i orta ve % 67.4'ü yüksek düzeyde akademik erteleme eğilimi göstermektedir.

Plan yapan öğrencilerin akademik erteleme eğilimleri üzerinde etkisi en yüksek olan değişken, öğrencilerin derslere devam durumları olmuştur. Elde edilen sonuçlara göre, sevdiği derslerden devamsızlık yapmayan veya devamsızlık hakkını sonuna kadar kullanan öğrencilerin % 6.8'i düşük, % 43.6'sı orta ve %49.6'sı yüksek düzeyde akademik erteleme eğilimi gösterirken, zorunlu olmadıkça devamsızlık yapmayan öğrencilerin %17.3'ü düşük, % 58.4'ü orta ve %24.3'ü yüksek düzeyde akademik erteleme eğilimi göstermektedir.

Zorunlu olmadıkça devamsızlık yapmayan öğrencilerin akademik erteleme eğilimleri üzerinde etkisi en yüksek olan değişken; öğrencilerin öğrenim gördükleri bölümleridir. Elde edilen sonuçlara göre; zorunlu olmadıkça devamsızlık yapmayan PDR, SBO veya BOTE bölümlerinde öğrenim gören öğrencilerin % 28.6'sı düşük, % 44'ü orta, % 27.0'ı yüksek düzeyde akademik erteleme eğilimi gösterirken, zorunlu olmadıkça devamsızlık yapmayan SO, DKAB, OOO veya ZEO bölümlerinde öğrenim gören öğrencilerin % 11.5'i düşük, % 65.6'sı orta ve % 27.0'ı yüksek düzeyde akademik erteleme eğilimi göstermektedir.

Derslere ilişkin erteleme eğilimi göstermeyen öğrencilerin, akademik erteleme eğilimleri üzerinde etki düzeyi en yüksek olan değişken öğrencilerin ‘‘planlı/düzenli bir insan olma’’ durumlarıdır. Araştırma kapsamındaki verilerden elde edilen sonuçlara göre; araştırmaya katılan öğrencilerden, kendilerini planlı/düzenli bir insan olarak görenlerin % 56.8'i düşük, % 36.7'si orta ve % 6.6'sı yüksek düzeyde akademik erteleme eğilimi gösterirken, kendilerini planlı/düzenli bir insan olarak görmeyenlerin % 22.1'i düşük, %61.4'ü orta ve %16.6'sı yüksek düzeyde akademik erteleme eğilimi göstermektedir.

Hem kendilerini planlı/düzenli bir insan olarak görenlerin hem de kendilerini planlı/düzenli bir insan olarak görmeyenlerin akademik erteleme eğilimleri üzerinde etkisi en yüksek olan değişken öğrencilerin ‘‘başarı’’ durumlarıdır. Kendilerini planlı ve düzenli bir insan olarak gören öğrencilerden, genel akademik başarı ortalaması 3.325'e eşit veya düşük olan öğrencilerin % 52.7'si düşük, %39.6'sı orta ve % 7.7'si yüksek düzeyde akademik erteleme eğilimi gösterirken, genel akademik başarı ortalaması 3.325'ten büyük olan öğrencilerin % 80.8'i düşük ve % 19.2'si orta düzeyde akademik erteleme eğilimi göstermektedir.



Şekil 7. 2 Üniversite öğrencilerinin Akademik Erteleme eğilimleri üzerinde etkili olan yor dayıcılara ilişkin ağaç yapısı (Twoing algoritmasına göre)

Şekil 7.2'deki ağaç yapısı incelendiğinde, araştırmaya katılan öğrencilerin % 30.7'si düşük düzeyde akademik erteleme eğilimi, % 46.6'sı orta düzeyde akademik erteleme eğilimi ve % 22.7'si yüksek düzeyde akademik erteleme eğilimi göstermektedir. Araştırmaya katılan öğrencilerin akademik erteleme eğilimleri üzerinde başat etkiye sahip olan değişken, öğrencilerin kendilerini planlı/düzenli bir insan olarak görme durumları olmuştur. Kendilerini planlı/düzenli bir insan olarak görenlerin % 48,9'u düşük, % 40.2'si orta ve % 10.9'u yüksek düzeyde akademik erteleme eğilimi gösterirken, kendilerini planlı/düzenli bir insan olarak görmeyenlerin % 13.7'si düşük, % 52.6'sı orta ve % 33.7'si yüksek düzeyde akademik erteleme eğilimi göstermektedir.

Kendilerini planlı/düzenli bir insan olarak gören öğrencilerin, akademik erteleme eğilimleri üzerinde etki düzeyi en yüksek olan değişken öğrencilerin ''derslere ilişkin erteleme'' durumlarıdır. Elde edilen sonuçlara göre, araştırmaya katılan öğrencilerden derslere ilişkin erteleme eğilimine sahip olanların % 14,2'si düşük, % 55,8'i orta ve % 30,1'i yüksek düzeyde akademik erteleme eğilimi gösterirken, derslere ilişkin erteleme eğilimine sahip olmayanların % 56,8'i düşük, % 36,7'si orta ve % 6,6'sı yüksek düzeyde akademik erteleme eğilimi göstermektedir.

Elde edilen sonuçlara göre; derslere ilişkin erteleme eğilimi göstermeyen öğrencilerin, akademik erteleme eğilimleri üzerinde etki düzeyi en yüksek olan değişken öğrencilerin ''başarı'' durumlarıdır. Genel akademik başarı ortalaması 3.305'e eşit veya düşük olan öğrencilerin % 52,6'sı düşük, % 39,7'si orta ve % 7,7'si yüksek düzeyde akademik erteleme eğilimi gösterirken, genel akademik başarı ortalaması 3.305'ten büyük olan öğrencilerin % 80,3'ü düşük ve % 19,7'si orta düzeyde akademik erteleme eğilimi göstermektedir.

Kendilerini planlı/düzenli bir insan olarak görmeyen öğrencilerin, akademik erteleme eğilimleri üzerinde etki düzeyi en yüksek olan değişken öğrencilerin ''derslere ilişkin erteleme'' durumlarıdır. Elde edilen sonuçlara göre, araştırmaya katılan öğrencilerden derslere ilişkin erteleme eğilimine sahip olmayanların % 22.1'i düşük, % 61.4'ü orta ve % 16.6'sı yüksek düzeyde akademik erteleme eğilimi gösterirken, derslere ilişkin

erteleme eğilimine sahip olanların % 7.1'i düşük, % 45.7'si orta ve % 47.3'ü yüksek düzeyde akademik erteleme eğilimi göstermektedir.

Elde edilen sonuçlara göre; derslere ilişkin erteleme eğilimi gösteren öğrencilerin, akademik erteleme eğilimleri üzerinde etki düzeyi en yüksek olan değişken öğrencilerin ''plan yapma'' durumlarıdır. Plan yapan öğrencilerin % 11.4'ü düşük, % 50.2'si orta ve % 38.4'ü yüksek düzeyde akademik erteleme eğilimi gösterirken, plan yapmayan öğrencilerin % 0.7'si düşük, % 38.9'u orta ve % 60.4'ü yüksek düzeyde akademik erteleme eğilimi göstermektedir.

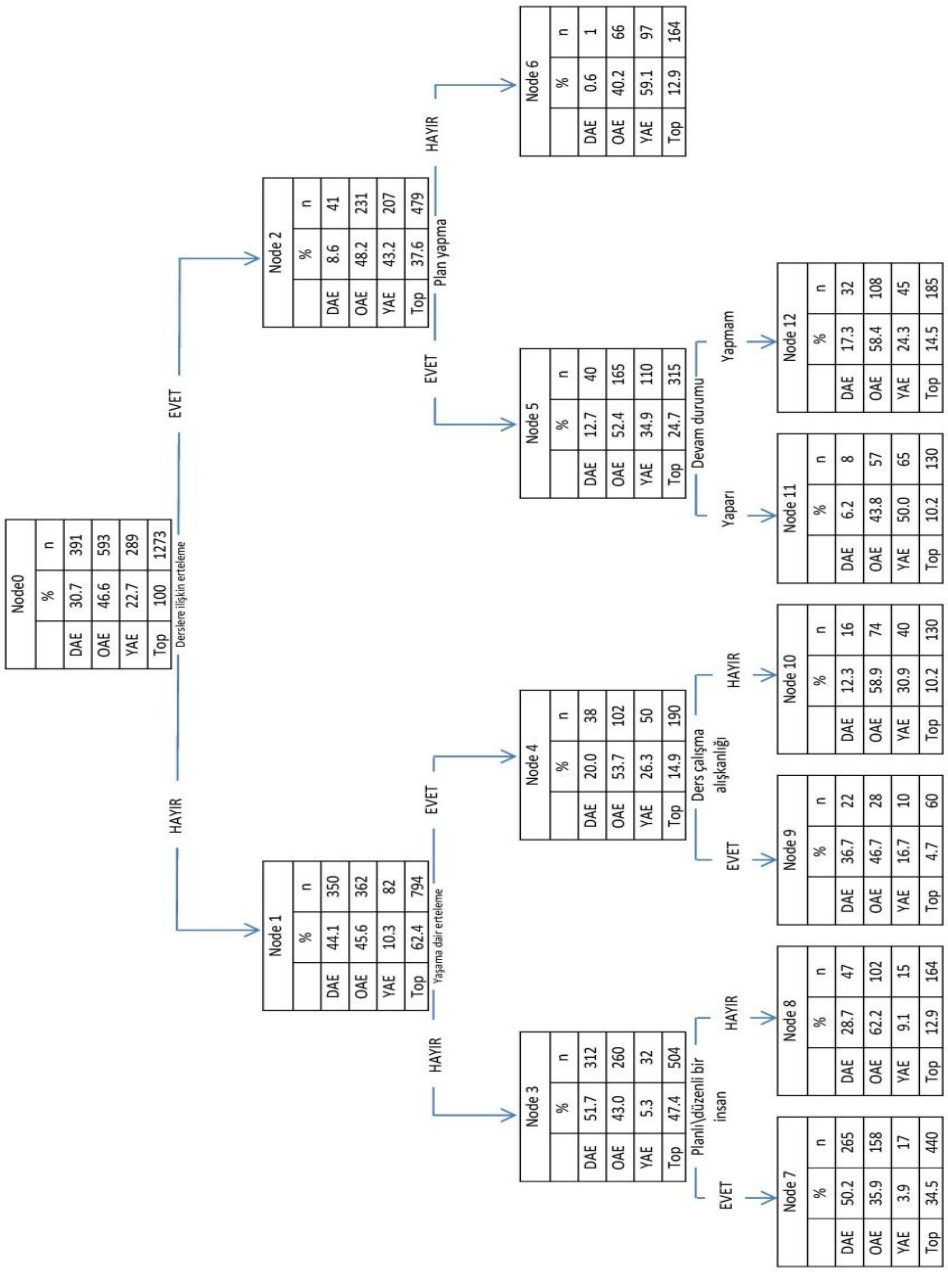
Plan yapan öğrencilerin akademik erteleme eğilimleri üzerinde etkisi en yüksek olan değişken, öğrencilerin derslere devam durumları olmuştur. Elde edilen sonuçlara göre, sevdiği derslerden devamsızlık yapmayan veya devamsızlık hakkını sonuna kadar kullanan öğrencilerin % 6.1'i düşük, % 42.9'u orta ve % 50'si yüksek düzeyde akademik erteleme eğilimi gösterirken, zorunlu olmadıkça devamsızlık yapmayan öğrencilerin %15.7'si düşük, % 56.2'si orta ve % 28.1'i yüksek düzeyde akademik erteleme eğilimi göstermektedir.

Elde edilen sonuçlara göre; derslere ilişkin erteleme eğilimi göstermeyen öğrencilerin, akademik erteleme eğilimleri üzerinde etki düzeyi en yüksek olan değişken öğrencilerin ''başarı'' durumlarıdır. Genel akademik başarı ortalaması 2.94'e eşit veya düşük olan öğrencilerin % 16.7'si düşük, % 63.6'sı orta ve % 19,7'si yüksek düzeyde akademik erteleme eğilimine gösterirken, genel akademik başarı ortalaması 2.94'ten büyük olan öğrencilerin % 41.9'u düşük, % 53.2'si orta ve % 4.8'i yüksek düzeyde akademik erteleme eğilimi göstermektedir.

Genel akademik başarı ortalaması 2.94'e eşit veya düşük olan öğrencilerin akademik erteleme eğilimleri üzerinde en yüksek düzeyde etkiye sahip olan değişken, öğrencilerin ''işlerini son dakikaya bırakma'' durumları olmuştur. Elde edilen sonuçlara göre; öğrencilerden işlerini son dakikaya bırakanların % 6.9'u düşük, % 61.8'i orta ve % 31.4'ü yüksek düzeyde akademik erteleme eğilimine gösterirken, işlerini son dakikaya

bırakmayanların % 24.6'sı düşük, % 65.1'i orta ve % 10.3'ü yüksek düzeyde akademik erteleme eğilimi göstermektedir.

7.2 Akademik Erteleme Eğiliminin CHAID Yöntemiyle İncelenmesi



Şekil 7. 3. Üniversite öğrencilerinin Akademik Erteleme eğilimleri üzerinde etkili olan yor dayıcılara ilişkin ağaç yapısı (Chaid algoritmasına göre)

Ölçek öğrencilere uygulandıktan sonra, her öğrenci için ölçekten elde edilecek toplam puan belirlenmiştir. Elde edilen veriler CHAID Analizi ile incelenirken, öğrencilerin ölçekten aldıkları toplam puan bağımlı değişken, öğrencilerin cinsiyetleri, yaşları, bölümleri, sınıf düzeyleri, okudukları bölümü tercih etme sıraları, genel akademik başarı durumları, öğrenim gördükleri bölüm, öğretmenlik yapma isteği, genel akademik başarılarını değerlendirme durumları, lisansüstü eğitim yapmayı isteme durumları, herhangi bir işte çalışma durumları, derslere devam durumları, sınavlara hazırlanırken bireysel ya da grupla ders çalışmayı tercih etme durumları, gün içinde ders çalışmak için çoğunlukla tercih edilen zaman aralığı, zamanı etkili kullanıp kullanmadıkları hakkındaki görüşleri, herhangi bir işe başlamadan plan yapıp yapmama durumları, öğrenmeyi tanımlama durumları, derslere/akademik görevlere ilişkin sorumlulukları erteleme durumları, ders çalışma alışkanlıklarından memnun olma durumları, günlük yaşamda yapılması gereken işleri erteleme durumları, çevredeki insanların onlar hakkındaki işlerini son dakikaya bırakıp bırakmama durumuna göre görüşleri, kendilerini planlı/düzenli bir insan olarak tanımlayıp tanımlamama durumları, arkadaşlarının okul başarısı üzerindeki etkisini değerlendirme durumları ve barınma durumları da bağımsız değişken olarak regresyon modeline dahil edilip model kurulmuştur.

Şekil 7.3'teki ağaç yapısı incelendiğinde, araştırmaya katılan öğrencilerden 593'ü (%46.6) orta düzeyde erteleme eğilimine, 391'i (%30.7) düşük akademik erteleme eğilimine ve 289'u (%22.7) da yüksek akademik erteleme eğilimine sahiptirler. Buna göre üniversite öğrencilerinin akademik erteleme eğilimleri üzerinde en önemli etkiye sahip olan değişkenin "derslere ilişkin erteleme durumları" olduğu görülmektedir ($\chi 2=265.514$; $p=0.000$). Derslere ve akademik görevlere ait işleri erteleyen bireylerin %8.6'sı düşük, %48.2'si orta ve %43.2'si de yüksek akademik erteleme eğilimine sahipken, derslere ve akademik görevlere ait işleri ertelemeyen bireylerin %44.1'i düşük, %45.6'sı orta ve %10.3'ü de yüksek akademik erteleme eğilimine sahiptir.

Derslere ilişkin işleri erteleyen bireylerin akademik erteleme eğilimleri üzerinde etki düzeyi en yüksek olan değişken "plan yapma durumu" olmuştur ($\chi 2=36.354$; $p=0.000$). Yapacağı işlere yönelik önceden plan yapmayan bireylerin %0.6'sı düşük, %40.2'si orta

ve %59.1'i de yüksek akademik erteleme eğilimine sahipken, yapacağı işlere yönelik önceden plan yapan bireylerin ise %12.7'si düşük, %52.4'ü orta ve %34.9'u da yüksek akademik erteleme eğilimine sahiptir.

Derslere ait işlerini ertelemeyen bireylerin akademik erteleme eğilimleri üzerinde başat etkiye sahip olan değişken "yaşama dair işleri erteleme durumu" olmuştur ($\chi 2$=98.267; p=0.000). Buna göre yaşama dair işlerini erteleyen bireylerin %20'si düşük, %53.7'si orta ve %26.3'ü de yüksek akademik erteleme eğilimine sahipken, yaşama dair işlerini ertelemeyen bireylerin ise %51.7'si düşük, %43'ü orta ve %5.3'ü de yüksek düzeyde akademik erteleme eğilimine sahiptir.

Yapacağı işlere yönelik önceden plan yapan bireylerin akademik erteleme eğilimleri üzerinde en önemli etkiye sahip olan değişken "derslere devam durumu" dur ($\chi 2$=24.958; p=0.000). Bu değişken bağımlı değişkeni iki kategori halinde etkilemiştir. Buna göre sevdiği derslerde devamsızlık yapmayan bireyler ile devamsızlık hakkını sonuna kadar kullanan bireyler benzer karakter sergilemiştir. Bu bireylerin, %6.2'si düşük, %43.8'i orta ve %50'si de yüksek akademik erteleme eğilimine sahiptir. Zorunlu olmadıkça devamsızlık yapmam diyen bireylerin ise %17.3'ü düşük, %58.4'ü orta ve %24.3'ü de yüksek akademik erteleme eğilimine sahiptir.

Yaşama dair işlerini erteleyen bireylerin akademik erteleme eğilimleri üzerinde başat etki gösteren değişken "öğrencilerin ders çalışma alışkanlıkları" olmuştur ($\chi 2$=16.086; p=0.001). Buna göre ders çalışma alışkanlığından memnun olan bireylerin %36.7'si düşük, %46.7'si orta ve %16.7'si de yüksek akademik erteleme eğilimine sahiptir. Ders çalışma alışkanlığından memnun olmayan bireylerin ise %12.3'ü düşük, %56.9'u orta ve %30.8'i de yüksek akademik erteleme eğilimine sahiptir.

Yaşama dair işlerini ertelemeyen bireylerin akademik erteleme eğilimleri üzerinde başat etki gösteren değişken "kendilerini planlı/düzenli bir insan olarak tanımlama durumları" olarak bulunmuştur ($\chi 2$=48.519; p=0.000). Kendilerini planlı/düzenli bir insan olarak görmeyen üniversite öğrencilerinin %28.7'si düşük, %62.2'si orta ve %9.1'i de yüksek

akademik erteleme eğilimine sahiptir. Kendilerini planlı/düzenli bir insan olarak gören üniversite öğrencilerinin de %60.2'si düşük, %35.9'u orta ve %3.9'u da yüksek akademik erteleme eğilimine sahiptir.

7.3 CART ve CHAID Yöntemlerinin Karşılaştırılması

Chaid algoritması 1980 yılında Kass tarafından geliştirilmiş bir algoritmadır. Böl yönet ve sonlandır olmak üzere üç adımda gerçekleşen bir algoritmadır. Oluşturulan ağaç başlangıç adımından itibaren adımlar ilerledikçe büyümeye devam eder ta ki ağaç yapısı sonlandırma şartlarını yerine getirene dek devam eder. CART algoritması bir üst dala oranla daha saf bir olay oluşturması için tasarlanmış bir algoritmadır. CART'ta bölme kriteri değişkenler arasında ki ilişkiye göre değişkenlik göstermektedir. Kategorik değişkenler için sınıflama ağaçları, sürekli değişkenler için regresyon ağaçları kullanılır. Sürekli değişkenler için CHAID ve CART algoritmaları kullanılabilir. CART ikili ağaç yapısı kullanırken CHAID ikili ağaç yapısı kullanmaz bu yüzden oluşan ağaç CHAID'te daha geniş olmaktadır. Bu yüzden yorumlama CHAID daha büyük kolaylık sağlamaktadır (Ruimin vd. 2010).

CART sınıflandırma yönteminde iki farklı ayırma yöntemi kullanılmıştır. Gini ve Twoing ayırma kurallarına göre oluşturulan ağaç yapılarında ayırma kriteri farklılığından kaynaklı farklı sonuçlar oluşmaktadır. Akademik erteleme eğilimleri düzeyi toplam puanda farklılık göstermemekle birlikte düzeyi etkileyen bağımsız değişkenlerin etki oranlarında farklılıklar tespit edilmiştir.

Gini ayırma metoduna göre, akademik erteleme eğilimi üzerindeki en büyük etkili değişken derslere ilişkin erteleme durumları olmuştur. İkici önemli etmen planlı ve düzenli olma durumudur. Twoing ayırma kuralına göre ise planlı ve düzenli olma durumu etkili olmuştur. İkinci önemli değişken derslere olan erteleme eğilimi olarak gerçekleşmiştir. CART algoritmasına göre akademik erteleme eğilimini etkileyen en önemli nedenler derslere olan erteleme eğilimi ve planlı düzenli olma durumlarıdır. Farklı ayırma kurallarında yakın sonuçlar oluşmasına rağmen şekil 7.1 ve şekil 7.2'de görüldüğü gibi yüzde 1 ile 5 arasında farklı sonuçlar gerçekleşmiştir.

Sınıflama ve regresyon ağaçlarında, veri setinin büyüklüğü daha doğru sonuçlar üretmektedir. örnek sayısı artıkça CART, CHAID' e göre daha doğru sonuçlar doğurmaktadır. Çizelge 7.2'de görüldüğü gibi Örnek sayısı artıkça hata oranı CART lehine değişmektedir(Ruimin vd. 2010).

Çizelge 7. 2 Cart ve Chaid algoritmalarının hata oranları (Ruimin vd. 2010)

Örnek Sayısı	Ortalama Hata Oranı %	
3590	CART	28.83
	CHAID	30.45
228	CART	32.57
	CHAID	31.76
454	CART	32.22
	CHAID	29.34
334	CART	32.64
	CHAID	31.25
3394	CART	28.41
	CHAID	31.24

Yapılan analizler sonucu elde edilen ağaç yapılarında CART'ın daha fazla düğümden oluşan ağaç yapısı oluşturmaktadır. Safsızlık ölçüsü daha düşük bir sonuç çıkardığından kaynaklı derinliği daha fazla oluşmaktadır (Wang vd. 2009).

Terminal düğüm, homojenliği en yüksek olan veya ağacın daha fazla bölünmeyen düğümüne denir. CART algoritması maksimum homojenliğe ulaşmak için gerekli bölme yöntemini kendisi hesapladığından daha fazla terminal düğüme sahip olur (Haughtan ve Oulabi 1997).

Çizelge 7.3 Algoritmalara ait bulguların hata ve estimate değerlerinin karşılaştırılması

	Estimate	Std. Eror	Doğruluk Oranı
CART	0.42340926944226240	0.013848406625262955	0.986151593
CHAID	0.41948153967007070	0.013830893660179621	0.986169106

Yapılan çalışmada elde edilen veriler Çizelge 7.3 görülmektedir. CART ve CHAID algoritmalarında tahmin etme oranları, hata oranları, doğruluk oranları tespit edilmiştir. Doğruluk oranı = 1 – Hata oranı' dır. Çizelgeye göre CART ile CHAID arasında tahmin yürütmede ve doğruluk oranında CART'ın daha avantajlı olduğu belirlenmiştir. tahmin etmede CART'ın yüzde 1 daha avantajlı olduğu, hata oranın da binde 1 daha iyi sonuçla analiz gerçekleştirdiği tespit edilmiştir.

Çizelge 7.4 Algoritmalara ait bulguların düğüm (node) sayısı derinlik (depth) ve terminal düğüm sayısı değerlerinin karşılaştırılması

	Düğüm (node)	Derinlik (depth)	Terminal düğüm (terminal node)
Chart	17	4	9
Chaid	13	3	7

Akademik erteleme eğilimi verileri ile elde edilen sonuçlar tablodaki gibi gerçekleşmiştir. Literatüre paralel olarak çalışmamızda cart'ın daha fazla düğüm oluşturduğu ve derinliğinin daha fazla olduğu görülmektedir. Homojenliğin Cart algoritması için daha detaylı olduğu daha önceki örneklerde açıklanmıştı. Çizelge 7.4'de görüldüğü gibi terminal düğüm sayısı CART'ta daha fazladır.

8. SONUÇ

Bu çalışmada, veri madenciliği yöntemlerinden CART ve CHAID algoritmaları uygulamalı olarak karşılaştırılmıştır. Öğrencilerin akademik erteleme eğilimleri incelenmiş sonuçlar iki farklı algoritmaya göre yorumlanmıştır.

Teknolojinin hızla gelişmesi ile veriler de artmaktadır. Verilerden anlamlı sonuçlar elde etmek için veri madenciliği yöntemleri sıkça kullanılmaktadır. Sosyal mühendislik, reklam verme, istatistik vb. gibi popüler alanlarda kullanılmaktadır. Veri madenciliği yöntemleri, kullanım alanları ve veri setine uygun yöntemler belirleme yöntemleri belirtilmiştir.

Toplanan verilerden anlamlı sonuçlara ulaşmak için gerekli adımlar gerçekleştirildikten sonra CART ve CHAID yöntemleri uygulanmıştır. CART algoritması Gini ve Twoing gibi iki ayırma kuralı kullanılarak ayrı ayrı çözümlenmiş, veri setine uygun yöntem ve ayırma kuralı tercih edilmesini belirleme kriterleri yedinci bölümde tartışılmıştır. Öğrencilerin akademik erteleme eğilimlerini etkileyen değişkenler incelenmiştir.

Yapılan çalışma ile eğitim bilimleri için öğrencilerin akademik erteleme eğilimi tutumu konusunda bilgiler analiz sonuçları ile tespit edilmiştir. Eğilimi engelleme konusunda yapılması gerekenler için ön çalışma niteliği taşımaktadır. Analizde yer alan her bir değişkenin etki oranı SPSS programı yardımı ile yorumlanmıştır. Günümüzde sıkça kullanılan CART ve CHAID algoritmaları karşılaştırılmış ve kategorik değişkenler için CART'ın daha doğru sonuçlar çıkardığı literatüre paralel olarak tespit edilmiştir.

Yapılan çalışmada, kullanılacak veri madenciliği yöntemi önemli olduğu kadar kullanılacak ayırma yönteminde önemli olduğu belirlenmiş sonuçlarda ve analiz gerçekleştirmede hata oranını düşürdüğü tespit edilmiştir.

KAYNAKLAR

Adrians, P. and Zantinge, D. 1997. Data Mining. Addison- Wesley, İngiltere

Akpınar, H. 2000. Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği. İ.Ü.İşletme Fakültesi Dergisi, Sayı:1, s. 1-22, İstanbul

Alpaydın, E. 2000. Zeki Veri Madenciliği: Ham Veriden Altın Bilgiye Ulaşma Yöntemleri. Bilişim 2000 Eğitim Semineri

Anonymous. 2012. Web Sitesi: spss.com.tr Erişim Tarihi: 05.08.2014.

Balkıs M., Duru, E., Bulus, M. ve Duru, S. 2006. Üniversite Öğrencilerinde Akademik Erteleme Eğiliminin Çeşitli Değişkenler Açısından İncelenmesi, Ege Eğitim Dergisi, cilt: 7, Sayı: 2, ss: 57-73

Bandura, A. 1995. Self-efficacy in changing societies. In A. Bandura (ed.) Exercise of personal and collective efficacy in changing societies, Cambridge University Pres. pp. 1–45, New York

Bandura, A. 1997. Self-efficacy: The exercise of control. New York: W. H. Freeman and Company.

Baykal, N. 2003. Veri tabanı ve Veri Madenciliği, Tıp Bilişimi Güz Okulu, s.61

Berry, M.J.A. and Linoff, G. 1998. Data Minig Solutions. Wiley Computer Publishing, 1998, p. 25, ABD

Berry, M.J.A. and Linoff, G. 2000. The Art and Science of Customer Relationship Management. Wiley Computer Publishing, pp. 7-8, ABD

Berry, M.J.A. and Linoff, G. 2000. Mastering Data Mining. Wiley&Sons, p. 146, ABD

Bevilacqua, M. Braglia, M. and Montanari, R. 2003. The classification and regression tree approach to pump failure rate analysis. Reliability Engineering and System Safety, pp. 59-67

Bounsaythip, C. and Rinta Runsala, E. 2001. Overview of Data Mining for Customer Behaviour Modeling. VTT Information Technology Research Report TTEI, p.18

Breimen, L. Freidman, J. H. Olshen, R.A. and Stone, C.J. 1993. Classification and regression trees. Chapman&Hall, pp. 32-104.

Bremner, A.P. and Taplin, R. 2003. Modified classification and regression tree splitting criteria for data with interactions. Australian Statistical & Data Analysis, pp.1-11

Brownlow, S. and Reasinger, R.D. 2000. Putting off until tomorrow what is better done today: Academic procrastination as a function of motivation toward college work. Journal of Social Behavior and Personality, 15(5), pp. 15-34.

Bryan B. 2002. Bioinformatics Computing. Prentice Hall PTR, s. 120, ABD
Chipman, H.A. George, E.I. and McCulloch, R.E. 2000. Hierarchical priorsforbayesian CART shrinkage. Statisticand Computing, pp. 17-24

Chye, K.H. and Gerry, C.K.L. Data Mining and Customer Relationship Management in the Banking Industry. Singapore Management Review, Volume 24, Issue 2, pp. 1-27

Çelikkale, Ö. ve Akbay, S.E. 2013. Üniversite Öğrencilerinin Akademik Erteleme Davranışı, Genel Yetkinlik İnancı ve Sorumluluklarının İncelenmesi, Ahi Evran Üniversitesi Kırşehir Eğitim Fakültesi Dergisi, Cilt: 14, Sayı: 2, ss. 237-254

De'ath, G. and Fabricius, K. 2000. Classification and regressiontrees: a powerfull yet simple technique for eqological data analysis. Ecology, p. 81

Doğan, N. ve Özdamar, K. 2003. Chaid analizi ve aile planlaması ile ilgili bir uygulama. Temel Klinik Tıp Bilimleri, 23, ss. 392-397

Öztemel, E. 2003. Yapay Sinir Ağları, İstanbul, Papatya Yayıncılık, ss. 17-22

Fu C. 2003. Combingloglinear model withCARTanalysis. An application to birth data. Computational and Statistic&Data Analysis, pp. 1-11

Giovinazzo, W.A. 2002. Internet Enabled Business Intelligence. Prentice Hall PTR, pp. 330-333, ABD

Gregory, P.S. 1989. Knowledge Discovery in Real Databases: A Report on the IJCAI89 Workshop. Al Magazine, 11(5), pp. 68-70

Hair, J.F.R. Anderson, R.L and Black, W.C. 1998. Multivariate Data Analysis, ABD, Prentice Hall, pp. 13-17

Hand, D. 1999. Statistics and Data Mining: Intersecting Disciplines. ACM SIGKDD, Volume 1, Issue 1, pp. 16-19

Haycock, L.A., McCarthy, P. and Skay, C.L. 1998. Procrastination in college students: The role of self-efficacy and anxiety. Journal of Counseling and Development, 76, pp. 317-324.

Helberg, C. 2002. Data Mining with Confidence. SPSS, 2.nd edition, inc., Chigago

Irmak, S., Köksal, C.D. ve Asilkan, Ö. 2012. Hastanelerin Gelecekteki Hasta Yoğunluklarının Veri Madenciliği Yöntemleri ile Tahmin Edilmesi, Uluslar arası Alanya İşletme Dergisi, cilt: 4, sayı: 1, ss.101-114

Jiawei, H. and Micheline, K. 2001. Data Mining Concepts and Techniques. Morgan Kaufman Publishers, p.106, ABD

Kachgal, M.M., Hansen, L.S. and Nutter, K.J. 2001. Academic procrastination prevention/intervention: Strategies and recommendations. Journal of Developmental Education, 25, pp. 14-21.

Koç, S. 2005. Türkiye'de İllerin Sosyo-Ekonomik Özelliklere göre Sınıflandırılması. Çukurova Üniversitesi Sempozyum Bildirisi, 27.06.2005

Koyuncugil, A.S. ve Özgülbaş, N. 2009. Veri Madenciliği: Tıp ve Sağlık Hizmetlerinde kullanımı ve uygulamaları, Bilişim teknolojileri dergisi, cilt:2, sayı: 2

Lewis r. 2000. An introduction to classification and regression tree analysis Academic Emergency Medicine. pp. 1-14, California

Li R., Zhao X., YU X., Li J., Cheng N. and Zhang J. 2010. Incident Duration Model on Urban Freeways Using Three Different Algorithms of Decision Tree" 2010 International Conference on Intelligent Computation Technology and Automation

Marian, M. Nestler, I. and Bernd H. 2002. GIS-based regionalization of soil profiles with classification and regression trees. J Plant Nutr. Soil. Sci,

McCown, W., Petzel, T. and Rupert, P. 1987. An experimental study of some hypothesized behaviors and personality variables of college student procrastinators. Personality and Individual Differences, 8(6) , pp. 781-786.

Moss, L.T. 2003. Business Intelligence Roadmap: The Complete Project Lifecycle for Decision- Support Applications, Addison Wesley, pp. 307-310

Orhunbilge, N. 1996. Uygulamalı Regresyon ve Korelasyon Analizi, Universitesi İsletme Fakultesi Yayını, Yayın No:267, İstanbul

Özdamar, K. 2004. Paket Programlar ile İstatistiksel Veri Analizi 2, Kaan Kitapevi, ss. 280-281, Eskişehir

Özer, A. ve Altun, E. 2011. Üniversite Öğrencilerinin Akademik Erteleme Nedenleri, Mehmet Akif Ersoy Üniversitesi Eğitim Fakültesi Dergisi, Sayı: 21, ss. 45-72

Özkan Y. 2013. Veri madenciliği yöntemleri, Papatya Yayıncılık, ss. 36-73, Ankara

Özmen, Ş. 2003. Veri Madenciliği Süreci. Veri Madenciliği ve Uygulama Alanları Konferansı, İstanbul Ticaret Üniversitesi, 2003, İstanbul

Özmen, Ş. 2004. İş Hayatı Veri Madenciliği ile İstatistik Uygulamalarını Yeniden Keşfediyor. Marmara Üniversitesi İ.İB.F, (Çevrimiçi), http://idari.cu.edu.tr/sempozyum/bil38.htm, 04/05/2014

Pehlivan, Z. 2006. Resmi Genel Liselerde Öğrenci Devamsızlığı ve Buna Dönük Okul Yönetimi Politikaları

Rothblum, E.D., Solomon, L.J. and Murakami, J. 1986. Affective, cognitive, and behavioral differences between high and low procrastinators. Journal of Counseling Psychology, 33(3), pp. 387-394.

Show-Jane, Y. and Yue-Shi, L. 2006. An Efficient Data Mining Approach for Discovering Interesting Knowledge from Customer Transactions. Expert Systems with Applications, Volume 30, Issue 4, pp. 650-657.

Speybroeck, N. Berkvens, D. Mfoukuo-Ntsakala, A. Aert, M. Hens, N. Huylenbroeck, G. and Thys E. 2003. Classificationtreesversusmultinomialmodels in theanalysis of urban farmingsystems in centralAfrica. Agricultural Systems, pp. 1-17

Umman Tuğba Ş. 2004. Veri Madenciliği Ve Müşteri İlişkileri Yönetiminde Bir Uygulama, Doktora Tezi, İstanbul Üniversitesi, Sosyal Bilimler Enstitüsü, İşletme Anabilim Dalı, İstanbul

Yaakub, N.F. 2000. Procrastination Among Students in Institutes of Higher Learning: Challenges for K-Economy. The School of Languages and Scientific Thinking, University of Utara, Malaysia (01 Aralık 2014) http://mahdzan.com/papers/procrastinate/

Yarımağan, Ü. 2000. Veritabanı Sistemleri, Ankara, Akademi & Türkiye Bilişim Vakfı Yayını, s.1

ÖZGEÇMİŞ

Adı Soyadı : Yasin İNAĞ

Doğum Yeri : Batman

Doğum Tarihi : 29/10/1986

Medeni Hali : Bekar

Yabancı Dili : İngilizce

Eğitim Durumu (Kurum ve Yıl)

Lise : Batman Fen Lisesi(2005)

Lisans : Eskişehir Osmangazi Üniversitesi Mühendislik Mimarlık Fakültesi Bilgisayar Mühendisliği Bölümü(2011)

Yüksek Lisans: Ankara Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı (Eylül 2011 – Ocak 2015)

Çalıştığı Kurum/Kurumlar ve Yıl

Muş Alparslan Üniversitesi, Mühendislik Mimarlık Fakültesi, Bilgisayar Mühendisliği Yazılım Anabilim Dalı (2011-)