MBD 2012, 1(1): 18 – 37

MESLEKİ BİLİMLER DERGİSİ
ISSN: 2146 - 7420
www.meslekibilimler.com

MAKALE HAKKINDA

Geliş : Ocak 2012

Kabul: Mart 2012

VERİ MADENCİLİĞİ YÖNTEMİ İLE DİVRİĞİ NURİ DEMİRAĞ MESLEK YÜKSEKOKULU ÖĞRENCİLERİNİN TEMEL BİLGİSAYAR DERSİNE AİT BAŞARI ANALİZİ UYGULAMASI

THE SUCCESS ANALYSIS OF BASIC COMPUTER COURSE OF DIVRIGI NURI DEMIRAG VOCATIONAL COLLEGE'S STUDENTS BY USING DATA MINING METHOD

Namık İçeli[a]

ÖZ

Günümüzde eğitim, sigortacılık, pazarlama, elektronik ticaret, bankacılık, sağlık gibi farklı alanlarda kullanılmaya başlanan veri madenciliği yöntemlerinden yararlanarak, öğrenciler üzerinden anket yoluyla elde edilen ham verilerin anlamlı bilgilere dönüştürülmesi bu makalenin amacını oluşturmaktadır. Bu amaçla Cumhuriyet Üniversitesi Divriği Nuri Demirağ Meslek Yüksekokulunda 2008 - 2010 eğitim-öğretim dönemleri arasında eğitim gören Temel Bilgisayar Bilimleri dersini almış 102 öğrenciye güvenirliği SPSS paket programı ile analiz edilmiş bir anket formu uygulanmıştır. Anket sonucunda elde edilen veriler ile öğrencilerin yılsonunda o dersten geçme/kalma başarı durumları veri madenciliği paket programı olan WEKA'ya girilerek J48 adında Entropiye dayalı Sınıflandırma algoritmasından geçirilmiş ve sonuç olarak karar ağacı şeklinde anlamlı bilgiler elde edilmiştir. Elde edilen bilgiler doğrultusunda gelecekte Temel Bilgisayar Bilimleri dersini alacak diğer öğrencilerin(25 öğrenci) sadece ankete verecekleri cevaplar ile ders başarı durumlarının tahmin edilmesi uygulanmıştır.

Anahtar Kelimeler: Veri Madenciliği, Karar Ağacı, Öğrenci, Başarı, Algoritma

ABSTRACT

In this paper, it is aimed that the data, obtained from students' survey, were transformed to meaningful information by using data mining methods, which are using different areas such as, education, marketing, electronic commerce, banking, health care industry and insurance trade. In order to this purpose, the questionnaire was applied to 102 basic computer course's students who are studying at Cumhuriyet University Divrigi Nuri Demirag Vocational College between the years of 2008 and 2010. And data were analyzed by using SPSS program. Results of questionnaire, which pass/fail situations of students, were entered to WEKA program with the name of J48 Entropy algorithm and meaningful information were obtained as decision tree. The success conditions of students, who will take this course in future, were predicted according to their answers to this questionnaire.

Keywords: Data Mining, Decision tree, Student, Success, Algorithm

[a]Öğr. Gör., Cumhuriyet Üniversitesi, Divriği Nuri Demirağ MYO, niceli@cumhuriyet.edu.tr

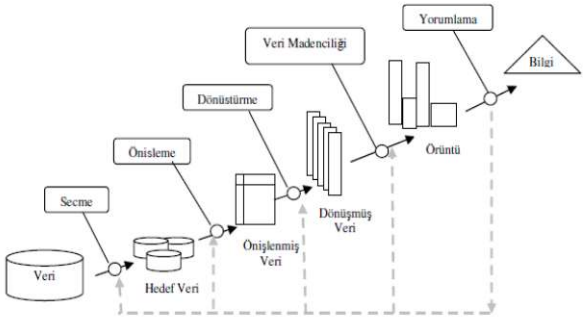
GİRİŞ

Dijital teknolojilerinin gelişmeleri sonucunda yüksek kapasiteli işlem yapabilme gücü ucuzlamış, depolama ünitelerinin veri saklama kapasiteleri hem çok daha gelişmiş hem de daha kolay hale gelmiştir. Bu durum verinin ucuzlamasına neden olmuştur. Uzun olmayan bir zaman öncesine kadar karar vericilerin, yöneticilerin karşılaştığı temel problemlerden biri olarak görülen veri kıtlığı, yerini aşırı bolluğa bırakmıştır. Bilgiye erişim endişesinin yerini artık erişilebilinen miktarla başa çıkma endişesi almıştır(Fayyad ve Symth, 1997).

Yıllar önce verilerin sayılarının az olması nedeniyle Veritabanları verileri saklamanın en uygun yolu iken zamanla verilerin artması ve bu verilerin organizasyonlarında sorunların oluşması nedeniyle artık veritabanları yerine Veri Ambarı kavramı ortaya çıkmıştır.

Veritabanlarında, veri ambarlarında depolanan verilerde gizli bulunan öz bilgiyi keşfedebilmek amacıyla insanlara yardımcı olacak yeni nesil hesaplama tekniklerine ve araçlarına ihtiyaç duyulmaktadır. Bu teknikler ve araçlar, veriyi anlamlı hale getirmek amacıyla yapılan değişik faaliyetlerin bütünü olarak tanımlanabilen Veritabanlarında Öz bilgi Keşfi (VTÖK)'nin konusudur(Jeffry, 2004).

Veri Madenciliği (VM) ise aşağıdaki şekilde de görüldüğü gibi Veritabanı Bilgi Keşfi(VTBK) değil VTBK sürecinin bir adımıdır.



Şekil 1. Veri tabanlarında bilgi keşfi süreci(Fayyad vd., 1996)

Veriden bilgiye ulaşma sürecindeki bu adımlarda;

- Veri Seçimi: Birden fazla veri kümesi içerisinden, üzerinde sorgu yapılmasına uygun örnek bir veri kümesi oluşturma aşamasıdır. Veri toplama ve farklı kümelerdeki verilerin birleştirilmesi işlemleri de bu aşamada gerçekleştirilir.
- Veri Önişleme: Veri seçimi ile elde edilen örnek veri kümesinde yer alan hatalı ve eksik değerlerin düzenlendiği ve çıkarıldığı aşamadır. Veri temizleme ve veri dönüştürme veri önişleme işlemlerindendir. Veri temizleme ile ilgisiz ve gürültülü veriler veri setinden çıkarılarak verilerin güvenilir olması sağlanır.
- Veri Dönüştürme: Kaynak veri seti içindeki farklı tip verileri ortak bir veri tipine dönüştürme sağlanmaktadır. Ayrıca veri setindeki bazı verilerin eksik olması durumunda da eksik olan yerdeki veriler yerine ortalama değer alınarak bu eksiklikler tamamlanabilir. Veri indirgeme ile seçilen örnek veri kümesindeki ilgisiz nitelikte ve tekrarlı verilerin çıkarıldığı aşamadır. Böylece veri boyutu azaldığından sorgu hızı da artmaktadır.
- Veri Madenciliği: Veri madenciliği(VM) algoritmalarının ve yöntemlerinin uygulandığı süreçtir. VM; veritabanı sistemleri, verilerin depolanması, istatistik, makine öğrenimi gibi alanların kombinasyonundan oluşan disiplinler arası bir yöntemdir. Her ne kadar VM istatistiğin bir alt kümesi olarak kabul edilse bile VM, veritabanı teknolojisi ve makine öğrenimi gibi diğer alanlara ait fikirleri, araçları ve yöntemleri de kullanılır(Yalçıntaş, 2003).
- Yorumlama (Değerlendirme) : Bilgi keşfi sürecinde bu aşamadan önceki aşamalar sonucunda elde edilen bilginin geçerlilik, yenilik, yararlılık ve basitlik kıstaslarına göre değerlendirilmesi aşamasıdır(Fayyad vd., 1996).

VERİ MADENCİLİĞİ

Veri madenciliği; daha önceden bilinmeyen, geçerli ve uygulanabilir verilerin geniş veri tabanlarından elde edilmesi, gelecekle ilgili tahminler yapılabilmesi ve bu bilgiler ışığında kararlar alınabilmesidir. Gartner Group

tarafından yapılan diğer bir tanımda veri madenciliği; istatistik ve matematik tekniklerle birlikte örüntü tanıma (pattern recognition) teknolojilerini kullanarak, depolama ortamlarında saklanmış bulunan veri yığınlarının elenmesi ile anlamlı yeni korelasyon, örüntü ve eğilimlerin keşfedilmesi sürecidir(Akpınar, 2000). Diğer bir tanımda ise veri madenciliği; büyük miktarda veri içinden, gelecekle ilgili tahmin yapmamızı sağlayacak bağıntı ve kuralların bilgisayar programları kullanılarak aranmasıdır. Veri analizi yapılarak, 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 onların davranışları ile ilgili tahminler yapılabilir(Gürsoy, 2009).

"Daha önce bilinmeyen" ya da tahmin edilemeyenle ilgili en ünlü örnek ise, artık klasikleşmiş, kulaktan kulağa anlatılan ve veri madenciliğinin "bilinmeyenini" çarpıcı bir şekilde önümüze koyan bira-çocuk bezi örneğidir(Silahtaroğlu, 2008):

Bir perakende mağazalar zincirinin yaptığı veri madenciliği araştırmasının sonuçlarına göre bira ile çocuk bezi satışları arasında, özellikle Cuma günler, güçlü bir ilişki vardır. Çocuk bezi satın alan kişilerin büyük çoğunluğu aynı zamanda bira da satın almaktadırlar. Daha doğrusu, Cuma günleri çocukları için alışverişe çıkan babalar arada kendileri için de alışveriş yapmaktadırlar(Cabena, 1998).

Veri Madenciliğinin Amaçları

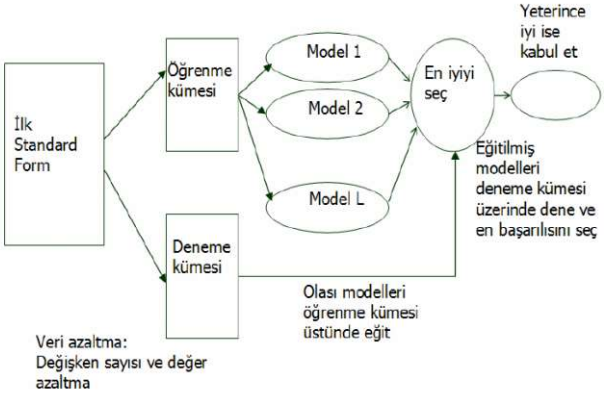
Veri madenciliğinin amaçları öngörü, tanıma, sınıflandırma ve en iyileme olarak dört başlık altında toplanabilir(Yarımağan, 2000).

- Öngörü; hangi ürünlerin hangi dönemlerde, hangi koşullarda, hangi miktarda satılacağına ilişkin yada eğitimde hangi davranışlar sonucunda hangi olayların gerçekleşebileceği gibi kestirimlerde bulunmak gibi tanımlanabilir.
- Tanıma; aldığı ürünlerden bir müşterinin tanınması veya kullanıldığı programlar ve yaptığı işlemlerden bir kullanıcının tanınması gibi ifade edilebilir.
- Sınıflandırma; birçok parametrenin birleşimi kullanılarak, örneğin ürünlerin, müşterilerin yada öğrencilerin sınıflandırılması olarak tanımlanabilir.
- En iyileme; belirli kısıtlamalar çerçevesinde zaman, yer, para yada ham madde gibi sınırlı kaynakların kullanımını en iyileme ve üretim miktarı, satış miktarı yada kazanç gibi değerleri büyütme olarak tanımlanabilir(Aydoğan, 2008).

Veri Madenciliği Metodolojisi

Aşağıdaki şekilde bir veri madenciliğinde kullanılan metodolojiyi gösterilmektedir. Standart form içinde verilen veri, öğrenme ve denem olmak üzere ikiye ayrılır. Her uygulamada kullanılabilecek birden çok teknik vardır ve önceden hangisinin en başarılı olacağını kestirmek olası değildir. Bu yüzden öğrenme kümesi üzerinden L kadar değişik teknik kullanılarak L tane model oluşturulur. Sonra bu L model deneme kümesi üzerinden denenerek en başarılı olanı, yani deneme kümesi üzerindeki tahmin başarısı en yüksek olanı seçilir.

Eğer bu en iyi model yeterince başarılıysa kullanılır, aksi takdirde başa dönerek çalışma tekrarlanır. Tekrar sırasında başarısız olan örnekler incelenerek bunlar üzerindeki başarının nasıl artırılabileceği araştırılır. Örneğin standart forma yeni alanlar ekleyerek programa verilen bilgi arttırılabilir veya olan bilgi değişik bir şekilde kodlanabilir veya amaç daha değişik bir şekilde tanımlanabilir(Alpaydın, 2000).



Şekil 2. Veri madenciliği metodolojisi

Veri Madenciliği Modelleri

Veri madenciliğinde kullanılan modeller; tanımlayıcı ve tahmin edici olarak iki grupta incelenir.

Tanımlayıcı modellerde; karar vermeye yardım edebilecek, mevcut verilerdeki örüntülerin tanımlanması sağlanmaktadır. X/Y aralığında geliri ve iki veya daha fazla arabası olan çocuklu aileler ile çocuğu olmayan ve geliri X/Y aralığından düşük olan ailelerin satın alma örüntülerinin birbirlerine benzerlik gösterdiğinin belirlenmesi örneği tanımlayıcı modeller grubuna girer. Kümeleme (clustering), birliktelik kuralı (association rule) ve ardışık örüntü (sequential pattern) madenciliği tanımlayıcı tekniklerden bazılarıdır.

Tahmin edici modellerde ise; gizli kalmış bilgilerin keşfine dayalı modellerdir. Sonuçları bilinen verişlerden hareket edilerek bir model geliştirilmesi ve kurulan bu modelden yararlanılarak sonuçları bilinmeyen veri kümeleri için sonuç değerlerin tahmin edilmesi amaçlanmaktadır(Akpınar,2000).

Sınıflama(classification), gerileme (regression) ve sapma (deviation) madenciliği tahmin edici tekniklerden bazılarıdır.

Veri madenciliği modelleri işlevlerine göre 3 temel grupta toplanır:

- Sınıflama ve Gerileme (classification and regression)
- Kümeleme (clustering)
- Birliktelik kuralları ve Sıralı örüntüler (association rules and sequential patterns)

Şekil 3. Veri madenciliği modelleri(Şen, 2008)

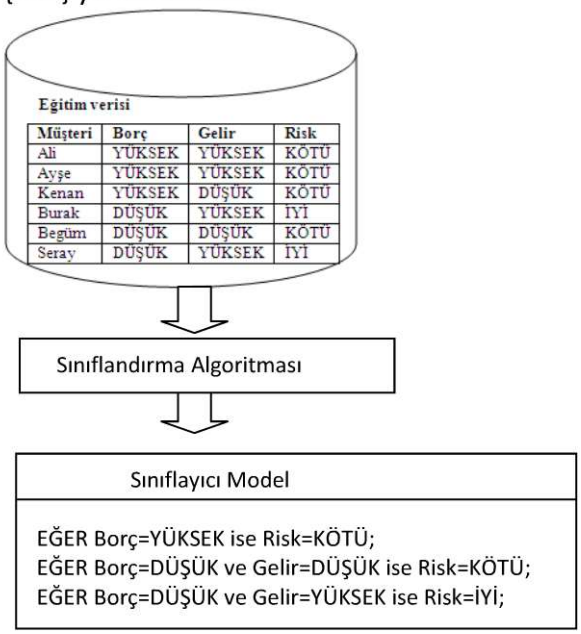
Sınıflama ve gerileme

Verilerin içerdiği ortak özelliklere göre ayrıştırılmasına sınıflandırma denmektedir. Örneğin bir sınıftaki öğrencileri; cinsiyetlerine, hangi burca sahip olduklarına, yaşadıkları evlerin kira mı yoksa kendilerinin mi olduklarına gibi farklı kriterlere göre sınıflandırma yapılabilinir.

Sınıflandırma bir öğrenme algoritmasına dayanır. Tüm veriler kullanılarak eğitme işi yapılmaz. Bu veri topluluğuna ait bir örnek veri üzerinde gerçekleştirilir. Öğrenmenin amacı bir sınıflandırma modelinin yaratılmasıdır(Özkan, 2008).

Sınıflandırma süreci iki aşamadan oluşmaktadır(Han ve Kamber, 2006):

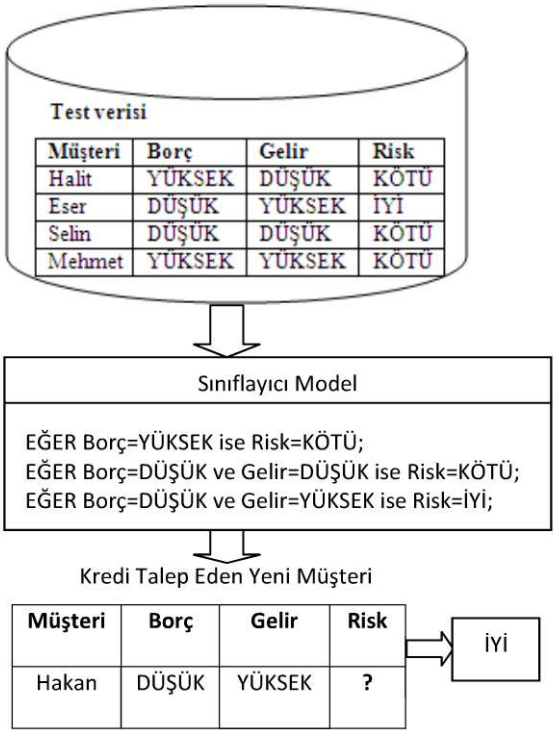
- İlk adım, veri kümelerine uygun bir modelin ortaya konulmasıdır. Söz konusu model, veritabanındaki kayıtların nitelikleri (attribute) veya bir başka deyişle alan isimleri(sütun isimleri) kullanılarak gerçekleştirilir. Sınıflandırma modelinin elde edilmesi için veri tabanının bir kısmı eğitim verileri olarak kullanılır. Bu veriler veri tabanından rastgele seçilir. Şekil 4 de görüldüğü gibi eğitim verileri üzerinde bir algoritma uygulanarak sınıflama modeli elde edilir. Şekil üzerinde {Müşteri, Borç, Gelir} niteliklerinin yanı sıra sınıf niteliği olarak da {Risk} yer almaktadır.



Şekil 4. Sınıflandırmada model kurma süreci

- Test verileri üzerinde sınıflandırma kuralları belirlenir. Ardından söz konusu kurallar bu kez test verilerine uygulanarak sınanır. Örneğin, Şekil 5 de "Hakan" isimli yeni bir müşterinin kredi talebinde bulunduğunu varsayalım. Bu müşterinin risk durumunu

belirlemek için örnek verilerden elde edilen karar kuralı doğrudan uygulanır. Bu müşteri için Borç=DÜŞÜK, Gelir=YÜKSEK olduğu biliniyorsa Risk=İYİ olduğu anlaşılır.

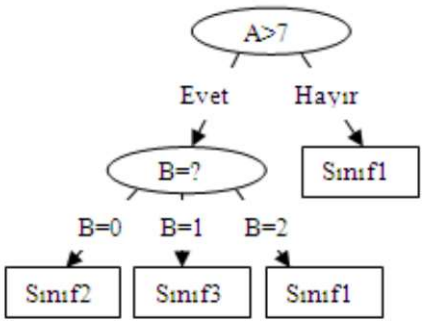


Şekil 5. Sınıflandırmada modelin uygulama süreci

Yukarıda test sonucunda elde edilen modelin doğru olduğu kabul edilecek olursa, bu model diğer veriler üzerinde uygulanır. Elde edilen sonuç model mevcut ya da muhtemel müşterilerin gelecekteki kredi risklerini belirlemede kullanılır(Özkan, 2008).
Sınıflama ve regresyon modellerinde kullanılan başlıca teknikler şunlardır(Akpınar, 2000):

Karar ağaçları

Karar ağaçları akış şemalarına benzeyen yapılardır. Her bir nitelik bir düğüm tarafından temsil edilir. Dallar ve yapraklar ağaç yapısının elemanlarıdır. En son yapı "yaprak", en üst yap "kök" ve bunların arasında kalan yapılar ise "dal" olarak adlandırılır(Quinlan, 1993).
Karar ağacı yapılarında, her düğüm bir nitelik üzerinde gerçekleştirilen testi, her dal bu testin çıktısını, her yaprak düğüm ise sınıfları temsil eder. En üstteki düğüm kök düğüm olarak adlandırılır. Karar ağaçları, kök düğümden yaprak düğüme doğru çalışır(Chiu ve Wei, 2002). Şekil 6 da sadece iki niteliğe(A-B) bağlı basit bir karar ağacı görülmektedir.



Şekil 6. A ve B niteliklerine bağlı bir karar ağacı

Şekil 6 da görülen basit bir karar ağacı örneğinde elips şeklinde gösterilen A>7 niteliği "kök", B=? gibi elips şeklinde gösterilen B niteliğinin alabileceği 0, 1 ve 2 değerleri "dal" ve dikdörtgenler ile gösterilen Sınıf1, Sınıf2 ve Sınıf3 nitelikleri ise "yaprak" olarak sınıflandırılır.
Sınıflandırmanın yapıldığı karar ağaçlarında önemli konulardan birisi de kökten itibaren dallanmanın(bölümlenmenin) hangi niteliğe göre olacağıdır. Bu amaçla;

- Entropiye dayalı algoritmalar
- Sınıflandırma ve regresyon ağaçları(CART) algoritmaları
- Bellek tabanlı sınıflandırma algoritmaları tercih edilebilir.

Entropiye dayalı bölümlemede "ID3, C4.5 ve J48" algoritmaları, Sınıflandırma ve regresyon ağaçlarında (CART) "Twoing ve Gini" algoritmaları, Bellek tabanlı sınıflandırma yöntemlerine ise "k-en yakın komşu" algoritmaları sayılabilir(Özkan, 2008). Bu algoritmalar birbirlerinden kök, düğüm ve dallanma kriteri seçimlerinde izledikleri yol açısından ayrılırlar(Silahtaroğlu, 2008).

ID3 algoritması

ID3 algoritması; makine öğrenmesi ve bilişim teorisine dayanarak verilen örnekler içinde en ayırıcı özelliğe sahip olan değişkeni bulan bir algoritmadır(Mitchell, 1997). Bunun için de entropi kavramından yararlanır.

Entropi beklentisizliğin maksimumlaşmasıdır(Fiske, 1998). Diğer bir tanımda ise entropi kavramı, eldeki bilgilerin sayısallaştırılmasıdır(Dunham, 2003). Dunham entropinin bir veri kümesi içindeki belirsizlik, şaşkınlık ve rasgeleliği ölçmek için kullanıldığını söyler. Eldeki bütün veriler tek bir sınıfa ait olsaydı, örneğin herkes aynı yaşta olsaydı, herhangi bir kişiye ise yaşını sorduğumuzda alacağımız yanıt bizi şaşırtmazdı; bu durumda entropi sıfır (0) olacaktır. Entropi sayısal olarak sıfır ile bir (0 – 1) aralığında bir değere sahiptir. Tüm olasılıklar($p_1,p_2,\ldots,p_i$) eşit olduğunda ise entropi en yüksek (1) değerine sahip olur.

S isimli bir kaynak veri seti olsun, p_i olasılık dağılımına bağlı S kaynağının tamamının entropi hesabı H(P) şu şekildedir(Shannon, 1948):

$$H(P) = -\sum_{i=1}^{n} p_i \log_2(p_i)$$

Aynı zamanda S veri kaynağındaki tüm T alt niteliklerinde kendi aralarında entropileri hesaplanır.

$$H(X,T) = \sum_{i=1}^{n} \frac{|T_i|}{|T|} H(T_i)$$

Veritabanı bölünmeden önce doğru sınıflandırma yapmak için gelen bilgiyle H(P), veritabanı bölündükten sonra doğru sınıflandırma için gelen bilgi H(X,T) arasındaki farka kazanım adı verilir.

$$Kazanç(X,T) = H(P) - H(X,T)$$

Burada bölümleme(dallanma) kriter seçimi yapılırken kazanç niteliklerinden en büyüğü öncelikli nitelik olarak seçilir ve seçilen nitelik üzerinden diğer niteliklere ait tekrar H(P) toplam entropi ve H(X,T) alt nitelik entropi hesabı yapılarak yeni kazançlar elde edilir. Bulunan kazançlardan en büyüğü seçilerek yeni alt bölümlemeler(dallanmalar) gerçekleştirilir.

Uygulamada Kazanç ölçütü adı verilen yukarıdaki formül yerine daha iyi sonuçlar veren Kazanç oranı adı verilen formül daha çok kullanılmaktadır(Shannon, 1948). T kümesi için X niteliğinin değerini belirlemek için gereken bilgi miktarını ortaya koymak için bu yol bulunmuştur. Söz konusu bilgi $H(P_{X,T})$ ile ifade edilir. $P_{X,T}$ ifadesi X değerlerinin olasılık dağılımıdır ve şu şekilde hesaplanır:

$$P_{X,T} = \left(\frac{|T_1|}{|T|}, \frac{|T_2|}{|T|}, \ldots, \frac{|T_k|}{|T|}\right)$$

Burada $H(P_{X,T})$ miktarı T kümesindeki X niteliği için bilgi bölünmesidir. Bu değer ise:

$$H(P_{X,T}) = H\left(\frac{|T_1|}{|T|}, \frac{|T_2|}{|T|}, \ldots, \frac{|T_k|}{|T|}\right)$$

veya

$$H(P_{X,T}) = -\sum_{i=1}^{k} \frac{|T_i|}{|T|} log_2\left(\frac{|T_i|}{|T|}\right)$$

formülleriyle bulunur. Yukarıda bulunan $H(P_{X,T})$ değeri ve Kazanç(Kazanım) ölçütü yardımıyla da kazanç oranı aşağıdaki gibi hesaplanır:

$$Kazanç\ Oranı(X,T) = \frac{Kazanç(X,T)}{H(P_{X,T})}$$

C4.5 (C5 ve J48) algoritmaları

C4.5 algoritması da ID3 algoritmasında olduğu gibi Quinlann tarafından geliştirilen bir sınıflama algoritmasıdır. Diğer ID3 algoritmasından en önemli üstünlükleri olarak, sayısal değerlere sahip nitelikler ve bilinmeyen-kaybolmuş değerlere sahip nitelikler için de geliştirilmiş bir karar ağacı oluşturmasıdır. Karar ağacı oluştururken kayıp verileri hesaba katmayarak, kazanım(kazanç) oranı hesaplanırken sadece verileri eksik olmayan diğer kayıtlar kullanılır. Böylece daha duyarlı ve daha anlamlı kurallar çıkartabilen bir ağaç üretmiş olur(Quinlan, 1986).

Kategorik olmayan sayısal değerli niteliklere ilişkin C4.5 karar ağacı modeli oluşturulurken değerleri iki aralığa bölmek için rastgele eşikler bulunmaktadır. En uygun t eşik değerini hesaplamak için birçok yöntem bulunmaktadır. Eşik değerinin belirlenmesi amacıyla, en büyük bilgi kazancını sağlayacak biçimde bir eşik değer belirlenir. Bu amaçla nitelik değerleri sıralanır ve $\{v_1, v_2, \ldots, v_n\}$ biçimini alır. Eşik değeri kullanarak nitelik değer iki parçaya ayrılır. Eşik değeri olarak $[v_i, v_{i+1}]$ aralığının orta noktası yada aritmetik ortalaması alınabilir. Veri setindeki sayısal bilgileri kategorik biçime

dönüştürdükten sonra ID3 algoritmasındaki gibi önce tüm veri setinin entropi hesabı sonra her nitelik için ayrı entropi hesabı yapılır. Yine her nitelik için ayrı kazanç oranları hesaplandıktan sonra elde edilen kazanç oranlarından en küçük değerli nitelik(değişken) kök ya da bir sonraki bölümleme niteliği olarak atanır(Silahtaroğlu, 2008).

Veri tabanındaki bilgilerde kayıplar olduğunda C4.5 algoritması bu sorunu gidermek için iki çözüm sunmaktadır.

Birincisi eğer kayıp veriler veri setindeki verilerin çoğunluğunu kapsamıyorsa o zaman kayıp verilerin olduğu kayıtlar veri tabanından çıkarılır ve algoritma geriye kalan veri tabanı üzerine uygulanır. Ancak veri tabanındaki kayıp verilerin sayısı fazla ise o zaman ikinci çözüm kullanılır.

İkinci çözüm ile kayıp verilerle de çalışacak bir algoritma uygulanır. Algoritma uygulanmadan önce kayıp verilere sahip örneklerde kazanç ölçütünün hesaplamak için bir F düzeltme faktöründen yararlanılır. Bu amaçla ilk olarak kayıp veriler çıkarılarak H(P) toplam entropi ve H(X,T) sınıflar için entropiler ID3 algoritmasındaki gibi hesaplanır. F faktörü kullanılarak kazanç ölçütü düzeltilir:

$$F = \frac{\text{Veri tabanında değeri bilinen niteliğe sahip örneklerin sayısı}}{\text{Veri tabanındaki tüm örneklerin sayısı}}$$

Yeni kazanç ölçütü ise(Han ve Kamber, 2006):

$$Kazanç(X) = F\big(H(T) - H(X,T)\big)$$

olarak hesaplanır.

Karmaşık görünümlü karar ağaçlarında bir alt ağacı atarak yerine bir yaprak (sınıf niteliği) yerleştirmeye karar ağacının budanması adı verilir. Alt ağacın yerine yaprak yerleştirmekle, algoritma öngörülü hata oranını azaltmayı ve sınıflandırma modelinin kalitesini artırmayı amaçlar(Kantardzic, 2003).

C4.5 algoritmasının ID3 algoritmasına göre diğer bir üstünlüğü de doğruluk ölçütü kullanarak karar ağacını budamaktır. Ağaçtaki her düğüm için U_{cf} üst güven sınırı iki terimli dağılımların istatistiksel tablolarını kullanarak elde edilebilir(Kantardzic, 2003). Verilen düğümde U_{cf} parametresi T_i ve E'nin bir fonksiyonudur. C4.5 algoritması %25 güven sınırı kullanır ve verilen her bir düğümdeki T_i, $U_{\%25}(|T_i|/E)$ düğüm yapraklarının güven aralığı ile karşılaştırılır. Her bir yaprakta ağırlıklar durumların toplam sayısıdır. Eğer alt ağaçtaki kök düğümün beklenen hatası, yapraklardaki $U_{\%25}$ toplam ağırlıktan daha küçük ise(alt ağacın beklenen hatası), o zaman alt ağaç yok edilir, yerine kök düğüm konulur. Böylece budanmış ağaçta yeni bir yaprak olarak yer alır(Özkan, 2008).

C5 algoritması da C4.5 algoritmasına dayalı geliştirilmiş bir karar ağacı algoritmasıdır. J48 algoritması ise C4.5 karar ağacı algoritmasının WEKA (Waikato Environment for Knowledge Analysis) açık kaynak kodlu paket programına uyarlanmış versiyonudur. J48 algoritması; aynı veri seti üzerinden WEKA paket programında Naive Bayes, Lojistik Regresyon, ID3, JRIP, PART ve Sinir Ağları gibi bilinen sınıflandırma algoritmalarına göre doğruluğu en yüksek sınıflandırma algoritmasıdır(Aydoğan vd., 2008).

CART algoritması

CART (Classification and Regression Trees) tekniği ID3 algoritmasında olduğu gibi en iyi dallara ayırma kriterini seçmek için entropiden yararlanır(Fiske, 1998). En iyi ayırma kriterini belirlemek için ise ID3 ve C4.5 algoritmalarından farklı bir formül kullanır. CART tekniğinde kullanılan en iyi dallara ayırma kriteri için(Webb ve Yohannes, 1999):

Herhangi bir t düğümündeki s dallara ayırma kriteri Ψ(s/t) olarak gösterilirse:

$$\Psi\left(\frac{s}{t}\right) = 2\,P_L P_R \sum_{J=1}^{M} \left| P\left(\frac{C_j}{t_L}\right) - P\left(\frac{C_j}{t_R}\right) \right|$$

Formüldeki;

t: dallanmanın yapılacağı düğümü

C: kriteri

L: ağacın sol tarafını

R: ağacın sağ tarafını

P_L ve P_R: öğrenim kümesindeki bir kaydın solda veya sağda olma olasılıklarını

$P(C_J/t_L)$ ve $P(C_J/t_R)$: C_J sınıfındaki bir kaydın sağda veya solda olma olasılıklarını ifade etmektedir.

CART tekniğinde Ψ(s/t) hesabı yapılırken; dallanmalar en büyük uygunluk ölçütüne göre gerçekleştiriliyorsa Twoing algoritması, dallanmalar en küçük uygunluk ölçütüne göre gerçekleştiriliyorsa Gini algoritması kullanılıyor denilebilir(Özkan, 2008).
CART tekniğinde dallara ayırma kriterleri hesaplanırken kayıp veriler önemsenmez(Silahtaroğlu, 2008).

SLIQ algoritması

SLIQ (Supervised Learning In Quest) algoritması IBM Almaden araştırma merkezinde (Agrawal vd., 1996) tarafından önerilmiştir. SLIQ algoritması hem sayısal hem de kategorik verilerin sınıflandırılmasında kullanılabilir. Sayısal verilerin değerlendirilmesindeki maliyeti azaltmak için ağacın oluşturulması sırasında önceden verileri sıralama tekniği kullanılır.
ID3 ve C4.5 gibi algoritmalar 'önce derinlik' ilkesine göre çalışırken SLIQ algoritması 'önce genişlik' düşüncesiyle hareket ederek aynı anda birçok yaprağı oluşturur. Bu durumda mevcut ağacın yapraklara ayrılma işlemi verinin üzerinden bir kez geçilmesiyle tamamlanmış olur. Dallara ayırma işleminde gini indeksi kullanılır(Agrawal vd., 1996).
Ayrıca SLIQ algoritması kategorik verileri alt kümelere ayırmada ID3 ve C4.5 algoritmalarına göre daha hızlıdır. Ağacın budanması işlemi için de; verileri en iyi temsil edecek modelin tanımlanma ve oluşturulma maliyeti en düşük olan model kavramı olan MDL[2] ilkesine dayanan bir strateji izlemektedir(Rissanen, 1989).

SPRINT algoritması

ID3, CART ve C4.5 gibi algoritmalar önce derinlik ilkesine göre çalışırlar ve en iyi dallara ayırma kriterine sahip olabilmek için her düğümde sürekli olarak verileri sıraya dizerler. SLIQ ise her bir değişken için ayrı bir liste kullanarak bu sıraya dizme işlemini sadece bir kez yapar. SPRINT algoritması ise bu yönüyle SLIQ algoritmasına benzer ve sözü edilen diğer algoritmalardan ayrılır. Ancak farklı veri yapıları kullanarak SLIQ algoritmasından ayrılır(Shafer, 1996).

SPRINT ilk olarak her bir nitelik için aşağıdaki tablolarda görüldüğü gibi ayrı bir nitelik listesi hazırlar.

Çizelge 1. Veri setindeki her bir niteliğin ayrılması

Yaş	Sınıf	Sıra No
17	Y	2
21	Y	1
22	Y	3
36	D	4

Araç Tipi	Sınıf	Sıra No
Sedan	Y	1
Spor	Y	2
Spor	Y	3
Sedan	D	4

Her tabloda kullanılacak olan değişken, sınıf ve sıra no bulunacaktır. Bu durumda veri setindeki nitelik sayısı kadar tablo oluşacaktır. Sayısal değerleri taşıyan tablolar sayısal değer değişkenine göre sıraya dizilirken, kategorik verileri taşıyan tablolar ise sıra numarasına göre sıralı olarak kalacaktır. Eğitim kümelerinden elde edilen ilk listeler sınıflandırma ağacının köküyle ilişkilendirilir. Ağaç büyüyüp düğümler yeni dallara bölündükçe her düğüme ait değişken listeleri de bölünerek yeni dallarla ilişkilendirilir. Bir liste bölündüğünde ise içindeki kayıtların sıralaması değiştirilemez; böylece bölünme suretiyle oluşturulmuş yeni listelerin bir daha kendi içlerinde sıraya dizilmesine gerek kalmaz(Silahtaroğlu, 2008).
Bölünme aşamasına gelmiş düğümler için $C_{üst}$ ve C_{alt} adı verilen ve düğümdeki sınıf dağılımlarını elde etmek için kullanılan histogramlar belirlenir. Düğümlerden alt dallara ayırma kriteri için ise SLIQ algoritmasında olduğu gibi Gini indeksi kullanılır. Herhangi bir K kümesinin gini(K) indeksi aşağıdaki gibi hesaplanır(Shafer, 1996; Brieman, 1984):

$$gini(K) = 1 - \sum p_j^2$$

Burada p_j, K kümesi içinde j sınıfının sıklığıdır. Eğer K kümesi K_1 ve K_2 gibi alt kümelere bölünürse bölünmüş K kümesinin $gini_{bölünmüş}(K)$ değeri:

$$gini_{bölünmüş}(K) = \frac{n_1}{n_1+n_2} gini(K_1) + \frac{n_2}{n_1+n_2} gini(K_2)$$

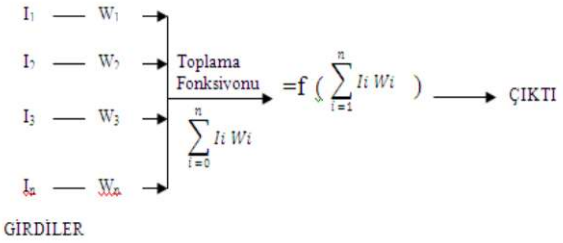
şeklinde hesaplanır.

Yapay sinir ağları

Sinir ağları, tanımlayıcı ve tahminci veri madenciliği algoritmalarındandır. İnsan beyninin fizyolojik yapısını taklit ederler.

Komplike ve belirsiz veriden bilgi üretirler. Keşfettikleri örüntü ve trendler, insanlar yada bilgisayarlarca kolay keşfedilmez. Bu tür karmaşık problemlerde birbirleriyle etkileşimli yüzlerce değişken bulunur(Aryeetey,2003). Bu teknik, veritabanındaki örüntüleri, sınıflandırma ve tahminde kullanılmak üzere genelleştirilir. Sinir ağları algoritmaları sayısal veriler üzerinde çalışırlar.

Aşağıdaki şekilde basit bir yapay sinir hücresi gösterilmiştir(Gürsoy, 2009):



Şekil 7. Basit bir yapay sinir hücresi

Şekilde de görüldüğü gibi n adet I girdi değeri, kendi W ağırlık değeri ile çarpılarak toplanır ve çıktıyı elde etmek için aktivasyon fonksiyonu ile işlem yapılır.

Genel olarak yapay sinir ağı modelleri; ağın yapısına ve öğrenme türüne göre; İleri beslemeli (Feed forward) ve Geri beslemeli(Feed back) ile Denetimli öğrenme (Supervised learning) ve Denetimsiz öğrenme (Unsupervised leraning) olmak üzere ikiye ayrılır(Gürsoy, 2009).

Çizelge 2. Yapay sinir ağlarının avantaj ve dezavantajları

Avantajları	Dezavantajları
• Gürültülü, hatalı veriler ile çalışabilirler. • Sayısal tahmin, sınıflandırma ve kümeleme problemlerinde kullanılabilirler. • Karmaşık problemlerin çözümünde iyi sonuçlar verebilmektedirler.	• Elde ettikleri çözümlerin gerekçelerini açıklayamazlar. • Optimum sonuca ulaşacaklarının bir garantisi yoktur. • Ezberleyebilirler.

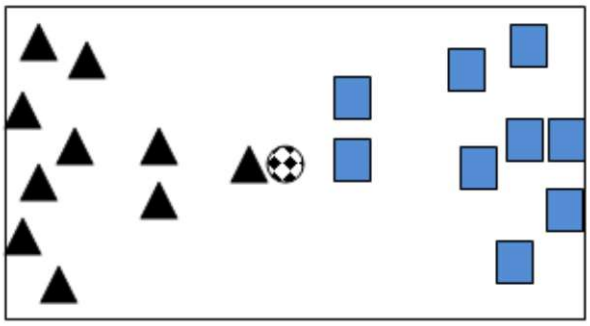
K en yakın komşu algoritması

Sınıflandırma yöntemlerinden birisi de K-en yakın komşu algoritmasıdır. Bellek tabanlı ve mesafeye dayalı algoritmalardan bir tanesidir. Sınıflandırma yapılırken eldeki verilerin birbirlerine olan uzaklığı veya benzerliği kullanılarak sınıflamanın gerçekleştirilmesi tekniğidir.

Veriler arası mesafe ölçülürken en çok kullanılan mesafe öklit (Euclid) mesafesidir. Ancak, bir kayıt için diğer kayıtlardan sadece k adedi göz önüne alınır. Bu k adet kayıt, diğer bir deyişle veri tabanındaki nokta, mesafesi hesaplanan noktaya diğer kayıtlara nazaran en yakın olan kayıtlardır(Beyer, 1999).

Algoritmada k değeri önceden seçilir; değerin yüksek olması birbirlerine benzemeyen noktaların bir araya toplanmasına, çok küçük seçilmesiyle birbirine benzediği, yanı aynı sınıfın noktaları oldukları halde, bazı noktaların ayrı sınıflara konmasına ya da o tür noktalar için ayrı sınıfların açılmasına neden olur. Tipik k değerleri 3.5 ve 7'dir(Khan, 2002).

Şekil 8 de verilen örneğe göre k'nın alacağı değerlere göre ortadaki yuvarlak nesne farklı sınıflarda olabilir. Örneğin k=2 seçilirse yani ortadaki yuvarlak noktaya en yakın 2 nesne nedir diye sorulursa, cevabı karelerin olduğu sınıftır diyebiliriz. Ancak k=3 seçilirse o zaman yuvarlak noktaya en yakın seçilen 3 nesnenin üçgenler sınıfı olduğu görülür.



Şekil 8. Ortadaki yuvarlak k=2 için karelere, k=3 için üçgenlere en yakındır.

Doğrusal regresyon sınıflandırıcılar

Herhangi bir bağımlı değişkenin bir veya birden fazla bağısız açıklayıcı değişken ile arasındaki ilişkinin matematik bir fonksiyon şeklinde yazılmasına regresyon denklemi adı verilir. Bağımlı bir değişkenin (y) tek bir bağımsız – açıklayıcı değişken (x) ile arasındaki ilişkinin doğrusal fonksiyonla ifade edilmesine doğrusal regresyon ile ifade edilir. Örneğin bir grup üniversite öğrencisinin başarı durumları ile bu

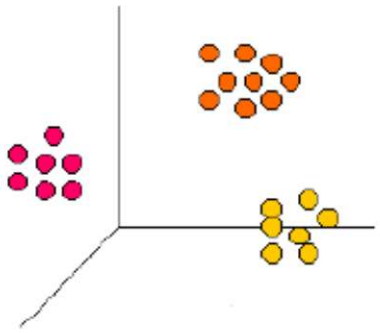
öğrencilerin üniversite giriş puanları arasındaki ilişkinin araştırılmasında basit doğrusal regresyon analizi kullanılır. Doğrusal regresyon denklemi ise y = a+bx şeklinde ifade edilir. Denklemde yer alan a parametresi regresyon doğrusunun y eksenini kestiği noktayı, b parametresi ise doğrunun eğimini açıklamaktadır. b parametresinin işareti pozitif ise artan, b parametresinin işareti negatif ise azalan bir eğim söz konusudur. Bazı alanlarda tek bir değişkeni başka bir değişkenle açıklamak mümkün değilse ve bunu birden fazla değişkenle açıklamak gerekiyorsa o zaman çoklu doğrusal regresyon analizleri yapılabilir(Gürsoy, 2009).

Lojistik regresyon sınıflandırıcılar

Lojistik regresyon analizi, bağısız değişkenlerde çoklu normallik ve gruplar arası varyans kovaryans matrislerinin eşitliği varsayımları sağlanamadığından diskriminant analizine alternatif olarak geliştirilmiş bir analiz tekniğidir. Lojistik regresyon analizinin çoklu regresyon analizine benzemesi, basit, kolay ve açık istatistik testlere sahip olması, metrik ve metrik olmayan değişkenleri birlikte analize dahil edebilme ve detaylı sonuç verme özelliklerine sahip olduğu için iki gruplu diskriminant analizinden daha güçlü sonuçlar verdiği söylenebilir(Gürsoy, 2009).

Kümeleme

Kümeleme analizinde; veri tabanındaki kayıtların hangi kümelere ayrılacağı veya kümelemenin hangi değişken özelliklerine göre yapılacağı, konunun uzmanı olan bir kişi tarafından belirtileceği gibi veri tabanındaki kayıtların hangi kümelere ayrılacağını geliştiren yazılımlar da yapabilmektedir. Kümeleme tekniğinde; sınıflama tekniğinde olan veri sınıfları yoktur. Sınıflama tekniğinde, verilerin sınıfları bilinmekte ve yeni bir veri geldiğinde bu verinin hangi sınıftan olabileceği tahmin edilmektedir(Han ve Kamber, 2006).
Kümeleme yöntemi, danışmansız sınıflama modeli olarak da bilinir(Pryke, 1998). Kümeleme heterojen veri kümelerini veri karakteristikleri homojen sayılabilecek gruplara bölme bir başka deyişle diğerlerinden çok farklı ancak üyeleri çok benzer olan grupları bulma işidir. Kümeleme modelinde; veri tabanındaki kayıtların hangi kümelere ayrılacağı veya kümelemenin hangi değişken özelliklerine göre yapılacağı, konunun uzmanı bir kişi tarafından belirlenebilir.
Kümeleme algoritması, veri tabanını alt kümelere ayırır. Her bir kümede yer alan elemanlar dâhil oldukları grubu diğer gruplardan ayıran ortak özelliklere sahiptirler. Kümeleme modellerinde amaç şekil 9 da görüldüğü gibi küme üyelerinin birbirlerine çok benzediği ancak özellikleri birbirlerinden çok farklı olan kümelerin bulunması ve veri tabanındaki kayıtların bu farklı kümelere bölünmesidir.



Şekil 9. Örnek kümeleme sorgusu

Kümeleme yöntemlerinin birçoğu, gözlem değerleri arasındaki uzaklıkların hesaplanması esasına dayanmaktadır. İki nokta arasındaki uzaklığı hesaplayan çeşitli bağıntılar vardır. Kümeleme analizlerinde en çok kullanılanlar ise:

- *Öklit uzaklığı;* iki boyutlu uzayda Pisagordaki hipotenüs bağıntısını kullanarak uzaklığı hesaplar ve kümeleme analizlerinde en çok tercih edilen uzaklık ölçüsüdür.

$$d(i,j) = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}$$

- *Manhattan uzaklığı;* gözlemler arasındaki mutlak uzaklıkların toplamı ile elde edilir.

$$d(i,j) = \sum_{k=1}^{p} \left(|x_{ik} - x_{jk}| \right)$$

- *Minkowski uzaklığı;* p adet değişken göz önüne alınarak gözlem değerleri arasındaki uzaklığın hesaplanmasıdır.

$$d(i,j) = \left[\sum_{k=1}^{p}\left(\left|x_{ik} - x_{jk}\right|^{m}\right)\right]^{1/m}$$

Kümeleme yöntemleri hiyerarşik ve hiyerarşik olmayan olmak üzere ikiye ayrılır:
Hiyerarşik kümeleme yöntemlerinde; önce tüm gözlemler arasındaki uzaklıklar hesaplanır. Birbirine yakın olan gözlemler birleştirilerek bir küme elde edilir. İşlemlere yeniden başlanarak kümeleme işlemi devam edilir. Eğer iki kümenin birbirine olan uzaklığı söz konusu ise, birbirine en yakın uzaklıklar yani en küçük uzaklık değerleri iki küme arasındaki uzaklık olarak belirlenir. Bu işlem en yakın komşu algoritması ile sağlanır. En uzak komşu algoritmasında ise iki kümenin arasındaki uzaklık olarak en büyük uzaklık seçilir.
Hiyerarşik olmayan kümeleme yönteminde ise k-ortalamalar algoritması uygulanabilir. Bu yöntem, küme içindeki ortalama hatayı en aza indirme amacını taşır(Özkan, 2008).

Birliktelik kuralları

Birliktelik kuralları ile bir ilişkide yer alan niteliklerin değerleri arasındaki bağımlılıklar, anahtarda yer almayan diğer niteliklerin gruplandırılması ile bulunur. Bu kurallar ilk olarak Agrawal tarafından 1994'te geliştirilmiştir(Altınışık, 2006).
Çok sayıda verinin depolandığı bir veri tabanı içinde çeşitli nitelikler arasında hemen fark edilmeyen bir takım ilişkiler mevcut olabilir. Bu tip ilişkilerin ortaya çıkartılması stratejik kararların alınmasına yardımcı olabilir. Ancak, bu ilişkilerin çok sayıda verinin içinden elde edilmesi basit bir süreç değildir. Bu süreç birliktelik kuralı madenciliği (association rule mining) olarak adlandırılmaktadır. Veriler arasındaki ilişkiler eğer – sonra ifadeleri ile aşağıdaki gibi gösterilmektedir(Ghosh ve Nath, 2004).
Eğer <bazı şartlar sağlanırsa> sonra <bazı niteliklerin değerlerini tahmin et>
Birliktelik kurallarının analizi süreci market sepeti analizi olarak da adlandırılır. Market sepeti analizinde müşteri ile ilgili veri hareketlerinden gelecekte müşterinin nasıl bir tercih yapacağına dair sonuçlar tahmin edilmektedir(Altınışık, 2006).
Market sepet çözümlerinde satılan ürünler arasındaki ilişkileri ortaya koymak için "destek" ve "güven" gibi iki ölçütten yararlanılır. Bu ölçütlerin hesaplanmasında "destek sayısı" adı verilen bir değer kullanılır. "kural destek ölçütü" bir ilişkinin tüm alışverişler içinde hangi oranda tekrarlandığını belirler. Kural güven ölçütü, A ürün grubunu alan müşterilerin B ürün grubunu da alma olasılığını ortaya koyar. A ürün grubunu alanların B ürün grubunu da alma durumu, yani birliktelik kuralı $(A \rightarrow B)$ biçiminde gösterilir. Bu durumda kural destek ölçütü şu şekilde ifade edilebilir(Özkan, 2008):

$$destek(A \rightarrow B) = \frac{sayı(A,B)}{N}$$

Burada sayı(A,B) destek sayısı A ve B ürün gruplarını birlikte içeren alışveriş sayısını, N ise tüm alışverişlerin sayısını göstermektedir. A ve B ürün gruplarının birlikte satın alınması olasılığını ifade eden kural güven ölçütü ise şu şekilde hesaplanır(Kumar vd., 2006):

$$güven(A \rightarrow B) = \frac{sayı(A,B)}{sayı(A)}$$

Birliktelik kuralları belirlenirken destek ve güven ölçütlerinin yanı sıra, bu değerleri karşılaştırmak üzere eşik değere gereksinim vardır. Hesaplanan destek veya güven ölçütlerinin *destek(eşik)* ve *güven(eşik)* değerlerinden büyük olması beklenir. Hesaplanan destek veya güven ölçütleri ne kadar büyük ise birliktelik kurallarının da o derece güçlü olduğuna karar verilir(Özkan, 2008).
"Çocuk maması alanların %40'ı makarna da satın alır" örneğinde bir bağıntı bulunmaktadır. Çocuk maması alanların çocuk bezi, makarna alanların ketçap alacağını tahmin etmek kolay olmaktadır. Ancak örnek incelendiğinde, sonucu çıkarmak için bütün olasılıkları göz önüne alarak kolayca aklımıza gelmeyen ürün birliktelikleri ortaya çıkartılmaktadır(Altınışık, 2006).
Zaman içinde sıralı örüntüler örneği ise "ilk üç taksitinden iki veya daha fazlasını geç ödemiş olan müşteriler %60 olasılıkla kanuni takibe gidiyor" şeklinde olabilir. Davranış skoru (behavioral score), başvuru skorundan farklı

olarak kredi almış ve taksitleri ödeyen bir kişinin sonraki taksitlerini ödeme/geciktirme davranışını notlamayı amaçlar(Alpaydın, 2000). En yaygın birliktelik kuralı algoritmaları arasında Aprioari ve GRI (The Generalized Rule Induction) sayılabilir(Yılmaz, 2008).

Veri Madenciliği Uygulama Alanları

Veri madenciliği birçok konuyu kapsayan disiplinler arası bir yaklaşımdır. Bu nedenle yeni bir disiplin olmasına rağmen uygulama alanı oldukça geniştir. Bunlar:

Web uygulamalarında

- Kullanıcı taraflı bilgiler (tarayıcı, dil, vb) ışığında alt yapı düzenlemelerine gidilebilir(Christensen vd., 2000).
- Kullanıcıların profilleri çıkarılabilir ve zaman içindeki değişimleri takip edilebilir, sitedeki beğenilen ya da beğenilmeyen köşeler tespit edilebilir(Bing, 2007).
- Kullanıcı profillerine göre site perspektifi düzenlenebilir.
- Site haritası, linkler, vs. düzenlemeleri yapılabilir.
- Kullanıcıların gezinti şekli/hızı sitenin içerik, yapılandırma ve alt-yapı açısından performansı hakkında fikir verir(Bing, 2007).
- Kullanıcı profillerine uygun ürünlerin reklam kampanyaları en çok ziyaret ettikleri sayfalara koyulabilir(Güvenç, 2001).
- En sık beraber ziyaret edilen çift sayfalar belirlenebilir(Güvenç, 2001).
- Farklı web şablonları, temaları arasında kullanıcı istekleri değerlendirilebilir.
- Form verilerinin toplanmasındaki zorlukları en aza indirme yöntemleri geliştirilebilir.
- Kötü niyetli kullanıcı istekleri belirlenip bunlara karşı alınması gereken önlemler belirlenebilir(Christensen vd., 2000).

İşletme alanında

Çeşitli işletme alanlarındaki uygulamalar(Hudairy, 2004):

- Bir işletme kendi müşterisiyken rakibine giden müşterilerle ilgili analizler yaparak rakiplerini tercih eden müşterilerinin özelliklerini elde edebilir ve bundan yola çıkarak gelecek dönemlerde kaybetme olasılığı olan müşterilerin kimler olabileceği yolunda tahminlerde bulunarak onları kaybetmemek, kaybettiklerini geri kazanmak için strateji geliştirebilir.
- Ürün veya hizmette hangi özelliklerin ne derecede müşteri memnuniyetini etkilediği, hangi özelliklerinden dolayı müşterinin bunları tercih ettiği ortaya çıkarılabilir.
- Müşterilerin kredi riskleri hesaplanarak hangi müşterilerin kredi riskinin yüksek olduğu, hangi müşterilerin geri ödemesini zamanında yapamayabileceği kestirilebilir.
- Kredi kartı ödemelerini aksatan, gecikmeli olarak yapan veya hiç yapmayanların özelliklerinden yola çıkılarak bundan sonra aynı duruma düşebilecek muhtemel kişiler saptanabilir.
- Ürün talebi bazında müşteri görünümlerini belirleyerek, müşteri segmentasyonuna gitmek ve çapraz satış olanakları yaratmakta kullanılabilir.
- Piyasada oluşabilecek değişikliklere mevcut müşteri portföyünün vereceği tepkinin firma üzerinde yaratabileceği etkinin tespitinde kullanılabilir.
- En kârlı mevcut müşteriler saptanarak, potansiyel müşteriler arasından en kârlı olabilecekler belirlenebilir. Kârlı müşteriler tespit edilerek onlara özel kampanyalar uygulanabilir. En masraflı müşteriler daha masrafsız müşteri haline dönüştürülebilir. Örneğin en çok bankacılık işlemi yapanlar ortaya çıkarılıp bunlar şube bankacılığı yerine daha masrafsız internet bankacılığına yönlendirilebilir.
- Bir ürün veya hizmetle ilgili bir kampanya programı oluşturmak için hedef kitlenin seçiminden başlayarak bunun hedef kitleye hangi kanallardan sunulacağı kararına kadar olan süreçte veri madenciliği kullanılabilir.
- Kurum teknik kaynaklarının en uygun şekilde kullanılmasını sağlamakta kullanılabilir.
- Geçmiş ve mevcut yapı analiz edilerek geleceğe yönelik tahminlerde bulunabilir. Özellikle ciro, kârlılık, pazar payı gibi analizlerde de veri madenciliği kullanılabilir.

Perakendecilik alanında

- Satış noktası veri analizlerinde
- Alış-veriş sepeti analizlerinde
- Tedarik ve mağaza yerleşim optimizasyonunda

Borsa alanında

- Hisse senedi fiyat tahmininde
- Genel piyasa analizlerinde
- Alım-satım stratejilerinin optimizasyonunda

Telekomünikasyon alanında

- Kalite ve iyileştirme analizlerinde
- Hisse tespitlerinde
- Hatların yoğunluk tahminlerinde
- İletişim desenlerinin belirlenmesinde
- Kaynakların verimli kullanılmasında
- Servis kalitesinin arttırılmasında

Sağlık alanında

- Test sonuçlarının tahmininde
- Ürün geliştirmelerinde
- Tıbbî teşhislerde
- Tedavi sürecinin belirlenmesinde
- Semptomlara göre hastalık tespitinde
- Magnetik Rezonans (MR) verileri ile sinir sistemi bölge ilişkilerinin belirlenmesinde

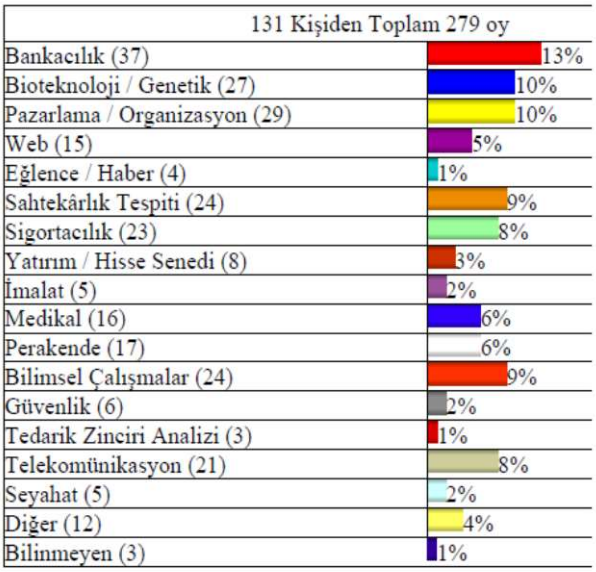
Endüstri alanında

- Kalite kontrol analizlerinde
- Lojistikte
- Üretim süreçlerinin optimizasyonunda

Eğitim alanında

- Kütüphane kullanıcılarının erişim örüntülerinin keşfi(Soğukpınar ve Takçı, 2002).
- Öğrenci Seçme Sınavına giren öğrencilerin profillerinin ve tercihlerinin öğrenci başarılarına etkisi(Delioğlu vd., 2007).

Aşağıdaki şekil 10 da ise 2003 yılında yapılan bir araştırma sonucuna göre veri madenciliğinin sektörler bazında ilişkin sonuçlar görülmektedir(web sayfası, 11.07.2011):



Şekil 10. Veri madenciliği uygulama alanları

1. UYGULAMA

Bilgisayardan eğitim alanında yararlanma konusunda rol oynayan etkenler ve bilgisayarı öğretim amaçlı kullanmanın sağladığı yararlar hakkında çeşitli çalışmalar yapılmıştır(Bindak ve Çelik, 2005; Ertekin vd., 2010). Ancak meslek yüksek okulu gibi ön lisans seviyesinde eğitim öğretim gören öğrencilerin bilgisayar hatta bilgisayar laboratuar ortamlarının başarılarına etkisinin incelenmesine de ihtiyaç duyulmaktadır. Üniversitelerin ön lisans seviyesinde eğitim gören birçok öğrencinin mezun olduklarında iş piyasasında ara eleman olacakları düşünüldüğünde, bu kişilerin temel bilgisayar ve Microsoft Office paket programlarını kullanma becerileri önem kazanmaktadır. Bu çalışmanın amacı da öğrencilerin bilgisayar ve bilgisayar laboratuar ortamlarında karşılaştıkları durumların ilgili derse hangi düzeyde başarı olarak geri yansıyacağı ve gelecekte karşılaşılacak sorunlar sonucunda öğrencilerin ders başarıları tahmin edilebilecektir. Bu amaçla da elde edilen anlamlı bilgiler çerçevesinde bilgisayarlar ve bilgisayar laboratuar ortamlarının geliştirilmesi sağlanabilir.

Öğrenciler anketi yanıtladıklarında bilgisayar laboratuar ortamlarında karşılaştıkları durumları

ve o derse ait geçme durumlarını ortaya koymuşlardır. Ancak araştırma 2008-2010 eğitim-öğretim yılları arasında Sivas Cumhuriyet Üniversitesi Divriği Nuri Demirağ Meslek Yüksekokulunda bilgisayar laboratuarlarında Temel Bilgisayar Bilimleri adlı dersi görmüş "eğitim (102) ve test (25) amaçlı" toplam 127 kişilik bir öğrenci grubundan oluşmaktadır.

Araştırmada, öğrencilere doldurmaları amacıyla (Bindak ve Çelik, 2005) tarafından geçerlik ve güvenirliği yapılmış bilgisayar tutum ölçeği ile birlikte 10 soruluk ve dersten geçme/kalma durumları ile ilgili bir anket uygulanmıştır.

Ölçekteki maddeler veri madenciliği uygulama adımlarından olan "veri dönüştürme" adımı ile 0 ile 4 arasında sayısal formda kodlanmışlardır. "hiç katılmıyorum" için 0, "katılmıyorum" için 1, "kararsızım" için 2, "katılıyorum" için 3 ve "tamamen katılıyorum" için de 4 kodları tanımlanmıştır. Dersten geçen öğrenciler için "G", kalan öğrenciler için de "K" biçiminde alfa nümerik olarak kodlanmıştır. Ayrıca karar ağacında daha düzgün görüntülenebilmesi için de aşağıdaki sorular kodlanmıştır:

Çizelge 3. Anket üzerindeki soru ve kodlamaları

Soru No ve Soru	Soru Kodu
1. Uygulama yapılan konuları içeren kaynak bir kitabın olması başarımı daha da artırır.	S1
2. Bilgisayar üzerinde uygulama yapıldığında, anlatılan konuları daha iyi anlıyorum.	S2
3. Bilgisayardaki yazılım eksiklikleri (eksik ya da hatalı kurulmuş program parçaları, vs) yüzünden uygulama yapamıyorum.	S3
4. Bilgisayardaki donanım eksiklikleri (mause un olmaması, klavye, monitör ya da kasanın çalışmaması, vs) yüzünden uygulama yapamıyorum.	S4
5. Bilgisayar laboratuarının fiziksel altyapı (bilgisayar sayısının azlığı, elektriksel güç sorunları, masa-sandalye yetersizliği gibi) nedenlerinden dolayı derse karşı motive olamıyorum.	S5
6. Anlatılan ders içeriği bilgisinin uygulamaya yönelik olmadığını düşünüyorum.	S6
7. Haftada 80 dakikalık ders saati uygulama yapmama yeterli olmuyor.	S7
8. Ders içinde uygulama yapılırken öğrencilerin bireysel hız farklılıkları dersten kopmama neden oluyor.	S8
9. Değerlendirme biçiminin yazılı olarak değil de, bilgisayar üzerinde uygulama olarak yapılması başarımı daha da artırır.	S9
10. Okul laboratuarı haricinde uygulama yapma şansım olmadığı için konuları daha çabuk unutuyorum.	S10

Öğrencilerin ankete verdikleri cevaplar şekil 11 de SPSS paket programında güvenirlik yönünden analiz edildiğinde Cronbach's Alfa katsayısı 0,654 (0,6<Alfa katsayısı<0,8) çıktığından anket oldukça güvenilir olarak kabul edilebilir.

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
,654	,624	10

Şekil 11. Verilerin SPSS programında cronbach's alfa güvenirlik analiz sonucu

WEKA paket programında eğitim amaçlı olmak üzere 102 öğrencinin anket bilgileri girilerek en doğru ve güvenilir bir karar ağacı oluşturmak için J48 algoritması uygulanmıştır(Akbulut vd., 2008).

Elde edilen analiz sonucunda ağaç görünümünde hem grafiksel hem de metinsel olarak bilgiler elde edilmiştir. Elde edilen bilgilere göre toplam 21 adet yaprak ve 41 adet de ağaç dalları oluşmuştur.

Çizelge 4'de ise test verileri ve sonuçları gösterilmiştir.

Çizelge 4. Test verileri ve Sonuçları

Sıra_no	s1	s2	s3	s4	s5	s6	s7	s8	s9	s10	Sonuc

1	3	4	1	1	0	0	0	2	2	2	*K*
2	1	1	3	1	0	0	0	1	1	1	*K*
3	3	3	2	2	2	2	2	3	3	3	*K*
4	3	3	2	1	2	4	1	3	3	2	*K*
5	3	3	2	1	3	4	2	3	3	2	*K*
6	2	3	2	0	0	0	1	3	4	4	*K*
7	2	4	2	3	2	1	1	3	2	3	*K*
8	4	4	2	1	2	3	2	3	1	4	*K*
9	4	2	2	2	1	2	1	2	1	0	K
10	0	4	1	1	0	0	0	3	4	0	*G*
11	4	3	0	1	0	0	3	1	3	3	*K*
12	2	4	3	1	0	0	0	0	4	2	*G*
13	4	4	0	0	0	0	0	0	4	0	*K*
14	4	4	0	0	0	0	0	0	0	0	G
15	2	3	1	1	0	0	2	0	3	1	*G*
16	4	3	1	0	0	0	1	1	2	3	K
17	3	2	0	2	0	0	0	0	0	0	*K*
18	4	3	2	1	0	0	4	4	4	1	*G*
19	3	2	2	1	1	2	3	3	0	3	*K*
20	4	3	1	3	1	1	2	2	3	2	K
21	1	4	1	1	2	0	2	1	3	2	*K*
22	4	4	1	1	2	0	2	3	4	2	K
23	4	4	0	2	0	0	0	0	4	0	*K*
24	3	4	4	1	1	1	2	3	4	0	*G*
25	3	3	1	1	2	2	2	3	2	3	*K*

```
|   |   |   |   |   |   |   |   |   s8 > 1: K (2.0)
|   |   |   |   |   |   |   |   s10 > 2: K (3.0)
|   |   |   |   |   |   |   s1 > 3: G (3.0)
|   |   |   |   s6 > 1: G (6.0/1.0)
|   |   |   s7 > 3: G (3.0)
|   |   s3 > 1
|   |   |   s8 <= 2
|   |   |   |   s9 <= 3: K (5.0/1.0)
|   |   |   |   s9 > 3
|   |   |   |   |   s7 <= 2: G (6.0)
|   |   |   |   |   s7 > 2
|   |   |   |   |   |   s10 <= 3: K (3.0)
|   |   |   |   |   |   s10 > 3: G (2.0)
|   |   |   s8 > 2: K (4.0)
s3 > 2
|   s2 <= 2: K (3.0/1.0)
|   s2 > 2: G (19.0/2.0)

Number of Leaves  :     21

Size of the tree :      41
```

Classifier output

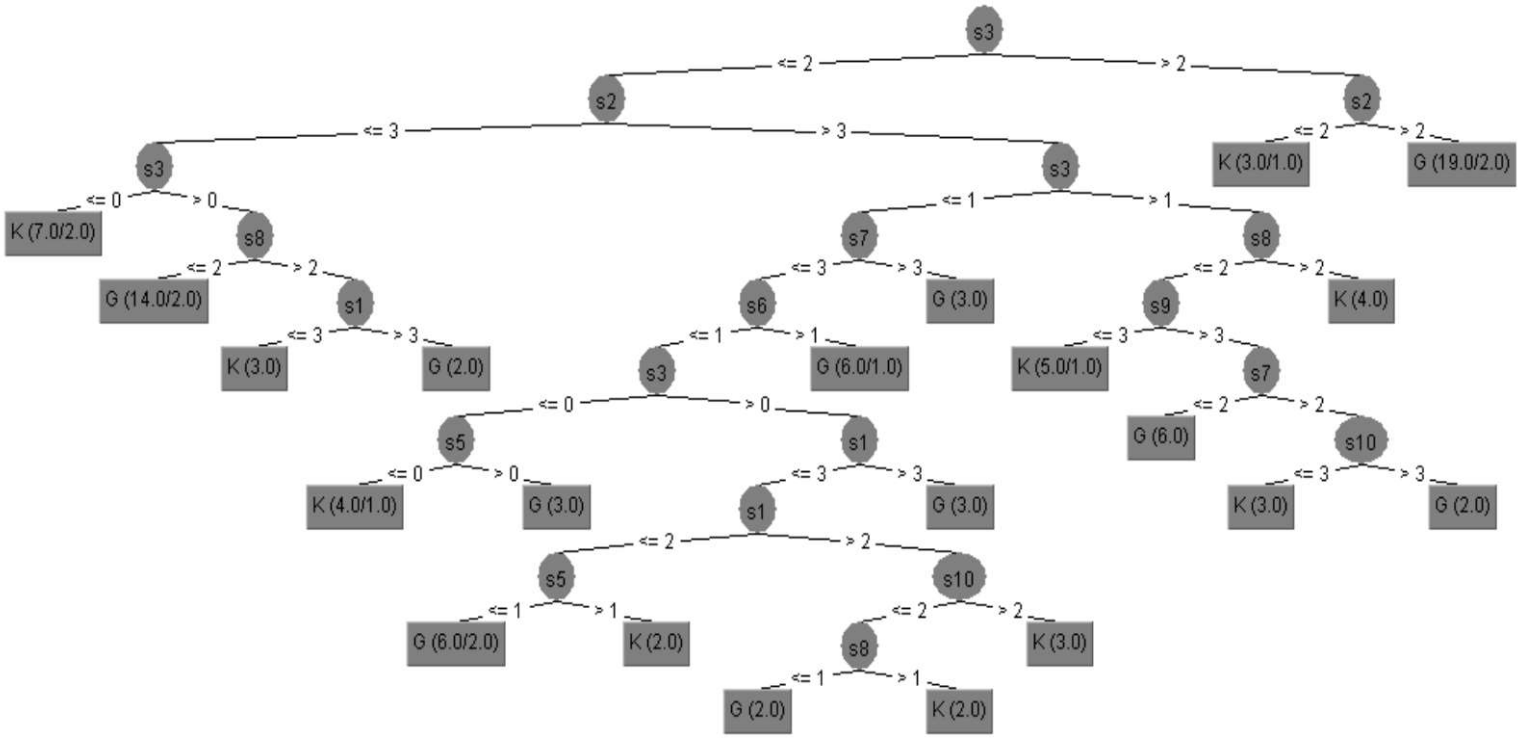
```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     egtmverileri_soncnominal
Instances:    102
Attributes:   11
              s1
              s2
              s3
              s4
              s5
              s6
              s7
              s8
              s9
              s10
              sonuc
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
s3 <= 2
|   s2 <= 3
|   |   s3 <= 0: K (7.0/2.0)
|   |   s3 > 0
|   |   |   s8 <= 2: G (14.0/2.0)
|   |   |   s8 > 2
|   |   |   |   s1 <= 3: K (3.0)
|   |   |   |   s1 > 3: G (2.0)
|   s2 > 3
|   |   s3 <= 1
|   |   |   s7 <= 3
|   |   |   |   s6 <= 1
|   |   |   |   |   s3 <= 0
|   |   |   |   |   |   s5 <= 0: K (4.0/1.0)
|   |   |   |   |   |   s5 > 0: G (3.0)
|   |   |   |   |   s3 > 0
|   |   |   |   |   |   s1 <= 3
|   |   |   |   |   |   |   s1 <= 2
|   |   |   |   |   |   |   |   s5 <= 1: G (6.0/2.0)
|   |   |   |   |   |   |   |   s5 > 1: K (2.0)
|   |   |   |   |   |   |   s1 > 2
|   |   |   |   |   |   |   |   s10 <= 2
|   |   |   |   |   |   |   |   |   s8 <= 1: G (2.0)
```

Şekil 12. Verilerin WEKA programında J48 karar ağacı metinsel analiz sonucu



Şekil 13. Verilerin WEKA programında J48 karar ağacı grafiksel analiz sonucu

2. SONUÇ VE ÖNERİLER

Ön lisans düzeyinde bilgisayar laboratuarlarında eğitim gören öğrencilerin ders başarı durumları "G/K" analizi sonucu önemli bilgiler elde edilmiş ve şu şekilde özetlenebilir:

- Eğitim verilerinden elde edilen J48 sınıflandırma algoritmasına göre; eğitim verisi olan 102 öğrenci haricinde, 25 kişilik test verisi öğrenci grubunda 20 öğrencinin (Çizelge 4 de sınıf niteliği olan sonuç alanında italik biçimdeki kayıtların) dersten geçme durumu doğru tahmin edilmiştir.
- Eğitim veri setindeki 102 kişilik öğrenci grubundan azınlığı (25 öğrenci) anketteki üçüncü soruya "Bilgisayardaki yazılım eksiklikleri (eksik yada hatalı kurulmuş program parçaları,vs) yüzünden uygulama yapamıyorum " katılıyorum/tamamen katılıyorum demiştir. Bu da bilgisayar laboratuarındaki bilgisayar yazılım eksikliklerinin çok fazla olmadığını gösterir.
- Eğitim veri setinde, anketteki sekizinci soruya "Ders içinde uygulama yapılırken öğrencilerin bireysel hız farklılıkları dersten kopmama neden oluyor" kararsızım/katılıyorum/tamamen katılıyorum şeklinde 2 veya üzeri işaretleyenlerin çoğunluğu desten kalmıştır. Bu da bilgisayar laboratuarlarında uygulamalı derslerde bireysel hız farklılıklarının öğrencilerin dersten geçmelerinde önemli olduğunu ortaya çıkarır.
- İkinci soruya "Bilgisayar üzerinde uygulama yapıldığında, anlatılan konuları daha iyi anlıyorum" verilen cevapların çoğunluğu 96(%94), 3 "katılıyorum" ve üzeri olduğundan öğrenciler derste uygulamalara daha ağırlık verilmesinin başarılarında önemli olduklarını düşünmektedirler. Bu 96 öğrenciden 61 öğrenci (%63)'ü ise dersten geçmiştir.
- Altıncı soruda "Anlatılan ders içeriği bilgisinin uygulamaya yönelik olmadığını düşünüyorum" 0 yani "hiç katılmıyorum" seçeneğini işaretleyen 49 öğrencinin 47 tanesi (%96) ikinci soruya 3 ve üzeri cevap vererek ders müfredatının uygulamaya yönelik olduğunu ve ders içinde daha çok uygulama yapılmasının başarılarını arttırtacağını belirtmişlerdir.
- Eğitim verisinde ankete katılan öğrencilerin hemen hemen yarısı (%47) yedinci soru "Haftada 80 dakikalık ders saati uygulama yapmama yeterli olmuyor" için 0 veya 1 "hiç katılmıyorum/katılmıyorum" işaretleyerek ders süresinin uygulama yapılması için yeterli olduğu görüşündeler.
- Dokuzuncu soruda "Değerlendirme biçiminin yazılı olarak değil de, bilgisayar üzerinde uygulama olarak yapılması başarımı daha da artırır" 2 ve üzeri seçeneği işaretleyen %70'lik bir öğrenci grubu içinden %63'lük kesimi dersten başarılı olarak mevcut klasik sınavın yerini uygulama sınavının alması istemektedirler.
- Ankette yer alan mevcut bilgisayar ve bilgisayar laboratuarının teknik-yazılım ve donanım anlamında yetersizliğinin söz konusu olmadığı eldeki verilerden ortaya çıkmıştır. Üçüncü "Bilgisayardaki yazılım eksiklikleri (eksik yada hatalı kurulmuş program parçaları,vs) yüzünden uygulama yapamıyorum", dördüncü "Bilgisayardaki donanım eksiklikleri (mause un olmaması, klavye ,monitör yada kasanın çalışmaması,vs) yüzünden uygulama yapamıyorum" ve beşinci soruları "Bilgisayar laboratuarının fiziksel altyapı (bilgisayar sayısının azlığı, elektriksel güç sorunları, masa-sandalye yetersizliği gibi) nedenlerinden dolayı derse karşı motive olamıyorum" 3 "katılıyorum" ve üzeri işaretleyenlerin oranları ortalaması %23 civarıdır.
- Ankete katılan öğrencilerin verilerine dayanarak; bilgisayar laboratuarlarında derslerin daha çok uygulamaya dayalı olmasının öğrenci başarılarını arttıracağı ve bunun için de 80 dakikalık normal ders süresinin yeteceği kanısı ortaya çıkmaktadır.
- Test veri setinde ortaya çıkan %20'lik hata payının giderilmesi için de veri setinde bulunan 102 öğrenci grubunun daha da artırılması ve karar ağacının daha da büyütülmesi gerekmektedir.

KAYNAKLAR

Fayyad U., Symth P., (1997), "Data Mining and KDD: Promise and Challenges," Future Generation Computer Systems, No:13, s.102.

Jeffry W. S., (2004), "Data Mining and the Search for Security: Challenges for Connecting the Dots and Databases," Government Information Quarterly, No:21, s.463.

Fayyad U., Piatetsky- Shapiro G., Smyth P., (1996), "From Data Mining to Knowledge Discovery in Databases", Al Magazine, cilt 17, sayı 3, s.41.

Yalçıntaş G., (2003), "Veri Madenciliği", Master tezi, Gazi Üniversitesi Fen Bilimleri Enstitüsü, Ankara.

Fayyad U., Piatetsky- Shapiro G., Smyth P., (1996), "The KDD process for extracting useful knowledge from volumes of data.", Communications of ACM, 39(11), s. 27-34.

Akpınar H., (2000), "Veri tabanlarında bilgi keşfi ve veri madenciliği", İ.Ü. İşletme Fakültesi Dergisi, C:29, sayı:1/ s.1-22 .

Gürsoy U.T.Ş., (2009), "Veri madenciliği- Veri madenciliği modelleme yöntemlerine genel bir bakış-İstatistiksel tahmin modelleri", Veri madenciliği ve bilgi keşfi, Pegem akademi, Ankara, s.27-90, 91-112.

Silahtaroğlu G., (2008), "Veri madenciliği – Sınıflandırma teknikleri ve algoritmalar", Kavram ve algoritmalarıyla temel veri madenciliği, Papatya yayın, İstanbul, s.9-82.

Cabena P., (1998), Discovering data mining:From concept to implementation, International business machines corporation, USA, s.12.

Yarımağan Ü., (2000), Veri tabanı sistemleri, Akademi&Türkiye bilişim vakfı, Ankara, s.7-9.

Aydoğan E.K., (2008), Veri madenciliğinde sınıflandırma problemleri için Evrimsel algoritma tabanlı yeni bir yaklaşım:Rough-Mep algoritması, Doktora tezi, Gazi Üniversitesi, Fen Bilimleri Enstitüsü, Ankara.

Alpaydın E., (2000), Zeki veri madenciliği: Ham veriden altın bilgiye ulaşma yöntemleri, Bilişim 2000 Eğitim semineri, İstanbul, s.1-10.

Şen F., (2008), "Veri madenciliğine genel bir bakış ", Veri madenciliği ile birliktelik kurallarının bulunması, Master tezi, Sakarya Üniversitesi Fen Bilimleri Enstitüsü, Sakarya.

Özkan Y., (2008), "Karar ağaçları ile sınıflandırma-Sınıflandırma ve regresyon ağaçları-Kümeleme-Birliktelik kuralları", Veri madenciliği yöntemleri, Editör: R.Çölkesen, Papatya yayın, İstanbul, 51-166.

Han J., Kamber M., (2006), "Data mining:Concepts and techniques", Morgan Kaufmann Publishers.

Quinlan J.R., (1993), "C4.5":,Programs for machine learning, Morgan Kaufmann.

Chiu T., Wei C., (2002), "Turning telecommunications call details to churn prediction: a data mining approach", Expert systems with applications, 23, s.103-102.

Mitchell T., (1997), Machine learning, McGraw-Hill International, London, s.52-81.

Fiske J., (1998), Introduction to communication studies, y.y., Routledge.

Dunham M.H., (2003), Data mining ıntroductory and advanced topics, Pearson education inc., Prentice Hall, New Jersey, s.8.

Shannon C.E., (1948), A Mathematical theory of communication, The Bell system technical journal, vol.27, s.379-423, 623-656.

Quinlan J.R., (1986), "Induction of decision trees", Machine learning , C:1, s.81-106.

Kantardzic M., (2003), "Concepts,methods and algorithms", Data mining, Wiley.

Akbulut S., Aydoğan E.K., Gencer C., (2008), "Veri madenciliği teknikleri ile bir kozmetik markanın ayrılan müşteri analizi ve müşteri bölümlenmesi", Sigma Mühendislik ve Fen Bilimleri Dergisi, cilt:26, sayı:1, s.42-56.

Webb P., Yohannes Y., (1999), "Classification and Regression Trees CART", A User manuel for indentifying indicators of vulnerability to famine and chronic food insecurity, y.y., International food policy research institute, s.15.

Agrawal R., Manish M., Rissanen J., (1996), "SLIQ:A fast scalable classifier for data mining", 5th international extending database technology conf., Aningnon France.

Rissanen J., (1989), Stochastic complexity in statistical inquiry, World Scientific Publication.

Shafer J.C., (1996), "SPRINT: A Scalable Parallel Classifier for Data Mining", 22th international conference of very large databses, Bombay India.

Brieman L., (1984), Classification and Regression Trees, Wadsworth, Belmont.

Aryeetey K., (2003), "Data analysis and predictive modelling using the variable precision rough set approach", Master tezi, Faculty of Graudate Study and Research of University of Regina, Canada, s.28-33,.

Beyer K., (1999), "When is 'Nearest Neighbor' meaningful?", 7th international database theory conference,(ICDT'99), Israel, s.217-235.

Khan M., (2002), "K-Nearest Neighbor Classification on Spatial Data Streams Using P-Trees",6th Pacific Asia Knowledge discovery and Data Mining conferance (PAKKDD'02), Taiwan, s.517-518.

Pryke A.N., (1998), "Data mining using genetic algorithms and interactive visualization", PhD ThesisFaculty of Sicence, University of Birmingham, Birmingham, s.187.

Altınışık U., (2006), "Öğrenci bilgi sisteminde veri madenciliğinin uygulanması", yayınlanmamış master tezi, Kocaeli Üniversitesi Fen Bilimleri Enstitüsü, Kocaeli, s.12-17.

Ghosh A., Nath B., (2004), "Multi objective rule mining using genetic algorithms", Information sciences, cilt:163, sayı1-3, s.123-133.

Kumar V., Steinbach M., Tan P.N., (2006), "Introduction to data mining", Addison Wesley.

Yılmaz Ş.K., (2008), "Veri madenciliği: İstanbul Menkul Kıymetler Borsası Örneği", Master tezi, Zonguldak Karaelmas Üniversitesi Sosyal Bilimler Enstitüsü, Zonguldak , s.21.

Christensen M., Hermiz K., Manganaris S., Zerkle D., (2000), "A data mining

analysis of RTID alarms", Computer Networks34, s.571-577.

Bing L., (2007), "Web data mining: Exploring Hyperlinks, Contents and Usage Data", ISBN-10 3-540-37881-2, Springer-Verlag Berlin Heidelberg.

Güvenç E., (2001), "Student performance assesment in higher education using data mining", Master tezi, Boğaziçi Üniversitesi, İstanbul, s.5-14.

Hudairy H., (2004), "Data mining and decision making support in the governmental sector", Faculty of Graduate School of The University of Louisville, Kentucky.

Soğukpınar İ., Takçı H., (2002), "Kütüphane kullanıcılarının erişim örüntülerinin keşfi", Bilgi dünyası dergisi, Cilt:3, sayı:1, s.12-26.

Delioğlu S., Dolgun M.Ö., Özdemir T.G., (2007) , "Öğrenci Seçme Sınavında (ÖSS) öğrencilerin tercih profillerinin veri madenciliği yöntemleriyle tespiti", Bilişim 07 kongresi, Ankara.

İnternet(Web) sayfası: Polls : Current data mining applications/industries http://www.kdnuggets.com/polls/2003/data_mining_applications_industries.htm, Erişim tarihi: 11.07.2011.

Bindak R., Çelik H.C., (2005), "İlk öğretim okullarında görev yapan öğretmenlerin bilgisayara yönelik tutumlarının çeşitli değişkenlere göre incelenmesi", İnönü Üniversitesi Eğitim Fakültesi Dergisi, Cilt 6, sayı 10, s.27-38.

Ertekin A.R., Tekindal B., Tekindal M.A., (2010), "Meslek liselerinde eğitim gören Öğrencilerin bilgisayara yönelik tutumlarının değerlendirilmesi (Yozgat ili Yerköy ilçesi örneği)", Gazi Üniversitesi Bilişim Teknolojileri Dergisi, cilt:3, sayı:1, s.23-30.