VERİ MADENCİLİĞİ VE ANADOLU ÜNİVERSİTESİ UZAKTAN EĞİTİM SİSTEMİNDE BİR UYGULAMA

Sinan AYDIN

DOKTORA TEZİ

İşletme Anabilim Dalı

Danışman: Prof. Dr. Ali Ekrem ÖZKUL

Eskişehir

Anadolu Üniversitesi Sosyal Bilimler Enstitüsü

Eylül 2007

DOKTORA TEZ ÖZÜ

VERİ MADENCİLİĞİ VE ANADOLU ÜNİVERSİTESİ UZAKTAN EĞİTİM SİSTEMİNDE BİR UYGULAMA

Sinan AYDIN

İşletme Anabilim Dalı

Anadolu Üniversitesi Sosyal Bilimler Enstitüsü, Eylül 2007

Danışman: Prof. Dr. Ali Ekrem ÖZKUL

Veri madenciliğinin uygulandığı birçok alanda olduğu gibi eğitimde de anlamlı ilişkilerin araştırılabileceği ve faydalı bilginin türetilebileceği geniş veritabanları mevcuttur. Tez kapsamında Anadolu Üniversitesi Uzaktan Eğitim Sisteminde eğitim gören öğrencilere ilişkin farklı kaynaklardaki veriler bir araya getirilerek veri madenciliği uygulaması gerçekleştirilmiştir. Uzaktan Eğitim Sisteminin planlama faaliyetlerine katkı sağlayabilecek öğrenci performansını tahmin etmeye yönelik model geliştirilmiş ve mezun olan öğrencilerin profillerini belirlemeye yönelik kümeleme çalışması yapılmıştır. Öğrenci başarısını tahmin etmeye yönelik çalışmada C5.0 karar ağacı algoritmasının kullanıldığı bir tahmin modeli önerilmiştir. Önerilen modelin karar kuralları sisteme entegre edilerek öğrenci başarı tahmini amacıyla kullanılabileceği öngörülmektedir. Mezun olan öğrencilere yönelik çalışmada “K-means” algoritması kullanılarak beş küme elde edilmiştir. Kümeleme analizi ile elde edilen bilgilerin bilgisayar kullanımı ve öğrenci başarısı arasındaki ilişkiyi doğrular nitelikte olduğu görülmüştür.

Açıköğretim öğrencilerine ilişkin veritabanındaki kısıtlı veriler üzerinde yapılan çalışmalar sonucunda veri madenciliğinin internet üzerinden uzaktan eğitim sistemleri için önemli bir karar destek aracı olma özelliği kanıtlanmıştır.

ABSTRACT

Nowadays, like in the other fields of study where data mining is applied, large databases are also available in education, from which meaningful relationships could be discovered and useful knowledge extracted. Within the scope of this PhD thesis, data mining is conducted by conjoining data from different sources about the students in Anadolu University Distance Education System. A model has been prepared to predict the student performance which could contribute planning activities of Distance Education System and clustering study has been made to determine the profiles of the graduates. A prediction model has been proposed in which C5.0 decision tree algorithm is used for predicting student achievement. It has been foreseen that the system could be used for the prediction of student achievement by integrating the decision rules of the proposed model. In result of the clustering, five clusters have been obtained by the use of “K-means” algorithm. According to these clusters, it has been seen that they verify the relationship between the computer usage and student achievement.

In the result of studies performed on the limited data in distance education students’ database, it has been proven that data mining is an important decision support tool for internet-based distance education systems.

JÜRİ VE ENSTİTÜ ONAYI

Sinan AYDIN'ın Veri Madenciliği Ve Anadolu Üniversitesi Uzaktan Eğitim Sisteminde Bir Uygulama başlıklı tezi 25.01.2008 tarihinde, aşağıdaki jüri tarafından Lisansüstü Eğitim Öğretim ve Sınav Yönetmeliğinin ilgili maddeleri uyarınca, İşletme Anabilim dalında Doktora tezi olarak değerlendirilerek kabul edilmiştir.

		Adı Soyadı	**İmza**
Üye (Tez Danışmanı)	:	Prof. Dr. Ali Ekrem ÖZKUL	
Üye	:	Prof. Dr. Haldun AKPINAR	
Üye	:	Prof. Dr. Emel ŞIKLAR	
Üye	:	Doç.Dr. Ahmad BABANLI	
Üye	:	Yard.Doç.Dr. Servet HASGÜL	

Prof. Dr. Numan AYDIN
Enstitü Müdürü

ÖNSÖZ

Bu çalışmada danışmanlığımı yürüten ve tezin sonuçlandırılmasında yakın ilgi ve desteğini esirgemeyen hocam Sayın Prof. Dr. Ali Ekrem ÖZKUL'a saygı ve teşekkürlerimi sunarım.

Bu çalışma sırasında bana yol gösteren, eleştirileriyle katkıda bulunan Sayın Doç. Dr. Ahmad BABANLI ve Yard. Doç. Dr. Servet HASGÜL'e teşekkürlerimi sunarım.

Çalışmalarım süresince manevi desteği için sevgili eşime sonsuz teşekkürlerimi sunarım.

İÇİNDEKİLER

İKİNCİ BÖLÜM

EĞİTİMDE VERİ MADENCİLİĞİ

ÜÇÜNCÜ BÖLÜM

ANADOLU ÜNİVERSİTESİ UZAKTAN EĞİTİM SİSTEMİ ÖĞRENCİ VERİLERİNDE VERİ MADENCİLİĞİ UYGULAMALARI

TABLOLAR LİSTESİ

ŞEKİLLER LİSTESİ

TERİMLER DİZİNİ

Alıcı çalışma karakteristik grafiği	: Receiver operating characteristic
Aykırı değer	: Outlier
Basit geçerlilik	: Simple validation
Benzeşme analizi	: Affinity analysis
Biçimlendirici değerlendirme	: Formative evaluation
Birliktelik kuralları	: Association rules
Bölmeleme	: Binning
Çapraz geçerlilik	: Cross validation
Denetimli öğrenme	: Supervised learning
Denetimsiz öğrenme	: Unsupervised learning
Eksik veri	: Missing data
Enformasyon	: Information
Genelleştirme	: Generalization
Görselleştirme	: Visulation
Gürültülü veri	: Noisy data
Kaldıraç grafiği	: Lift chart
Karar ağacı	: Decision tree
Kestirim	: Prediction
Katışıklık ölçümü	: Impurity measure
Kümeleme	: Clustering
Makine öğrenmesi	: Machine learning
Metin madenciliği	: Text mining
Normalleştirme	: Normalization
OLAP	: On-line analytical processing
Örüntü tanımlama	: Pattern recognitions
Özellik oluşturma	: Attribute (feature) construction
Özetleme	: Summurazation
Pazar sepeti analizi	: Market basket analysis
Sınıflama	: Classification
Sıra örüntüleri	: Sequence discovery

Standartlaştırma	: Standardization
Tutarsız veri	: Inconsistent data
Veri bütünleştirme	: Data integration
Veri dönüştürme	: Data transformation
Veri madenciliği	: Data mining
Web içerik madenciliği	: Web content mining
Web kullanım madenciliği	: Web usage mining
Web yapı madenciliği	: Web structure mining
Yapay sinir ağları	: Artificial neural networks

GİRİŞ

Günümüzde bilgisayar kullanımı ve internet erişiminin yaygınlaşması hem çok çeşitli işlemlerin kaydedilmesini hem de bunların manyetik ortamda saklanmasını kolay ve ucuz hale getirmiştir. Teknolojinin sağladığı olanaklar yoluyla büyük miktar ve çeşitteki veriler farklı özelliklere sahip veritabanı yönetim sistemlerinde depolanmaktadır. Yönetsel kararlar için anlamlı olabilecek kaynakları oluşturan bu verileri inceleme ve analiz etmede kullanılan yöntemler ve basit araçlar verilerin bir çığ gibi büyümesi ve karmaşık hale gelmesiyle yetersiz kalmış, yeni yöntem ve teknolojilerin geliştirilmesi ihtiyacının ortaya çıkmasına neden olmuştur. Bu ihtiyaçlar doğrultusunda, bilgi teknolojileri, istatistik, makine öğrenmesi, veritabanı teknolojileri ve ilgili diğer disiplinleri bir araya getiren veri madenciliği yaklaşımı yeni bir veri analiz yöntemi olarak ortaya çıkmıştır.

Veri madenciliği "kullanıcılara yeni yöntemlerle anlaşılabilir ve faydalı olan verileri özetlemek ve aralarındaki beklenmeyen ilişkileri bulmak için genellikle büyük gözlemsel veri kümelerinin analiz edilmesi" olarak tanımlanmaktadır[1]. Tanımdaki "gözlemsel veri" kavramı "deneysel veri" kavramının tam tersini ifade etmektedir. Veri madenciliği, gerçekleştirilen işlemin doğası gereği toplanan verilerle ilgilidir. Veri madenciliğini istatistikten ayıran bir özellik, veri madenciliği uygulamalarının veri toplama stratejisi üzerinde rol oynamamasıdır.

Veri madenciliğinde uygulanan modelleme teknikleri tahmin etmeye ve tanımlamaya yönelik modeller olarak iki kategoriye ayrılır. Tahmin edici modeller, sonucu bilinen verilerden hareketle sonucu bilinmeyen veriler üzerinde tahmin üretmeyi amaçlarken tanımlayıcı modeller ise veritabanındaki verinin genel özelliklerini tanımlarlar.

Veri madenciliği uygulamalarında öncelikle incelenen sistemle ilgili problem tanımlanır. Tanımlanan probleme ilişkin veri madenciliği veritabanının

[1] David Hand, H. Manila ve P. Smyth**, Principles of Data Mining** (USA: MIT Press, 2001), s.1.

oluşturulması için birçok farklı kaynaktan toplanan verilerin bir araya getirilmesi gerekmektedir. Verilerin veri madenciliğine uygun hale getirilmesi, eksiklerinin ya da hatalarının araştırılarak giderilmesi için veri temizleme, veri bütünleştirme, veri dönüşümü ve veri azaltma gibi ön hazırlık süreçleri uygulanır.

Verinin özelliklerinin anlaşılması ve uygun veri analiz tekniğinin seçilmesinde özet istatistikler ve görselleştirme tekniklerinden faydalanılır. Veri madenciliği farklı görevleri yerine getirmek amacıyla pek çok farklı algoritmayı kullanır. Algoritmalar veriyi inceler ve incelenen verinin özelliklerine en uygun modeli belirler. Veri madenciliğinde geliştirilen modeller sınıflama, kümeleme, birliktelik kuralları oluşturma, örüntü tanımlama gibi görevlerin yerine getirilmesinde kullanılırlar.

Veri madenciliğinin uygulandığı birçok alanda olduğu gibi eğitimde de anlamlı ilişkilerin araştırılabileceği ve faydalı bilginin türetilebileceği geniş veritabanları mevcuttur. Eğitim alanındaki veri madenciliği çalışmaları eğitim sistemlerinde yer alan veritabanlarında öğrencilere, akademik sorumlulara ve eğitimcilere faydalı olabilecek henüz keşfedilmemiş bilginin mevcut olduğu olgusundan yola çıkmaktadır. Veri madenciliği çalışmalarının günümüzde giderek yaygınlaşan ve e-öğrenme olarak adlandırılan elektronik ortamlarda öğrenme alanında geniş uygulama potansiyeli bulmaktadır. Öğrencilerin e-öğrenme ortamlarındaki davranışları ve öğrenme faaliyetlerinin sonucunda oluşan web hareketleri, e-öğrenme sistemlerinin daha etkin bir yapıya kavuşturulmasını hedefleyen veri madenciliği çalışmalarının kaynak verilerini oluştururlar.

Bu çalışmada veri madenciliği konusu ele alınarak Anadolu Üniversitesi Uzaktan Eğitim Sisteminde eğitim gören öğrencilere ilişkin farklı kaynaklardaki verilerin bir araya getirildiği veri madenciliği çalışması yapılarak Uzaktan Eğitim Sisteminin planlama faaliyetlerine katkı sağlayabilecek öğrenci performansını tahmin etmeye yönelik bir model önerilmiştir. Ayrıca mezun öğrencilerin profillerini çıkarmaya yönelik kümeleme çalışması gerçekleştirilmiştir.

BİRİNCİ BÖLÜM

VERİ MADENCİLİĞİ

1. VERİ MADENCİLİĞİNE GİRİŞ

İlk bilgisayarların üretilmesiyle başlayan manyetik ortamda veri saklama süreci veri depolama teknolojilerindeki hızlı gelişim sonucunda günümüzde çok büyük miktar ve çeşitteki verinin depolanmasına olanak sağlar duruma gelmiştir. Market alışveriş verileri, kredi kartı kullanım verileri, telefon görüşme detayları ve bankacılık işlemleri gibi günlük yaşamda oluşturulan verilerin yanında uydu gözlemlerinden elde edilen veriler, tıbbi veriler ve bilimsel araştırmalarda depolanan çok çeşit ve özellikteki veriler farklı amaçlar için geliştirilmiş veritabanı yönetim sistemleri sayesinde manyetik ortamlarda saklanabilmekte ve yönetilmektedir. Diğer taraftan devlet kuruluşları ve özel işletmelerde oluşturulan veritabanları da kuruluşların faaliyetlerini sürdürmeleri için gerekli olan verileri depolamaktadırlar.

Depolanan verileri inceleme ve analiz etmede kullanılan yöntemler ve basit araçlar, verilerin bir çığ gibi büyümesi ve karmaşık hale gelmesiyle yetersiz kalmış, yeni yöntem ve teknolojilerin geliştirilmesi ihtiyacı ortaya çıkmıştır. Enformasyon teknolojileri, istatistik, makine öğrenmesi, veritabanı teknolojileri ve ilgili diğer disiplinlerdeki teknikleri bir araya getiren veri madenciliği bu ihtiyacı gideren yeni bir veri analiz yöntemi olarak ortaya çıkmıştır. Bu yeni yöntem büyük veritabanlarından anlamlı örüntülerin otomatik olarak keşfedilmesi amacıyla kullanılmaktadır[2].

1.1. Veri Madenciliğinin Tanımı

"Veri tabanlarında bilgi keşfi" ve "veri madenciliği" literatürde birbirine yakın anlamlarda kullanılmaktadır. Usama M. Fayyad'a göre "veritabanlarında

[2] Pang-Ning Tan, M. Steinbach ve V.Kumar, **Introduction to Data Mining** (USA: Pearson Education, 2006), s.2.

bilgi keşfi" veriden faydalı bilgiyi keşfetme süreci, "veri madenciliği" ise veriden örüntülerin çıkarılması için algoritmaların uygulanması olarak tanımlanır[3]. Veri madenciliği hedeflenen sonuçları elde edebilmek için, analiz edilmek üzere hazırlanmış verilere algoritmaların uygulandığı bilgi keşif sürecinin adımı olarak görülmektedir. Bununla beraber endüstride, medya ve veritabanı araştırmalarında "veri madenciliği" terimi "veritabanlarında bilgi keşfi" teriminden daha yaygın olarak kullanılmaktadır. Bu nedenle sürecin tamamı genellikle veri madenciliği olarak anılmaktadır.

Bu alanda çalışan araştırmacılar tarafından veri madenciliğinin farklı tanımları yapılmıştır:

Veri madenciliği;

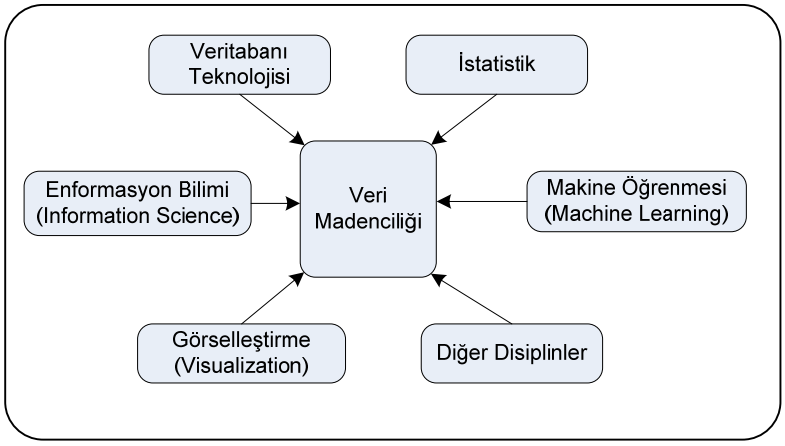
- kullanıcılara yeni yöntemlerle anlaşılabilir ve faydalı olan verileri özetlemek ve veriler arasındaki beklenmeyen ilişkileri bulmak için büyük ölçekli gözlemsel veri kümelerinin analiz edilmesidir[4],
- geçerli tahminler yapmak için kullanılan verilerdeki örüntüleri ve ilişkiyi açığa çıkarmak için çeşitli veri analiz araçlarını kullanan süreçtir[5],
- büyük veritabanlarında gizli ilişkilerin ve genel örüntülerin araştırılmasıdır[6],
- verilerden anlamlı örüntülerin otomatik veya yarı otomatik olarak keşfedilme sürecidir[7],
- veritabanında yer alan verilerden bilginin otomatik olarak çıkarılması ve analiz edilmesinde bir veya daha fazla bilgisayar öğrenme tekniklerinin uygulanması sürecidir[8].

---

[3] Usama M. Fayyad ve diğerleri, **Advances in Knowledge Discovery and Data Mining** (USA: MIT Press, 1996), s.4.
[4] Hand, **Ön.ver.**, s.1.
[5] Two Crows Corp., **Introduction to Data Mining and Knowledge Discovery** (Versiyon 3: www.trocrows.com, 1999), s.1.
[6] M. Holsheimer ve A. Siebes, **Data Mining: The Search for Knowledge in Databases** (CWI Technical Report, Amsterdam: 1994), s.2.
[7] Ian H. Witten ve E. Frank, **Data Mining** (USA: Elsevier Inc., 2005), s.5.

Genellikle büyük miktarlardaki verilerden faydalı ve gizli bilgilerin ortaya çıkarılması olarak tanımlanan veri madenciliği cevher elde etmek için yapılan madenciliğe benzetilmektedir. Örneğin altın madenciliğinde de tonlarca hammadde ayrıştırılarak saf altın elde edilmektedir. Veri madenciliğinde hammadde veri, maden yani ürün ise bilgi olmaktadır. Ürünün bilgi olmasına rağmen sürecin bilgi madenciliği değil de veri madenciliği ismini almasının nedeni verinin büyüklüğünü vurgulamaktır.

Veri madenciliği Şekil 1'de görüldüğü gibi veritabanı sistemleri, istatistik, makine öğrenmesi, görselleştirme ve enformasyon bilimini içeren disiplinler arası bir alandır.



Şekil 1. Veri Madenciliğini Oluşturan Disiplinler.

Jiawei Han ve M. Kamber, **Data Mining** (USA: Academic Press, 2001)den uyarlandı.

Veri madenciliği işlevlerinde teoriye dayalı modellerin oluşturulması, verideki gürültü ve eksik değer sorunlarının giderilmesi ve verinin anlaşılmasında istatistik bilimine dayalı tekniklerden faydalanılmaktadır. Veri madenciliği uygulamalarının birçoğunda veri kaynağı olarak veritabanı yazılımları kullanılmaktadır. Veritabanı yazılım sistemleri verileri depolamanın yanında verileri ilişkilendirmek, özetlemek, çok boyutlu verileri işlemek gibi fonksiyonları yerine getirirler. Veri madenciliği için verinin hazırlanmasında veritabanı teknolojilerinden faydalanılır. Verinin anlaşılmasında ve örüntülerin

[8] Richard J. Roiger ve M. W. Geatz, **Data Mining** (USA: Pearson Education, 2003), s.4.

tanımlanmasında faydalanılan bir alan da görselleştirmedir. Görselleştirme verinin tablolar ve grafikler halinde görüntülenmesini sağlayan teknolojileri içerir. Veri madenciliği sistemleri analiz türüne ve verinin içeriğine bağlı olarak uzaysal veri analizi (spatial data analysis), örüntü tanımlama (pattern recognition), görüntü analizi (image analysis), sinyal işleme (signal processing), web teknolojisi, ekonomi, iş dünyası, biyoinformatik veya fizyoloji alanlarına ilişkin teknikleri bütünleştirebilir[9].

Veri madenciliği modellerinde diğer önemli bir kavram ise makine öğrenmesidir. Makine öğrenmesi aslında insanın öğrenmesine benzer bir yapıdadır. İnsanlar çocukluk döneminde temel kavram tanımlarını şekillendirmede tümevarım yöntemini kullanırlar. Hayvanlar, bitkiler, bina yapıları ve bunun gibi kavramları ifade eden örnekleri görürüz. Birey örneklere verilen isimleri işitir ve kavram özelliklerini tanımladığına inandığını seçerek sınıflama modelini oluşturur. Daha sonra bu modelleri benzer yapıdaki nesneleri tanımlamada kullanır. Bu tür öğrenme "tümevarıma dayalı denetimli kavram öğrenme" veya kısaca "denetimli öğrenme" (supervised learning) olarak tanımlanır[10]. Denetimli öğrenmeyle, girdi verilerinin değerleri kullanılarak çıktı değerleri tahmin edilmeye veya öğrenilmeye çalışılır. Bu süreçte öncelikle sonuçları bilinen veriler üzerinde bir sınıflama yapılır ve sonuçları bilinmeyen veri kümesi için sonuçlar tahmin edilmeye çalışılır. "Denetimsiz öğrenmede" (unsupervised learning) önceden tanımlanmış bir sınıfa ait olmayan verilerden model oluşturulur. Veri örnekleri, kümeleme sistemleri tarafından tanımlanan bir benzerlik taslağına göre gruplandırılır. Elde edilen kümelerin anlamı, bir veya birden çok değerlendirme tekniğinin yardımıyla kullanıcı tarafından belirlenir. Denetimsiz öğrenme modellerinde bir çıktı alanı söz konusu değildir.

Veri madenciliğinde farklı disiplinlerin kullanılması, veri madenciliği sistemlerinde özelleştirme gerektirmektedir. Bu nedenle veri madenciliği sistemlerinin sınıflandırılması yerinde olacaktır.

---

[9] Jiawei Han ve M. Kamber, **Data Mining** (USA: Academic Press, 2001), s.28.
[10] Roiger, **Ön.ver.**, s.8.

1.2. Veri Madenciliği Sistemlerinin Sınıflandırılması

Veri madenciliği sistemlerinin sınıflandırılması, potansiyel kullanıcıların kullanılabilecek yazılımları ve sistemleri ayırt etmelerini ve yeterli bir şekilde tanımlamalarına yardımcı olacaktır. Veri madenciliği sistemleri çeşitli ölçütlere göre sınıflandırılabilir[11].

- Veritabanına göre: Veritabanı yönetim sistemleri; veri modelleri, veri tipleri veya uygulama alanları gibi farklı özelliklere göre kendi içlerinde sınıflandırılırlar ve kendilerine özel veri madenciliği tekniklerinin uygulanmasını gerektirirler. Örneğin veri madenciliği sistemleri veritabanı modellerine göre sınıflandırıldığında; ilişkisel, harekete dayalı, nesneye dayalı, nesne-ilişkisel veya veri ambarı kategorileri ortaya çıkar. İşlenecek verilerin özel türde olması durumunda veri madenciliği sisteminin; uzaysal, zaman serileri, metin, çoklu ortam veya web madenciliği şeklinde sınıflandırılması gerekir.
- Bilgi türüne göre: Veri madenciliği sistemleri kümeleme, sınıflama, aykırı değer analizi gibi veri madenciliği işlevlerine göre sınıflandırılabilir. Kapsamlı bir veri madenciliği sistemi birden fazla işlevi gerçekleştirdiği gibi birden fazla işlevin bütünleştirildiği teknikleri de sunabilmektedir.
- Tekniklere göre: Veri madenciliği sistemlerini uygulanan belirli veri madenciliği tekniklerine göre sınıflamak mümkündür. Bu teknikler makine öğrenmesi, istatistik, örüntü tanımlama, yapay sinir ağları gibi uygulanan pek çok veri analiz metotlarına veya kullanıcının müdahale düzeyine göre tanımlanabilir. Kapsamlı bir veri madenciliği sistemi çoğu zaman çoklu veri madenciliği tekniklerini sağlayabilmeli veya bireysel yaklaşımları etkin bir şekilde sistemle bütünleştirebilmelidir.
- Uygulama alanına göre: Veri madenciliği sistemleri aynı zamanda uyarlandıkları alana göre de sınıflandırılabilir. Özellikle finans, iletişim, DNA, borsa, e-posta gibi alanlar için hazırlanmış sistemler mevcuttur. Bu

[11] Han, **Ön.ver.**, s.29.

nedenle genel amaçlar için tasarlanmış veri madenciliği sistemi özel bir alanda gerçekleştirilen madencilik çalışmasına uygun olmayabilir.

Veri madenciliği çalışmalarının ve sistemlerinin çok geniş bir alana yayılmasının ve farklılaşmasının en temel nedeni enformasyon teknolojilerinin hemen hemen tüm uygulamalarda kullanılması ve bunun sonucunda oluşan veri dağlarıdır.

1.3. Veri Madenciliği Uygulama Alanları

Veri madenciliği yeni bir disiplin olmasına karşın oldukça geniş bir alanda uygulanmaktadır. Bu yeni disiplin organizasyonlarda ve bilimsel araştırmalarda oluşturulan veriler üzerinde bilgi keşfi sürecini gerçekleştirdiğinden iş dünyası ve bilimin bazı alanlarında birçok problemin çözülmesinde etkin rol oynamıştır. Veri madenciliği aracılığıyla finans ve ekonomi, sağlık hizmetleri, güvenlik hizmetleri, sosyal hizmetler, e-devlet, eğitim, telekomünikasyon ve nakliye gibi alanlarda gerçekleştirilmiş başarılı uygulamalar bulunmaktadır. Veri madenciliğinin kullanılabileceği alanlar ve problem örnekleri Tablo 1'de özetlenmiştir[12].

Tablo 1. (a) Veri Madenciliğinin Uygulandığı Alanlar

Bankacılık ve Finans
Kredi kartı dolandırıcılığının belirlenmesi
En iyi müşterilerin belirlenmesi
Kredi kartı harcamalarına göre müşteri gruplarının belirlenmesi
Kredi kartını değiştirmesi muhtemel müşterilerin belirlenmesi
Farklı finansal göstergeler arasındaki gizli korelasyonun bulunması
Benzer davranışlar gösteren müşterilerin sınıflandırılması
Müşteri kredi taleplerinin değerlendirilmesi
Döviz fiyatlarındaki değişikliklerin önceden tahmin edilmesi
Vergi dolandırıcılığı vakalarının tespit edilmesi
Pazarlama
Çapraz satış amacıyla müşteri satın alma alışkanlıklarının belirlenmesi
Müşteri profillerinin belirlenmesi
Kaybedilen müşterilerin benzer özelliklerinin ortaya çıkarılması
Müşteri davranışlarındaki değişkenliklerin fark edilmesi
Karlı müşterilerin elde tutulması amacı ile profillerin belirlenmesi
Müşteri ihtiyaçlarının belirlenmesi

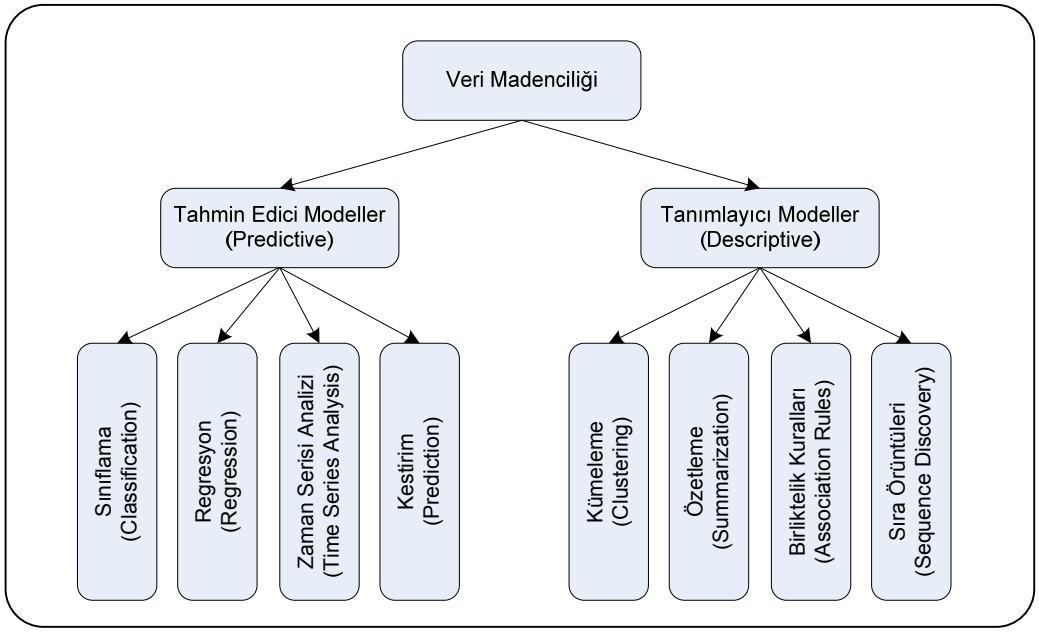
[12] Michael J. A. Berry ve G. Linoff, **Data Mining Techniques** (USA: John Wiley&Sons, 1997), s.10; Hamid R. Nemati ve C.D. Barko, **Organizational Data Mining** (USA: Idea Group, 2004), s.4; Parag C. Pendharkar, **Managing Data Mining Technologies in Organizations** (USA: Idea Group, 2003), s.27.

Tablo 1. (b) Veri Madenciliğinin Uygulandığı Alanlar (Devam)

Mühendislik çalışmaları
Örüntü tanımlama
Benzetim
Sinyal işleme
Sigortacılık
Riskli müşterilerin davranış örüntülerinin belirlenmesi
Sigorta dolandırıcılık vakalarının tespiti
Poliçelerini yenilemeyecek ve yeni poliçe alacak müşterilerin tahmin edilmesi
Sağlık
Cerrahi süreçlerin etkinliğinin belirlenmesi
Tedavi süresinin en aza indirilmesi
İlaç kullanım sahtekarlığının saptanması
Hastaların tüm tıbbi verileri kullanılarak hasta sağlık risklerinin tahmin edilmesi
Farklı hastalıklar için uygulanan tıbbi tedavi süreçlerinin etkinliğinin belirlenmesi
Biyomedikal ve DNA
DNA dizilimlerinin benzerliğini kıyaslama
Genetik veri ambarlarının oluşturulması
Birliktelik analizi ile gen gruplarının keşfi ve aralarındaki etkileşim ve ilişkilerin belirlenmesi
Genlerin hastalıkların farklı aşamalarındaki etkilerinin belirlenmesi
Biyomedikal verilerin anlaşılmasında görsel araçların kullanımı
İmalat
Ürün hataları, müşteri memnuniyet oranları gibi çıktı değişkenlerindeki sapmaların nedenlerinin belirlenmesi
Etkin kaynak kullanımı
Araştırma ve geliştirme faaliyetleri
İnternet
Web sayfalarını kullanan ziyaretçilerin sayfa içindeki davranışlarının analiz edilmesi
Ziyaretçilerin profilini belirleme
İnternet alışveriş siteleri kullanıcılarının satın alma alışkanlıklarının belirlenmesi
Web arama motorlarının web sayfaları içeriklerini kümelemek ve aralarındaki bağlantıları araştırmak için gerçekleştirdikleri analizler
Telekomünikasyon
Veri trafiği, sistem iş yükü, kaynak kullanımı, kullanıcı grup davranışları gibi verilerin tanımlanması ve karşılaştırılması
Arama zamanı, mekanı, süresi, aranılan bölge gibi boyutlara sahip olan telekomünikasyon verileri üzerinde örüntülerin keşfedilmesi
Dolandırıcılık yapan müşterilerin profil ve davranış örüntülerinin belirlenmesi
Görsel veri madenciliği tekniklerinin kullanımı
Eğitim
Öğrenci profillerine göre başarısını tahmin etme
Öğrencinin başka bir eğitim kurumuna geçme olasılığını tahmin etme
Zeki ölçme ve değerlendirme sistemleri için bilgi üretme
Benzer özelik gösteren örencileri gruplama
Öğrenme ortamlarının iyileştirilmesine yönelik gerekli araştırmaların yapılması
Daha etkin e-öğrenme ortamlarının tasarımı için web madenciliği uygulamaları
Diğer alanlar
Ulaşım güvenliği ve trafik düzenlemelerine ilişkin çalışmalar
Sosyal güvenlik alanlarında suç olay analizleri, terör faaliyetleri analizi

2. VERİ MADENCİLİĞİ GÖREVLERİ

Veri madenciliğinde farklı görevleri yerine getirmek için pek çok farklı algoritmalar kullanılır. Bu algoritmalar verilere uygun modeli bulmaya çalışır. Algoritmalar verileri inceler ve özelliklerine en uygun modeli seçer. Veri madenciliği görevleri "tahmin edici" (predictive) ve "tanımlayıcı" (descriptive) modeller olmak üzere iki kategoriye ayrılır[13]. Bu kategoriler ve modeller Dunham tarafından Şekil 2'de görüldüğü gibi özetlenmiştir.



Şekil 2. Veri Madenciliği Modelleri Ve Görevleri.

Margaret H. Dunham, **Data Mining** (New Jersey: Pearson Education, 2003)'den uyarlandı.

2.1. Tahmin Edici Modeller

Tahmin edici modeller, sonuçları bilinen verileri kullanarak ilgili unsurlar için bir tahmin modeli oluşturur. Elde edilen bu model, sonuçları bilinmeyen unsurların tahmin edilmesinde kullanılır. Örneğin bir hastanede bir hastalığa ilişkin veri setini düşünelim. Veri madenciliği teknikleri uygulanarak hastalığa ilişkin geçmiş olaylardan elde edilmiş tıbbi veriler ve hasta durumu verilerinden

[13] Fayyad, **Ön.ver.**, s.12.; Han, **Ön.ver.**, s.21.; Margaret H. Dunham, **Data Mining Introductory and Advanced Topics (**New Jersey: Pearson Education, Inc., 2003), s.5.

bir tahmin modeli oluşturulabilir. Bu model sayesinde, hastaneye yeni gelmiş bir hastanın hastalığına ilişkin tahmin testler sonrası oluşan tıbbi veriler kullanılarak yapılabilir. Tahmin edici modeller sınıflama, regresyon, zaman serisi analizi ve kestirim olmak üzere dört grup halinde incelenebilir.

2.1.1. Sınıflama

Sınıflama, veri sınıfı ve kavramlarını tanımlama ve ayırt etmeyi sağlayan bir model kümesini bulma sürecidir. Türetilen model, "eğitim veri kümesi" (sınıf adı bilinen veri nesneleri) analizine dayalıdır. Sınıflama modelleri, sınıflar önceden incelenen veriler vasıtasıyla oluşturulduğundan, denetimli öğrenme olarak da ifade edilir. Örüntü tanımlama da bir sınıflama modelidir ve bu modelde de bir girdi örüntüsü önceden tanımlanmış sınıflara benzer çeşitli sınıflardan birisine yerleştirilir. Örneğin örüntü tanımlama modeli bir güvenlik tarama istasyonunda yolcuların potansiyel bir suçlu olup olmadığını belirlemek için kullanılabilir. Bu işlem için her yolcunun yüzü taranarak göz, beden, ağız ve baş şekli gibi özellikleri tanımlanır. Elde edilen bu veriler daha önceden özellikleri tanımlanmış suçlu bilgilerinin yer aldığı veritabanındaki verilerle karşılaştırılarak kimliğini saklamış olan suçlunun tespiti yapılabilir.

2.1.2. Regresyon

Regresyon bağımsız değişkenler ile bağımlı değişken arasındaki ilişkiyi en iyi tanımlayan fonksiyonu elde etmek için uygulanan istatistiksel tekniklerdir[14]. Veri madenciliği uygulamalarında bağımsız değişkenler veritabanında yer alan tablolardaki özellikleri, bağımlı değişken ise hedef özellik veya sınıf etiketi olarak adlandırılır. Diğer denetimli öğrenme uygulayan veri madenciliği tekniklerinde olduğu gibi regresyon analizinde de sonucu bilinen veriler kullanılarak ilişkiyi tanımlayan bir model oluşturulur. Yapay sinir ağları gibi makine öğrenmesi tekniklerinde oluşturulan modellerin aksine regresyon analizinde ilişkinin net olarak gösterilebildiği bir fonksiyon modeli temsil eder.

[14] Stephan Kudyaba ve R. Hoptroff, **Data Mining and Business Inteligence** (USA: Idea Group, 2001), s.8.

Verinin özelliklerine göre doğrusal regresyon, çoklu regresyon, lojistik regresyon gibi farkı regresyon modelleri kullanılabilir.

2.1.3. Zaman Serisi Analizleri

Zaman serisi analizi, zaman içinde değişiklik gösteren verilerin tahmin edilmesi problemidir[15]. Zaman serisi analizlerinin kullanıldığı en yaygın alan borsa işlemleridir. Bir hisse senedinin veya borsa endeksinin gelecek değeri tahmini zaman serisi problemlerine örnek oluşturur. Zaman serisi problemlerinin çözümünde istatistiksel ve istatistiksel olmayan birçok veri madenciliği algoritması kullanılmaktadır. Tahmin modellerinin oluşturulmasında geçmiş verilerden yararlanılması nedeniyle bu modeller denetimli öğrenme modeli olarak nitelendirilirler.

2.1.4. Kestirim

Pek çok veri madenciliği uygulaması geçmiş ve güncel verilere dayalı olarak gelecekteki veri değerlerini tahmin etme gayretindedir. Kestirim modelleri bir sınıflama modeli gibidir ancak bu modeli tahmin ve sınıflama modellerinden ayıran özellik gelecekteki verilerin tahmin etmesidir. Kestirim modellerini bu anlamda teknik özellik değil de uygulamanın bir özelliği olarak tanımlamak yerinde olacaktır. Veri madenciliği uygulamalarında kestirim modellerine örnek olarak su baskını tahmini, konuşma sesinden sözcükleri tanımlama, örüntü tanımlama problemleri verilebilir. Gelecek değerlerin zaman serisi analizi veya regresyon modelleri kullanılarak tahmin edilebilmesine rağmen farklı yaklaşımlar da kullanılabilmektedir[16]. Örneğin su baskını tahmini oldukça güç bir problemdir. Bir yaklaşıma göre nehrin farklı noktalarına yerleştirilen alıcılar nehrin su seviyesini, yağmur miktarı, zaman, nem gibi verileri toplayarak su baskınına ilişkin tahmin modeli oluşturulabilmektedir.

---

[15] Roiger, **Ön.ver.**, s.328.
[16] Dunham, **Ön.ver.**, s.7.

2.2. Tanımlayıcı Modeller

Tanımlayıcı modeller verilerdeki örüntü veya ilişkileri tanımlarlar. Bu modeller tahmin edici modellerin aksine analiz edilen verilerin özelliklerini incelemek için kullanılan modellerdir. Örnek olarak sigorta poliçesini yenilememiş müşterilerin benzer özelliklerini belirleyecek bir kümeleme çalışması verilebilir. Kümeleme, özetleme, birliktelik kuralları, sıra örüntüleri keşfi modelleri tanımlayıcı modeller olarak nitelendirilir.

2.2.1. Kümeleme

Kümeleme, verileri anlamlı ve/veya kullanışlı kümelere (gruplara) ayırır. Eğer amaç anlamlı kümeler oluşturmaksa o zaman kümeler verilerin doğal yapısını yansıtmalıdır. Bazı durumlarda ise kümeleme veri özetleme gibi farklı amaçlar için kullanışlı bir başlangıç noktası oluşturmaktadır[17].

Kümeleme analizi bir hedef değişken içermediğinden sınıflama analizinden farklı bir yaklaşımdır. Kümeleme analizinde hedef değişkenin değerini belirlemeye yönelik sınıflama, tahmin etme veya kestirim yapılmaya çalışılmaz. Bunun yerine verinin tamamını bölümlere ayırmak için homojen alt gruplar veya kümeler araştırılır. Bu işlem gerçekleştirilirken kümeler içindeki verilerin benzerliği göz önüne alınır[18]. Oluşturulan kümeler önceden tanımlanmadığından ve verinin özelliklerine göre belirlendiğinden kümelerin anlamı konuyla ilgili bir alan uzmanı tarafından yorumlanmalıdır. Kümeleme analizi denetimsiz öğrenme veya kesimleme olarak da ifade edilir.

Örneğin ulusal bir mağaza zincirinin, müşterilerinin fiziksel özellikleri (yaş, boy, kilo vb.), geliri ve ikamet yeri gibi özelliklere dayalı olan çeşitli demografik grupları hedef alan kataloglar oluşturmak istediği varsayılsın. Bu amaçla şirket, çeşitli katalogların hedef müşterilerini belirlemek ve yeni daha özel gruplara

---

[17] Tan, **Ön.ver.**,s.487.
[18] Daniel T. Larose, **Discovering Knowledge in Data** (USA: John Wiley&Sons, 2005), s.16.

hitap eden katalogların yaratılmasına yardımcı olmak için belirlenmiş özelliklerin değerlerine dayalı potansiyel müşteri kümeleri oluşturabilir.

2.2.2. Özetleme

Karekterizasyon veya genelleştirme olarak da adlandırılan özetleme, verileri basit tanımları yapılmış alt gruplar içine yerleştirme işlemidir[19]. Özetleme veritabanı hakkında betimleyici bilgileri ortaya çıkarır ve verilerden elde edilen ortalama veya standart sapma gibi tüm veriyi temsil eden göstergelerin hesaplanmasını ifade eder. Özet bilgiler, veritabanı fonksiyonları ve tanımlayıcı veri madenciliği teknikleri kullanılarak elde edilebilir.

2.2.3. Birliktelik Kuralları

Birliktelik kuralları verideki güçlü birliktelik özelliklerini tanımlayan örüntüleri keşfetmek için kullanılan bir analiz yöntemidir. Keşfedilmiş örüntüler özel olarak çıkarılan kurallar veya özellik alt grupları şeklinde temsil edilebilir. Araştırma uzayının üssel büyüklüğünden dolayı birliktelik analizinin amacı önemli örüntüleri etkin bir şekilde çıkarmaktır[20]. İş dünyasında birliktelik analizi, pazar sepeti veya benzeşme (affinity) analizi olarak da anılır. Birliktelik kuralları müşteriler tarafından birlikte satın alınan ürünleri bulmada yaygın olarak kullanılmaktadır. Örneğin bebek bezi satın alan müşterilerin bebek maması da satın almaya meyilli olduklarını birliktelik analizi ile keşfedebiliriz. Bu tür birliktelik kuralları ilgili ürünler arasındaki potansiyel çapraz satış olanaklarını tanımlamak için kullanılır. Birliktelik kuralları aynı zamanda işlevlerine göre gen gruplarının bulunması, ulaşılan web sayfalarının ilişkilerin tanımlanmasında, dünya iklim sisteminin farklı bileşenleri arasındaki ilişkilerin araştırılmasında veya telekomünikasyon ağlarındaki arızaların tahmin edilmesi gibi pek çok benzer problemlerin çözümünde kullanılır.

---

[19] Dunham, **Ön.ver.**, s.8.
[20] Tan, **Ön.ver.**,s.9.

2.2.4. Sıra Örüntüleri

Sıra örüntülerinde olayların zaman sıralarıyla ilgilenilir ve birbiriyle ilişkili olan verilerdeki birliktelik kurallarına benzer bir yapıdadır[21]. Fakat burada veriler arasındaki ilişki zamana bağlıdır. Pazar sepeti analizinde ürünler müşteri tarafından aynı anda alınmasına karşın sıra örüntüleri analizinde belirli bir zaman periyodunda satın alınmış ürünlerin ilişkileriyle ilgilenilir. Telekomünikasyon ağları, bilgisayar ağları gibi fiziksel izleme sistemlerinden veya bilimsel deneylerden toplanan olay-tabanlı verilerde sistemin doğası gereği olaylar arasında sıralı bir ilişki mevcuttur. Bu tür zamana dayalı olayların sıra örüntülerinin keşfedilmesinde tanımlayıcı model olan sıra örüntüleri analizleri kullanılır.

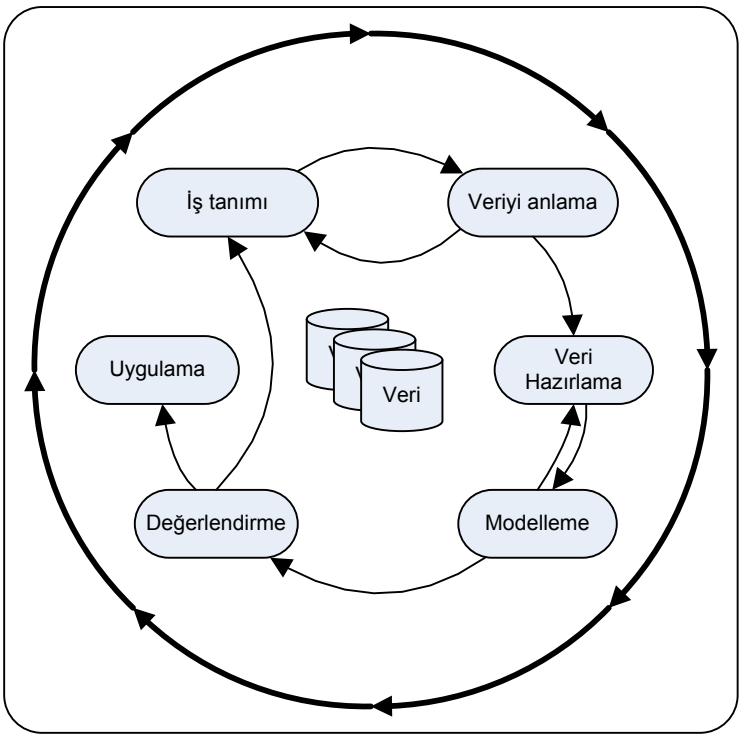
3. VERİ MADENCİLİĞİ UYGULAMA ADIMLARI

Pek çok veri madenciliği sistem yazılımı geliştiren kuruluş, kullanıcılara yol göstermek amacıyla bir uygulama süreç modeli önerirler. Bu modeller genellikle ardışık adımların yürütülmesiyle kullanıcıları hedefe ulaştırmayı amaçlar. CRISP-DM (CRoss-Industry Process For Data Mining) uygulama süreci, veri madenciliği uygulamalarında başarılı sonuçlar alan şirketlerin ve veri madenciliği araçlarını geliştiren bir başka şirketin oluşturduğu grup tarafından geliştirilmiş yaygın olarak kullanılan bir modeldir. Bu uygulama süreç modeli kullanıcıların gerekli adımları anlamasına yardımcı olan iyi bir başlangıçtır[22]. Uygulama süreci, yerine getirilmesi gereken görevler ve bu görevler arasındaki ilişkileri içerir. CRISP-DM tarafından önerilen uygulama süreç adımları Şekil 3'de gösterilmiştir. Bu sürecin her adımında uygulanan görev sonucu üretilen çıktı, sıradaki adımın girdisini oluşturur. Bazı durumlarda farklı aşamalar arasında ileri geri hareket etmek gerekebilir. Şeklin dışındaki daire veri madenciliğinin döngüsel doğasını sembolize eder. Süreç içinde elde edilen sonuçlar çalışılan konuyla ilgili yeni problemleri tetikleyebilir. Bir sonraki veri

[21] Dunham, **Ön.ver.**,s.9
[22] CRISP-DM Consortium, **CRISP-DM 1.0** (www.crisp-dm.org), s.1.

madenciliği süreci önceki süreçlerde elde edilen tecrübelerden faydalanmaktadır.

CRISP-DM'in önerdiği sürecin ilk adımı çalışma hedefleri ve gereksinimlerinin belirlenerek veri madenciliği probleminin tanımlandığı "iş tanımı" adımıdır. Veriyi anlama aşaması ilk adımda tanımlanan problemin çözümünde kullanılacak verinin bir araya getirilmesi, veri kalite problemlerinin çözülmesi, verinin incelenmesi ve gizli enformasyona ulaşmak için veri alt kümelerinin tespit edilmesi faaliyetlerini içerir. Veri hazırlama aşamasında başlangıç veri kümesinden modelde kullanılacak veri kümesini oluşturmak için dönüşüm ve temizleme işlemleri uygulanır. Modelleme adımında problem ve veri özelliklerine uygun modelleme teknikleri seçilir ve model parametrelerinin en iyi değerleri belirlenir. Bu adımda uygulanan veri madenciliği teknikleri veri hazırlama adımına dönülmesini gerektirebilir. CRISP-DM uygulama sürecinin son iki adımında modelin değerlendirilmesi ve uygulamasına ilişkin görevler yer alır.



Şekil 3. CRISP-DM Veri Madenciliği Uygulama Süreci.

CRISP-DM Consortium, **CRISP-DM 1.0** (www.crisp-dm.org)'den uyarlandı.

Bir diğer veri madenciliği uygulama süreci Two Crows şirketi tarafından önerilmiştir[23]. Two Crows şirketi bankacılık, sigortacılık, telekomünikasyon, perakendecilik, devlet uygulamaları, danışmanlık ve enformasyon sistemleri için veri madenciliği uygulama adımlarını tanımlayan raporun üçüncü sürümünü 1999 yılında yayınlamıştır. Bu teknik rapora göre uygulama adımları aşağıdaki gibi sıralanmıştır.

1. Problemin tanımlanması
2. Veri madenciliği veritabanın oluşturulması
3. Verinin incelemesi
4. Model için veri hazırlama
5. Modelin oluşturulması
6. Modelin değerlendirilmesi
7. Modelin uygulanması ve sonuçların izlenmesi

Veri madenciliği uygulama adımları literatürde farklı adlarla isimlendirilse de gerçekte benzer işlemler uygulanarak gerçekleştirilir. Bu çalışmada veri madenciliği uygulama adımlarında Two Crows'un önerdiği süreç adımları takip ederek tanımlanmıştır. Her adımda kullanılan tekniklerin özellikleri izleyen bölümler halinde yapılandırılmıştır.

3.1. Problemin Tanımlanması

Birçok sistem çalışmasında olduğu gibi veri madenciliği uygulamalarında da problemin tanımlanması ilk adım olup en önemli adımlardan biridir. Gerçekte, işletmeler ya da organizasyonlar amaçlarını açık bir şekilde ifade edebilmekte ancak önemli problemlerin detaylı amaçlara dönüştürülmesi güç olabilmektedir[24]. Problem ve amaçların açık olarak ifade edilmesi analizin doğru olarak yapılandırılması için önkoşuldur. Problemin ve amaçların ifade edilmesi, veri madenciliği uygulama aşamalarından en zor adım olmasının nedeni bu

[23] Two Crows Corp, **Ön.ver.**, s.22.
[24] Paolo Giudici, **Applied Data Mining** (England: John Wiley&Sons, 2003), s.7.

adımda yapılan tanımlamaların, çalışmanın nasıl yapılacağının belirlenmesidir. Bu nedenle amaçlar şüphe ve belirsizlikten uzak olmalıdır.

3.2. Veri Madenciliği Veritabanının Kurulması

Veri madenciliği veritabanının kurulması aşaması, ilk adımda tanımlanan problemin çözümüne ilişkin analizin ihtiyaç duyduğu özellik ve nitelikteki verinin hazırlanması olarak ifade edilebilir. Bu aşama veri kaynaklarının belirlenmesi, veri tanımlama, veri seçme, veri kalitesi ve ön hazırlık süreçleri, veri madenciliği veritabanının yüklenmesi ve bakımı görevlerinin yerine getirilmesi ile tamamlanır. Bu adımları uygulamak diğer tüm adımların uygulanmasından daha fazla zaman ve çaba gerektirir. Veri hazırlama adımına, model geliştirme adımı gerçekleştirilirken geri dönmek gerekebilir. Bunun nedeni model oluşturma adımında modelden öğreneceğimiz herhangi bir enformasyonun veride değişiklik yapmamızı gerektirmesidir. Veri hazırlama adımları tüm bilgi keşfi süreci için harcanan zaman ve çabanın %50 ile %90 arası bir kısmını oluşturur. İzleyen bölümlerde veri madenciliği veri kaynakları ve veri hazırlık süreçleri anlatılacaktır.

3.2.1. Veri Kaynaklarının Belirlenmesi

Veri kaynaklarının belirlenmesi, madenciliği yapılacak olan veri kaynaklarının tanımlanmasıdır. Veriler pek çok farklı kaynaktan elde edilebilir. Verilerin saklanmasında ve yönetilmesinde basit dosya sistemlerinden veritabanları ve veri ambarlarına uzanan birçok farklı yöntem ve teknoloji bulunmaktadır. Her yöntem işleyen sistemin verisini tutmak, tanımlamak ve depo yapısını oluşturmak için kendine özgü sistematik bir perspektife sahiptir[25]. Veri depolama ve veri sistemleri aşağıda yer alan bölümlerde açıklanmıştır. Günümüzde veri depolama ve yönetim sistemlerinin uygulandığı yazılımların listesi EK 1'de verilmiştir.

---

[25] Nong Ye, **The Handbook of Data Mining** (USA: Lawrence Erlbaum, 2003), s.304.

3.2.1.1. Metin Dosyaları Ve İşlem Tabloları

Metin dosyası veri saklamada ve yönetiminde en basit ve en temel yöntemdir. Bilgisayarlardaki en yaygın metin formatı ASCII'dir (American Standard Code for Information Interchange). Bir ASCII formatlı metin dosyasında her bir karakter 7 bit ikili sistem numarası ile gösterilir ve toplamda 128 farklı karakter tanımlanabilir. UNIX ve DOS işletim sistemleri metin dosyaları için ASCII kod sistemini kullanır. Windows, NT ve 2000 işletim sistemleri 16 bit karaktere kadar genişletilmiş ve yine ASCII'ye dayalı olan Unicode formatını kullanır. Formatlar arası dönüşümü sağlayan çeşitli dönüşüm yazılımları mevcuttur.

Veri saklama ve yönetmede kullanılan diğer bir basit yöntem de işlem tablolarıdır. İlk olarak 1978 yılında geliştirilen "VisiCalc" işlem tablosu yazılımından sonra Excel, Lotus 1-2-3, Apple works, Filemaker ve Corel Quatro-Pro gibi işlem tablosu yazılımları geniş kullanıcı grupları tarafından yaygın olarak kullanılmıştır[26]. İşlem tablolarında veriler, sütunlarda özelliklerin satırlarda ise veri kayıtlarının bulunduğu iki boyutlu bir tabloda saklanır. Excel gibi işlem tablosu araçları verilerin görüntülenmesi, yüklenmesi, düzenlenmesi, saklanması veya başka bir dosya sisteminden işlem tablosu formatına dönüşümü gerçekleştirecek fonksiyonların kullanılmasını kolaylaştırır.

Metin dosya ve işlem tabloları, veri kayıtlarının oluşturulmasında daha az çaba gerektiren düz dosya olarak da adlandırılır. Düz dosyalar, birliktelik kurallarının kullanıldığı bazı veri madenciliği uygulamalarında kullanılır. Çoklu tabloların özellikleri arasındaki ilişkilerin yakalanması kolay olmadığından düz dosyalar pek çok veri madenciliği tekniklerinde kullanılmamaktadır. Bu sınırlamalar yüzünden düz dosyalar küçük boyut ve hacimdeki veri kümelerini içeren veri madenciliği uygulamalarında kullanışlıdır.

---

[26] **Aynı.** s.397.

3.2.1.2. Veritabanı Sistemleri

Veritabanı sistemleri günümüzde veri saklama ve yönetiminde önemli bir yere sahiptir. Bir veritabanı sistemi veritabanı dosyaları ve veritabanı yönetim sisteminden oluşur. Veritabanı mantıksal olarak ilişkilendirilmiş veri topluluğu, veritabanı yönetim sistemi ise kullanıcılara veritabanı oluşturmaya ve bakımını yapmaya olanak sağlayan programlar topluluğudur[27]. 1960 ve 1970'lerde ilk kez ticari veritabanı sistemlerinde kullanılan hiyerarşik modeller ve ağ modellerini diğer veritabanı sistemleri izlemiştir.

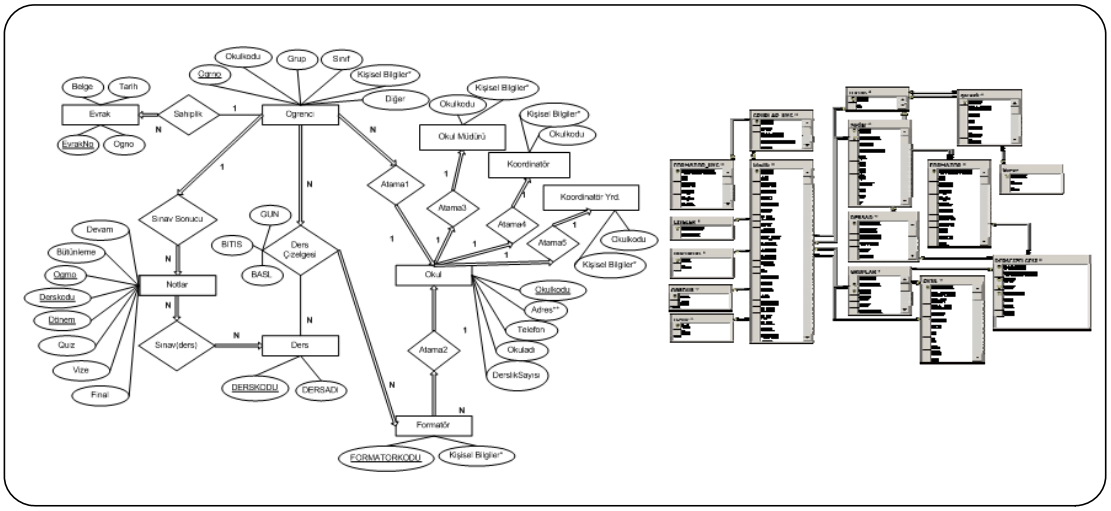
3.2.1.2.1. İlişkisel Veritabanları

İlişkisel veritabanı her biri benzersiz isime sahip tabloların toplamından oluşur. Her tablo özellikler kümesi ve geniş bir nesne topluluğundan oluşur. Veritabanı terminolojisinde veri, gerçek dünya nesnelerine ilişkin enformasyonu sakladığından her satır veya kayıta nesne, sütun veya alanlar ise özellik olarak adlandırılır. Bu tanımlar ayrıca veri madenciliğinde verinin tanımlanmasında kullanılan terimlerdir. Şekil 4'te örnek bir ilişki-varlık şeması ve ilişkili veritabanı tabloları görülmektedir. İlişkisel tablodaki her sıra benzersiz bir anahtarla temsil edilen ve bir özellik değer kümesiyle tanımlanmış bir nesneyi temsil eder. İlişkisel veritabanlarının tasarlanmasında varlıklar ve aralarındaki ilişkileri modelleyen varlık-ilişki veri modeli kullanılır.

İlişkisel veritabanlarındaki tüm veriler tablolarda depolanır ve veri ile ilgili tüm işlemler ya tablonun kendi üzerinde ya da işlemin sonucu olarak başka tablolara aktarılır. İlişkisel tabloları yönetmek amacıyla çeşitli sorgulama diller yaratılmıştır ve bunların en yaygın olanı SQL (Structured Query Language) adıyla anılan yapılandırılmış sorgu dilidir[28]. SQL sorguları bir veritabanından veriyi getirmek veya güncellemek gibi görevleri uygulamak için kullanılır.

---

[27] Ramez Elmasri ve S.B. Navathe, **Fundamentals of Database Systems** (2. Basım. USA: Benjamin/Cummings Publishing, 1994), s.2.
[28] Ye, **Ön.ver.**, s.399.



Şekil 4. İlişki-Varlık Şeması ve İlişkisel Veritabanı Tablo Görünümleri.

Veritabanı yönetim yazılımlarında istenilen veri kümelerini elde etmek için SQL komutları yazılır ya da kullanıcı grafik ara-yüzünde sorgular oluşturulur. Bu istekler bağlama, seçme ve özetleme gibi ilişkisel işlemler kümesine dönüştürüldükten sonra verimli bir süreç için en iyileniler ve hedef küme elde edilerek kullanıcıya sunulur.

İlişkisel veritabanlarında veri madenciliği uygulandığında eğilimler veya veri örüntüleri araştırılarak daha fazla bilgi elde etmek mümkün olur. Örneğin yeni müşterilerin kredi riskini tahmin etmek için müşterilerin gelir, yaş ve önceki kredi bilgilerine dayanarak veri madenciliği analizi gerçekleştirilebilir[29]. İlişkisel veritabanları en popüler ve en zengin enformasyon depolarından biridir ve bu nedenle veri madenciliğinde kullanılan en yaygın veri kaynaklarıdır.

3.2.1.2.2. Hareket Veritabanları

Her bir hareketin ya da işlemin bir satır veya kayıtla ifade edildiği veritabanlarıdır. Hareket veritabanlarında her bir işlem benzersiz bir tanımlayıcı ile ifade edilirken hareketi oluşturan nesnelere ait tablolar bulunmaktadır. Hareketi oluşturan nesnelerin özelliklerini barındıran bu tablolar da aynı veritabanında saklanmaktadır. Bir markette yapılan satışların depolandığı veritabanı bu tür bir veritabanına örnektir. Yapılan her satış işlemi bir hareketi

[29] Han, **Ön.ver.**, s.12.

ifade eder ve her hareket için benzersiz bir işlem numarası üretilir. Müşteriler ve satılan ürüne ait veriler de bu hareket kaydının diğer alanlarını oluşturur. Birlikte satılan ürünleri belirlemeye ilişkin bir analiz muhtemelen bir hareket veritabanını kaynak veri olarak kullanılacaktır.

3.2.1.2.3. İleri Düzey Veritabanı Uygulamaları

Veritabanı teknolojilerindeki ilerlemeler doğrultusunda farklı gereksinimlere cevap vermek amacıyla ileri düzey veritabanı sistemleri geliştirilmiştir. Yeni veritabanı uygulamaları uzaysal verileri (haritalar), mühendislik tasarım verilerini (bina tasarımı, sistem bileşenleri, bütünleştirilmiş devreler), bağlantılı web belgeleri ve çoklu ortam verilerini (metin, resim, video ve ses), zamana bağlı verileri (tarihsel kayıtlar, döviz verileri) ve internet sayfalarını (internetten elde edilen devasa dağıtık enformasyon deposu) kapsamaktadır. Bu uygulamalar karmaşık nesne yapılarını, değişken uzunluktaki kayıtları, yarı yapılandırılmış veya yapılandırılmamış verileri, metin ve çoklu ortam verilerini ve dinamik değişkenli veritabanı şemalarını yönetmek için etkin veri yapılarını ve ölçeklenebilir yöntemleri içermelidir[30]. İleri düzey veritabanları ve özel uygulama yönelimli veritabanları nesne yönelimli ve nesne–ilişkisel veritabanlarını, uzaysal veritabanlarını, geçici ve zaman serileri veritabanlarını, metin ve çoklu ortam veritabanlarını, heterojen veritabanlarını ve web'e dayalı global enformasyon sistemlerini içine alır. İleri düzey veritabanları veri madenciliğinde yeni araştırma alanları için uygun veri kaynakları sağlarlar.

3.2.1.3. OLAP Ve Veri Ambarları

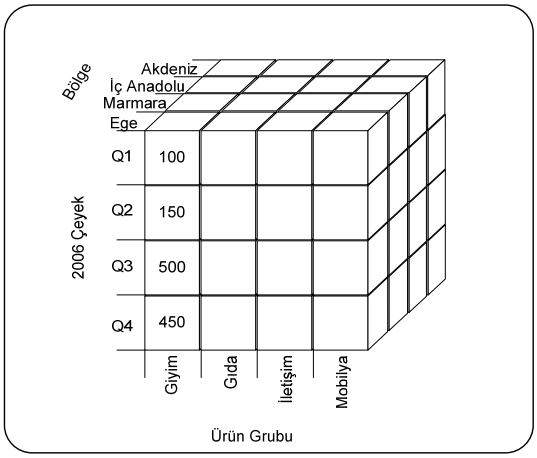
OLAP (Online Analytical Proccessing) olarak kısaltılan çevrimiçi analitik işleme, kullanıcılara problemin gerçek boyutunu yansıtan ve ham veriden dönüştürülmüş enformasyonun çeşitli açıdan görünüşlerine hızlı ve etkileşimli ulaşımı sağlayan bir yazılım teknolojisidir[31]. OLAP uygulamaları genellikle gerçek verilerin analizini içerir. Bu uygulamalar SQL'deki mevcut temel

[30] **Aynı**, s.16.
[31] Ye, **Ön.ver.**, s.402.

gruplama işlevlerinin geliştirilmiş hali olarak düşünülebilir. OLAP'ın birincil amacı, karar destek sistemlerinde ihtiyaç duyulan özel amaçlı (ad-hoc) sorgulamaları sağlamaktır. OLAP veri kaynağı olarak genellikle bir veri ambarı kullanır.

Veri ambarları bir veri kümesinin önemli kısımlarını veya tümünü saklamak ve analiz etmek için tasarlanan yapılandırılmış bir karar destek sistemidir. Bir veri ambarında veriler çoklu kaynak uygulamalarında fiziksel ve mantıksal olarak dönüştürülür ve belirli zaman aralıklarıyla güncellenir. Veri ambarı çoklu heterojen kaynaklardan elde edilen verilere veri birleşme, veri temizleme ve veri bütünleştirme süreçleri uygulanarak oluşturulur. Karar destek sistemlerinde kullanılmayacak verilerin veri ambarına aktarılması gereksiz zaman ve kaynak israfına yol açar.

Veri ambarları ve OLAP araçları çok boyutlu veri modeline dayalıdır. Çok boyutlu veri modeli, verileri bir küp şeklinde ele alır. Bir veri küpü verilerin çok boyutlu olarak yapılandırılmasını ve görüntülenmesini sağlar. Örneğin bir şirketin satışlarına ilişkin veri küpü oluşturulduğunu düşünelim. Veri küpünün ana temasının satış tutarları olduğunu, boyutlarının ise satış bölgesi, zaman ve ürün grubu olduğunu varsayalım. Bu durumda Şekil 5'de görüldüğü gibi bir veri küpü oluşturulabilir. Bir veri küpünün ihtiyaca göre üçten fazla boyutu da tanımlanabilir. Oluşturulan veri küpleri ihtiyaç duyulan bilgiye göre OLAP sistemi tarafından sorgulanır.

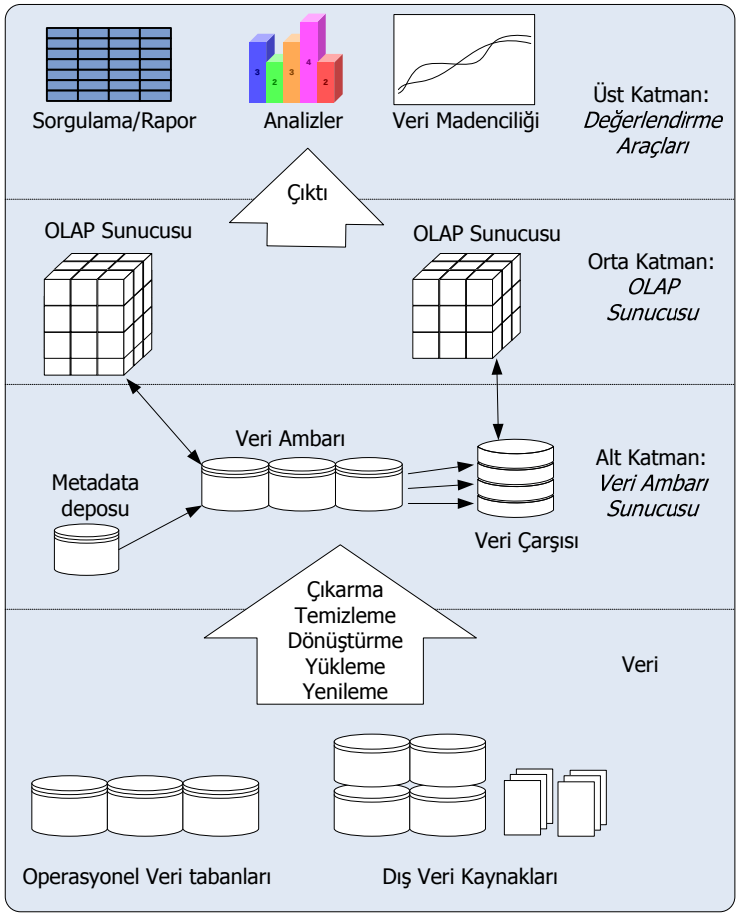


Şekil 5. Veri Küpü Örneği.

OLAP ve Veri Küpünü tanımladıktan sonra veri ambarı mimarisi ve kullanım şeklini tanımlamak faydalı olacaktır. Şekil 6'de veri madenciliği için bir veri ambarı ve OLAP teknolojisinin nasıl yapılandırılacağı gösterilmiştir. Veri ambarının mimarisinde alt katman genellikle ilişkisel veritabanı sistemi olan veri ambarı verilerinin depolandığı veritabanı sunucusudur. Alt katmanda veriyle ilgili enformasyonun yer aldığı metadata deposu ve kurumların işlerine özgü hareket veritabanları olan veri çarşıları (data marts) yer almaktadır. İşlemsel veritabanları ve dış veri kaynaklarından elde edilen veriler SQL komutlarını kullanan uygulamalar tarafından veri ambarı veritabanına yüklenirler. Orta tabaka ilişkisel OLAP (ROLAP) modeli ya da çok boyutlu OLAP (MOLAP) modelinin kullanıldığı OLAP sunucusudur. Son tabaka ise sorgulama ve raporlama araçları, analiz araçları veya veri madenciliği araçlarını barındıran istemcidir[32].

Veri madenciliğinin ihtiyaç duyduğu ve veri ambarında depolanan veriler farklı yapılardaki veri kaynaklarının bir araya getirilmesinden oluşturulur. Bu nedenle veri ambarlarında depolanan veriler veri madenciliği kaynağı olarak kullanılabilir. Veri madenciliği ve veri ambarları birbirini tamamlayıcıdırlar. Örneğin yönetim, bir reklam kampanyasının hedef kitlesini belirlemeye yardımcı olmak amacıyla, müşteri verilerinin kullanıldığı sınıflama veya birliktelik kuralları uygulamasının sonucunu kullanabilir. Veri madenciliği faaliyetleri bir veri ambarındaki verileri kullanarak fayda sağlayabilir fakat zorunlu değildir. Birbiriyle ilişkili olan veri ambarı ve veri madenciliği benzer görülse de birbirinden farklıdır ve biri diğeri olmaksızın kullanılabilir.

[32] Han, **Ön.ver.**, s.67.



Şekil 6. Üç Katmanlı Veri Ambarı Mimarisi.

Jiawei Han ve M. Kamber, **Data Mining** (USA: Academic Press, 2001)'den uyarlandı.

3.2.2. Veri Tanımlama

Madenciliği yapılacak verinin içeriği bu aşamada tanımlanır. Veri kaynağında yer alan tablo, dosya, alanların özellikleri raporlanır. Mevcut veritabanında yer alan her tablo veya dosya için raporlanması gereken bazı özellikler aşağıda verilmiştir[33].

- Tabloda yer alan alanların sayısı
- Eksik değerler yer alan kayıtların sayısı ve yüzdesi
- Alan isimleri
- Veri türü
- Açıklaması

[33] Two Crows Corp., **Ön.ver.**, s.24.

- Tanımı
- Alan kaynağı
- Ölçü birimi
- Benzersiz değerler sayısı
- Değerler listesi
- Değer aralıkları
- Eksik değerlerin sayısı ve yüzdesi
- Enformasyonun toplandığı kaynak, toplanma sıklığı ve veri güncellenme özelliği
- Birincil anahtar ve yabancı anahtar ilişkileri

3.2.3. Seçim

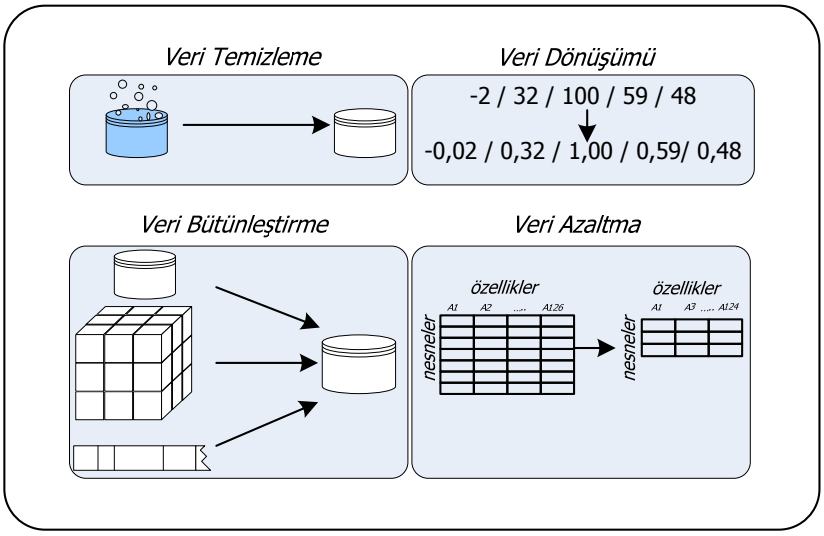
Veri madenciliği veritabanı hazırlamada veri tanımlama aşamasından sonra madenciliği yapılacak verinin alt kümesi seçilir. Bu aşamada veritabanını örnekleme veya tahmin edici değişkenleri seçme işlemi değil gereksiz veya ihtiyaç duyulmayan verinin analiz dışı bırakılmasıdır. Kaynakların yetersizliği, maliyet, veri kullanım kısıtlamaları veya kalite problemleri gibi sınırlamalar da bazı verilerin analiz dışında bırakılmasını gerektirebilir.

3.2.4. Veri Kalitesini İyileştirme Ve Ön Hazırlık Süreçleri

Veritabanlarında yer alan verilerin mükemmel olması çoğu zaman mümkün değildir. Veri madenciliği tekniklerinin çoğu verilerdeki kusurları göz ardı edebilmesine rağmen veri kalitesini anlamak ve iyileştirmek konusuna odaklanmak veri madenciliği çıktı kalitesini arttırır. Veri kalitesi kavramı verideki gürültü ve aykırı değerler, eksik, tutarsız veya tekrarlı verilerin varlığı ile ölçülebilir. Veri kalitesinin düşük olması verinin bizi yanıltmasına yani hedeflenen sonuca ulaşamamamıza neden olur[34]. Verilerin veri madenciliğine uygun hale getirilebilmesi, kusurlarının araştırılarak giderilmesi gerekmektedir. Verilerdeki kusurların giderilmesi için birtakım ön hazırlık süreçleri uygulanır.

[34] Tan, **Ön.ver.**, s.19.

Veri temizleme, veri bütünleştirme, veri dönüşümü ve veri azaltma süreçleri Şekil 7'de simgesel olarak gösterilmiştir.



Şekil 7. Veri Ön Hazırlık Süreci.

Jiawei Han ve M. Kamber, **Data Mining** (USA: Academic Press, 2001)'den uyarlandı.

3.2.4.1. Veri Temizleme

Veri madenciliğinde veri kalite problemlerini engellemek için önce veri kalitesi problemlerinin farkına varılarak doğrulanması ve zayıf veri kalitesini göz ardı edebilen algoritmaların kullanılması üzerinde odaklanılır. Veri kalitesi problemlerinin farkına varılması ve doğrulanması veri temizleme olarak adlandırılır[35]. Veri temizleme yoluyla eksik değerler tamamlanarak, gürültülü veri düzeltilerek, aykırı değerler tanımlanarak veya çıkarılarak ve tutarsızlıklar giderilerek veri kalitesi arttırılmaya çalışılır.

Veri temizleme için temel yöntemler eksik değerler, gürültülü veri ve tutarsızlık olmak üzere üç temel başlıkta gruplanabilir[36].

[35] **Aynı**, s.36.
[36] Han**, Ön.ver.**, s.109.

3.2.4.1.1. Eksik Değerler

Madenciliği yapılacak verinin bazı özellik değerleri boş yani eksik olabilir. Özellik değerlerinde eksik veya boş değer olmasının birçok nedeni vardır. Veritabanında yer alan verilerin anket verisi olması ve bilgisi toplanan bireyin bilgi vermek istememesi, yanlış anlama veya veri giren personelin hatası, diğer veri özellikleriyle tutarsızlığı yüzünden silinmesi gibi nedenler eksik veri oluşmasına neden olabilir. Bazı durumlarda değerin boş olması eksik veri değil her nesne için uygulanabilir bir özellik olmamasından kaynaklanabilir. Bir kimlik tablosunda bayanlara ait kayıt alanlarında askerlik bilgisinin yer almaması bu duruma örnek verilebilir. Bu durumda benzer verilerin eksik değer olarak algılanması ve giderilmesi hataya neden olabilecektir. Eksik değer ile ilgili stratejileri üç ana grupta toplayabiliriz[37]. Bu stratejiler aşağıdaki bölümlerde incelenmiştir.

3.2.4.1.1.1. Veri Nesne Veya Özelliklerini Elemek

Eksik değerlerle ilgili nesneleri yani eksik değer olan kayıtları çıkartmak basit ve etkili bir stratejidir. Ancak eksik değerlere sahip olan nesnelerin çıkartılması nesnelerin diğer özelliklerinde yer alan enformasyonun kaybına neden olacağından analizin güvenirliğini azaltır[38]. Bununla birlikte bir veri kümesi sadece birkaç eksik değere sahip nesne içeriyorsa bu nesneleri çıkarmak uygun olabilir. Diğer bir strateji de eksik değerlere sahip özelliklerin analizden çıkartılmasıdır. Çıkartılacak özellikler analiz için önemli olabileceğinden bu stratejinin uygulanmasına dikkat edilmelidir.

3.2.4.1.1.2. Eksik Değerlerin Tahmin Edilmesi

Bazı durumlarda eksik değer güvenilir bir şekilde tahmin edilebilir. Örneğin birkaç tane geniş alana yayılmış eksik değere sahip zaman serisini düşürelim. Bu durumda eksik değerler diğer veriler kullanılarak tahmin edilebilir.

---

[37] Pendharkar, **Ön.ver.**, s.40.
[38] Tan, **Ön.ver.**,s.41.

Eksik değerin tahmin edilmesi için kullanılan başlıca stratejiler aşağıda verilmiştir[39].

- *Eksik değerlerin el ile doldurulması:* Bu strateji zaman alıcıdır ve eksik değerlerin fazla olduğu büyük veri kümelerinde kullanılması uygun değildir.
- *Eksik değerlerin tamamlanmasında genel bir sabitin kullanılması:* Tüm eksik değerlerin belirlenecek bir sabit değer ile değiştirilmesidir. Bu değişiklik uygulandığında veri madenciliği algoritmalarını olumsuz etkileyebilir. Bu nedenle basit bir strateji olmasına rağmen tercih edilmez.
- *Eksik değerlerin verinin özelliğin diğer değerleri dikkate alınarak tamamlanması:* Bu stratejide eksik değerler, aynı özelliğin eksik olmayan kayıtları göz önüne alınarak ortalama, medyan, mod gibi verinin tamamını temsil eden tek bir değer ile değiştirilir.
- *Eksik değerin kendi sınıfında yer alan değerlerin ortalaması ile tamamlanması:* Eksik değerlerin tamamlanması öncesinde veri üzerinde bir sınıflama çalışması yapılarak eksik değerlerin ait olduğu sınıflar belirlenir. Her eksik değer bulunduğu sınıfın eksik olmayan özellik değerlerinin ortalaması ile tamamlanır.
- *Eksik değerlerin tamamlanmasında en uygun değerin kullanılması:* Eksik değerlerin bulunduğu özelliğin en uygun değeri regresyon yönteminin kullanıldığı sonuç çıkarmaya dayalı araçlar veya karar ağaçları kullanılarak belirlenebilir. Diğer stratejilere kıyasla bu strateji eksik değerleri tahmin etmede mevcut enformasyondan en fazla faydalanan yöntemdir. Bu nedenle en sık kullanılan stratejidir.

3.2.4.1.1.3. Eksik Değerlerin Göz Ardı Edilmesi

Birçok veri madenciliği yaklaşımı eksik değerleri göz ardı edecek şekilde düzenlenebilir. Örneğin, kümeleme analizinde nesne çiftleri arasındaki benzerlik hesaplamalarını düşünelim. Eğer bir çift nesnenin biri veya her ikisi bazı

[39] Han, **Ön.ver.** s.109.

özellikler için eksik değerlere sahipse o zaman benzerlik sadece eksik değerler içermeyen özellikler kullanılarak hesaplanabilir. Özelliklerin toplam sayısı az veya eksik değerlerin sayısı çok olmadığı sürece benzerlik hesaplaması hemen hemen doğru olacaktır. Bundan başka pek çok sınıflama yaklaşımlarında eksik değerleri göz ardı ederek düzenlemeler yapılabilir[40].

3.2.4.1.2. Gürültülü Veri

Gürültü, veri madenciliği tekniği ile analiz etmek istediğimiz verilerdeki beklenen değerlerden sapan aykırı değerler veya hatalardır. Gürültülü veri büyük veritabanları ve veri ambarlarında karşılaşılan yaygın problemlerdendir. Ölçülen bir değerdeki hata veya hatalı veri toplama, veri girişi problemleri, teknolojik kısıtlar gibi yanlış nitelik değerleri gürültülü verinin olası nedenleridir. Veri madenciliği uygulanmadan önce bu değerlerin neden olduğu gürültü düzeltilmelidir[41]. Verideki gürültünün belirlenip giderilmesi için bölmeleme, kümeleme, bilgisayar ve insan denetiminin birleştirilmesi ve regresyon yöntemleri kullanılabilir.

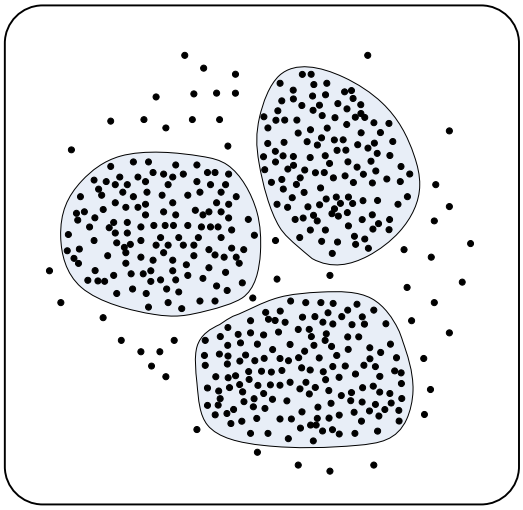
Bölmeleme yöntemlerinde öncelikle veriler artan sırada sıralanır. Bölme sayısı belirlenerek veriler eşit sayıda bölmelere ayrılır. Farklı düzeltme seçenekleri kullanılarak her bölmedeki veriler düzleştirilir. Örneğin her bölmedeki sayılar ilgili bölmenin ortalaması ile değiştirilir. Bölmelemede verilerin sıralanmasıyla yapılan düzeltme, komşu değerlere yakınlık sağlayacağından yerel bir düzleştirme sağlar.

Aykırı değerler kümeleme analizi ile ortaya çıkarılabilir. Kümeleme analizinde benzer değerler gruplar veya kümeler halinde bir araya getirildiğinden aykırı değerler Şekil 8'deki gibi belirlenir[42].

Aykırı değerler bilgisayar ve insan denetiminin birleşimi ile belirlenebilir. Örneğin bir el yazısı karakter tanıma uygulamasında aykırı değer örüntüleri

[40] Tan, **Ön.ver.**, s.41.
[41] Dunham, **Ön.ver.**, s.15.
[42] Han, **Ön.ver.** s.111.

karakterin tahmin edilmesinde yardımcı olabilir. Bu örnekte aykırı değer örüntülerinin faydalı ya da işe yaramaz olup olmadığı insan tarafından daha kolay ayırt edilebilir.



Şekil 8. Aykırı Değerlerin Kümeleme Analizi ile Belirlenmesi.

3.2.4.1.3. Tutarsız Veri

Bazı veritabanı kayıt işlemlerinde verilerde tutarsızlıklar oluşabilir. Bazı tutarsızlıklar dış veri kaynakları kullanılarak elle düzeltilebilir. Örneğin bir veri girişinde yapılan hata verinin girildiği kaynak belgelerden kontrol edilerek düzeltilebilir. Bilgi mühendisliği araçları bilinen veri sınırlamalarını bozan verileri ortaya çıkaran araçlara sahiptir[43]. Örneğin özellikler arasındaki işlevsel bağımlılıkların bu hataları bulabildiği bilinmektedir.

3.2.4.2. Veri Bütünleştirme

Veri bütünleştirme, çoklu kaynaklardan gelen verinin uygun bir veri ambarına birleştirilmesidir. Çoklu veri kaynakları; veritabanları, veri küpleri veya dış dosyalardan oluşabilir. Veri bütünleştirmede şema birleştirmesi, fazla veri sorunları ve veri değer karmaşalarının belirlenmesi ve çözümlenmesi olmak üzere üç temel konu ön plana çıkar.

---

[43] **Aynı**, s.112.

Şema bütünleştirme iki farklı kaynaktan gelen verilerin eşleştirilmesi için aynı varlıklar belirlenerek veriler şemalar yardımıyla birleştirilir. Şema birleştirme işleminde hataları engellemek için meta veri kullanılabilir. Veritabanları ve veri ambarlarında yer alan meta veri kavramı veri hakkında depolanan veri olarak tanımlanır.

Veri bütünleştirmede ikinci önemli konu olan veri fazlalığı, bir varlığın özelliklerinin birden fazla kaynaktan toplanması durumunda ortaya çıkar. Bazı veri fazlalığı korelasyon analizi ile ortaya çıkarılabilir. Korelasyon analizi iki değişken arasındaki ilişkinin yönünün, büyüklüğünün ve önemini gösteren istatistiksel bir yöntemdir.

Veri bütünleştirmede üçüncü önemli konu veri değer karmaşıklığının belirlenmesi ve çözümlenmesidir. Farklı veri kaynaklarından gelen özellik değerleri ölçekleme, birim sistemi veya gösterimdeki farklılıklar yüzünden birbirlerinden farklı olabilirler. Örneğin ağırlık özelliği farklı kaynaklarda farklı birim sistemiyle depolanmış olabilir. Veri bütünleştirme işlemlerinde verinin bu tür heterojenliği dikkate alınmalıdır.

3.2.4.3. Veri Dönüştürme

Bazı durumlarda orijinal veri kümelerindeki özellikler gerekli enformasyonu içerdiği halde veri madenciliği algoritmaları için uygun yapıda olmayabilirler. Bu durumda orijinal özelliklerinden oluşturulan bir veya daha fazla yeni özellik orijinal özelliklerden daha faydalı olabilir[44]. Veri dönüşümünde verilerin veri madenciliği için uygun formlara dönüştürülmesi düzeltme, bir araya getirme, genelleme, normalleştirme ve özellik oluşturma işlemleriyle gerçekleştirilir.

- *Düzeltme:* Bölmeleme, kümeleme ve regresyon gibi teknikler kullanılarak verilerdeki gürültünün temizlenmesidir.

[44] Tan, **Ön.ver.**, s.57.

- *Bir araya getirme:* Veriler bir araya getiren gruplama fonksiyonları kullanılarak gerçekleştirilir. Günlük temelde bulunan bir veri özelliğinin aylık temele dönüştürülmesi örnek verilebilir.
- *Genelleme:* Düşük düzeydeki verinin kavram hiyerarşisi kullanılarak daha yüksek seviyeye dönüştürülmesidir. Örneğin yaş gibi sayısal verilerin kategorik olan genç, orta yaşlı veya yaşlı gibi değerlere dönüştürülmesi ya da cadde isimlerinden oluşan kategorik verilerin şehir veya ülke şeklinde daha yüksek kavramlara dönüştürülmesidir.
- *Normalleştirme veya standartlaştırma:* Bir değişkenin standartlaştırılması veya normalleştirilmesi yaygın olarak kullanılan veri dönüşüm tekniğidir. Veri madenciliği terminolojisinde her iki terim birbiri yerine kullanılmaktadır. Ancak buradaki normalleştirme terimi, istatistikte kullanılan bir değişkenin normal dağılmış bir değişkene dönüştürülmesi ile karıştırılmamalıdır[45]. Standartlaştırma veya normalleştirmenin amacı sayısal veri değerlerinin küçük bir bölgede yer alması için ölçeklenmesidir. Normalleştirilmiş veriler sınıflama için kullanılan yapay sinir ağları algoritmalarının öğrenme aşamasının hızlanmasına yardım edecektir. Kümeleme gibi mesafe ölçümlerine dayalı algoritmalarda normalleştirilmiş verilerin kullanılması faydalı olacaktır.
- *Özellik oluşturma:* Yeni özellikler madencilik sürecine yardımcı olmak için verilen özellikler kümesinden oluşturulur ve düzenlenir. Özellik oluşturma karar ağacı algoritmaları sınıflama için kullanıldığında bölümleme problemini azaltmaya yardımcı olabilir. Yükseklik ve genişlik özelliklerinden alan özelliğinin oluşturulması bu duruma bir örnek olarak verilebilir.

3.2.4.4. Veri Azaltma

Oldukça karmaşık olan ve çok büyük veri kümelerinin madenciliğinin yapılması çok uzun zaman aldığından bu tür verilerin olduğu gibi alınarak analiz

[45] **Aynı**, s.64.

edilmesi uygulanabilir ve pratik olmamaktadır. Bu nedenle veri azaltma yöntemleri çok daha küçük hacimde azaltılmış veri kümelerinin oluşturulması için kullanılır. Veri azaltma işlemi sonrası elde edilen veri seti üzerinde uygulanan madencilik sonucu, verinin tamamından elde edilen sonuçtan çok farklı olmamalıdır. Veri azaltma yöntemleri aşağıdaki bölümlerde açıklanmıştır.

3.2.4.4.1. Veri Küpü Birleştirme

Veri madenciliğinin veri kaynağının bir OLAP sistemi olması durumunda ihtiyaç duyulan verilerin ön hesaplama ve özetlenmesi daha hızlı gerçekleştirilebilir. Veri küpleri çok boyutlu birleştirilmiş verileri saklar. Bazı durumlarda tüm verinin veri madenciliği algoritmalarında işlenmesi yerine özet bilgilerin kullanılması gerekebilir. Bu durumda OLAP küplerinin sağladığı özetleme fonksiyonlarından faydalanılabilir. Aylık satış fiyatlarının yıllık temelde daha küçük veri seti haline dönüştürülmesi örnek olarak verilebilir.

3.4.4.4.2. Boyut Azaltma

Veri kümeleri, analizle ilgisi olmayan veya gereksiz yüzlerce özellik içerebilir. Gereksiz olan özelliklerin azaltılması bir başka deyişle boyut azaltma pek çok veri madenciliği algoritmasının daha verimli çalışmasını, daha anlaşılabilir bir modelin oluşturulmasını, verilerin daha kolay görselleştirilmesini ve veri madenciliği algoritmaları için gerekli olan işlemci süresi ve hafızasını azaltır. İyi bir özellik alt kümesi asıl özelliklerden seçilir. Asıl özelliklerin sayısı "d" ise olası alt küme sayısı "2^d" olmaktadır. En iyi (veya en kötü) özellikler istatistiksel anlamlılık testleri kullanılarak belirlenir. Bu testler özelliklerin birbirinden bağımsız olduklarını kabul eder. Asıl özelliklerin sayısı fazla olduğunda en iyi alt kümenin belirlenmesi için yapılacak araştırma maliyetli olabilecektir. Bu nedenle azaltılmış özellik uzayını araştıran sezgisel yöntemler yaygın olarak kullanılır[46]. Özellik alt küme seçiminde "ileriye doğru seçme", "geriye doğru eleme" ve "ileriye doğru seçme ve geriye doğru eleme birleşimi" gibi tekniklerin uygulandığı sezgisel yöntemler kullanılır.

[46] Han, **Ön.ver.**, s.119.

Diğer bir yöntem de sınıflama için karar ağaçlarının oluşturulmasında kullanılan enformasyon kazanma (information gain) gibi ölçümlerin kullanılmasıdır. Eğer madencilik görevi sınıflama ise ve madencilik algoritması özellik alt kümesini belirlemede kullanılıyorsa bu özellik seçme yöntemine "sarmalama" (wrapper) yaklaşımı denilir[47]. Aksi takdirde bu bir süzme yaklaşımı olarak ifade edilir. Sarmalama yaklaşımı ile özellikler çıkartılırken algoritmanın değerlendirme ölçümünü en iyilediği için daha geçerli sonuçlara ulaşılır. Ancak süzme yaklaşımından çok daha fazla hesaplama gerektirir.

3.2.4.4.3. Veri Sıkıştırma

Veri sıkıştırmada veri kodlama veya dönüşümleri asıl verinin azaltılmış veya sıkıştırılmış gösterimini elde etmek için uygulanır. Asıl veri herhangi bir enformasyon kaybı olmaksızın sıkıştırılmış veriden tekrar elde edilebiliyorsa o zaman veri sıkıştırma işlemi "kayıpsız" (lossless) olarak nitelendirilir. Bundan başka asıl verinin gerçeğe yakın bir değeri oluşturulabilirse o zaman veri sıkıştırma kayıplı (lossy) olarak nitelendirilir. Metin verilerin sıkıştırılmasında kullanılan algoritmalar kayıpsız sıkıştırma yöntemleri olmalarına rağmen verinin sınırlı olarak işlenmesine neden olurlar. Bu nedenle daha yaygın ve etkili olan kayıplı yöntemler tercih edilir.

3.2.4.4.4. Büyük Sayıların Azaltılması

Verilerde yer alan büyük sayların daha küçük şekilleri seçilerek veri hacminin azaltılması için uygulanan yöntemlerdir. Veri hacmi parametrik veya parametrik olmayan yöntemler kullanılarak azaltılır. Parametrik yöntemlerde gerçek veri yerine sadece veri parametreleri saklanır ve sıkıştırılan veriyi tahmin etmek için bir model kullanılır. Parametrik olmayan veri indirgeme yöntemlerine histogramlar, kümeleme ve örnekleme gösterilebilir.

Parametrik olan regresyon ve logaritmik doğrusal regresyon modelleri verinin parametrelere dayalı gösterimini oluşturarak verinin azaltılmasında

[47] **Aynı**, s.121.

kullanılabilir. Doğrusal regresyon veriyi bir düz doğruya uydurarak modellerken çoklu regresyon birden fazla özellik vektörü kullanılarak veriye modeller. Logaritmik doğrusal regresyon kesikli çok boyutlu olasılık dağılımları yaklaşımlarını uygular ve kesikli özellikler kümesi için çok boyutlu veri küplerinin her hücresinin tahmin edilmesinde kullanılır.

Parametrik olmayan veri azaltma yöntemlerinde en yaygını histogram yöntemidir. Histogram yöntemi verileri farklı yöntemlerle aralıklara bölerek veriye ilişkin dağılımı elde eder. Diğer bir yöntem ise kümelemedir. Kümeleme veri azaltmada kullanılan verilerin kümelenerek daha küçük bir aralığa indirgenmesiyle gerçekleştirilir. Örnekleme de veri azaltma için kullanılan parametrik olmayan yöntemlerden biridir. Örnekleme geniş bir veri kümesinin çok daha küçük bir alt kümesi ile gösterilmesini sağlayabilir[48]. Veri azaltmada örnekleme yöntemi bir gruplama sorgusunun cevabını tahmin etmek için yaygın olarak kullanılır.

3.2.4.5. Benzerlik Ve Farklılıkların İncelenmesi

Benzerlik ve farklılık hesaplamaları; kümeleme, en yakın komşu sınıflaması, anomal değerlerin ortaya çıkarılması gibi veri madenciliği teknikleri tarafından kullanılır. Benzerlikler ve farklılıklar veri hazırlık aşamasında hesaplanırsa madencilik sürecinde doğrudan kullanılabilir. Bu tür yaklaşımlarda verilerin bir benzerlik (ya da farklılık) uzayına dönüştürüldükten sonra analiz gerçekleştirilir. Benzerlik ya da farklılığı ifade etmek için genellikle “yakınlık” (proximity) terimi kullanılır. İki nesne arasındaki yakınlık bu iki nesnenin uygun özellikleri arasındaki yakınlık fonksiyonu ile belirlenir.

İki nesne arasındaki benzerlik, nesnelerin birbirine ne kadar benzediğinin sayısal ölçümüdür. İki nesne arasındaki farklılık ise nesnelerin birbirinden ne kadar farklı olduğunun sayısal ölçümüdür. Bir benzerlik ölçümünü farklılık ölçümüne çevirmede dönüşümler uygulanır. Bir yakınlık ölçümünü [0-1]

[48] Berry, **Ön.ver.**, s.114.

aralığında ifade etmek bu dönüşüm işlemini kolaylaştırmaktadır. Bu durumda benzerlik, "1-farklılık" şeklinde hesaplanabilir[49].

3.2.4.5.1. Tek Özelliğe Sahip Nesneler Arasındaki Benzerlik ve Farklılıklar

Tek bir özelliğe sahip nesneler arasındaki yakınlık kullanılarak farklı tipteki özellikler için benzerlik ve farklılıklar hesaplanabilir. Nominal, ordinal veya aralık türünde bir özelliğe sahip x ve y gibi iki nesne için benzerlik ve farklılık hesaplamaları Tablo 2'de verilmiştir. Bu tabloda yer alan "Nominal" veri türü var ya da yok veya 0 ve 1 gibi değerler alan değişkenlerdir. "Ordinal" veri türü sıralı değerleri ifade ederken "aralık" veri türü belirli aralıktaki sayısal değerleri ifade eder.

Tablo 2. İki Nesne Arasındaki Bir Özeliğe İlişkin Benzerlik ve Farklılık Hesaplamaları.

Veri türü	***Farklılık***	***Benzerlik***
Nominal	$d = \begin{cases} 0 & x = y \\ 1 & x \neq y \end{cases}$	$s = \begin{cases} 1 & x = y \\ 0 & x \neq y \end{cases}$
Ordinal	$d = \frac{\lvert x - y \rvert}{n - 1}$	$s = 1 - d$
Aralık veya Oran	$d = \lvert x - y \rvert$	$s = -d, s = \frac{1}{1+d}, s = e^{-d},$ $s = 1 - \frac{d - enk(d)}{enb(d) - enk(d)}$

Tan, 2006, s.69.

Nominal değişkenlerin benzerlik ve farklılık hesaplamaları oldukça basittir. Ordinal değerler için öncelik "zayıf, orta, uygun, iyi, mükemmel" gibi değerler alan özellikler 0,1,2,3,4 gibi sıralı sayılara dönüştürülür. Elde edilen değerlerin 0 ile 1 arasında yakınlık ölçümü yapıldığında farklılık $d = \frac{\lvert x - y \rvert}{n - 1}$ şeklinde hesaplanır. Burada n nesnenin alabileceği değer sayısıdır. Benzerlik ise 1-d olarak bulunabilir. Aralık veya oran değerleri alan iki nesne arasındaki farklılık ölçümü nesnelerin değerleri arasındaki mutlak farktır ($d = \lvert x - y \rvert$). Bu

[49] Tan, **Ön.ver.**, s.66.

durumda farklılık değeri 0 ile 1 arası değil 0 ile ∞ arasındaki bir bölgede yer alır. Benzerlik ölçümü ise bir farklılığın bir benzerliğe dönüşümü kullanılarak hesaplanabilir. Benzerli 0 ile ∞ arasında hesaplamak için $s = -d$ veya 0 ile 1 arasında bir değer olarak ifade etmek istersek o zaman $s = \frac{1}{1+d}$, $s = e^{-d}$ ve $s = 1 - \frac{d - enk(d)}{enb(d) - enk(d)}$ dönüşümleri uygulanır.

3.2.4.5.2. Çoklu Özelliği Sahip Nesneler Arasındaki Faklılık ve Benzerlik Hesaplamaları

Birden fazla özelliğe sahip olan nesneler arasındaki farklılığın hesaplanmasında nesneler arasındaki uzaklık hesaplanmaktadır. Bu bağlamda uzaklık terimi farklılık yerine kullanılabilir. Uzaklık hesaplamaları için farklı yöntemler mevcuttur. Minkowski uzaklık hesaplamaları en yaygın olarak kullanılan yaklaşımlardır. $d_p(\boldsymbol{a},\boldsymbol{b}) = \|\boldsymbol{a} - \boldsymbol{b}\| = \left(\sum_j |a_j - b_j|^p \right)^{1/p}$ olarak ifade edilen Minkowski uzaklık hesaplamaları yaklaşımında J boyutları, a_j ve b_j ise sırasıyla **a** ve **b** vektörlerinin j. özelliklerinin değerlerini, p ise Minkowski uzaklık mertebesini gösterir[50].

Minkowski uzaklıkları p uzaklık mertebesinin aldığı değerlere göre özelleştirilebilir ve uzaklıklar farklı isimler ile ifade edilir. Buna göre

p=1 ise Hamming (Simetrik) uzaklığı $d_1(\boldsymbol{a},\boldsymbol{b}) = \sum_j^J |a_j - b_j|$

p=2 ise Öklid uzaklığı $d_2(\boldsymbol{a},\boldsymbol{b}) = \sqrt{\sum_j^J |a_j - b_j|^2}$

[50] Hervé Abdi ve D. Valantin, **Encyclopedia of Measurement and Statistics** - **Cilt1: Distance** (USA: Sage,2007), s.7.

p=∞ ise Supremum uzaklığı $$d_{\infty}(\boldsymbol{a},\boldsymbol{b}) = \lim_{p\to\infty}\left(\sum_{j}^{J}\left|a_j - b_j\right|^p\right)^{1/p}$$

olarak isimlendirilir. Minkowski hesaplamaları, pozitiflik, simetri ve üçgen eşitsizliği özelliklerine sahiptir. Veri nesneleri arasındaki benzerlikleri hesaplamada veri tipine göre farklılaşan birçok hesaplama tekniği kullanılır.

3.2.5. Veri Madenciliği Veritabanın Yüklemesi Ve Bakımı

Pek çok uygulamada veriler kendi veritabanlarında saklanmalıdır. Büyük miktardaki veya karmaşık veriler için veriler düz dosya yerine genellikle bir veritabanı yönetim sisteminde depolanır. Verilerin toplanıp, bütünleştirilip ve temizlendikten sonra veri madenciliği sisteminin ulaşacağı veritabanına yüklenmesi gereklidir. Veri madenciliği veritabanının yüklenmesi veritabanı yönetim yazılımına ve donanımına, veri miktarına, veritabanı tasarımının karmaşıklığına bağlı olarak enformasyon sistemi uzmanlarının deneyimini gerektirir.

Özellikle sistemle bütünleşik olarak kullanılacak veri madenciliği uygulamaları için veri kaynağını saklayan, düzenleyen ve güncelleyen veritabanlarının performansı gözlenmelidir. Veritabanının performansını iyileştirmek için veri şemalarının yeniden organize edilmesi gerekebilir. Ayrıca her veritabanının temel gereksinimi olan yedekleme faaliyetlerinin bu veritabanları için de periyodik olarak organize edilmesi gerekir.

3.3. Verinin İncelenmesi

Verinin incelenmesi verinin özelliklerinin daha iyi anlaşılmasını sağlar ve uygun veri analiz tekniğinin seçilmesine ve verinin model için hazırlanmasına yardımcı olur. Aynı zamanda veri madenciliği analizi tarafından cevaplanacak bazı sorulara ilişkin net ipuçları elde edilebilir. Örneğin, örüntüler görsel olarak verinin incelenmesi ile bulunabilir. Veri incelemesinde kullanılan görselleştirme gibi bazı teknikler de veri madenciliği sonuçlarını anlamada ve yorumlamada kullanılabilir. Özet istatistikleri ve görselleştirme veri incelemesinde yaygın

olarak kullanılan standart yöntemlerdir. Genellikle veri ambarlarında yer alan çok boyutlu verilerin incelenmesinde ise çok boyutlu veri analizinden faydalanılır. OLAP verinin ve verideki önemli örüntülerin anlaşılması için kullanıcılara çok boyutlu veritabanlarında inceleme yapmasına olanak sağlar. OLAP görselleştirme gibi sadece veri madenciliği için tasarlanmış bir araç değildir[51]. Çok boyutlu veri analizi geçmişi çok gerilere dayanmayan çok boyutlu değerler dizilerini incelemek için kullanılan teknikler kümesidir.

3.3.1. Özet İstatistikler

Özet istatistikleri büyük bir değerler kümesinin çeşitli özelliklerini yansıtan ortalama ve standart sapma gibi nicelikler ile tek bir sayı veya küçük sayılar kümesinden oluşur[52]. Bu istatistikler tek bir özelliğe ilişkin bilgiyi ve özellikler arasındaki ilişki hakkında bilgi sağlamak amacıyla kullanılırlar. Veri hakkında bilgi sağlayan istatistiksel ölçümler aşağıda özetlenmiştir.

- Frekans ve mod: Frekans sıralı olmayan kategorik değerler kümesinde her değerin kaç kere gözlendiğini gösterir. Mod ise gözlenen en büyük frekansa sahip olan değerdir.
- Yüzdelikler: Değerleri küçükten büyüğe dizilmiş ordinal veya sürekli değerlerin oluşturduğu bir küme dağılımını yüz eşit parçaya ayıran ölçülerdir.
- Aritmetik ortalama ve ortanca (yer ölçüleri): Aritmetik ortalama bir nesneler kümesindeki değerlerin toplamının nesne sayısına oranı olarak hesaplanan bir yer ölçüsüdür. Ortanca sıralı bir değerler kümesini birim olarak iki parçaya ayıran değerdir. Kümedeki nesne sayısının çift olması durumunda ortanca iki değerin ortalaması ile ifade edilir.
- Aralık ve varyans (dağılım ölçüsü): Sürekli veriler için değerler kümesinin dağılım ölçüleridir. Özellik değerlerinin ortalama gibi tek bir nokta etrafında toplandığını ya da geniş bir alana yayıldığını gösteren bir

[51] Berry, **Ön.ver.**, s.123.
[52] Tan, **Ön.ver.**, s.98.

ölçümdür. Aralık, değerler kümesinin en küçük ve en büyük değerini ifade eder. Varyans ise değerlerin ortalamadan olan farklarının karelerinin toplamının değer sayısının bir eksiğine oranıdır.

- Heterojenlik ölçümü: Kategorik değerler için belirlenen frekans dağılımının heterojenliğinin ölçülmesidir. Her bir kategorik değerin frekans dağılım olasılıklarının kareler toplamı temel alınarak hesaplanır.
- Çarpıklık ölçümü: Değerler kümesinin çarpıklığı değerlerin ortalama etrafında simetrik olup olmadığını belirleyen ölçümdür. Genellikle değerlerin dağılımının simetrisi normal dağılıma göre ölçülür.
- Basıklık ölçümü: Değerler kümesinin normal dağılıma göre basık ya da sivri olup olmadığını belirleyen bir ölçüdür. Bu tür ölçüler histogram grafiklerinin çizilmesi ile de görüntülenebilir.

Özet istatistikleri tek bir özellik için hesaplanabildiği gibi bir dizi özellikler kümesi (çoklu değişken kümesi) için de hesaplanabilir[53]. Örneğin $\bar{x} = (\bar{x}_1, \bar{x}_2, ..., \bar{x}_n)$ olarak ifade edilen dizi n adet özelliğe sahip bir veri kümesinin her özelliği için hesaplanmış aritmetik ortalama değerlerini ifade eder. Birden fazla özelliğe sahip veri kümlerinin özellikleri arasındaki ilişki ve bağımlılıkları incelemek için birçok analizin uygulanması gerekebilir. Örneğin özellikler arasındaki açıklanabilir ilişki miktarı korelasyon matrisleri oluşturularak incelenebilir.

3.3.2. Görselleştirme

Veri görselleştirme verinin tablolar veya grafikler halinde görüntülenmesidir. Başarılı bir görselleştirme verinin özelliklerinin ve veri parçaları veya özellikleri arasındaki ilişkilerin analiz edilmesi veya raporlanabilmesi için verinin görsel duruma dönüştürülmesini gerektirir. Grafik ve tablolar gibi görsel teknikler hava, ekonomi ve politik seçim sonuçlarını açıklamada kullanılan yaklaşımlardır. Algoritmaya dayalı tekniklerin veya

[53] **Aynı**, s.104.

matematiksel yaklaşımların veri madenciliği gibi birçok teknik disiplinde önemli olmasına rağmen görsel teknikler veri analizinde anahtar rol görevi görürler. Veri madenciliğinde görselleştirme tekniklerinin kullanılması görsel veri madenciliği olarak da ifade edilir. Görsel veri madenciliğinin temel amacı veriyi görsel şekilde sunmak, kullanıcının veriyi anlamasını sağlamak, sonuç çıkarmak ve doğrudan veri ile etkileşimini sağlamaktır[54].

Veri görselleştirmede verilere, özelliklerin veya boyutların farklı kombinasyonları olarak bakılabilir. Veriler sütun grafikleri, üç boyutlu küpler, veri sütun grafikleri, eğrileri, yüzeyleri, bağlantı grafikleri gibi çeşitli görsel şekillerde sunulabilir. Grafiklerde renk ve animasyon kullanımı da verinin anlaşılmasına büyük katkı sağlar. Görselleştirme aynı zamanda bilgisayar grafikleri, çoklu ortam sistemleri, insan-bilgisayar ara-yüzleri, örüntü tanımlama ve yüksek performanslı işleme konularıyla yakından ilişkilidir. Görselleştirme teknikleri az sayıdaki özelliklerin, uzaysal ve zaman bağımlı özelliklerin ve çok boyutlu verilerin görselleştirilmesi şeklinde üç grupta sınıflandırılabilir.

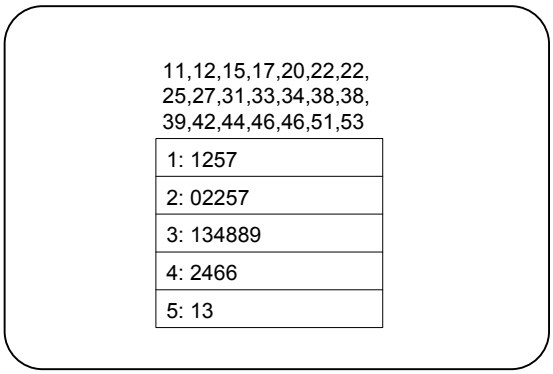
3.3.2.1. Az Sayıda Özelliğin Görselleştirilmesi

Histogramlar gibi görselleştirme teknikleri tek bir özellik için gözlenen değerlerin dağılımı hakkında fikir sahibi olmayı, saçılma grafikleri gibi diğer teknikler de iki özelliğin değerleri arasındaki ilişkiyi göstermeyi amaçlar. Özellik sayısı fazla olmayan veri kümelerinin veri kümelerinin görselleştirilmesinde kullanılan grafik türleri aşağıda sıralanmıştır.

- Gövde ve yaprak grafikleri bir boyutlu tamsayı veya sürekli verinin dağılımının gözlenmesini sağlar. Bu tür grafiklerin oluşturulmasında değerler gruplara ayrılır ve her grup gövdeyi gösterirken verinin son rakamları ise yaprakları oluşturur. Şekil 9'da örneği verilen gövdeler

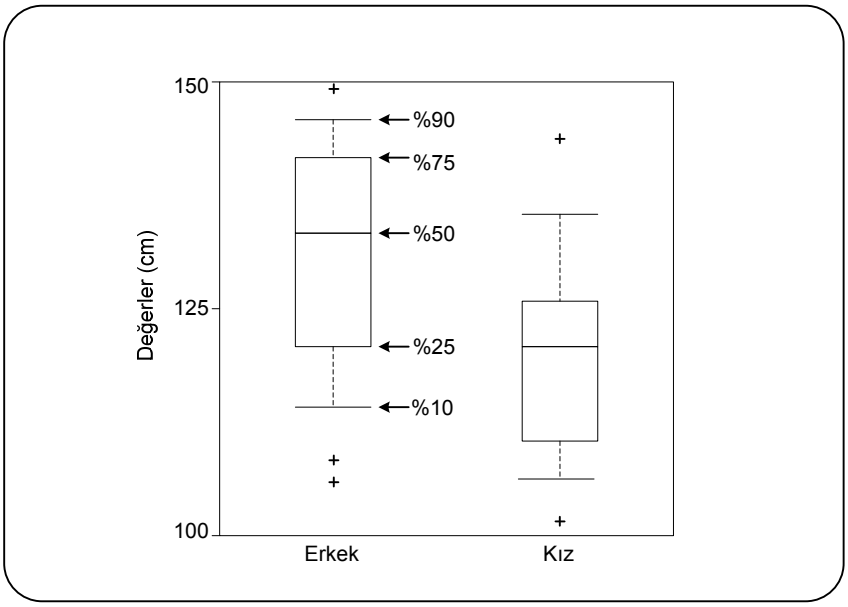
[54] Michael Berthold ve D. Hand, **Intelligent Data Analysis. (**2. basım. Berlin: Springster, 2003), s.404.

düşey, yapraklar ise yatay olarak yerleştirilir ve verinin dağılımı görselleştirilebilir.



Şekil 9. Gövde Yaprak Grafiği Örneği.

- Histogramlar özellik değerlerinin dağılımını göstermek için olası değerleri bölümlere ayırarak her bölümde yer alan nesne sayısını gösterir. Kategorik değerlerden oluşan özelliklerin histogramlarında her değer bir bölüm olarak ele alınır. Farklı kategorik değerlerin sayısının çok fazla olması durumunda değerler uygun bir şekilde birleştirilir. Bağıl histogram ve pareto grafikleri özelleştirilmiş histogram grafikleridir. İki boyutlu histogram ise iki özelliğin değerlerinin histograma yansıtıldığı grafiklerdir.
- Kutu grafikleri tek bir sayısal özelliğe sahip değerlerin dağılımını göstermek için kullanılır. Şekil 10'da bir ilköğretim okulunda öğrencilerin boylarının dağılımına ilişkin bir grafik verilmiştir. Bu grafikte erkek ve kız öğrencilerin boylarının uzunlukları iki ayrı özellik olarak ele alınmıştır. Artı işaretler aykırı değerleri, kesikli çizgiler %10'ar dilime kadar olan dağılımları ifade ederken kutuların bulunduğu kısım %50'lik dilimleri göstermektedir.
- Pasta grafikleri histograma benzer ancak özellikle az sayıda değerlere sahip kategorik özellikler için kullanılır.
- Yüzde grafikleri veri dağılımının daha net ifade edildiği deneysel birikimli dağılım fonksiyon grafikleridir. Her gözlenen değer için deneysel birikimli dağılım fonksiyonunun aldığı değerler bu grafiklerde gösterilir.



Şekil 10. Bir Okuldaki Öğrencilerin Boy Uzunluğu Ölçümlerine İlişkin Bir Kutu Grafiği.

- Saçılma grafikleri iki özelliğin değerleri kullanılarak her veri nesnesinin düzlemde bir nokta olarak gösterildiği grafikleridir. Saçılma grafikleri iki özellik arasındaki ilişkiyi gösterir ve doğrusal olmayan ilişkilerin fark edilmesinde kullanılır. Saçılma grafikleri şekil, renk, boyut ve tarama kullanılarak ek özellikleri göstermek için üç veya dört boyutlu grafiklere genişletebilir.

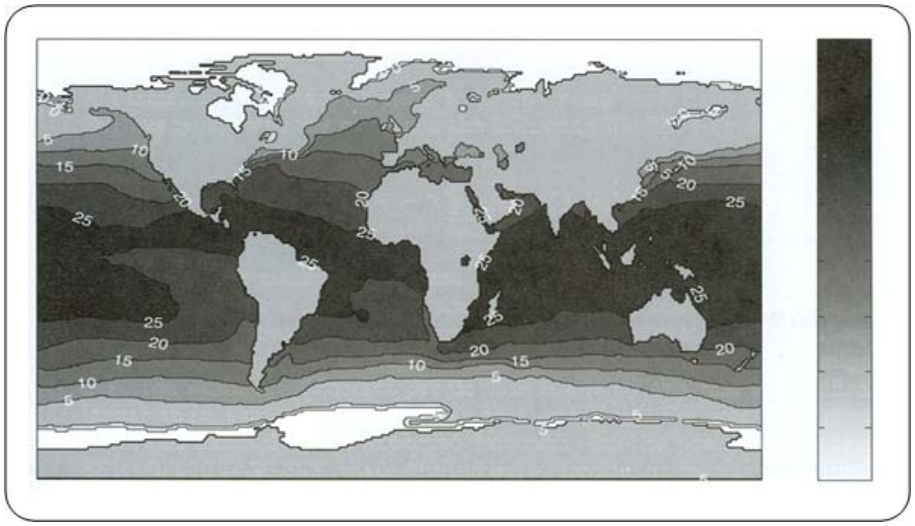
3.3.2.2. Uzaysal Ve Zaman Bağımlı Özelliklerin Görselleştirilmesi

Uzaysal veya zaman bağımlı özelliklere sahip veriler bir gözlemler kümesinden oluşur. Bu tür gözlemlere örnek olarak dünya yüzeyindeki hava basınç gözlemleri veya bir fiziksel sistemin benzetiminde çeşitli noktalardan ölçülen sıcaklık gözlemleri verilebilir. Ayrıca veriler günlük hisse senedi fiyatlarından oluşan zaman serisi verileri gibi tek bir zamana bağlı özellik içerebilir[55]. Bu tür verilerin görselleştirilmesinde özel grafik teknikleri kullanılır.

- Düzeç (contour) grafikleri üç boyutlu veriler için bir düzlemi bölgelere ayırır. Şekil 11'deki gibi iki özellik bir düzlemdeki bir konumu tanımlarken üçüncü özellik sıcaklık gibi sürekli bir değeri gösterir. Harita üzerinde yer

[55] Tan, **Ön.ver.**, s.119.

alan bölgelere ait yüksekliğin tanımlanması ile düzeç grafiğinin görünümü üç boyutlu hale getirilebilir. Bu şekilde bir boyut daha görselleştirilmiş olur.

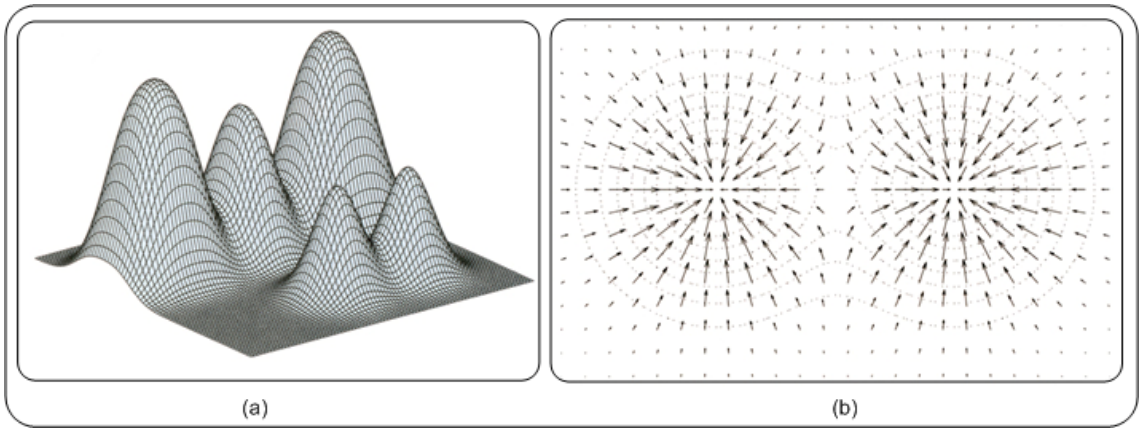


Şekil 11. Ortalama Deniz Yüzey Sıcaklığı (Aralık 1998).

Pang-Ning Tan, M. Steinbach ve V.Kumar, **Introduction to Data Mining** (USA: Pearson Education, 2006)'den uyarlandı.

- Yüzey grafiklerinde iki özellik düzlemi tanımlamak, üçüncü özellik ise düzlemin yüksekliğini göstermek için kullanılır. Yüzey grafikleri ilk iki özellik değerlerinin tüm kombinasyonları için tanımlanabilen üçüncü bir özellik değerini gerektirir. Eğer yüzey çok keskin hatlara sahipse o zaman grafiğe etkileşimli olarak bakılmadıkça tüm veriyi görmek zor olabilir. Bu nedenle yüzey grafikleri genellikle matematiksel fonksiyonları veya fiziksel yüzeyleri görselleştirmek için kullanılır (Şekil 12a).
- Vektör alan grafikleri hem büyüklük hem de yön özelliğine sahip olan verileri görselleştirmede kullanılır. Yoğunluğun konumla değişimi yön ve büyüklük özelliğine sahip vektör alan grafikleri ile görselleştirilebilir (Şekil 12b).
- Dört boyutlu bir veri kümesinin görselleştirilebilmesi için üç boyutlu veriyi içeren grafik serileri kullanılır. Şekil 11'de görülen aralık ayına ait grafiğin yılın her ayı için tekrarlandığını düşünebiliriz. Genellikle zamana dayalı veriler için çoklanan bu grafiklerde animasyon tekniği de kullanılabilmektedir. Animasyon kullanarak veri değişiminin incelenmesi

mümkün olabilmekte ancak bu tür görselleştirmeler sadece ekran ile sınırlı olabilmektedir.



Şekil 12. (a) Yüzey Grafikleri, (b) Vektör Alan Grafikleri.

Pang-Ning Tan, M. Steinbach ve V.Kumar, **Introduction to Data Mining** (USA: Pearson Education, 2006)'den uyarlandı.

3.3.2.3. Çok Boyutlu Verilerin Görselleştirilmesi

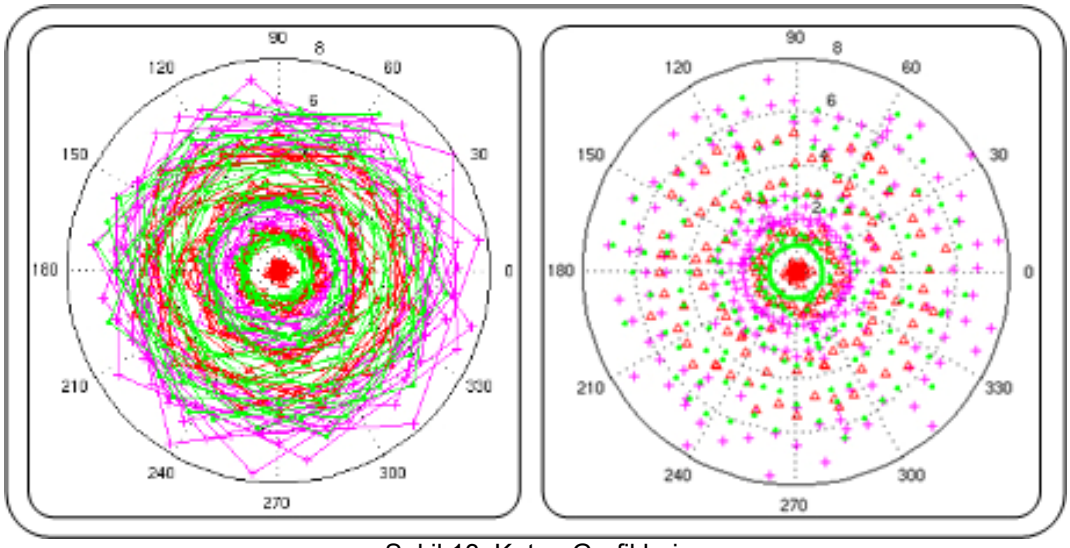
Diğer grafik tekniklerinin uygulanamayacağı kadar çok özelliğe sahip olan verilerin görselleştirilmesi için kullanılan grafik teknikleri verinin anlaşılmasında görsel değerlendirme olanağı sağlar. Matris, paralel koordinatlar, kutup grafikleri, ikon grafikleri çok boyutlu verilerin görselleştirilmesinde kullanılan tekniklerdir.

Matrisler renk ve parlaklıkla temsil edilen piksellerin dikdörtgen bir dizisidir. Bir veri matrisi her bir değerin resimdeki bir pikselle ilişkilendirildiği resim olarak görselleştirilebilir. Pikselin parlaklığı veya rengi matrisin ilgili değeri ile belirlenir. Veri sınıf etiketlerinin bilinmesi durumunda sınıftaki tüm nesnelerin bir araya getirilmesi için veri matrisinin tekrar düzenlenmesi kolaylık sağlar. Sınıf etiketlerinin bilinmemesi durumunda birbirine benzer özellik gösteren nesne ve özellik gruplarının bir araya getirilebilmesi ve görsel olarak tanımlanabilmesi için veri matrislerinin sıra ve sütunları yeniden düzenlenebilmelidir. Bu basit bir kümeleme işlemidir.

- Paralel koordinatlar her özellik için bir koordinat eksenine sahiptir ve farklı eksenler birbirine paraleldir. Ayrıca her bir nesne nokta değil bir

doğru ile gösterilir. Bir nesnenin her özelliğinin değeri özellikle ilgili koordinat üzerinde bir nokta olarak işaretlenir. Bu noktalar nesneyi gösteren doğruyu oluşturmak için birbirine bağlanır. Her nesnenin bir doğru ile ifade edildiği paralel koordinatlar grafiği karmaşık bir görünüm oluşturabilir. Bununla birlikte nesneler küçük bir grup içinde toplanma eğilimindedirler. Veri nesnelerin sayısı çok fazla olmazsa elde edilen paralel koordinatlar ilginç örüntüleri ortaya çıkarabilir.

- Kutup grafikleri açı ve yarıçap özellikleri kullanılarak verinin iki boyutlu yüzey üzerinde haritasının oluşturulması suretiyle çizilir. Bu grafikler paralel koordinat grafiklerinin dairesel gösterimi olarak düşünülebilir. Kutup grafikleri ile daha fazla boyut merkeze yakınlığa bağlı olarak görselleştirilebilir[56]. Çok boyutlu veri kümesinin kutup-çizgi ve kutup nokta grafikleri Şekil 13'de görülmektedir.



Şekil 13. Kutup Grafikleri.

Georges Grinstein, M. Trutschl ve U. Civek, "High-dimensional Visulations," **KDD-2001'de sunulan bildiri** (San Francisco. 2001)'den uyarlandı.

- Yıldız koordinatları ve Chernoff yüzleri çok boyutlu verilerin ikon veya sembollerle kodlanarak görüntülenmesidir. Bir nesnenin her özelliği, bir ikonun farklı uygun görünümü ile ifade edilir. Özelliğin değeri ikonun görünümünü yansıtacak özellikte olmalıdır. Yıldız koordinatları tekniği her

[56] Georges Grinstein, M. Trutschl ve U. Civek, "High-dimensional Visulations," **KDD-2001'de sunulan bildiri** (San Francisco. 2001), s.28.

özellik için bir eksen kullanır. Bu eksenlerin tümü bir merkez noktadan yayılır ve tüm özellik değerleri [0-1] aralığına ölçeklenir. Chernoff yüzleri tekniğinde ise her özellik bir yüz ikonunun görünümü ile ilişkilendirilir ve özellik değeri yüz şeklinin nasıl ifade edileceğini belirlemekte kullanılır.

3.3.3. Çok Boyutlu Veri Analizi

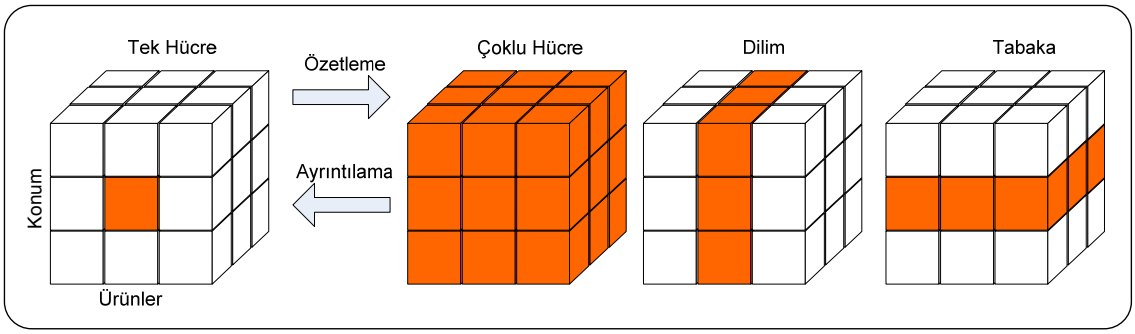
OLAP sistemleri etkileşimli veri analizi üzerine odaklanmıştır ve özellikle veri görselleştirme ve özet istatistiklerin oluşturulmasında üstün özelliklere sahiptir[57]. Bu bölümde çok boyutlu veri analizinde yaygın olarak kullanılan teknikler yer alacaktır.

Bir veri küpünün boyutu üçten fazla veya daha az olabilir. Bir veri küpü istatistik terminolojisinde çapraz tablo olarak bilinen yapının genelleştirilmiş halidir. Kısaca verinin incelenmesi amacıyla farklı özet düzeylerinde verilerin getirilmesi sağlanarak görselleştirilebilir ve istatistiksel özetleme tekniklerinin yardımıyla da verinin yapısıyla ilgili fikir sahibi olunabilir.

OLAP sistemleri ile ilgili dilim, tabaka, özetleme ve ayrıntılandırma işlemleri Şekil 14'te gösterilmiştir. Şekilde konum, ürün ve zaman boyutlarına sahip satış verilerinin oluşturduğu bir veri küpü görülmektedir. Tek hücre olarak tanımlanan işlem belirli konum, zaman ve ürüne ait toplam satışı ifade etmektedir. Dilim belirli bir ürüne ait tüm konum ve zamanlarda yapılan satış veri kümesine ulaşmayı sağlayan bir grup hücrenin seçilmesidir. Tabaka ise bir konumdaki tüm ürün ve zaman boyutundaki satış değerlerini sağlayan hücrelerin alt kümesinin seçimidir. OLAP sistemleri çoklu boyutlar üzerinde sınırlamalar yapılarak verinin istenen grubuna ulaşmak için gerekli işlevlere sahiptir. Bir konum kıta, ülke, şehir gibi çeşitli özelliklere sahip olabilir. Ürünler de mobilya, elektronik eşya, giyim gibi kategorilere bölünebilir. Bu tür kategoriler hiyerarşik bir ağaç veya örgü şeklinde düzenlenebilir. Bu hiyerarşik yapı özetleme ve ayrıntılama işlemlerinde verilerin farklı özet ve ayrıntı düzeylerinde elde edilmesinde kullanılır.

---

[57] Ye, **Ön.ver.**, s.402.



Şekil 14. OLAP Sorgu Türleri.

Margaret H. Dunham, **Data Mining** (New Jersey: Pearson Education, 2003)'den uyarlandı.

3.4. Model Oluşturma

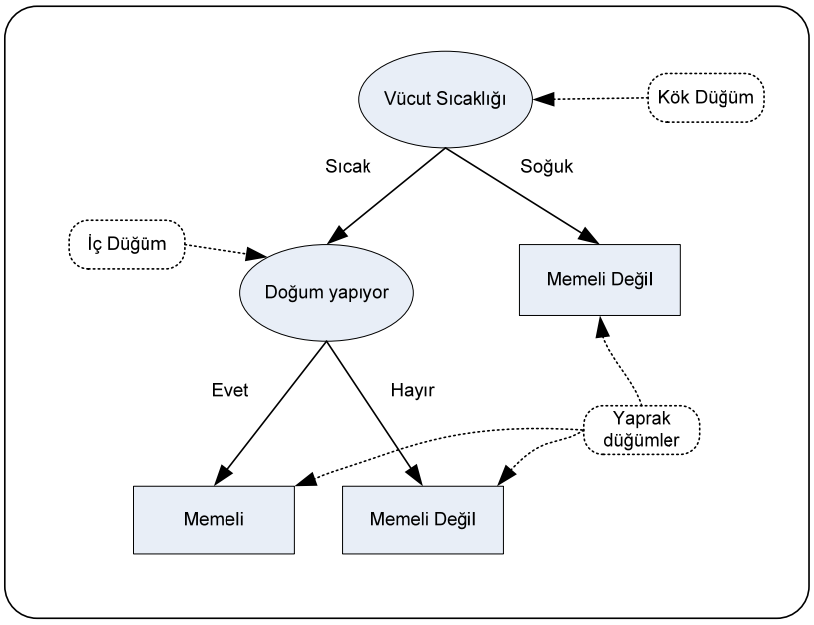
Veri madenciliği büyük hacimli verilerin işlenmesi için geliştirilmiş algoritmalar ile geleneksel veri analiz yöntemlerinin karması olan bir teknolojidir. Veri madenciliğinde büyük hacimlerde gözlenen verilerin analiz edilmesi diğer bir deyişle veriye en uygun hipotezlerin bulunması ile ilgilenilir. Tahmin edici ve tanımlayıcı veri madenciliği görevlerinin başarılmasında istatistik disiplininden örnekleme, tahmin ve hipotez testlerinden faydalanırken yapay zeka, makine öğrenmesi, örüntü tanımlama disiplinlerinden de arama algoritmaları, modelleme teknikleri ve öğrenme teorileri kullanılır.

Veri madenciliği farklı görevleri yerine getirmek amacıyla pek çok farklı algoritmayı kullanır. Algoritmalar veriyi inceler ve incelenen verinin özelliklerine en uygun modeli belirler. Verinin ve problemin özelliklerine göre uygulanabilecek birçok farklı algoritma sınıflama, kümeleme, birliktelik kuralları, örüntü tanımlama gibi görevlerin yerine getirilmesinde kullanılır. Veri madenciliğinde sıklıkla kullanılan teknikler aşağıda açıklanmıştır.

3.4.1. Karar Ağaçları

Karar ağaçları sınıflama, kümeleme ve tahmin görevlerinde kullanılan yaygın bir tekniktir. Bir karar ağacı düğümler ve düğümleri birleştiren bağlantılardan oluşan hiyerarşik bir yapıdır. Şekil 15'de örnek bir karar ağacı yapısı görülmektedir. Hayvanları sınıflamaya yönelik bu örnekte üç tür düğüm vardır. İlk sorunun cevaplandığı ve cevaba göre diğer düğümlere bağlanan en

üst düğüm kök düğüm olarak adlandırılır. Üst ve alt bağlantılara sahip düğümler iç düğüm, en alt noktada yer alan ve sınıf etiketi olarak belirlenen düğümler yaprak düğümleri veya uç düğümler adını alır. Şekilde en basit haliyle temsil edilen karar ağacı, veri kümesinde yer alan özellikler ve belirlenen çıktı alanı olan sınıf etiketi özelliği kullanılarak çeşitli algoritmaların uygulanmasıyla elde edilir.



Şekil 15. Karar Ağacı Yapısı.

Pang-Ning Tan, M. Steinbach ve V. Kumar, **Introduction to Data Mining** (USA: Pearson Education, 2006)'den uyarlandı.

Algoritmalar sınıflama problemlerini genellikle bir eğitim verisi üzerinde modeller ve daha sonra oluşturulan modeli test için ayrılan veri üzerinde doğrularlar. Karar ağacı düğüm ve bağlantılarını üreten algoritmaların genel yaklaşımı iki konu üzerine odaklanır. Bunlardan ilki veri kümesindeki kayıtları daha küçük alt kümelere bölmek için bir özelliğin seçilmesidir. Bu seçim veri tipine ve algoritmanın bölme için öngördüğü koşullara göre farklılık gösterir. Diğeri ise kayıtları alt kümelere bölme sürecinin sonlandırılmasıdır. Bu konuda olası bir strateji bir düğümü tüm kayıtların aynı sınıfa veya tüm kayıtların kendi özellik değerlerine kadar genişletilmesidir. Ancak algoritmaların veri kümesini alt düğümlere bölme işleminde durduracakları koşullar bulunmaktadır. Karar ağaçlarında amaç sonucu bilinen veriler üzerinde model oluşturulduktan sonra

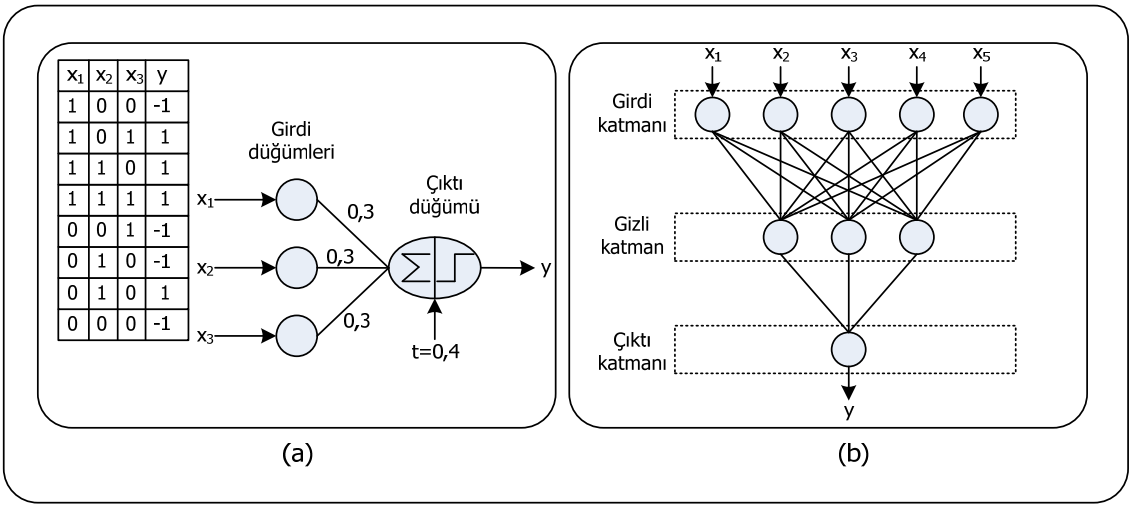
sonucu bilinmeyen veriler üzerinde özellikler kümesinin aldığı değerlere göre sonuç değişkeninin tahmin edilmesidir. Veri madenciliği yazılımlarının karar ağacı oluşturmak için yaygın olarak kullanıldığı algoritmalar arasında ID3, C4.5, C5.0, CART, QUEST, SPRINT, SLIQ algoritmaları yer almaktadır.

3.4.2. Yapay Sinir Ağları

Yapay sinir ağları karmaşık hesaplamaları gerçekleştiren biyolojik sinir sistemlerinin simülasyonudur. Biyolojik sinir sistemlerinde öğrenme, sinir hücreleri arasındaki etkileşim ile gerçekleşir. Biyolojik sinir sistemlerinin öğrenme özelliği, tanımlanan görevden bağımsız olarak esnek yapıda karmaşık verilerin işlenmesinde hesaplamaya dayalı modellerin oluşturulmasına esin kaynağı olmuştur[58]. Yapay sinir ağları örüntü tanımlama, konuşma tanımlama ve sentezi, tıbbi uygulamalar, hata tespiti, problem teşhisi ve robot kontrolü gibi alanlarda uygulama alanları bulmuştur.

Yapay sinir ağları veri madenciliğinde denetimli ve denetimsiz öğrenme amacıyla kullanılmaktadır. Sınıflama problemlerinin çözümünde ve bazı kümeleme problemlerinde faydalanılan yapay sinir ağları en basit haliyle Şekil 16(a)'da gösterilmiştir. Bu basit yapı giriş düğümleri ve bir çıktı düğümünden oluşur. Aradaki bağlantılar sinir hücreleri arasındaki sinaptik bağlantıları temsil eder ve her girdi özelliğinin ağırlığını gösterir. T ise bir öngörü değeridir ve sonucun hesaplanmasında aktivasyon fonksiyonu olarak adlandırılan fonksiyonun bir parametresidir. Bu örnek için aktivasyon fonksiyonu $\hat{y} = \text{sign}\,(0{,}3x1 + 0{,}3x2 + 0{,}3x3 - 0{,}4)$ olarak yazılabilir. Örnekte yer alan katsayılar modelin öğrenme sürecinde tekrarlı bir algoritma ile en iyi değerleri hesaplanmaya çalışılır. Her aşamada yapılan hatalar hesaplanarak ağırlık ve öngörü değeri yeniden belirlenir. Şekil 16 (b)'de beş değişken için çok katmanlı ileri yayılımlı bir yapay sinir ağı modelinin mimarisi verilmiştir.

[58] Berthold, **Ön.ver.**, s.269.



Şekil 16. (a) Basit Bir Yapay Sinir Ağı Modeli, (b) Çok Katmanlı İleri Yayımlı Bir Yapay Sinir Ağı Örneği.

Pang-Ning Tan, M. Steinbach ve V.Kumar, **Introduction to Data Mining** (USA: Pearson Education, 2006)'den uyarlandı.

Sınıflama problemlerini yapay sinir ağları ile çözmede dikkat edilmesi gereken konular aşağıda özetlenmiştir[59].

- Her sayısal veya ikili veri için bir giriş düğümü oluşturulmalıdır. Eğer kategorik veri mevcutsa sayısal dönüşüm uygulanmalıdır.
- Çıktı katmanında sınıflama problemlerinin tahmin edilecek özellik sayısında düğüm oluşturulmalıdır.
- Gizli katman sayısı, gizli düğüm sayısı, yayılım biçimi gibi ağ topolojisi belirlenmelidir. Doğru topolojiyi bulmak kolay değildir ancak en zordan basite doğru bir yaklaşım önerilebilir.
- Ağırlıklar ve öngörü değerlerinin başlangıç değerleri verilmelidir. Rasgele değerlerden oluşması uygun olmaktadır.
- Eğitim verisinde yer alan eksik değerler çıkarılmalı ya da doldurulmalıdır.

Yapay sinir ağları denetimsiz öğrenme modeli olarak kümeleme görevinde de kullanılmaktadır. "Kohonen" ağları olarak adlandırılan yapay sinir

[59] Tan, **Ön.ver.**, s.255.

ağı mimarisi her girdi özelliği için bir düğümü olan giriş katmanı ve genellikle iki boyutlu bir yüzeyde tanımlaman çıktı katmanı mimarisinden oluşur[60].

3.4.4. İstatistiğe Dayalı Teknikler

İstatistik veri madenciliğinin dayandığı temel disiplinlerden en önemlisidir. Veri madenciliği görevlerini başarmada kullanılan analiz tekniklerinin bir kısmı yapay zeka algoritmaları gibi hesaplamaya dayalı algoritmalar olurken istatistik hipotezlerinden türetilen birçok istatistiğe dayalı teknikler yaygın olarak kullanılmaktadır. Regresyon ve korelasyona dayalı teknikler, Bayes teoremine dayalı sınıflama teknikleri ve veri özetlemede kullanılan tanımlayıcı istatistik yöntemleri veri madenciliğinde kullanılan başlıca tekniklerdir.

Regresyon ve korelasyon iki değişken arasındaki ilişkinin değerlendirilmesinde kullanılabilir. Regresyon genellikle noktalar kümesini bir eğriye uydurarak gelecek değerleri geçmiş değerlere dayalı olarak tahmin etmede başvurulan bir tekniktir. Korelasyon ise iki değişken arasındaki benzerliği ölçmek ve sınıflama veya kümelemede benzerlik ölçümünü yapmak için kullanılabilmektedir. Regresyon tekniği sınıflama problemlerinin çözümünde kullanıldığında bağımsız değişkenler veritabanı özelliklerini, bağımlı değişken ise tahmin edilmesi gereken sınıf etiketini ifade eder.

Verinin özelliklerine göre verinin modellenebileceği farklı regresyon modelleri mevcuttur. Doğrusal regresyon, çoklu regresyon, doğrusal olmayan regresyon teknikleri sürekli değişkenlerin modellenmesinde kullanılırken lojistik regresyon kesikli veya kategorik verilerin modellenmesinde kullanılabilmektedir[61].

Sınıflama problemlerini çözmede kullanılan istatistiğe dayalı bir teknik de Bayes teoremidir. Bayes teoreminden faydalanan Bayes sınıflayıcıları verilen bir örneğin özel bir sınıfa ait olma olasılığı gibi sınıf üyelik olasılıklarını tahmin

---

[60] Roiger, **Ön.ver.** s.253.
[61] Olivia P. Rud, **Data Mining Cookbook** (New York: John Wiley & Sons, 2001) s.15.

edebilirler. Veri kümesindeki her özelliğin sınıflama problemine eşit katkıda bulunduğu ve katkıların birbirinden bağımsız olduğu varsayıldığında basit bir sınıflama olan “Naive Bayes” sınıflayıcısı kullanılabilir. “Naive Bayes” sınıflamada her bağımsız özelliğin katkısı analiz edilerek bir koşul olasılığı belirlenir. Sınıflama farklı özelliklerin etkileri birleştirilerek gerçekleştirilir. Algoritma test verisinden elde ettiği olasılıkları, sonucu bilinmeyen sınıfların etiketlenmesinde kullanır. Bayes sınıflayıcılar büyük veritabanlarında uygulandığında karar ağacı ve yapay sinir ağları sınıflayıcıları ile kıyaslanabilecek başarı gösterebilmişlerdir. “Naive Bayes” sınıflayıcılar gürültülü verinin etkilerini gidermede başarılıdırlar. Ayrıca eksik değerler bu yaklaşımda hesap dışı bırakılabilir[62].

Fakat “Naive Bayes” yaklaşımının özelliklerin bağımsız olmadığı durumlarda kullanılması tatmin edici sonuçlar veremeyebilir. Bu durumda özelliklerin alt kümeleri arasındaki ilişkilerin grafiksel model olarak gösterildiği “Bayes Belief” ağlarının kullanılması uygun olacaktır.

3.4.4. Genetik Algoritmalar

Genetik algoritmalar kesinlikle bir veri madenciliği modeli olmamasına karşın herhangi bir madencilik modelinde kullanılabilen bir eniyileme yöntemidir. Genetik algoritmalar da yapay sinir ağları gibi biyolojik mekanizmalardan esinlenerek geliştirilmiş algoritmalardır[63].

Genetik algoritmalar doğada gözlenen evrim sürecine benzer bir yapıda ele alınan problemi sanal olarak evrimden geçirerek çözmektedir. Problemin çözümü için öncelikle popülasyon olarak ifade edilen bir çözüm seti belirlenir. Bir popülasyondan alınan sonuçlar bir öncekinden daha iyi olacağı beklenen yeni bir popülasyonu oluşturmak için kullanılır. Yeni popülasyonların seçiminde her yeni bireyin problem için çözüm olup olmadığına uygunluk fonksiyonları kullanılarak karar verilir.

---

[62] Dunham, **Ön.ver.**, s.86.
[63] Giudici, **Ön.ver.**, s.199.

Genetik algoritmalar sınıflama, kümeleme ve birliktelik kurallarını içeren veri madenciliği problemlerini çözmede yardımcı araç olarak kullanılabilmektedir. Örneğin farklı sınıflayıcıların kombinasyonunu eniyilemek için genetik algoritma yaklaşımını uygulayan çalışmalar bulunmaktadır[64]. Problemlerin çözümünü parametre değerleriyle değil kodlarıyla arayan genetik algoritmalar parametreler kodlanabildiği takdirde çözüm üretebilir. Genetik algoritmalar çözümü noktalar kümesinden aramaya başladığı için genellikle yerel en iyi çözümde sıkışmazlar.

Bu avantajlara rağmen genetik algoritmaların uygulanmasında uzun kodlama süreleri, büyük hesaplama kaynağına ihtiyaç duyulması, sonuçların kolay yorumlanabilir olmaması ve uygunluk fonksiyonunun belirlenmesindeki güçlükler dezavantajlar olarak gösterilmektedir[65].

3.4.5. Model İçi Değerlendirme Süreci

Çözülmesi gereken problem için en uygun modelin bulunabilmesi çok sayıda modelin oluşturularak test edilmesi ile mümkündür. Bu nedenle bir veri kümesi için model oluşturma en iyi modele ulaşıncaya kadar tekrarlanan bir süreçtir[66]. Model oluşturma süreci denetimli ve denetimsiz öğrenme modellerine göre farklılık gösterir. Denetimsiz öğrenmede genellikle veriler analiz edilerek sınıfların tanımlanması amaçlanır. Bu tür modellerin geçerlilik yöntemleri modelin tipine göre farklılık gösterir.

Sınıflama, regresyon gibi tahmin edici modeller genellikle denetimli öğrenme modellerdir. Denetimli öğrenme modellerinde öncelikle model, bir miktar veri üzerinde çalıştırılır. Bu aşama modelin eğitilmesi olarak da adlandırılır. Daha sonra model verinin kalan kısmında test edilerek doğrulanır. Eğitim ve test döngüsü tamamlandıktan sonra model oluşturulur. Test kümesi

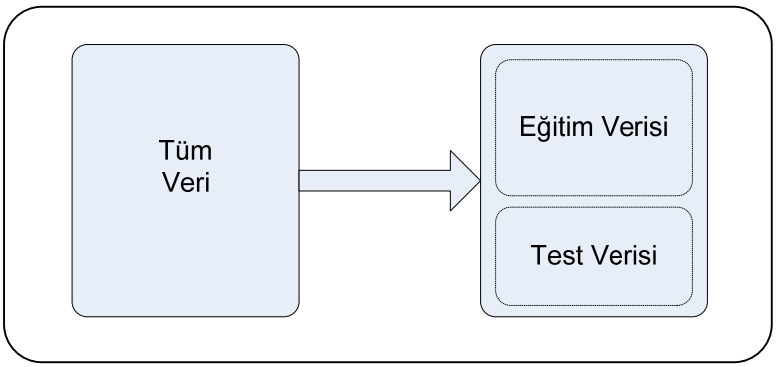
---

[64] Behrouz Minaei-Bidgoli ve W. F. Punch. "Using Genetic Algorithms for Data Mining Optimization in an Educational Web-Based System," **Genetic and Evolutionary Computation Conference'da sunulan bildiri** (Chicago, IL, USA: 12-16 Temmuz 2003), s.2252.
[65] Dunham, **Ön.ver.**, s.70.
[66] Two Crows Corp., **Ön.ver.**, s.27.

ile modelin doğruluk derecesi belirlenir. Bir modelin doğruluğunun test edilmesinde çeşitli geçerlilik yöntemleri kullanılır. Geçerlilik yöntemleri sınıflama modelleriyle bütünleşmiştir.

3.4.5.1. Basit Geçerlilik

Bir sınıflama modelinin doğruluğunun belirlenmesinde kullanılan en temel yöntemdir. Bu yöntemde veritabanının %5 ile %33 arasında bir kısmı test verisi olarak ayrılır ve modelin eğitilmesinde herhangi bir şekilde kullanılmaz[67]. Doğrulanması gereken tüm hesaplamalar için verinin bu şekilde iki gruba ayrılmasında seçimi tesadüfi olarak yapan yöntemler kullanılmalıdır. Böylece eğitim ve test veri kümleri modeli oluşturan veriyi temsil edecektir. Şekil 17'de basit geçerlilik hesaplamasında veri organizasyonu görülmektedir.



Şekil 17. Basit Geçerlilik Ölçümünde Veri Organizasyonu.

Nong Ye, **The Handbook of Data Mining** (USA: Lawrence Erlbaum, 2003)'den uyarlandı.

Model, eğitim verisine dayalı olarak oluşturulduktan sonra sınıfları veya test veritabanı değerlerini tahmin etmek için kullanılır. Test verisi üzerinde modelin çalıştırılmasından sonra elde edilen sonuçlar değeri bilinen gerçek verilerle karşılaştırılır. Yanlış tahmin edilen veya sınıflanan örnek sayısının toplam test örnek sayısına oranı modelin hata oranını verir. Benzer olarak doğru sınıflanan veya tahmin edilen verinin toplam test örnek sayısına oranı da doğruluk oranını verir. Bir başka şekilde ifade edilirse, "doğruluk oranı=1-(hata

[67] Haldun Akpınar, "Veritabanlarında Bilgi Keşfi ve Veri Madenciliği," **İşletme Fakültesi Dergisi**, (Cilt No: 29, Sayı No:1: 1-22, Nisan 2000). s.6.

oranı)" şeklinde yazılabilir[68]. Bir regresyon modeli için korelasyon katsayısının karesi (r^2) genellikle bir doğruluk tahmini olarak kullanılır.

3.4.5.2. Çapraz Geçerlilik

Çapraz geçerlilik yöntemi veri miktarının sınırlı olması halinde tercih edilen bir yöntemdir. Modelin oluşturulmasında tüm verinin kullanılmasına olanak sağlar. Çapraz geçerlilikte veri kümesi tesadüfî olarak iki eşit kümeye ayrılır. İlk olarak alt kümelerden birisi eğitim diğeri ise test için seçilir. Model oluşturularak test için veri üzerinde basit gereçlilik yönteminde olduğu gibi hata ve doğrulama oranı hesaplanır. Daha sonra test ve eğitim kümelerinin rolleri değiştirilerek aynı işlemler tekrarlanır. Elde edilen iki bağımsız doğrulama değerinin ortalaması alınarak modelin doğruluk oranı hesaplanır[69].

Her verinin bir kez eğitim ve bir kez de test olarak kullanıldığı çapraz geçerlilik yönteminin genelleştirilmiş hali n-katlı çapraz geçerlilik yöntemidir. Bu yöntemde veri kümesi tesadüfi olarak eşit n adet gruba ayrılır. Verilerin n gruba bölünmesinin ardından bir grubun test olarak belirlendiği ve kalan n-1 grubun da eğitim için kullanıldığı n adet basit geçerlilik süreci tekrarlanır. Böylece her grup bir kez test için kullanılmış olur. Elde edilen n adet bağımsız hata oranının ortalaması oluşturulan modelin hata oranı olarak kullanılır.

3.4.5.3. Bootstrap

Bir modelin hata oranının tahmin edilmesinde kullanılan ve çapraz geçerlilik yönteminde olduğu gibi tüm veri kümesinin model oluşturmada kullanıldığı bir yöntemdir. "Bootstrap" denilen örnek veri kümeleri asıl veri kümesinden örneklenerek oluşturulur. "Bootstrap" veri kümeleri modelin oluşturulması için eğitim verisi olarak kullanılır ve "Bootstrap" veri kümesinin dışında kalan veriler test verisi olarak kullanılır. Tesadüfi olarak örneklenen Bootstrap veri kümesinin büyüklüğünün tespitinde modelin hata oranını en iyi

[68] **Aynı.** s.6.
[69] Berthold, **Ön.ver.**, s.58.

temsil edecek oran 0,632 olarak belirlenmiştir[70]. Test veri kümesinin büyüklüğü $\left(1-\frac{1}{n}\right)^n$ ile hesaplanır ve bu oran n yeteri kadar büyütüldüğünde $e^{-1}=0{,}368$ değerine yaklaşır. Asıl verinin %63,2 gibi bir oranını temsil eden "Bootstrap" veri kümesi her yinelemede test kümesindeki verilerle yer değiştirir. Böylece her veri en az bir kez eğitim verisinde yer alır. Bazen bu işlem binin üzerinde tekrarlanır. Oluşturulan modelin hata oranı her "Bootstrap" örneğinin hata oranının ortalaması ile hesaplanır.

3.5. Modelin Değerlendirilmesi Ve Yorumlanması

Model oluşturulduktan sonra modelin sonuçlarının değerlendirilmesi ve elde edilen sonuçların öneminin yorumlanması gereklidir. Veri kümeleriyle ile ilgili olarak elde edilen modellerin değerlendirilmesi için sadece modellerin kendi aralarında karşılaştırılması değil aynı zamanda tercih edilen bir modelin uygulanması ile sağlanacak faydaların da karşılaştırılması gereklidir[71]. Bir sınıflama probleminin değerlendirilmesinde Risk Matrisi, Kaldıraç Grafiği ve ROC Grafiği yaygın olarak kullanılır.

3.5.1. Risk Matrisi

Risk matrisi sınıflama problemleri için, sonuçların anlaşılmasında yararlı bir araçtır. Bir risk matrisinde tahmin edilen sınıf değerleri satırlarda, gerçek değerler ise sütunlarda yer almaktadır. Bu yüzden matrisin köşegeni doğru tahmin edilen sınıf sayısını, diğer alanlar ise hata sayılarını gösterir. Tablo 3'de verilen örnek risk matrisinde 45 adet C sınıfına ait verinin 40 tanesi doğru tahmin edilmiş, 5 tane yanlış tahmin edilen verinin 3 tanesi A, 2 tanesi ise B olarak tahmin edilmiştir. Bir risk matrisi kullanarak modelin doğruluğunun ifade edilmesi modelin doğruluk oranının %82 gibi bir rakamla ifadesinden çok daha fazla bilgilendirici olmaktadır. Modelin değerlendirilmesinde bir başka boyut da maliyetlerdir. Bir sınıfın doğru tahmin edilmesinden sağlanacak gelirle bir sınıfın

[70] Standford University, Cross-Validation and the Bootstrap (Standford:1995), s.3.
[71] Giudici **Ön.ver.**, s.200.

yanlış tahmin edilmesi nedeniyle katlanılacak maliyetler hesaplanabilir. Bu durumda modelin doğruluk oranı yerine sağlayacağı marjinal fayda değerlendirilebilir.

Tablo 3. Bir Risk Matrisi Örneği.

Gözlenen / Tahmin	Sınıf A	Sınıf B	Sınıf C
Sınıf A	45	2	3
Sınıf B	10	38	2
Sınıf C	4	6	40

Giudici, 2003, s.201.

3.5.2. Birikimli Kazanç Eğrisi Ve Kaldıraç Grafiği

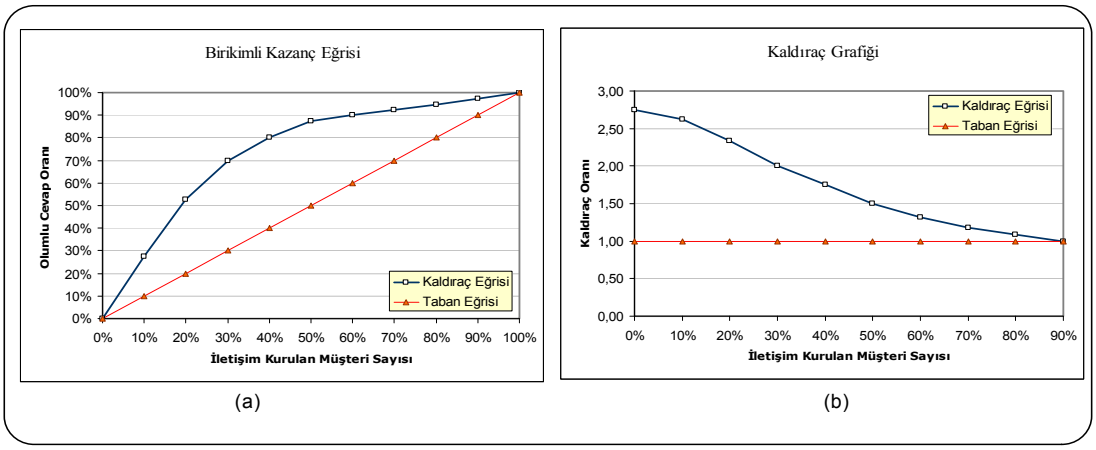
Kaldıraç Grafiği bir modelin sağladığı faydanın değerlendirilmesinde kullanılan diğer bir araçtır. Kaldıraç grafiği eğitim veri kümesine dayalı olarak gerçekleştirilen modelin test veri kümesinde başarılı olarak yaptığı tahminlerin oranını artan ya da azalan sırada grafiğe yansıtmasıdır. Kaldıraç grafikleri tahmin modeli olmadan gerçekleştirilen uygulama ile modelin uygulandığı durumda kazancın oranlarını grafiğe yansıtır. Örneğin bir şirketin müşterilerine tanesi 1 YTL ye mal olan posta atarak bir kampanya planlanmaktadır. Mevcut uygulamaya göre 100.000 müşteriden 20.000'nin geri dönmesi beklenmektedir. Uygulanan veri madenciliği çalışması ile bu müşteriler sınıflanarak Tablo 4'teki tahminler belirlenir. Bu durumda Şekil 18 (a)'da görülen birikimli kazanç eğrisi, yatay eksende posta atılan müşteri sayısı, düşey eksende olumlu cevap alınması beklenen müşteri yüzdeleri olmak üzere modelin sağladığı yüzdeler ve mevcut durumdaki yüzdeler işaretlenerek oluşturulur.

Şekil 18 (b)'de gösterilen kaldıraç grafiği, tahmin modelinin sağladığı olumlu cevap oranlarının mevcut durumdaki oranlara bölünmesiyle elde edilen kaldıraç değerlerinin düşey eksende, posta atılan müşteri oranlarının ise yatay eksende gösterilmesi ile elde edilir. Çizilen bu iki grafik modelin değerlendirilmesine yardım eden görsel araçlardır. Kaldıraç eğrisi ile taban eğrisi arasındaki alan ne kadar büyükse model o kadar iyidir. Alternatif

modellerden elde edilen tahmin değerleri aynı grafik üzerine çizilerek karşılaştırılabilir.

Tablo 4. Bir Şirketin Reklam Kampanyası Verileri.

Maliyet	***İletişim kurulan toplam müşteri sayısı***	***Tahmin edilen olumlu cevap sayısı***
10.000	10.000	5500
20.000	20.000	10500
30.000	30.000	14000
40.000	40.000	16000
50.000	50.000	17500
60.000	60.000	18000
70.000	70.000	18450
80.000	80.000	18950
90.000	90.000	19500
100.000	100.000	20000



(a) (b)

Şekil 18. (a) Birikimli Kazanç Eğrisi (Kaldıraç Eğrisi) (b) Kaldıraç Grafiği

3.5.3. Alıcı Çalışma Karakteristik Grafiği (ROC)

ROC (Receiver Operating Characteristic) grafikleri bir modelin doğruluğunu ölçmek için kullanılan diğer bir yöntemdir. Risk matrislerine dayalı olarak çizilen ROC grafiklerinde çıktı özelliği ikili (binary) ve parametrik olmayan sınıflama modellerinin değerlendirilmesinde kullanılır. Bu modellerde ikili çıktı alanı pozitif veya negatif olarak tahmin edilir. Modelin risk matrisi yanlışlıkla pozitif (YP), doğru olarak pozitif (DP), yanlışlıkla negatif (YN) ve doğru olarak negatif (DN) değerlerinden oluşur. Bir ROC grafiği YP oranını yatay eksende, DP oranını düşey eksende gösterir. Bu grafiğin (0,1) noktası tüm pozitif ve

negatif durumların doğru olarak tahmin edildiği mükemmel bir sınıflamanın gerçekleştiğini ifade eder. (0,0) noktası tüm durumların negatif, (1,1) noktası tüm durumların pozitif ve (1,0) noktası tüm durumların hatalı tahmin edildiğini ifade eder. Bir ROC grafiği sınıf dağılımından veya hata maliyetlerinden bağımsızdır. ROC grafiği bir sınıflama algoritmasının pozitif durumları doğru olarak belirleme yeteneği ile hatalı olarak sınıflanmış negatif durumların sayısı arasındaki ödünleşmeyi incelemek için kullanılan görsel bir araçtır.

3.6. Modelin Uygulanması ve İzlenmesi

Bir veri madenciliği modeli oluşturulduktan ve geçerliliği kabul edildikten sonra uygulama aşamasına geçilir. Veri madenciliği sonuçları modelin özelliğine göre iki şekilde uygulanabilir[72]. Bunlardan ilki modelin sonuçlarına göre faaliyetlerin önerilmesidir. Örneğin madencilik modelinin oluşturduğu kümelere veya modeli tanımlayan kurallara bakılarak faaliyet planları oluşturulabilir. Ayrıca kaldıraç ve ROC grafikleri kullanılarak faaliyetlerin sağlayacağı faydalar vurgulanabilir. Veri madenciliği sonuçların modeli farklı veri kümelerine uygulamakta kullanılabilir. Model, verinin sınıflanmasına dayalı olarak bazı nesneleri ön plana çıkarabilir. Bu verilere ilişkin OLAP sistemi aracılığıyla daha ayrıntılı analizler yapabilir.

Diğer uygulama şekli ise elde edilen modelin mevcut sistem içine konumlandırılmasıdır. Çoğu zaman veri madenciliği modelleri risk analizi, kredi değerlendirme veya dolandırıcılık tespiti gibi iş süreçlerinin parçasıdır. Bu durumlarda model iş sürecinde kullanılmak üzere bir yazılım haline getirilebilir. Örneğin tahmin edici bir model bir mortgage kredi uygulaması ile birleştirilebilir. Bu durumda model, bir kredi uzmanının müşterisini değerlendirmede kullanabileceği bir araç haline getirilebilir. Model bir envanter sipariş sistemi gibi bir uygulama içine gömülebilir. Bu durumda model çeşitli girdi parametreleriyle otomatik olarak sipariş verilmesini sağlayan bir modül olarak ortaya çıkar. Çoğu veri madenciliği uygulaması elde edilen modeli diğer uygulamalarda kullanılabilmeleri için farklı formatlarda kaydedebilirler.

[72] Two Crows Corp, **Ön.ver.**, s.33.

Model uygulandıktan sonra sistemin ne kadar iyi çalıştığının ölçülmesi gerekir. Model ne kadar iyi çalışıyor olsa da modelin performansının sürekli olarak izlenmesi gerekir[73]. Zaman içerisinde tüm sistemler değişime uğrar. Örneğin enflasyon oranı gibi dış etmenlerin değişmesi insanların davranış şekillerini değiştirebilir. Zamanla oluşturulan modelin değişen koşullara uyum sağlaması için test edilmesi, tekrar eğitilmesi ve gerekiyorsa yeniden oluşturulması gerekebilir. Tahmin edilen değerlerle gözlenen değerler arasındaki farkların grafikleri model sonuçlarının izlenmesinde mükemmel bir yoldur. Hesaplamanın yoğun olmadığı bu grafikleri kullanmak, anlamak kolaydır ve modeli uygulayan yazılımın içine yerleştirilmesi de mümkündür. Böylece sistem kendini izleyebilecektir.

[73] CRISP-DM Consortium, **Ön.ver.**, s.55

İKİNCİ BÖLÜM

EĞİTİMDE VERİ MADENCİLİĞİ

1. EĞİTİM VE VERİ MADENCİLİĞİ

Veri madenciliğinin uygulandığı birçok alanda olduğu gibi eğitimde de anlamlı ilişkilerin araştırılabileceği ve faydalı bilginin türetilebileceği geniş veri tabanları mevcuttur. Luan günümüzde yükseköğretim kurumları önündeki en önemli konulardan birisinin öğrenci ve mezunların takip ettikleri yolun kestirimi olduğunu belirtmekte, kurumların örneğin belirli derslere hangi öğrencilerin kayıt olacağı, hangi öğrencilerin mezun olabilmek için desteklenmesi gerektiği, başka kurumlara geçiş olasılığı yüksek öğrencilerin kim oldukları, katkıda bulunabilecek mezunların belirlenmesi gibi soruların cevaplarını bilmek istediklerini belirtmektedir[74]. Bunların yanında yükseköğretim kurumlarında geleneksel olarak kayıtların yönetimi, ortalama mezuniyet süresi gibi hususlarda daha iyi çözüm arayışları devam etmektedir. Verilerin analizi ve sunumu, başka bir deyişle, veri madenciliği bu sorunların çözümü için uygun yaklaşımlardan birisidir. Veri madenciliği yoluyla kurumlar mevcut raporlama yeteneklerini kullanarak geniş veri tabanları içersinde bilinmeyen örüntüleri ortaya çıkarıp anlayabilmektedirler. Bu örüntüler daha sonra veri madenciliği modelleri yoluyla bireysel davranışları yüksek bir doğruluk oranında kestirmekte kullanılmaktadır. Bunun sonucunda da kurumlar kaynaklarını çok daha etkin kullanabilmektedirler.

Delavari ve diğerleri yükseköğretim sistemlerinde karşılaşılan sorunların bilgi boşluğundan ortaya çıktığı olgusundan hareketle veri madenciliğine dayalı yeni bir model önermektedirler[75]. Bilgi boşluğu; planlama, değerlendirme ve danışmanlık gibi eğitim süreçlerinde yeterli miktar ve derinlikte bilgiye sahip olunmamasından kaynaklanmaktadır. Veri madenciliği yoluyla gizli örüntülerin,

---

[74] Jing Luan, **Data Mining Applications in Higher Education** (SPSS Inc.: http://www.spss.com/home_page/wp114.htm, 2004),s1.

[75] Naeimeh Delavari, M. R. Beikzadeh ve S. Phon-Amnuaisuk, "Application of Enhanced Analysis Model for Data Mining Process in Higer Educational System," **ITHET 6th da sunulan bildiri** (Juan Dolio. 8 Haziran 2005), s.1.

ilişkilerin veya anormalliklerin ortaya çıkarılmasıyla bu bilgi boşluğu kapatılabilir. Modelin eğitim kurumlarında organizasyonel bir iyileşme için veri madenciliğinden yararlanılmasında rehber ya da yol haritası olabileceği belirtilmektedir. Modelde öncelikle ana süreç ve alt süreçler ortaya konmakta, bunlarla ilgili veri madenciliği teknikleri yoluyla keşfedilecek bilgilerin neler olduğu belirlenerek her kategoride süreçlerin iyileştirilmesi ya da yeni süreçlerin uygulanması sağlanmaktadır.

Vranić ve Skoćır veri madenciliği algoritmaları ve tekniklerinin akademik ortamlarda eğitsel kalitenin bazı yönlerini nasıl iyileştireceğini belirli bir dersin öğrencilerini hedef kitle alarak incelemektedirler[76]. Bu öğrencilere ilişkin yararlı olabilecek ancak henüz keşfedilmemiş bilgilerin mevcut olduğu olgusundan yola çıkılmaktadır. Çalışma yoluyla öğrencilerin davranışlarını ve çeşitli konuları öğrenme becerilerini anlamak ve öğrencilerinin elde ettiği başarı doğrultusunda bir sonraki yılın öğrencilerinin başarısını kestirmek amaçlanmaktadır.

Günümüzde giderek yaygınlaşan elektronik ortamlarda öğrenme olarak adlandırılan e-öğrenmede geleneksel yöntemle öğrencinin izlenmesi mümkün değildir. Bu nedenle eğitimciler öğrencinin öğrenme süreci içindeki davranışlarını izlemek için farklı yöntemler aramalıdır. Uzaktan eğitim organizasyonları, web sunucuları tarafından otomatik olarak oluşturulmuş veya öğrenme yönetim sistemleri güncelerinde depolanmış büyük hacimli verileri biriktirirler. Web'e dayalı öğrenme ortamları öğrencilerin pek çok öğrenme davranışlarını kaydedebilir ve böylece öğrenme profili hakkında bilgi sağlayabilir.

Web'e dayalı öğrenme, herhangi bir mekanda yer alan bir donanımdaki içeriğin mekandan bağımsız olarak öğrenciye ulaştırılmasını içerir. Son yıllarda binlerce ders uzaktan eğitim organizasyonları tarafından web ortamında yayınlanmaktadır. Ancak çoğu web'e dayalı ders, öğrenci faklılığını hesaba katmayan statik öğrenme materyaline dayalıdır. Uyarlanabilir ve zeki web'e

[76] Mihaela Vranić, Damir Pintar ve Zoran Skoćır, "The Use of Data Mining in Education Environment," **ConTEL 2007'de sunulan bildiri** (Zagrep 13-15 Haziran 2007), s.243.

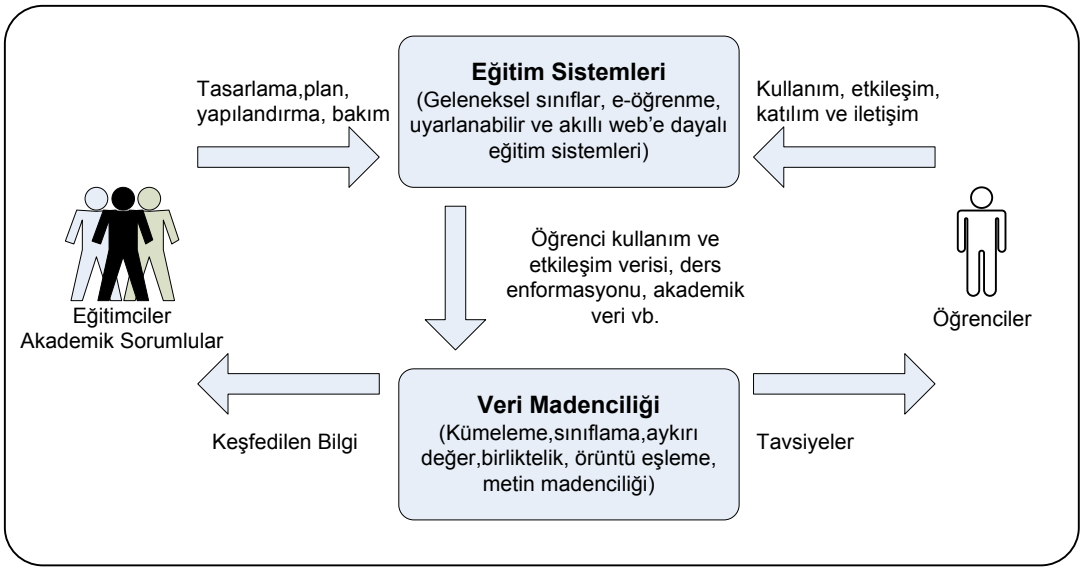
dayalı eğitim sistemleri bireysel olarak daha zengin öğrenme ortamları sunan bir çözüm olmaktadır. Bu sistemler öğrenenlere bireyin amaçları, tercihleri ve tecrübesine dayalı bir model oluşturarak kişisel eğitim olanağı sunar. Büyük veri birikimlerinden kesin ve ilginç örüntülerin otomatik olarak çıkarıldığı veri madenciliği, öğrenme süreci veya öğrenci davranışları hakkında bilgi sahibi olmak için kullanılabilir. Ayrıca bu bilgi ışığında e-öğrenme sistemlerinin değerlendirilmesi ve geliştirilmesi mümkündür[77].

Eğitim sistemlerinde öğrenmeyi geliştirmek için veri madenciliğinin uygulanması biçimlendirici değerlendirme (formative evaluation) yöntemi gibi görülebilir[78]. Öğrencinin sistemi nasıl kullandığını incelemek biçimsel olarak içerik tasarımını değerlendirmektir. Bu sayede eğitimciler daha mükemmel içerikleri tasarlamak için gerekli bilgiyi elde ederler. Veri madenciliği ile elde edilecek bilgiler, eğitimcilere öğrenme ortamını veya yaklaşımını tasarlarken ve güncellerken pedagojik temele dayalı kararlar vermelerine yardımcı olmak için biçimlendirici değerlendirmede kullanılabilirler. Şekil 19'da eğitim sistemlerinde uygulanan veri madenciliğinin tekrarlı döngüsü görülmektedir.

Veri madenciliği dersler, öğrenciler, kullanım ve etkileşim hakkında tüm mevcut enformasyondan başlanarak e-öğrenme sürecini geliştirmeye yardım eden faydalı bilgiyi keşfetmek için uygulanır. Veri madenciliğinin eğitim sistemlerinde kullanılması, farkı amaçlar için öğrencilere, eğitimcilere, akademik sorumlulara ve yöneticilere göre yönelim gösterir. Eğitimcilere yönelik veri madenciliği yoluyla ders içerik ve etkinliğinin değerlendirilmesi, öğrencilerin elde edilen örüntülere göre gruplandırılması, sık tekrarlanan hataların bulunması, daha etkili faaliyetlerin belirlenmesi, dersin kişiselleştirilmesinin sağlanması gibi amaçlar için ders hakkında daha nesnel geri bildirim sağlanır. Akademik sorumlular ve yöneticilere yönelik veri madenciliği çalışmalarının amaçları arasında e-öğrenme sisteminin etkinliğini arttırmak, kaynak kullanımı hakkında

[77] Osmar Zaïane ve J. Luo, "Web Usage Mining for a Better Web-based Learning Environment," **Advanced Technology for Education'da sunulan bildiri** (Banff, Alberta. 27-28 Haziran 2001), s.60.
[78] Cristobal Romero ve S. Ventura, "Educational Data Mining: A Survey from 1995 to 2005," **Expert Systems with Applications**. Cilt No 33, Sayı No 1: 135-146, (2007), s.136.

bilgi sağlamak, kurumsal kaynakları daha iyi organize etmek sıralanabilir. Öğrencilere yönelik çalışmalarda amaç genellikle öğrenme sürecine yardımcı olacak ve geliştirecek faaliyet, kaynak ve öğrenme deneyimleri ile ilgili öneriler sunmak, öğrencilerin geçmiş deneyimlerine ve başarılarına dayalı olarak e-öğrenme sistemi içinde öğrencilere kolay ulaşım olanağı sunmaktır.



Şekil 19. Eğitim Sistemlerinde Uygulanan Veri Madenciliğinin Tekrarlı Döngüsü.

Cristobal Romero ve S. Ventura, "Educational Data Mining: A Survey from 1995 to 2005," **Expert Systems with Applications**. Cilt No 33, Sayı No 1: 135-146, (2007)'den uyarlandı

Eğitim alanında gerçekleştirilen veri madenciliği uygulamaları geleneksel eğitim ve uzaktan eğitim olmak üzere iki kısımda incelenebilir. Farklı veri kaynaklarına sahip olan bu eğitim sistemlerinde farklı veri madenciliği modelleri uygulanmaktadır.

2. GELENEKSEL EĞİTİM SİSTEMLERİ

Eğitimci ile öğrencinin yüz-yüze iletişimde bulunduğu geleneksel eğitim ortamlarında veri madenciliği öğrenme sürecinde hem eğiticiye hem de öğrenciye yardımcı olabilir. Geleneksel eğitim kurumlarında öğrenci, ders, eğitimci, ders çizelgeleri gibi bilgiler eğitim enformasyon sistemlerinde saklanmaktadır. Ayrıca sistem hakkında bilgi sağlayan web sayfaları, öğrencilerin kullanımına sunulmuş çevrimiçi kütüphane ve çoklu ortam

veritabanları bulunabilmektedir[79]. Eğitim kurumunda bir eğitim dönemi planlanırken kayıt olacak ya da mezun olması muhtemel öğrenci sayıları tahmin edilmek istenir. Bunun amacı gelecek dönemle ilgili hazırlık faaliyetlerinin gerçekleştirilmesidir. Öğrenciler en iyi performansı sergileyecekleri dersleri belirlemek isterken eğitimciler de ders verdikleri öğrenci gruplarının performansını nasıl etkilediklerini bilmek isterler. Veri madenciliği bu soruların yanıtlanmasına katkıda bulunabilir. Geleneksel eğitimde uygulanan bazı örnek veri madenciliği çalışmalarına aşağıda yer verilmiştir.

- Eğitimde veri madenciliğinin uygulandığı ilk çalışmalardan birisi 1995 yılında Sanjeev ve Zytkow tarafından yayınlanmıştır[80]. Araştırmacılar bilgi keşfini "R aralığındaki veriler için P örüntüsü" şeklinde ifadeler halinde üniversite veritabanından elde etmişlerdir. Sonuçlar kurumsal politikalarla ilgili stratejik kararların verilmesi için üniversite yönetimine sunulmuştur.
- Veri madenciliğinin eğitimde uygulandığı diğer bir çalışma da Brezilya'daki bir üniversitede müfredat değişikliklerinin öğrenci üzerindeki etkilerinin tanımlanması ve anlaşılması için Becker ve diğerleri tarafından 2000 yılında yayınlanmıştır[81]. Araştırmacılar değişikliklerin nitel etkisini doğrulamışlardır ve özetleme, birliktelik kuralları ve sınıflama gibi veri madenciliği görevlerini kullanarak bu etkiyi değerlendirmişlerdir.
- Diğer bir uygulama ise Singapur Eğitim Bakanlığının özel bir eğitim programı için zayıf öğrencilerin seçilmesi ile ilgilidir. Ma ve diğerleri tarafından 2000 yılında gerçekleştirilen çalışmada birliktelik kurallarına dayalı olan bir puanlama fonksiyonu geliştirilmiştir[82]. Problemin çözümünün ilk aşamasında C4.5 sınıflama algoritması kullanılarak potansiyel zayıf öğrenciler belirlenmiştir. İkinci aşamada ise her zayıf

---

[79] Yiming Ma ve diğerleri, "Targeting the Right Students Using Data Mining," **The 7th ACM SIGKDD'de sunulan bildiri** (San Francisco. 26-29 Ağustos 2001), s.457.
[80] Arun P. Sanjeev ve J. M, Zytkow. "Discovering Enrollment Knowledge in University Databases," **1th Conference on KDD'de sunulan bildiri** (Montreal. 20-21 Ağustos 1995), s.246.
[81] Karin Becker, C. Ghedini ve E.L. Terra, "Using KDD to analyze the impact of curriculum revisions in a Brazilian university," **SPIE 14th Annual International Conference'da sunulan bildiri** (Orlando. Nisan 2000), s.412.
[82] Ma, **Ön.ver.**, s. 457.

öğrencinin alması gereken dersler birliktelik analizi kullanılarak belirlenmiştir.

- Yükseköğretimde öğrencilerin belirleyici özelliklerinin kullanıldığı öğrenci memnuniyetini ölçmeye yönelik bir veri madenciliği uygulaması 2002 yılından Luan tarafından gerçekleştirilmiştir[83]. Luan eğitim kurumlarının kaynak ve personel kullanımını daha verimli hale getirebilmeleri için C5.0 gibi tahmin edici denetimli öğrenme modelleri ve Kohonen ağları gibi kümeleyici denetimsiz öğrenme modellerini kullanmayı önermiştir.
- Maltepe üniversitesi öğrencilerinin belirleyici özelliklerinin "K-Means" algoritması kullanılarak kümelendiği bir çalışma 2005 yılında Erdoğan ve Timor tarafından yayınlanmıştır. 2003 yılına ait 722 öğrenci verisinin kullanıldığı çalışmada öğrencilerin üniversiteye giriş sınav sonuçları ile başarıları arasındaki ilişki incelenmiştir[84].

Geleneksel eğitim kurumlarında gerçekleştirilen veri madenciliği uygulamalarında genellikle eğitim uzmanlarından çok yöneticilere yönelik araştırmalar gerçekleştirilmiştir. Bunun olası nedeni geleneksel eğitim sisteminde depolanan veri özelliklerinin öğrenme davranışlarına ilişkin araştırmalara olanak vermemesidir. Geleneksel eğitim sistemlerinde uygulanan veri madenciliği, ulusal eğitim problemlerinin çözümünde de önemli faydalar sağlayabilecektir.

3. UZAKTAN EĞİTİM SİSTEMLERİ VE VERİ MADENCİLİĞİ

Uzaktan eğitim veya uzaktan öğrenme, zaman ve mekan olarak öğretenden uzak olan öğrencilerin eğitim programına ulaşmasını sağlayan yöntem ve teknikler olarak tanımlanabilir[85]. Uzaktan eğitim teknolojisinin gelişimine paralel olarak mektupla öğrenme, video-kaset eğitimi, bilgisayar destekli eğitim (çoklu ortam eğitimi, internet eğitimi ve web'e dayalı eğitim) gibi

---

[83] Jing Luan, "Data Mining, Knowledge Management in Higher Education, Potential Applications", **42nd Associate of Institutional Research International Conference çalıştayında sunulan bildiri** (Toronto,Canada: 2002), s.1.
[84] Şenol Erdoğan ve Mehpare TİMOR, "A Data Mining Application in a Student Database," **Havacılık ve Uzay Dergisi**. Cilt No 2,Sayı 2: 57-64, (Temmuz 2005), s.57.
[85] Romero, **Ön.ver.**, s.138.

farklı isim ve yöntemlerle uygulanmıştır. Günümüzde yaygın olarak kullanılan web'e dayalı eğitim ise öğrencilerin interneti kullanarak öğrenmelerini sağlamaktadır. Web'e dayalı eğitim, uzaktan eğitimin internet aracılığı ile gerçekleştirilmesini sağlarken e-öğrenme, e-eğitim, çevrimiçi ders gibi terimlerle de anılmaktadır.

Web'e dayalı eğitim sistemleri özelliklerine ve kullanım amaçlarına göre farklı sistem türleri ortaya çıkarmıştır. Web'e dayalı öğrenme sistemleri eşzamanlı (senkron) ve eşzamanlı olmayan (asenkron), işbirliğini destekleyen ve desteklemeyen, anonim erişim ve sınırlı erişim olmak üzere farklı türlere ayrılabilmektedir. Tipik bir web'e dayalı öğrenme ortamı ders içerik hazırlama araçları, eşzamanlı ve eşzamanlı olmayan konferans sistemleri, anket ve kısa sınav bileşenleri, kaynak paylaşımı için sanal çalışma ortamları, beyaz tahta, not raporlama sistemi, günce kitabı, ödev yayınlama bilişenlerini içermektedir[86]. Sanal sınıf uygulamalarında ise eğitimciler, öğrencilere eşzamanlı olarak metin, çoklu ortam ve simülasyon dosyalarını paylaşabilmekte ve tartışma ortamı oluşturabilmektedir. Öğrenciler bu araçlar sayesinde öğrenme faaliyetlerini sürdürebilmektedirler. Ancak eğitimcilerin öğrencilerin tüm bu araçlarla gerçekleştirdikleri eğitim faaliyetlerini izlemeleri ve değerlendirmeleri mümkün olamamaktadır. Her ne kadar gelişmiş web'e dayalı öğrenme sistemleri öğrencilerin faaliyetlerine ilişkin istatistiksel raporlar sunsalar da öğrencilerin öğrenme stillerini anlamaya yönelik anlamlı bilgilerin türetilmesini sağlayacak kadar gelişmiş araçlara sahip değildirler[87]. Ancak diğer taraftan hemen hemen tüm web'e dayalı öğrenme sistemleri öğrencilerin sisteme ulaşma ve sistem içersindeki davranışlarını güncellere kaydetmektedirler. En basit ve ilkel haliyle bir web sayfası kendisine ulaşan kullanıcı bilgisi sınırlı da olsa güncelere kaydetmektedir. Web sistemlerinin depoladığı günceler üç farklı yapıda toplanmaktadır[88].

---

[86] Zaïane, **Ön.ver.**, s.60.
[87] **Aynı**, s.60.
[88] Jaideep Srivastava ve diğerleri, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data," **SIGKDD Explor. Newsl.**. Cilt No 1, Sayı No 2: 12–23, (Ocak 2000), s.12.

- *Sunucu düzeyinde:* Veri madenciliği genellikle bu günceler üzerinde uygulanır. Sunucu yazılımın oluşturduğu kayıtlarda genellikle ulaşılan web sayfası, zaman, alınan veya gönderilen dosyanın boyutu, kullanıcıya ilişkin bir takım veriler yer almaktadır. Web sunucularının kaydettikleri güncelerin içerdikleri veri ve türleri standartlarla belirlenmiştir.
- *İstemci düzeyinde:* İstemci tarafında internet uygulamaları veya internet gezginleri tarafından oluşturulan günce dosyalarıdır. Bu güncelerin kullanıcı tarafında depolanması, kullanıcıların tekrar bağlanmada hızlı ulaşımını sağlar. Ancak bu verilere sunucu tarafından ulaşılamadığı için değerlendirilmesi mümkün olmamaktadır.
- *Proxy düzeyinde:* Sunucu ve istemci arasında web sayfalarının daha hızlı iletilmesini sağlayan sistemlerin oluşturduğu tampon veri ve güncelerdir.

Web sunucuları kullanıcılarına ait web kullanım bilgilerini güncelere belirtilen ayrıntılarda kaydederler. Bu bilgiler web'e dayalı öğrenme sistemlerinde öğreticilerin öğrenme faaliyetleri hakkında bilgi sağlayabilir. Eğitim amacıyla geliştirilmiş Blackboard, Virtual U vb. uygulamalar da sistem içindeki öğrenci hareketlerini kendi veritabanlarına kaydederler.

Web sistemlerince kaydedilen verilerin analiz edilmesinde kullanılan veri madenciliği, web madenciliği olarak adlandırılır. Web madenciliği web verilerinden bilgi keşfi için veri madenciliğinin uygulanmasıdır. Web madenciliği görevleri web içerik madenciliği (web content mining), web yapı madenciliği (web structure mining) ve web kullanım madenciliği (web usage mining) olmak üzere üç temel gruba ayrılır. Web içerik madenciliği web dokümanlarının içeriklerinden faydalı bilgiyi keşfetme sürecidir. Basit bir arama motorunun gerçekleştirdiği işin genişletilmesi gibi düşünülebilir. Web yapı madenciliği web sayfalarını sınıflandırmak veya dokümanlar arası benzerlik ölçümlerini yapmak için kullanılmaktadır. Bir başka deyişle web'den yapısal bilginin keşfedilme süreci olarak tanımlanabilir[89]. Web kullanım madenciliği ise web kullanım verileri

[89] Dunham, **Ön.ver.**, s.204.

veya web güncelerinden kullanım bilgisine ilişkin anlamlı örüntülerin keşfedilmesidir.

Ha ve diğerleri internet çağında dijital teknolojinin insanoğlu tarafından eğitime uyarlanmasının dört eğilim çerçevesinde gerçekleşmekte olduğunu belirtmektedirler[90]. Bu eğilimlerden birincisi zihinsel süreçlerin ve yapıların daha iyi anlaşılmasıdır. Bu bilgiyle teknoloji insanlara daha kolay uyarlanabilmektedir. Kitlesel standardizasyondan kitlesel bireyselleştirmeye yöneliş bir diğer eğilimdir, böylece eğitim ve teknolojide bireysel öğrenme ihtiyaçları ve tercihlerine göre şekillenmektedir. Üçüncü eğilim, kavrama (comphehension) ve öğrenme için bağlamın daha iyi anlaşılıyor olmasıdır. Diğer eğilim ise bilgi tabanının çok hızlı genişliyor olmasıdır. Bunlar arasında kitlesel bireyselleştirme, öğretimin birey ihtiyaçlarına göre uyarlanmasını öngörmektedir. Bu da bireylerin sisteme adaptasyonu olan kitlesel standardizasyondan uzaklaşma anlamını taşımaktadır. Web madenciliği eğitimde kitlesel bireyselleştirme çalışmalarına yardımcı olacaktır. Buna ilave olarak web tabanlı eğitimde sanal bilgi yapılarının tanımlanmasında yol gösterecektir. Bu aşamada web içerik madenciliği ve web kullanım madenciliğinden faydalanılabilir.

Öğrencilerin sistem içindeki davranışları ve erişim bilgileri web kullanım madenciliğinin konusunu oluşturmaktadır. Osmar R. Zaïane web'e dayalı eğitim sistemlerinde uygulanan web kullanım madenciliğini çevrimiçi ve çevrimdışı olarak iki şekilde ele almıştır[91]. Çevrimdışı web madenciliği eğitimcilere öğrenme modellerini doğrulamada ve web sayfalarını yeniden yapılandırmada yardımcı olacak örüntüleri ve faydalı bilgiyi keşfetmek için kullanılır. Çevrimiçi web madenciliği ise örüntüleri otomatik olarak kullanım anında keşfeder ve zeki öğretim sistemlerinde öğrencilerin çevrimiçi öğrenme süreçlerine yardımcı olurlar. Keşfedilen örüntüler de uygulamayı geliştirmek için sistem tarafından anında kullanılabilir. Ancak günümüzde birkaç örnek dışında uygulanabilen bu

[90] Sung Ho Ha, S. M. Bae ve S. C. Park, "Web Mining for Distance Education," **IEEE International Conference on Management of Innovation and Technology'de sunulan bildiri** (Singapore. 12-15 Kasım 2000), s.715.
[91] Zaïane, **Ön.ver.**, s.61.

tür araçlar henüz yaygınlaşmamıştır. Kullanım anında gerçekleştirilen web madenciliği örnekleri e-ticarette daha sık görülmektedir.

Web'e dayalı eğitim sistemlerinde gerçekleştirilen veri madenciliği çalışmalarını web'e dayalı dersler, öğrenme içerik yönetim sistemleri ve uyarlanabilir ve zeki web'e dayalı eğitim sistemleri başlıkları altında incelemek mümkündür.

3.1. Web Madenciliğinde Veri Hazırlama

Veri hazırlama, asıl verinin veri madenciliği algoritmalarına uygulanabilmesi için uygun bir şekle dönüştürülmesini sağlar. Daha önce ele alınan veri madenciliği veri hazırlama işlemlerinden farklı olarak bu bölümde web günce verilerinin hazırlanması ile ilgili işlemler açıklanacaktır. Web güncelerinin yapısı gereği madenciliği yapılacak verinin bir takım temizleme, tanımlama, dönüştürme gibi süreçlerden geçirilmesi gerekir. Bu işlemler aşağıdaki gibi kısaca açıklanabilir[92].

- *Veri temizleme:* Grafik, komut dosyaları (script) gibi madencilik sürecinde gerek duyulmayan geçersiz referans ve günce kayıtlarının temizlenmesi işlemi gerçekleştirilir.
- *Kullanıcı tanımlama:* Web sayfasına bağlanan kullanıcı ile web sayfasının ilişkilendirilmesidir.
- *Oturum tanımlama:* Bir web güncesindeki bir kullanıcı ve derse ait tüm sayfa referanslarının kullanıcı oturumlarına bölünmesidir. e-ticaret uygulamalarında oturum genellikle bir ürünün satın alınması veya bir e-kartın kontrolü sonrası veya zaman aşım süresi dolduğunda otomatik olarak sonlandırılır. Bu süreç çevrimiçi dersler için geçerli değildir. Bunun

[92] Martha Koutri, N. Avouris ve S. Daskalaki, "A Survey on Web Usage Mining Techniques for Web-based Adaptive Hypermedia Systems," **Adaptable and Adaptive Hypermedia Systems** (Hershey: IRM Press, 2005); Marta Elena Zorrilla ve diğerleri, "Web Usage Mining Project for Improving Web-Based Learning Sites," **EUROCAST 2005 sunulan bildiri** (Canary Islands. 7-11 Şubat 2005), s. 205.

nedeni açılan oturumu sonlandırmak için bir güvenlik sorununun olmamasıdır.

- *Sayfa referanslarının tamamlanması:* İnternet tarayıcıları veya vekil sunucularının ön belleklerinden kaynaklanan boş sayfa referanslarının tamamlanması işlemidir.
- *Veri dönüşümü ve zenginleştirme:* Mevcut özelliklerden yeni özelliklerin türetilmesi veya web güncelerindeki bazı alanların anlamlandırılması için yapılan dönüşümlerdir.
- *Veri bütünleştirme:* Birbirinden farklı yapıda olan veri kaynaklarının bütünleştirilmesi ve eşleştirilmesidir.
- *Veri azaltma:* Analizde yer almayacak web günce özelliklerinin elenmesi ya da gereksiz kayıtların çıkarılmasıdır.
- *Günce kayıtlarının öğrenme faaliyetlerine adreslenmesi:* Web sayfalarına ulaşım güncelerinin gerçek öğrenme faaliyetlerine adreslenmesi, web sayfa adresleri ve sayfaları çağıran web uygulama parametrelerinin faaliyet kodlarına dönüştürülmesidir[93]. Bu sayede karışık yapıdaki web günce verileri sıralı olarak çevrimiçi öğrenen faaliyetlerine dönüştürülür. Örneğin bir çevrimiçi öğrenen faaliyeti "Giriş→Alıştırmalistesi→ Kısa_sınav _gönderimi → Alıştırma_ listesi → Konferans_ mesajı_oku ma, ..." şeklinde sıralanabilir.

Web madenciliği veri hazırlama süreci, karmaşık yapıdaki web güncelerinin veri madenciliği algoritmaları için hazır hale getirilmesidir. Bu aşamada uygulanan işlemler genellikle bir veritabanı yönetim sistemi içerisinde gerçekleştirilir. Metin halindeki günceler veritabanı yönetim sistemine aktarılarak ilgili dönüşüm ve yapılandırma işlemleri gerçekleştirilir. SQL sorgulama ve programlama dili bu süreçte yoğun olarak kullanılır. Günceleri analiz edilen sistemin öğrenme için yapılandırılmış olması bazı özel durumları ortaya çıkarabilmektedir. Karşılaşılan özel durumlardan bazıları aşağıda sıralanmıştır:

[93] Zaïane, **Ön.ver.**, s.62.

- Çoğu sistem öğrenen etkileşim kayıtlarını sadece bir günce dosyasında değil aynı zamanda doğrudan bir veritabanına depolamaktadır. Veritabanları sıradan günce dosyalarından daha güçlüdür ve aynı zamanda daha az hata eğilimi içermesi ve daha esnek olması nedeniyle analiz kolaylığı sağlar[94].
- Bireysel ziyaret oturumlarının izlenen içerik sayfaları şekline dönüştürülmesi için ihtiyaç duyulan enformasyon ile ilgili olan alt oturum veya görevler tanımlanabilir. Bu sayede günce kayıtları yerine ziyaretçinin gezinme davranışlarını ifade eden gerçek içerik sayfalarının tanımladığı öğrenme görevleri modellenebilir. Verinin bu şekilde modellenmesinde tanımlanan sıralama gerçek içerik sayfalarına dayalıdır[95].
- Verinin anlamlı hale getirilmesi için deneme sayısı, tekrarlanan okuma sayısı, bilgi düzeyi gibi özel eğitim kavramları kullanılarak veri filtrelenebilir[96].

Böylece veri hazırlama aşamasında karmaşık yapıdaki web günce verileri, veri filtrasyonu, özellik türetme veya dönüşümü gibi işlemler ile öğrenenlerin öğrenme davranışları haline dönüştürülebilir

3.2. Web'e Dayalı Öğrenmede İzleme Sistemleri

Öğrenenlerin öğrenme amaçlı kullandığı web sistemleri üzerindeki kullanım faaliyetlerine ilişkin istatistikler, veri madenciliği algoritmalarının sağladığı bilgilere ulaştırmasa da e-öğrenme sistemlerinin değerlendirilmelerinde bir başlangıç noktası olarak düşünülebilir. Kullanım istatistikleri AcessWatch, Analog, NetTrack, Webtrends, SurfAid gibi web sunucu güncelerinin analiz edilmesi için geliştirilen standart araçlar kullanılarak

---

[94] Luis Talavera ve E. Gaudioso. "Mining Student Data to Characterize Similar Behavior Groups in Unstructured Collaboration Spaces," **16th ECAI 2004 - Workshop on Artificial Intelligence'da sunulan bildiri** (Valencia. 22-27 Ağustos 2004), s.17.

[95] Jia Li ve O. R. Zaïane, "Combining Usage, Content, and Structure Data to Improve Web Site Recommendation," **e-Commerce and Web Technologies, 5th International Conference'da sunulan bildiri** (Zaragoza. 31 Ağustos-3 Eylül 2004), s.305.

[96] **Aynı**, s.305.

çıkarılabilir. Bu araçlar ile "t süresi boyunca, P sayfası için n adet tıklama mevcuttur" gibi raporlar türetilebilir. Ancak bu araçlarla sağlanan sonuçlar kesin kullanım bilgisi ve gizli eğilimleri anlamaya yardımcı olacak nitelikte değildirler. Yeni web analiz yazılımları daha karmaşık analitik araçları sunabilmektedir. Fakat bu sistemler web sisteminin analizi için geliştirilmiştir ve veri analizinde kullanılabilmeleri için önemli müdahaleler gerektirir. Bu sistemler genellikle web kaynaklarını veya kullanıcının bu kaynaklara olan ilgisini değerlendirmek için sayfa erişim veya rastlama (hit) sayılarını kullanırlar. Ancak bu sayısal veriler web sisteminin değerlendirilmesinde hem yetersiz hem de hatalı sonuçlar verebilmektedir[97].

Web'e dayalı eğitim sistemlerinden elde edilen çok boyutlu öğrenen izleme verileri görselleştirilerek büyük miktardaki veriler hakkında fikir sahibi olunabilir. Örneğin bir web sisteminde öğrencilerin kullanım verilerini izlemek için "GISMO" (Graphic Interactive Student Monitoring System) aracı geliştirilmiştir. GISMO uzaktan öğrenim gören öğrencilerin davranışsal görünüşlerini incelemek için dersi yürüten tarafından kullanılabilen grafikleri oluşturur[98]. Şekil 20'de GISMO grafik aracının öğrencilerin derslere erişimlerini göstermek amacıyla oluşturulan ekran görünümü yer almaktadır. Yatay eksen üzerinde ders tarihlerinin, düşey eksen üzerinde öğrenci adlarının yer aldığı basit bir matris olan grafik, ders erişimlerini göstermektedir.

Görselleştirme aracı Avrupa Birliği tarafından desteklenmiş "Edukalibre" projesinde "CourseVis" araştırmasında elde edilen bilgiler ışığında geliştirilmiştir. Görselleştirme teknikleri bilgisayar destekli işbirliğine dayalı öğrenmede, bire bir sistemlerdeki topluluk ilişkilerinde ve çevrimiçi gruplar arası iletişimde sosyal davranışları görselleştirmek için kullanılmaktadır. Dersi veren öğreticiler bu sistem ile üretilen grafikler sayesinde öğrenme faaliyetleri hakkında fikir sahibi olabilirler.

---

[97] Osmar R. Zaïane, M. Xin ve J. Han, "Discovering Web Access Patterns and Trends by Applying OLAP and Data Mining Technology on Web Logs," **ADL'98 de sunulan bildiri** (Santa Barbara. 22-24 Nisan1998), s.19.

[98] Riccardo Mazza ve C. Milani, "Exploring Usage Analysis in Learning Systems: Gaining Insights from Visualisations," **12th International Conference on Artificial Intelligence in Education'da sunulan bildiri** (Amsterdam. 18 Temmuz 2005), s.66.

Şekil 20. Öğrencilerin Derse Erişimlerinin Gösterildiği GISMO Grafik Aracının Ekran Görüntüsü.

Riccardo Mazza ve C. Milani, "Exploring Usage Analysis in Learning Systems: Gaining Insights from Visualisations," **12th ICAI in Education'da sunulan bildiri** (Amsterdam. 18 Temmuz 2005)'den uyarlandı.

Web'e dayalı eğitim sistemlerinde özel bir görselleştirme aracı da Birleşik Devletler Hükümeti Ulusal Bilim Fonu tarafından desteklenen proje "LISTEN"'dır. Bu çalışmada çocukların bilgisayar ekranında yer alan metni okumalarını ve telaffuzlarını analiz etmek için teknolojinin izin verdiği ölçüde otomatik konuşma tanıma yazılımı geliştirilmiştir[99]. Geliştirilen sistem öğrencilerin okuma performansları hakkında bilgi sunan görsel raporlar oluşturabilmektedir.

Web'e dayalı eğitim sistemlerinde izleme amaçlı araçlar özellikle sistemin işletilmesinde öğreticilere, sistem yöneticilerine ve karar vericilere bilgi sağlamaktadır. Web madenciliği işlevlerini içermeyen bu araçlar özellikle verinin genel görünüşü hakkında bilgi verirken veri madenciliği için bir başlangıç noktası olabilmektedir.

[99] Jack Mostow, "Some Useful Design Tactics for Mining Its Data," **ITS2004 Workshops - Analyzing Student-Tutor Interaction Logs to Improve Educational Outcomes'da sunulan bildiri** (Maceió, Alagoas. 30 Ağustos 2004) s.1.

3.3. Web'e Dayalı Dersler

Web'e dayalı dersler standart HTML dili kullanılarak ders içeriğinin internet üzerinden öğrenciye ulaştırıldığı web'e dayalı eğitim sistemleridir. İnternet üzerinde bu yöntemle yayınlanan birçok program, ders ve özel amaçlı bilgi kaynağı bulunmaktadır. Eğitim amacı ile kullanılan bu web sistemleri üzerinde gerçekleştirilen web madencilik çalışmaları, genel amaçla oluşturulan web sayfaları üzerinde uygulanan web madencilik çalışmalarına benzerdir. Web madenciliği uygulamalarına kaynak oluşturacak verileri aşağıdaki gibi gruplamak mümkündür[100].

- *İçerik:* Web sayfalarındaki gerçek verileri kullanıcılara iletmek için oluşturulan veri grubudur. Bu veriler genellikle metin, grafik, video, ses ve diğer içeriklerden oluşur.
- *Yapı:* İçeriğin düzenlenmesini tanımlayan verilerdir. HTML veya XML etiketleri kullanılarak bir sayfadaki verinin biçimlendirilmesi sağlanır. Bu yapı <HTML> etiketinin gövde olarak gösterildiği bir ağaç yapı olarak tanımlanabilir. Sayfa içersinde diğer sayfalara ulaşımı sağlayan bağlantılar yer almaktadır.
- *Kullanım:* Veriler IP adresi, sayfa referansı ve ulaşım tarih ve saati gibi web sayfalarının kullanım bilgisini içerir. Temelde iki tür öğrenci verisi mevcuttur[101]. Bunlardan ilki öğrencinin faaliyetleri ve bağlanma verileri, diğeri ise dersle ilgili öğrenci faaliyetleridir.
- *Kullanıcı profili:* Web sayfasına ulaşan kullanıcılar hakkında demografik bilgi sağlayan verilerdir. Bu veriler kullanıcının sistemdeki kayıt verisi ile ilişkilendirilebilir.

---

[100] Jaideep Srivastava ve diğerleri, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data," **SIGKDD Explor. Newsl.**. Cilt No 1, Sayı No 2: 12–23, (Ocak 2000), s.13.

[101] Daniela R. Silva ve M. T. P. Vieira, "Using Data Warehouse and Data Mining Resources for Ongoing Assessment in Distance Learning," **IEEE International Conference on Advanced Learning Technologies'de sunulan bildiri** (Kazan,Russia. 9-12 Eylül 2002), s.41.

Öğrencinin web ortamında sunulan derse nasıl katıldığını, bir pedagojik stratejinin farklı öğrencileri nasıl etkilediğini öğrenmek için veri madenciliği kullanılabilir. Web sunucuları günceleri işlenerek öğrencinin alt konuları hangi sırayla izlediği, atladığı konuların hangileri olduğu ve öğrencinin tek bir sayfa, bölüm veya tüm derste ne kadar zaman harcadığı gibi verileri türetmek mümkündür.

Web'e dayalı eğitim sistemlerinde uygulanabilecek farklı web madenciliği çalışmalarını veri madenciliği görevine bağlı olarak üç grupta toplayabiliriz. Bunlar;

- Kümeleme ve sınıflama
- Birliktelik kuralları ve sıra örüntüleri analizi
- Metin madenciliği

Metin madenciliği dışında diğer veri madenciliği görevlerine ilişkin bilgiler önceki bölümlerde verilmişti. Metin madenciliği ise veri madenciliğinin metin verilerinin analizinde kullanılması olarak düşünülebilir ve web içerik madenciliği ile yakından ilişkilidir. Metin madenciliği herhangi bir etiket olmaksızın işlenmemiş belgelerin toplanması ile başlar. Başlangıçta belgeler, belgelerden doğrudan elde edilen sınıf, ifade veya varlıklar tarafından otomatik olarak etiketlenir. Sonra kavram ve ek üst düzey varlıklar belge üzerinde bilgi keşfi için kullanılır. Metin madenciliğinde veri madenciliğinde faydalanılan disiplinlere ek olarak doğal dil işleme disiplininden de faydalanılır[102]. Metin madenciliği metin belgeleri, HTML ve XML belgeleri ve e-posta belgeleri gibi yapılandırılmamış veya yarı yapılandırılmış veri kümeleri üzerinde uygulanabilir.

3.3.1. Kümeleme Ve Sınıflama Uygulamaları

Veri madenciliğinde kullanılan teknikler web'e dayalı eğitim sistemlerinde uygulanabilmektedir. Web'e dayalı derslerin yayınlandığı web sistemleri için literatürde yer alan başlıca çalışmalar aşağıda incelenmektedir.

[102] Ye, **Ön.ver.** s.482.

- Chen ve diğerleri karar ağaçları ve veri küpü teknolojilerini öğrencilerin davranışlarını gözlemek ve öğrencilerin öğrenme performansları ile ilgili pedagojik kuralları keşfetmek için web güncelerinde uygulamışlardır[103]. Tayvan'daki ulusal bir üniversite öğrencilerinin C++ programlama dilini öğrenmelerini desteklemek için web'e dayalı öğrenme sistemi tasarlanmış ve gerçekleştirilmiştir. Web sunucusunun oluşturduğu günceler MS SQL Server veritabanı yazılımına aktarılmıştır. Veriler öğrenci davranışlarının anlaşılabilmesi için SQL dili kullanılarak tekrar yapılandırılmış ve veri küpleri oluşturulmuştur. Oluşturulan veri küpleri sayesinde öğreticiler, ders sonrası öğrencilerin sorulan sorulara ilişkin davranışlarını izleyebilmişlerdir. Öğreticiye çevrimiçi izleme ile öğrencilerin sisteme girişi, materyal okuma ve gönderme davranışlarını zaman boyutuyla birlikte inceleme olanağı sunulmuştur. Veri küpleri ve karar ağacı analizleri birleştirilerek web güncelerinden elde edilen pedagojik stratejiler ve öğrenme performansı arasındaki ilişkileri analiz etmek için OLAP fonksiyonları oluşturulmuştur. Karar ağacı analizinde C5.0 algoritması kullanılarak öğreticilerin web güncelerinden potansiyel karar kurallarını çevrimiçi olarak keşfetmelerini sağlamıştır.
- Minaei-Bidgoli ve Punch web'e dayalı eğitim sisteminde kullanım verilerinden elde edilen özelliklere dayalı olarak Michigan State Üniversitesi (MSU) öğrencilerinin final notlarını tahmin etmek için bir sınıflama yaklaşımı önermişlerdir[104]. Üniversitenin çevrimiçi derslerinin sunumu için geliştirilen LON-CAPA (Learning Online Network with Computer-Assisted Personalized Approach) sistemi vasıtasıyla elde edilen iki büyük veri kümesinden faydalı bilginin keşfedilmesi için veri madenciliği uygulanmıştır. İlk veri kümesi web sayfaları, uygulamalı konu anlatımları, simülasyon ve ev ödevi, kısa sınav ve sınavlarda kullanılmak üzere tasarlanmış bireysel problemler gibi eğitim kaynaklarıdır. İkinci veri kümesi ise ilk veri kümesindeki kaynakları oluşturan, değiştiren,

[103] Gwo-Dong Chen ve diğerleri, "Discovering Decision Knowledge from Web Log Portfolio for Managing Classroom Processes by Applying Decision Tree and Data Cube Technology," **Journal of Educational Computing Research**. Cilt No 23, Sayı No 3: 305–332, (2000), s.305.
[104] Minaei-Bidgoli, **Ön.ver.**, s.2252.

değerlendiren kullanıcılar hakkındaki bilgilerdir. Bu iki veri kümesi kullanılarak çevrimiçi kaynakları benzer şekilde kullanan öğrencilerin ve öğrenciler tarafından çözülen problemlerin sınıflandırılması amaçlanmıştır. Bu çalışmada sınıflayıcıların kombinasyonlarını eniyilemek için araştırmacılar tarafından genetik algoritmalar kullanılmıştır. MSU'de 2002 bahar dönemine ait LON-CAPA'da sunulan fen bilimleri ve mühendislik fizik 1 dersi ile ilgili 261 öğrenci ve çevrimiçi olarak 184 soru içeren 12 ödev verisi çalışmanın veri kümesini oluşturmuştur. Dersi bırakan öğrenciler veri kümesinden çıkarılarak 227 öğrenci, final sınav notlarına göre sınıflandırılmıştır. Sınıflama algoritmalarında giriş özellikleri olarak LON-CAPA sisteminden sağlanan özellikler şunlardır:

1. Toplam doğru cevap sayısı (başarı oranı)
2. İlk denemede başarı
3. Doğru cevabı verene kadar gerçekleştirilen toplam deneme sayısı
4. Problem çözülünceye kadar harcanan süre
5. Öğrencinin doğru cevabı verip vermediğini dikkate almadan problem için harcanan toplam süre
6. Öğrencinin diğer öğrenciler ve öğretici ile çevrimiçi etkileşim durumu

Belirlenen özellikler kullanılarak parametrik olmayan örüntü sınıflayıcıları ile parametrik bir örüntü sınıflayıcı algoritması kullanılarak hata tahminlerine göre karşılaştırılmıştır. Uygulanan sınıflama algoritmaları "Kuadratik Bayesian" sınıflayıcı, en yakın komşu (1-NN), k en yakın komşu (k-NN), "Parze-window", "MLM" (Multi-Layer perceptron) ve karar ağaçlarıdır. "Bayesian" ve "Parzen-Window" sınıflayıcılarında özelliklerin normal dağılıma sahip olması gerektiğinden her özellik için veri normalleştirilmiştir. Altı farklı sınıflama algoritması kullanılarak öğrenci final notları sınıflanmış ve sınıflama performansını arttırmak için çoklu sınıflayıcıların kombinasyonu gerçekleştirilmiştir. Çoklu sınıflayıcıların kombinasyonunun performansını en büyüklemek için genetik algoritmalardan faydalanılmıştır. Çoklu sınıflayıcıların kombinasyonunun sınıflama performansını arttırdığı gözlenmiştir.

- Öğrencilerin düzensiz öğrenme süreçlerinin çevrimiçi aykırı değer tespit yöntemi kullanılarak belirlendiği bir çalışma 2004 yılında Ueno tarafından yayınlanmıştır[105]. Araştırmacı geleneksel aykırı değer tespit yöntemlerinin e-öğrenmede düzensiz öğrenme süreçlerinin belirlenmesinde yetersiz kaldığını belirtmiş ve Bayesian tahmin edici dağılımını kullanan yeni bir aykırı değer tespit yöntemi önermiştir. Bu uygulamada önerilen yöntemin üç üstünlüğü vurgulanmıştır. Bu üstünlüklerden ilki Bayesian yaklaşımının öğrenme sürecinin başında düzenli bir süreci, düzensiz bir süreç olarak algılamasından sakınmasıdır. Bir diğer üstünlüğü önerilen yöntemin görev zorluklarını ve öğrenci yeteneklerini göz önüne almasıdır. Ayrıca Bayesian aykırı değer tespiti bütünleştirilmiş bir istatistik test yöntemi türetir. Çalışmanın gerçekleştirildiği e-öğrenme sistemi içerik sunum sistemi, içerik veritabanı, öğrenme geçmişi veritabanı ve veri madenciliği sistemini içerir. Öğrenme geçmişi veritabanında saklanan günceler içerik_id, öğrenici_id, öğrenilen_konu_numarası, test_id, işlem_sırası_id, gerçekleştirilen_işlem_id, işleme başlama tarih ve saati, işlemin sonlandığı tarih ve saat verilerinden oluşmaktadır. Önerilen modelde $x_1,x_2,\ldots..,x_n$ gibi öğrenciye ait öğrenme süreçleri verilerinden yeni bir x_{n+1} verisinin Bayesian tahmin edici dağılımı türetilerek yeni verinin aykırı değer testi gerçekleştirilmiştir. Öğrencilerin düzensiz öğrenme süreçlerinin tespit edilmesi için önerilen yöntem yeni bir modül olarak sisteme eklenmiştir.

3.3.2. Birliktelik Kuralı Ve Sıra Örüntüleri Uygulamaları

En yaygın kullanılan veri madenciliği görevi veri kümesindeki bir veya daha fazla özelliği diğer bir özellikle birleştiren birliktelik kurallarıdır. Sıra örüntüleri madenciliği zamana göre sıralanmış bir oturum kümesindeki birbirini takip eden olaylar kümesi gibi oturumlar arası örüntüleri bulmaya çalışır. Bu veri madenciliği görevleri web'e dayalı eğitim sistemlerinde uygulanmaktadır. Web'e

[105] Maomi Ueno, "Online Outlier Detection System for Learning Time Data in e-learning and Its Evaluation," **7th IASTED'da sunulan bildiri** (Kauai, Hawaii. 16-18 Ağustos 2004), s.248.

dayalı eğitim sistemlerinde birliktelik kuralları ve sıra örüntüleri madenciliği ile ilgili önemli çalışmalar aşağıda verilmiştir.

- 2000 yılında Ha ve diğerleri tarafından yayınlanan çalışmada uzaktan eğitim sistemlerinde web madenciliğinin olası uygulamaları tartışılmış ve bireysel öğrenme ihtiyaç ve tercihlerini dikkate alan kişiselleştirilmiş bir web'e dayalı öğrenme ortamında aykırı yol analizi hakkında bilgi verilmiştir[106]. Yol analizi bir web sitesinin fiziksel düzeninde bulunan aykırı örüntülerin bulunmasını içerir. Yol analizi, sistemi kullanan tüm kullanıcıların izlediği yolların incelenmesi ve bireysel gezinti özelliklerinin incelenmesi olarak iki farklı konuya odaklanır. Çalışmada web kullanım madenciliğinin web'e dayalı eğitim sistemlerinde uygulanabilirliği gösterilmiştir.
- Shen ve diğerleri uzaktan eğitimde sıkça karşılaşılan iki problemi çözmek amacıyla veri madenciliğinden faydalanarak zeki uzaktan öğrenme ortamı tasarlamışlardır[107]. Uzaktan eğitimde karşılaşılan bu problemler geleneksel eğitim sistemine alışkın olan öğreticilerin derslerini çevrimiçi ortama aktarmada zorlanmaları ve özellikle yetişkin öğrencilerin kendilerini aşırı yük altında hissetmeleri olarak tanımlanmıştır. Shanghai Jiao Tong Üniversitesi Network Education College'de geliştirilen ve kullanılan zeki uzaktan öğrenme ortamı, öğrencilerin sorularını otomatik olarak cevaplayan bir sistem ile etkileşmelerini ve aynı zamanda öğreticilerin de öğrencilerin öğrenme örüntülerini analiz etmelerini sağlayarak bu iki probleme çözüm getirmiştir. Araştırmacılar kullanıcı örüntülerini ve davranışlarını belirlemede veri madenciliğini, soru-cevap sistemini oluşturmak için ise durum tabanlı çıkarsama tekniklerini kullanmışlardır. Öğrencilerin öğrenme davranışlarını, kişisel özelliklerini ve bilgi tecrübelerini analiz etmek için oluşturulan veri analiz merkezi, web günceleri ve sistem veritabanını kullanarak öğrencilerin öğrenme faaliyetlerine göre sınıflarını belirlemek için kümeleme, farklı bilgi

[106] Ha, **Ön.ver.**, s.718.
[107] Ruimin Shen ve diğerleri, "Data Mining and Case-based Reasoning for Distance Learning," **Journal of Distance Education Technologies**. Cilt No1, Sayı No 3: 46–58, (2003), s.46.

noktalarının elde edilmesi için de sıralı birliktelik kuralı analizlerini uygulamıştır. Sıralı birliktelik kuralları analizi sonucu "Bölüm 3'ü faydalı bulan aynı zamanda Bölüm 5'i de faydalı bulmaktadır" gibi bilgiler keşfedilmiş ve bu bilgi ışığında öğreticiler web kaynaklarını düzenleme olanağı bulmuşlardır. Aynı zamanda materyal seçimi daha iyi organize edilmiş ve verimli öğrenci gruplarının belirlenmesi bulgular kullanılarak gerçekleştirilmiştir.

- Minaei–Bidgoli ve diğerleri web'e dayalı eğitim sistemlerinde bir topluluğun farklı bölümlerinin ilginç özelliklerini tanımlayarak ilginç zıt kuralların keşfedilmesi için yöntem önerisinde bulunmuşlardır[108]. Web'e dayalı eğitim sisteminde zıt kurallar çeşitli öğrenci grupları arasında performans farkı örüntülerinin özelliklerini tanımlamaya yardımcı olur. Çalışmanın gerçekleştirildiği Michigan State Üniversitesi LON-CAPA web'e dayalı öğrenme sistemi, öğrenme faaliyetlerine ilişkin çok sayıdaki özelliği güncelerde saklar. Zıt özelliklerin keşfedilmesine ilişkin LON-CAPA güncelerinden seçilen özellikler dört grupta ele alınmıştır.

  1. Öğrenci özellikleri (genel not ortalaması, cinsiyet, üniversite öncesi başarısı)
  2. Problem özellikleri (zorluk derecesi, ayırt etme derecesi, ortalama deneme sayısı)
  3. Öğrenci-problem etkileşim özellikleri (başarı durumu, son cevap öncesi deneme sayısı, ilk denemeden son cevap verilinceye kadar geçen süre)
  4. Öğrenci / ders etkileşim özellikleri (not, geçme-kalma durumu)

  Zıt kuralların keşfedilmesinde hedef değişken olarak cinsiyet ve geçme-kalma özellikleri kullanılmıştır. Araştırmacılar zıt nesneler arasında gizli örüntüleri keşfetmek için MCR (Mining Contrast Rules) adını verdikleri algoritmayı önermişlerdir. Algoritmanın LON-CAPA güncelerinde uygulanması sonucu elde edilen kurallar aşağıda sıralanmıştır.

[108] Behrouz Minaei-Bidgoli, P. Tan ve W. Punch, "Mining Interesting Contrast Rules for a Web-Based Educational System," **ICMLA ' 04'da sunulan bildiri** (Louisville, Kentucky. 6 -18 Aralık 2004), s. 320.

Beklenen ve önceden bilinen: Ev ödevlerinde başarılı olan öğrencilerin final sınavında başarılı olmaları, düşük ortalama ile gelen öğrencilerin ise başarısız olmaları.

Beklenmeyen: Geçmiş not ortalamaları yüksek olan ve problemlerde başarısız çözüm denemeleri için fazla süre harcayanların bayan olması, genel not ortalaması 3,0 ile 3,5 arasında olan erkek öğrencilerin kimya ev ödevi problemlerini ilk denemede çözmeleri.

Bilinmeyen: Zorluk derecesi orta ve ayırt etme derecesi düşük olan soruları ilk denemede çözen öğrencilerin final sınavında başarılı olmaları.

Araştırmacılar elde edilen bulguların ilgili derslerin daha etkili tasarlanması ile öğrencilerin kaynakları daha verimli kullanabileceklerini vurgulamışlardır.

- Birliktelik kurallarının kullanıldığı bir diğer çalışma da Markellou ve diğerleri tarafından 2005 yılında yayınlanmıştır[109]. Çalışmada kişiselleştirilmiş öğrenme için ontolojiye (varlık bilimine) dayalı olarak bir yapı önerilmiştir. Semantik web, web içeriklerinin sadece doğal dillerde değil aynı zamanda diğer uygulamalar tarafından anlaşılabilir, yorumlanabilir ve kullanılabilir bir biçimde ifade edilmesidir. Web madenciliği ve semantik web'in birleştirilmesi semantik web madenciliği olarak ifade edilir ve yeni, hızlı bir araştırma alanı yaratmıştır. Yazarlara göre kişiselleştirilmiş web çalışmalarında semantik web'in kullanılması yeni semantik yapılardan faydalanılarak web madenciliğinin gelişmesine olanak sağlayacaktır. Semantik web madenciliğinin kullanılmasıyla web uygulamaları ve özellikle e-öğrenme sistemlerinin daha zeki olacağı ve gelişeceği ifade edilmiştir. Bu çalışmada semantik web teknolojilerinin ve özellikle ontolojilerin e-öğrenme sistemlerinde nasıl kullanılabileceği araştırılmıştır. e-öğrenme uygulamalarında ontoloji kişiselleştirilmiş bir e-öğrenme sisteminin oluşturulması ve faydalı bilginin elde edilmesi için kullanılan bir bilgi tabanı olarak düşünülebilir. Önerilen kişiselleştirilmiş e-

[109] Penelope Markellou ve diğerleri, "Using Semantic Web Mining Technologies for Personalized e-learning Experiences," **Web-based Education'da sunulan bildiri** (Grindelwald. 21-23 Şubat 2005), s.461.

öğrenme sisteminde web güncelerindeki veriler temizlenerek birliktelik kuralları keşfedilmekte ayrıca ontolojiye dayalı olarak bireysel ihtiyaçlara cevap verecek içeriklerin hazırlanması sağlanmaktadır. Önerilen e-öğrenme sistemi senaryosuna göre "Apriori" algoritmasının uygulanması ile kullanıcıların sayfalarda gezinme örüntülerine ilişkin kurallar çıkarılır ve içeriğin ontoloji yapısı ile birleştirilerek bir öneri motoru oluşturulur. Bu öneri motoru çevrimdışı verilerden faydalanarak öğrencinin kendisine özel bir öğrenme içeriği hazırlayabilecektir.

3.3.3 Metin Madenciliği Uygulamaları

Metin madenciliğinin genellikle geleneksel veri madenciliğinden daha zor olduğu düşünülür[110]. Bunun nedeni geleneksel veritabanlarının sabit ve bilinen bir yapıya sahip olmalarına rağmen metin belgelerinin yapılandırılmamış veya web dokümanlarında olduğu gibi yarı yapılandırılmış olmasıdır.

Metin madenciliği ile bilgi keşfi, metin ve web belgelerindeki özel bilgileri bulmaya yöneliktir. Bu alanda kullanılan yaklaşımlar belge çözümleme, analiz ve yeniden yapılandırmayı kapsar. Bu şekilde e-öğrenmede mevcut öğrenme materyalinin yeniden yapılandırılması sağlanabilir. Metin madenciliğindeki diğer yaklaşımlar yarı yapılandırılmış önemli bilgiyi tanımlamayı ve çıkarmayı, ifade indeksleme ve karşılaştırma kullanılarak belgelerdeki anahtar kelime ve ifadeleri ortaya çıkarmayı kapsar. Bu yöntemler faydalı bilgiyi otomatik olarak elde edebildikleri için e-öğrenmede yüksek potansiyele sahiptirler.

Bilgi organizasyonunda metin madenciliği tek tek belge içeriklerini okumadan büyük bir belge kümesindeki konu başlıklarının genel bir özetini elde etmek için kullanılabilir[111]. Bu görev kümeleme ve sınıflama veri madenciliği algoritmaları ile gerçekleştirilebilir. Belge içeriklerinin otomatik olarak analiz edilmesinde sınıflama ve kümeleme analizi, belgelerdeki anahtar kelime

---

[110] Khaled Hammouda ve M. Kamel, "Data Mining in e-learning," **e-learning Networked Environments and Architectures: A Knowledge Processing Perspective** (London: Springer, 2006), s.375.
[111] **Aynı**, s.376

dağılımına dayalıdır. Aynı zamanda kelime ve ifade eşleştirme, benzerlik hesaplamalarının kullanımını gerektirir. Sonuçta metin madenciliği ile e-öğrenmede, konu ve konu başlıkları ile etiketlenen belgelerin daha iyi yönetilebilir gruplanması sağlanabilir. Web'e dayalı derslerde uygulanan metin madenciliği ile ilgili çalışmalar özet olarak aşağıda verilmiştir.

- Ueno geçmiş e-öğrenme günce verilerini kullanan yeni veri madenciliği yöntemleri geliştirmiştir ve "Samurai" adını verdiği bir Zeki Öğrenme Yönetim Sistemi (ILMS - Intelligent Learning Management System) tasarlamıştır[112]. Çalışmada, işbirliğine dayalı öğrenmede madencilik teknolojilerinin kullanımı üzerine odaklanılmış ve geliştirilen ILMS sisteminin işlev ve performansı örneklenmiştir. Samurai'de kullanılan madencilik işlevleri aşağıda verilmiştir.
  - Düzensiz e-öğrenme süreci sergileyen öğrenenlerin tespiti
  - Çevrimiçi içerik analizi
  - Karar ağaçları kullanılarak öğrenenlerin geçmiş günce analizleri
  - "Bayesian Belief" ağları kullanılarak öğrenenlerin geçmiş günce analizi
  - Öğrenenlerin tartışma süreçlerinin Markov analizi ile incelenmesi
  - Tartışma verilerinin entropiye dayalı olarak analizi
  - Geliştirilmiş uyum analizinin kullanıldığı metin madenciliği ile tartışma verilerinin analizi

  Ueno, işbirliğine dayalı öğrenmede öğrenenlerin tartışma ortamında seçilen konular hakkında yaptıkları yorumları değerlendiren madencilik tekniklerini özetlemiştir.

- Tane ve diğerleri ontolojiye dayalı olarak geliştirilen "Wachdog" ders yazılımını ve çalışma prensiplerini konu alan çalışmalarını 2004 yılında

[112] Maomi Ueno, "Data Mining and Text Mining Technologies for Collaborative Learning in an ILMS 'Samurai'," **ICALT'04'da sunulan bildiri** (Joensuu. 30 Ağustos - 1 Eylül 2004), s.1052.

yayınlamışlardır[113]. e-öğrenme alanında standartların oluşması, öğrenme nesnelerinin tanımlanmasına yardımcı olmuş ve öğrenme materyallerinin yapılarının ve içeriklerinin birleştirilmesinde daha kapsamlı yaklaşımlara olan ihtiyacı ortaya çıkarmıştır. Wachdog ders yazılımının mimarisi beş adımla tanımlanan görevlerden oluşmaktadır. Bu görevler:

1. Ontolojinin anlaşılması ve içeriğin incelenmesi
2. Odaklanmış bir web sayfası yakalama aracı (crawler) vasıtasıyla ilgili materyalin getirilmesi
3. Kaynak ambarlarının semantik olarak sorgulanması
4. Belgelerin ontolojiye göre organize edilmesi ve kümelenmesi
5. Mevcut veriye göre bilgi tabanı ve ontolojinin güncellenmesi

Odaklanmış web sayfası yakalama bileşeni web sayfalarını ve bireysel paylaşım ağlarından (P2P-Peer-to-peer) ilgili belgeleri getirerek depolar. Sistem birkaç adımdan oluşan metin madenciliği hazırlık sürecini uygulayarak depolanan belgeleri "Bisection K-Means" algoritmasını kullanarak kümeler. Araştırmacılar Watchdog'un, hızlı değişen çalışma ortamlarında bireylerin öğrenme ihtiyaçlarını desteklemek ve güncel konular hakkında yeni dersler hazırlamak zorunda kalan öğreticiler için oldukça kapsamlı bir yaklaşım olduğu belirtmişlerdir. Watchdog sisteminin gelişme süreci halen tamamlanmamış ve farklı bileşenlerinin geliştirilme çalışmaları devam etmektedir.

3.4. Öğrenme Yönetim Sistemleri

Öğrenme Yönetim Sistemleri (LMS) ders katılımcıları arasında iletişimi ve bilgi paylaşımını sağlamak için çok çeşitli kanallar ve çalışma ortamı sunan platformlardır. Bu platformlar bilgi aktarımını, içerik materyali üretilmesini, ev ödevleri ve sınavların hazırlanmasını, tartışma ortamlarının yürütülmesini, uzak sınıfların yönetimini ve ayrıca sohbet, forum, dosya depolama alanı, haber servisleri gibi hizmetlerle işbirliğine dayalı öğrenmeyi destekler. LMS yazılımları;

[113] Julien Tane, C. Schmitz ve G. Stumme, "Semantic Resource Management for the Web: an e-learning Application," **13th WWW Conference'da sunulan bildiri** (New York. 17-22 Mayıs 2004), s.1.

e-öğrenme ortamları, içerik yönetim sistemleri, ders yönetim yazılımları, e-ders gibi isimlerle de telaffuz edilmektedir. Angel, BlackBoard, WebCT, Desire2Learn, Virtual-U en yaygın kullanılan ticari LMS yazılımlarına örnektir. Moodle, .LRN, Ilias, ATutor gibi yazılımlar ücretsiz ve açık kaynak kodlu LMS yazılımlarına örnek gösterilebilir. Bu sistemler öğrencilerin okuma, yazma, sınav gibi sistem içindeki tüm faaliyetlerini kaydedebilirler[114]. LMS veritabanları genellikle kullanıcı verileri, akademik sonuçlar, kullanıcı etkileşim verileri gibi tüm sistem enformasyonunu saklarlar.

Çoğu öğrenme yönetim sisteminde raporlama araçlarının bulunmasına rağmen öğrenci sayısı arttığında bir öğreticinin faydalı bilgi elde etmesi zorlaşır. Bu nedenle faydalı örüntülerin tanımlamada, veriyi incelemede, görselleştirme ve analiz etmede veri madenciliğinden faydalanılabilir[115]. Veri madenciliği aynı zamanda öğrenenlerin LMS'de nasıl öğrendiklerine ilişkin bilgi edinebilmek ve ders içerikleri hakkında daha fazla nesnel geribildirimler sağlamak için web faaliyetlerini değerlendirmede kullanılabilir[116].

3.4.1. Kümeleme Ve Sınıflama Uygulamaları

Öğrenme yönetim sistemleri eğitim için yapılandırılmış sistemler olduklarından öğrencilerin öğrenme davranışlarına ait daha fazla veri saklayabilme yeteneğine sahiptirler. Bu sistemlerin ilişkisel veritabanlarında sakladıkları veriler, veri madenciliği ve web madenciliği çalışmalarında gözde veri kaynakları olmuştur.

- Mor ve Mingoillon bir sanal yerleşke ile bütünleşen bir e-öğrenme ortamının kullanıcılarına ait gezinme davranışlarını çözümlemek için kullanışlı bir yapı önermişlerdir[117]. Çalışmanın gerçekleştirildiği sanal yerleşke olan İspanya'daki Catalonia Açık Üniversitesinde 26 binden

[114] Mostow, **Ön.ver.**, s.2.
[115] Talavera, **Ön.ver.**, s.17.
[116] Zaïane, **Ön.ver.**, s.60.
[117] Enric Mor ve J. Minguillon, "E-learning Personalization Based on Itineraries and Long-term Navigational Behavior," **13th WWW Conference'da sunulan bildiri** (New York. 17 - 22 Mayıs 2004), s. 264.

fazla öğrenci ve 1500'ü aşkın içerik tasarımcısı, öğretici ve akademik personeli bulunmaktadır. Sanal yerleşke öğrencilerine e-posta, ajanda, haber sistemi, sanal sınıflar, sayısal kütüphane ve e-öğrenme araçları sunmaktadır. e-öğrenmede paylaşılabilir içerik nesne referans modeli olan SCORM (Sharable Content Object Reference Model) standardı kullanılmaktadır. SCORM standardı öğrencilerin amaçlarına, tercihlerine, performansına ve benzer faktörlere dayalı olarak uyarlanabilir içeriği desteklemek için tekrar kullanılabilir öğrenme nesnelerine ait dinamik bir ortama sahiptir. Araştırmacılar SCORM standartlarının uygulandığı sanal yerleşkelerinde, öğrencilerin davranış örüntülerinin kümeleme çalışmasıyla elde edilebileceğini vurgulamışlardır. Tasarlanacak benzer sistemlerin içerik tasarımcılarına ve öğreticilere sağlayacağı katkıları tartışılmıştır.

- Talevera ve Gaudioso bir LMS veritabanında veri madenciliği uygulaması gerçekleştirmiş ve etkileşim örüntülerini özetlemek için analitik modeller önermişlerdir[118]. Web topluluklarını desteklemek amacıyla tasarlanmış açık kaynak kodlu web uygulaması olan ACS (ArsDigita Community System) üzerinde bir ders platformu oluşturulmuştur. Yazılım; kullanıcı veya grup yönetimi, içerik yönetimi, haber, sıkça sorulan sorular, takvim, forum gibi hizmetleri bir ilişkisel veritabanı desteğiyle sunmaktadır. Birçok LMS sisteminde olduğu gibi ACS kullanıcı bilgisi, ders içeriği ve işbirliğine dayalı tüm olayları izlemek ve yapılandırmak için gerekli fonksiyonlara sahip bir yazılımdır. Araştırmacılar öğrenci gruplarının davranış profillerini türetmek amacıyla sistem veritabanında yer alan veri tablolarından SQL sorgularıyla elde etmişlerdir. Ayrıca araştırmalardan elde edilen öğrencilere ait demografik ve geçmiş bilgiler, mevcut veri ile birleştirilmiştir. Veri kümesine eklenen bir diğer kaynak da yine araştırmalardan elde edilen öğrencilerin ilgi alanlarına ilişkin verilerdir. Verilerin kümelenmesinde yaygın olarak kullanılan EM (Expectation-Maximation) algoritması kullanılmıştır. EM algoritması model parametrelerini başlangıçta tahmin ederek her adımda bu tahminleri

[118] Talavera, **Ön.Ver.**, s.17.

geliştiren bir kümeleme analizidir[119]. Araştırmacılar gerçekleştirilen kümeleme analizi sonucu 6 küme belirlemişlerdir. Öğrencilerin forumdaki davranışlarına göre belirlenen bu örüntülerin, işbirliğine dayalı faaliyetlerde öğrencilerin gruplandırılmasına katkıda bulunacağı savunulmuştur.

3.4.2. Birliktelik Kuralları Ve Sıra Örüntüleri Uygulamaları

Web'e dayalı derslerde kullanıcı davranışları genellikle web sunucularının güncelerinden elde edilirken LMS'lerde bu veri kaynağına ek olarak veritabanlarında depolanan kayıtlar bulunmaktadır. Bu kayıtlar e-öğrenme için tasarlanmış LMS yazılımları tarafından amaca yönelik olarak yapılandırılmış ve çeşitlendirilmiş kullanım bilgilerini içerirler. Birliktelik kuralları ve sıra örüntülerinden genellikle öğrencinin sistem içinde kullanım davranışlarını belirlemek amacıyla faydalanılmaktadır. Elde edilen bilgiler sistemin ve içeriğin iyileştirilmesine yardımcı olurken ders yazılımı üreticilerine de önemli geribildirim kaynağı oluşturmaktadır.

- Wang ve diğerleri bir öğrencinin ait olduğu grubu tahmin etmek için bir karar ağacı oluşturmuş ve öğrenme özelliklerini keşfetmek amacıyla sıra örüntüleri madenciliği, kümeleme ve karar ağacı yöntemlerini içeren öğrenme portfoli madenciliği (LPM- Learning Portfolio Mining) yaklaşımını önermişlerdir[120]. LPM yaklaşımı kullanıcı modeli tanımlama, öğrenme örüntüsü çıkarma, karar ağacı oluşturma ve faaliyet ağacı üretimi aşamalarını içermektedir. Kullanıcı modeli tanımlama aşamasında öğrencinin cinsiyet, öğrenme tarzı ve öğrenme tecrübesi kullanılarak pedagojik teoriye dayalı öğrenci profili belirlenir. Cinsiyet, yaş, eğitim durumu, bilgisayar tecrübesi ve medya tercihi değerleri öğrenciler tarafından sisteme girilen verilerdir. Öğrenme motivasyonu, kavrama tarzı, öğrenme tarzı ve sosyal durum değerleri ise öğrencilere uygulanan anketlerden elde edilir. Öğrenme portfoli; dersi öğrenme, ders

[119] Tan, **Ön.ver.**, s.583.
[120] Wei Wang ve diğerleri, "Learning Portfolio Analysis and Mining in SCORM Compliant Environment," **FIE 2004 Conference'da sunulan bildiri** (Savannah. 20-23 Ekim 2004), s.17.

notu ve öğrenme zamanı gibi verileri içeren öğrenme davranış bilgisi olarak tanımlanmıştır. Kullanıcı modeli elde edildikten sonra öğrenme portfoli içindeki öğrenme sırasından en sık karşılaşılan öğrenme örüntülerini elde etmek için üç aşama uygulanmıştır. Bunlar sıra örüntüleri madenciliği, özellik dönüştürme ve kümeleme analizinden oluşmaktadır. İlk aşamada öğrenme portfolilerinden en sık karşılaşılan öğrenme örüntülerini çıkarmak amacıyla sıra örüntüleri madenciliği uygulanır. İkinci aşamada en sık karşılaşılan öğrenme örüntülerine dayalı olarak her öğrencinin asıl öğrenme sıraları küçük bir vektör ile temsil edilecek şekle dönüştürülür. Her vektörün değeri, öğrencinin site içersindeki öğrenme sırası elde edilen örüntünün alt kümesi ise 1 diğer durumlarda ise 0 değerini alır. Böylece her öğrenciye ait bir vektör elde edilmiş olur. Üçüncü aşamada ise öğrenciler elde edilen vektörlere göre K-means algoritmasına benzer ISODATA kümeleme algoritması ile gruplara ayrılır. Çalışmada öğrenciler, bu üç aşama sonucu öğrenme örüntülerine göre dört grupta kümelenmiştir. Yeni bir öğrencinin öğrenme özellikleri ve yeteneklerine göre uygun bir kümeye yerleştirilmesi için bir karar ağacı oluşturulmuştur. Sınıflamayı gerçekleştiren karar ağacı modeli oluşturulurken verilerin üçte ikisi eğitim, kalan veriler ise test verisi olarak kullanılmıştır. Aslında bu aşama, LPM yaklaşımının modeli uyguladığı aşama olarak düşünülebilir. Öğrencinin özelliklerine göre karar ağacı modeli kullanılarak öğrencinin öğrenme kümesi tahmin edilir ve ilgili kümedeki öğrenme özelliklerine uygun öğrenciye kılavuzluk edecek bir faaliyet ağacı oluşturulur. Wang ve diğerleri LPM yaklaşımı ile öğrencilerin özellik ve yeteneklerine göre özelleştirilmiş e-öğrenme sistemi oluşturmuşlardır.

- Li ve Zaïane bir web sitesindeki kullanım verileri, içerik verileri ve yapısal verileri kullanıcı gezinme modelleri oluşturmak için birleştiren bir öneri sistemi yapılandırmışlardır[121]. Çalışmada Canada'da Alberta Üniversitesinin bir eğitim programına ait Eylül 2002-Nisan 2003 tarih aralığında e-öğrenme sayfaları ve günceleri kaynak olarak kullanılmıştır.

[121] Li, **Ön.ver.**, s.305.

e-öğrenme sitesi, aralarında yaklaşık 150.000 bağlantı olan 40.000'den daha fazla web sayfası içermektedir. Her ay yaklaşık 200.000 ziyaretçi oturumu gerçekleşen e-öğrenme sisteminde araştırmacıların tasarladığı öneri sistemi, çevrimiçi ve çevrimdışı iki modülden oluşur ve kullanıcı gezinme modellerini oluşturmak için web güncelerini ön hazırlık sürecinden geçirir. Çevrimiçi bileşen gerçek zamanda çalışan bir öneri motorudur. Web sunucusundaki günceler kullanıcıları ve oturumları tanımlamakta kullanılırken sistemdeki web sayfaları veya kaynakları içeriklerine göre kümelenir. Kümelenen içerik ile kullanıcı tanımlamaları birleştirilerek kullanıcı gezinme örüntüleri alt görevlere bölünür. Bu bilgiler ışığında gezinme örüntüleri keşfedilir. Kümeleme algoritmaları ile elde edilen gezinme örüntüleri geçmiş oturum verilerinden elde edildiğinden sisteme eklenen yeni kaynaklar yer almamaktadır. Ancak çalışmada sisteme eklenen yeni kaynakları gezinme örüntüleri içine konumlayan bir yaklaşım geliştirilmiştir. Çevrimdışı verilerden keşfedilen gezinme örüntüleri öneri motoru için bilgi kaynağı oluşturmaktadır. Modelin tutarlılığını deneysel olarak ölçmek amacıyla üç adet ölçüt kullanılmıştır. Bu ölçütler; "öneri doğruluğu" tüm öneriler arasındaki doğru öneri oranı, "kısayol kazanımı" öneri sistemi sayesinde kullanıcının tasarruf ettiği tıklama sayısı ve "kaplama oranı" öneri sisteminin kullanıcıya ziyaret etmesi için önerdiği tüm sayfaların sistemdeki tüm sayfalara oranı olarak tanımlanmıştır. Çalışmada e-öğrenme web sitesindeki çevrimiçi öğrenme faaliyetlerine öneri sağlamak amacıyla geliştirilen öneri motoru, kullanıcıların bilgi gereksinimlerini tahmin ederek ilgili sayfalara hızlı ulaşımlarını sağlayacak kısayol önerisinde bulunmayı gerçekleştirmiştir.

3.4.3. Metin Madenciliği Uygulamaları

e-öğrenme sistemlerinde metin madenciliği, içeriğin analiz edilerek öğrenci özelliklerine göre içerik hazırlama faaliyetlerinde kullanılabilmektedir. Ancak içerik yönetim sistemlerinde bu işlevler yazılımın iç fonksiyonları tarafından gerçekleştirildiğinden bu alana çok fazla müdahale edilememektedir. Bu nedenle öğrenme yönetim sistemlerinde metin madenciliği daha çok öğrenci

etkileşim arşivlerinin incelenmesi ve faydalı bilgilerin keşfedilmesinde kullanılabilmektedir.

Dringus ve Ellis, Nova Southeasthern Üniversitesi, Bilgisayar ve Bilişim Eğitim Programında kullanılan Allaire's Cold Fusion yazılımı ile sağlanan bir çevrimiçi etkileşim ortamında bir öğreticinin forumdan faydalı bilgileri otomatik olarak elde etmesini amaçlayan bir model geliştirmişlerdir[122]. Çalışmada öğreticinin açılan bir tartışma konusunu değerlendirmesine yardımcı olabilecek veri ve metin madenciliğinin kullanıldığı bir sorgulama süreci geliştirilmiştir. Bu sorgulama süreci bir öğreticinin tartışma ortamında öğrencilerine anlamlı geribildirim sunmalarına destek olabilmek için öğrencilerin tartışma ortamındaki performansı hakkında bilgilerin türetilmesini sağlar. Ayrıca metin madenciliği ile öğreticinin tartışma forumunu değerlendirme yeteneğinin gelişeceği belirtilmiştir. Tartışma forum verilerin yer aldığı veritabanlarında veri madenciliğini uygulama dokuz adımda gerçekleştirilmiştir. Bu adımlar aşağıdaki gibi sıralanmıştır.

1. Görevi açıkça tanımlama
2. Mevcut veri özelliklerini inceleme
3. Verinin elde edilmesi
4. Bütünleştirme ve kontrol
5. Veri temizleme
6. Forum içerisinde soruların tespit edilmesi
7. Verinin madenciliğinin yapılması
8. Doğrulama
9. Yorumlama

Diringus ve Ellis bir forum yazılımının verilerini kullanarak bilgi elde etmek amacıyla metin madenciliği yaklaşımına dayalı bir strateji önermişlerdir. Bu stratejinin WebCT, Blackboard ve AltaVista gibi yazılımlar tarafından yönetilen forumlara da uygulanabilecek esneklikte olduğu vurgulanmıştır. Harici araçlar geliştirilerek bu sistemlerde yer alan tartışma ortamlarının

[122] Laurie P. Dringus ve T. Ellis, "Using Data Mining as a Strategy for Assessing Asynchronous Dicussion Forums," **Computer & Education Journal**. Cilt No 45: 141–160, (2005), s.141.

değerlendirilmesi sağlanacağı gibi ders yönetim sistemlerinin raporlama ve sorgulama araçlarına metin madenciliği yeteneklerinin kazandırılması mümkündür.

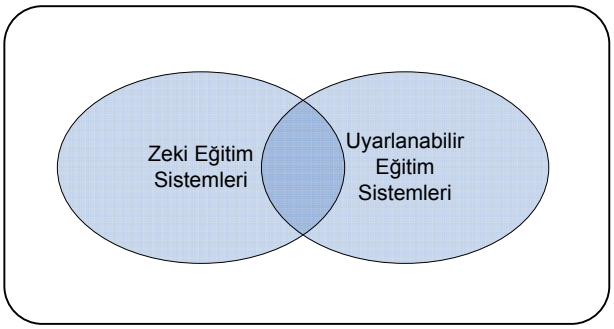
3.5. Uyarlanabilir Ve Zeki Web'e Dayalı Eğitim Sistemleri

Uyarlanabilir ve zeki web'e dayalı eğitim sistemleri (AIWBES) web'e dayalı eğitim sistemlerinde geleneksel ders sunumuna bir alternatif sağlar. AIWBES'de her öğrencinin amaçlarını, tercihlerini ve bilgisini içeren bir öğrenme modeli oluşturulur. Bu eğitim sisteminde öğrencinin öğrenme ihtiyaçlarının karşılanabilmesi için oluşturulan model ile öğrencinin etkileşimi kullanılarak uyarlanabilir web'e dayalı eğitim sistemi amaçlanır[123]. Web'e dayalı eğitim sistemlerinde uyarlanabilir ve zeki terimleri gerçekte eşanlamlı değildir. Uyarlanabilir sistemlerde, her bir öğrenci veya öğrenci grupları için bireysel veya grup öğrenci modellerinde toplanan bilgiler hesaba katılarak farklılık yaratılmaya çalışılır. Zeki sistemlerde ise kullanıcılara daha serbest ve gelişmiş destek sağlanması amacıyla yapay zeka teknikleri uygulanır. Uyarlanabilir ve zeki web'e dayalı sistemleri arasındaki ilişki Şekil 21'de gösterilmiştir. Web'e dayalı bir eğitim sistemi, sadece zeki veya sadece uyarlanabilir eğitim sistemi özelliklerini barındırabileceği gibi hem zeki hem de uyarlanabilir web'e dayalı eğitim sisteminin özelliklerini taşıyabilir.

Uyarlanabilir eğitim sistemlerinde öğrenme materyalinin hazırlanması öğrenci modeline göre kişiselleştirilir. Bununla beraber sistemde yer alan materyaller, sistem eğitmeni ve tasarımcısı tarafından belirlenir. Uyarlanabilir eğitim sistemlerinde kişiselleştirilmiş öğrenme materyalleri gelişmiş e-öğrenme sistemleri olarak adlandırılan sistemlerde web'den otomatik olarak bulunur ve

[123] Peter Brusilovsky ve C. Peylo, "Adaptive and Intelligent Web-based Educational Systems," **International Journal of Artificial Intelligence in Education**. Cilt no13: 156–169, (2003), s.156.

kullanıcının sistemle etkileşimi ve kişisel özelliklerine bağlı olarak kullanıcıya sunulur[124].



Şekil 21. Zeki Ve Uyarlanabilir Eğitim Sistemleri.

Peter Brusilovsky ve C. Peylo, "Adaptive and Intelligent Web-based Educational Systems," **International Journal of Artificial Intelligence in Education**. Cilt no13: 156–169, (2003)den uyarlandı.

AIWBES'de öğrencilerin sistemle etkileşiminin modellenmesi, web içeriklerinin öğrencinin özelliklerine göre seçilmesi, öğrencilere kişiselleştirilmiş öğrenme deneyimleri ve faaliyetleri hakkında öneri geliştirmek gibi ileri düzey analizlerde veri madenciliği tekniklerinden faydalanılmaktadır.

3.5.1. Kümeleme Ve Sınıflama Uygulamaları

Uyarlanabilir ve zeki web'e dayalı eğitim sistemleri hakkında birçok sistem önerisi ve pilot çalışma bulunmaktadır. Bu çalışmalarda, öğrenci davranışları ve öğrenme içeriklerinin çeşitli özelliklere göre gruplandırılmalarında kümeleme ve sınıflama veri madenciliği algoritmalarından faydalanılmaktadır.

- Arroyo ve diğerleri öğrencilerin öğrenmesini etkileyen gizli değişkenleri ortaya çıkarmak amacıyla anket ve günce verilerini birleştirerek bir "Bayesian Network" modeli oluşturmuşlardır[125]. Çalışma lise düzeyinde

[124] Tiffany Ya Tang ve G. McCalla, "Smart Recommendation for an Evolving e-learning System," **International Journal on E-Learning**. Cilt No 4, Sayı No 1: 105–129, (Haziran 2005), s.105.
[125] Ivon Arroyo ve diğerleri, "Inferring Unobservable Learning Variables from Students' Help Seeking Behavior," **ITS2004 Workshops - Analyzing Student-Tutor Interaction Logs to Improve Educational Outcomes'da sunulan bildiri** (Maceió, Alagoas. 30 Ağustos 2004), s.782.

matematik dersleri için tasarlanmış çoklu ortamla desteklenen web'e dayalı eğitim sistemi (Wayang Outpost) verileri kullanılarak gerçekleştirilmiştir. Sistem, öğrencilerin problemleri çözmelerine yardımcı olmak için veya öğrencinin yardım talep etmesi durumunda çoklu-ortam ve animasyonlarla desteklenen adım adım içeriği sunan bir yapıya sahiptir. Bu sistemde öğrencinin sistemle olan her etkileşimi ayrıntılı olarak bir ilişkisel veritabanına kaydedilir. Bireylerin gizli öğrenme değişkenlerini belirlemek amacıyla Massachusetts'de kırsal kesimde yer alan iki lisede eğitim gören 150 öğrenciye sistemi kullanım öncesi ve sonrasında anket uygulanmıştır. Öğrencilerin davranışları, yardım talepleri ve diğer değişkenlerin yer aldığı boyutlar arasındaki ilişki, anket ve sistem günce verileri kullanılarak analiz edilmiştir. Araştırmacılar, öğrencilerin sistemi kullanırken olumlu ve olumsuz tavırlarını ortaya çıkaran veriye dayalı bir model oluşturmuşlardır. Çalışmada ayrıca öğrencilerin sistemle etkileşimlerini gösteren verilerle, sistem kullanımı sonrası öğrencilere uygulanan anketlerden toplanan verilerin nasıl birleştirileceği tanımlanmıştır. Arroyo ve diğerleri, sistemde yardım için harcanan zaman gibi gözlenebilir değişkenlerle öğrencilerin motivasyon, tavır, algılama, inanç ve diğer gözlenemeyen davranışlarını belirleyen gizli değişkenleri birleştiren bir Bayesian öğrenci modeli oluşturmuşlardır.

- Muehlenbrock, veritabanı ve makine öğrenme tekniklerini kullanarak web'e dayalı öğrenme ortamları ile öğrencinin etkileşimini otomatik olarak analiz eden bir sistem tasarlamıştır[126]. Analiz sistemi matematik öğretiminin gerçekleştirildiği web'e dayalı etkileşimli öğrenme ortamı verileriyle test edilmiştir. Ayrıca sistem farklı web'e dayalı öğrenme sistemlerine uygulanabilecek şekilde tasarlanmıştır. Analiz sistemi, "ActiveMath" olarak adlandırılan web'e dayalı öğrenme ortamı verilerine "Java" ve "MySQL" teknolojilerini kullanarak ulaşmıştır. "ActiveMath" öğrenci amaçları, tercihleri, yetenekleri ve önceki bilgilerine uyarlanabilen dinamik olarak etkileşimli dersler oluşturan web'e dayalı öğrenme

[126] Martin Muehlenbrock, "Automatic Action Analysis in an Interactive Learning Environment," **12th AIED-2005 - workshop on Usage Analysis in Learning Systems'de sunulan bildiri** (Amsterdam. 18-22 Haziran 2005), s.73.

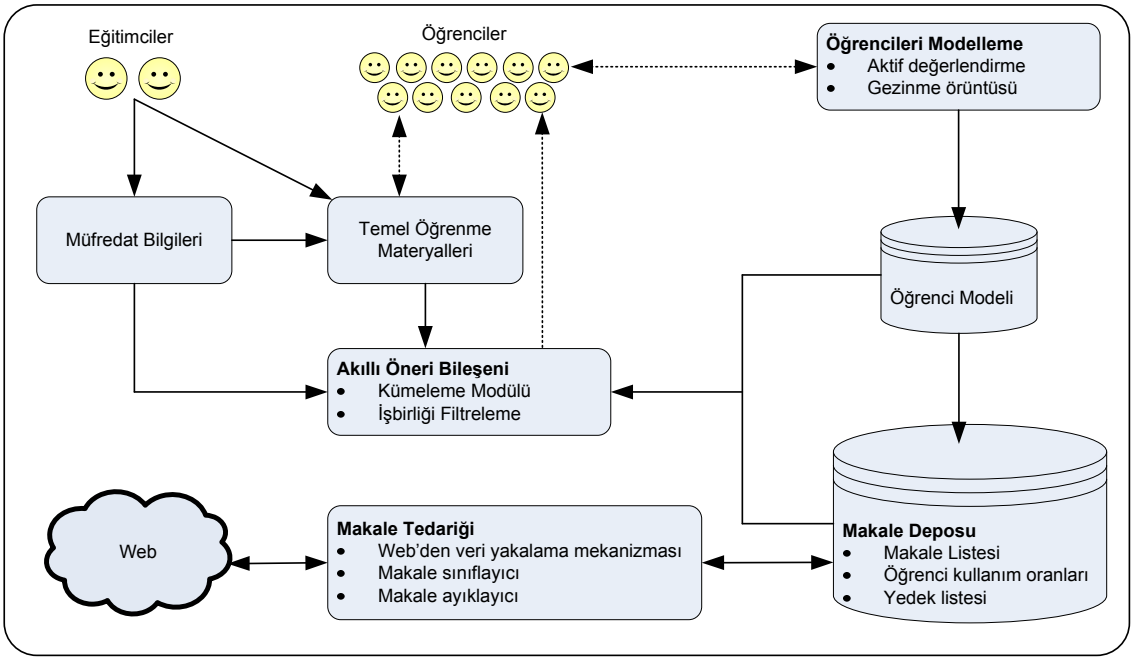
ortamıdır. Bu sistemde içerik XML enformasyon gösterimi formatında oluşturulur. Araştırmacının geliştirdiği analiz sisteminde, "ActiveMath" günce verilerine ulaşılarak veri madenciliği uygulanacak verinin hazırlanması için SQL komutları çalıştırılır. Analiz sisteminde C4.5 karar ağacı algoritmaları gibi farklı makine öğrenme yöntemleri uygulanarak veriler arasındaki ilişkiler ortaya çıkarılabilmektedir.

- Damez ve diğerleri acemi kullanıcıları deneyimli kullanıcılardan otomatik olarak ayırt etmede "fuzzy" karar ağaçlarını kullanmışlardır[127]. Çalışmada, insan-bilgisayar etkileşimlerini belirleyen bir öğrenme etmenini (learning agent) ve bir fuzzy karar ağacı üreticisini kullanan "TAFPA" (Tree Analysis for Providing Advices) adı verilen yazılım tanıtılmıştır. Kullanıcıların bilgisayarla olan etkileşim verilerinin kullanıldığı TAFPA yazılımında, kullanıcının deneyimli veya acemi olarak sınıflandırılmasında bir etmen kullanılır. Bu etmen aynı zamanda kullanıcının bilgisayarla olan etkileşiminin bilişsel özelliklerini öğrenmek için kullanılır. TAFPA yazılımında ilk aşamada günce verileri kullanılarak etmenin eğitilmesi gerçekleştirilir. İkinci aşamada ise ilk aşamada öğrenilen özellikler kullanılır. Yazılımda XML yapısı kullanılarak düğme tıklama, kayar çubuk kullanımı, klavye olayları, bağlantı tıklama gibi tüm kullanım davranışları günce veritabanına kaydedilir. Eğitim aşamasında sınıflama için gerekli bilişsel tanımlayıcılar elde edilir. Sınıflayıcı tarafından kullanıcının acemi ya da deneyimli olduğuna dair kararı verildiği anda TAFPA yazılımında "yardım görüntüle" kararı alınır ya da fuzzy karar ağacının yeniden oluşturulması için yeni bir "tanımlayıcı" bulunur. Fuzzy karar ağaçları, karar ağacı performansının arttırılması için fuzzy küme teorisi ile karar ağacı tekniğinin birlikte kullanılmasıdır. Geliştirilen yazılımın uyarlanabilir web'e dayalı eğitim sistemlerinde uygulanabileceği vurgulanmıştır.

[127] Mark Damez ve diğerleri, "Fuzzy Decision Tree for User Modeling from Human–Computer Interactions," **5th ICHSL'de sunulan bildiri** (Marrakech. 22-25 Kasım 2005), s.287.

- Tang ve McCalla gelişmiş bir e-öğrenme sistemi için “akıllı öneri” (smart recommendation) sistem mimarisi önermişlerdir[128]. Önerilen sistem uyarlanabilir web’e dayalı öğrenme sistemlerine ek olarak açık internet kaynaklarının da öğrenme içeriklerine dahil edildiği bir mimariye sahiptir. Önerilen sistemin yapısı ve bileşenleri Şekil 22’de verilmiştir. Sistemde temel öğrenme materyallerinin hazırlanılmasında, veri madenciliği ve web madenciliği tekniklerinden faydalanılmaktadır. Akıllı öneri bileşeni öğrencileri ilgi alanına göre kümeler ve öğrencilere önerilecek kaynakları belirler. Bu bileşende yer alan kümeleme analizleri her özel kullanıcı grup için genelleştirilmiş temsili ilgi alanlarını belirler. Böylece her öğrenci birden fazla kümede yer alabilir. Kümeleme analizi benzerliklerine göre kullanıcıları gruplamada başarılı olmasına rağmen kullanıcılara uygun kaynağı belirlemekte yetersizdir. Bu amaçla araştırmacılar her öğrenciye uygun kaynağın önerilmesi için “işbirliği filtreleme tekniğinin” (collaborative filtering technique) uygulanmasını önermişlerdir. İşbirliği filtreleme tekniğinde amaç, hedef kullanıcının komşularını oluşturmaktır. Bu teknikte “Pearson-korelasyonu” temelli ve “kosinüs” temelli benzerlik ölçümleri kullanılmaktadır. Çalışmada Pearson-korelasyonu temelli benzerlik ölçümü tercih edilmiştir. Akıllı öneri bileşeni, kümeleme analizi ve kullanıcılara önerilecek materyallerin seçiminde öğrencilerin sistemin seçtiği makaleleri oylama verilerinin de kullanıldığı işbirliği filtreleme tekniği önerilmiştir.

[128] Tang, **Ön.ver.**, s.105.



Şekil 22. Tang Ve McCalla Tarafından Önerilen e-Öğrenme Sisteminin Mimarisi.

Tiffany Ya Tang ve G. McCalla, "Smart Recommendation for an Evolving e-learning System," **International Journal on E-Learning**. Cilt No 4, Sayı No 1: 105–129, (Haziran 2005)den uyarlandı.

3.5.2. Birliktelik Kuralları Ve Sıra Örüntüleri Uygulamaları

Uyarlanabilir web'e dayalı eğitim sistemleri öğrenci gereksinimlerini karşılayacak ders içeriklerini belirlemede ilk olarak öğrenci modellerini oluştururlar. Bu aşamada öğrencilerin bir takım özellikleri, kümeleme ve sınıflama modellerinin girdi verisi olarak kullanılmaktadır. Birliktelik kuralları ise genellikle öğrenci gruplarının öğrenme materyalleri ile eşleştirilmelerinde başvurulan yöntemler olmuştur.

Zeki eğitim sistemlerinde birliktelik kurallarından, içerik ile öğrenci etkileşimi arasındaki ilişkileri analiz etmede faydalanılabilmektedir. Freyberger ve diğerlerinin geliştirdikleri transfer modelinde, zeki ders sunum ortamında yer alan sorular ile soruyu doğru yanıtlamak için gerekli yönlendirici bilgi, beceri, strateji gibi bilgi bileşenleri arasında konumlama yapan birliktelik kuralları

kullanılmıştır[129]. Problemler genellikle birden fazla beceri ile ilişkilidir. Bilgi transferi, bir öğrencinin bir problemi çözdükten sonra farklı bir problemi çözerken önceki problemde kazandığı beceri veya bilgiyi yeni problemde kullanması olarak tanımlanabilir. Transfer modelini oluşturmak bir problem için gerekli olan becerilerin değerlendirilmesinde kolaylık sağlar. Transfer modeli, satırlarında farklı problem türlerinin, sütunlarında ise her problemin çözülmesi için gerekli bilgi bileşenlerinin yer aldığı bir tablo yardımıyla gösterilebilir. Çalışmada etkin bir transfer modeli oluşturmak amacıyla çevrimiçi bir matematik dersinde oluşturulmuş öğrenci etkileşim günceleri kullanılmıştır. Transfer modellerinin oluşturulmasında birliktelik kurallarını kullanan yordamların daha başarılı olduğu ifade edilmiştir.

Merceron ve Yacef, Sydney Üniversitesi Enformasyon Teknolojileri Eğitim Programlarında kullanılan web'e dayalı zeki öğretim yardımcı sistemi Logic-ITA üzerinde birliktelik kurallarını ve sembolik veri analizini (symbolic data analysis) uygulamışlardır[130]. Çalışmada Logic-ITA sisteminin öğrenme üzerindeki etkisi öğrencilerin sistem kullanım verilerinin sembolik veri analizi kullanılarak incelenmiştir. Logic-ITA sisteminin öğretme üzerindeki etkisi ise öğrencilerin soru çözerken yaptıkları hataların birliktelik analizi kullanılarak incelenmesi ile ortaya çıkarılmıştır.

Sembolik veri analizi ile "sadece alıştırma1'i çözen öğrenciler", "sadece alıştırma 2'yi çözen öğrenciler", … gibi sembolik nesneler yaratılmıştır. Bu nesneler sistemi etkin kullanmayan öğrencileri belirlemek için kullanılmıştır. Logic-ITA veritabanlarına öğrenci ile ilgili "ortalama hata oranı", "ortalama doğru oranı" ve "başarılı tamamlanan alıştırma sayısı" şeklinde yeni değişkenler eklenmiştir. Elde edilen nesne ve eklenen değişkenler SODAS grafik aracı kullanılarak beş eksenli bir histogram ile gösterilmiştir. Bu analiz sonucu,

---

[129] Jonathan Freyberger, N. T. Heffernan ve C. Ruiz, "Using Association Rules to Guide a Search for Best Fitting Transfer Models of Student Learning," **ITS2004 Workshops - Analyzing Student-Tutor Interaction Logs to Improve Educational Outcomes'da sunulan bildiri** (Maceió, Alagoas. 30 Ağustos 2004), s.1.

[130] Agathe Merceron ve K.Yacef, "Mining Student Data Captured From a Web-based Tutoring Tool: Initial Exploration and Results," **Journal of Interactive Learning Research**. Cilt No 15, Sayı No 4: 319–346, (2004), s.319.

öğrenci başarısı ve çözülen alıştırma sayısı arasında doğru orantı olduğu gözlenmiştir.

Logic-ITA zeki web'e dayalı öğretim sisteminin öğretme üzerindeki etkisi öğrencilerin alıştırma çözümünde yaygın olarak yaptıkları hatalardan yola çıkılarak belirlenmeye çalışılmıştır. Mevcut sistemde öğreticiler, çeşitli raporlama araçlarını kullanarak öğrencilerin öğrenme davranışlarını izleyebilmektedir. Bu araçlar "yaygın hatalar", "hatalara neden olan alıştırmalar", "öğrenci gelişme seviyesine göre yapılan hatalar" şeklinde öğreticiler tarafından sorgulanabilmektedir. Öğreticiler SQL sorgularından elde ettikleri betimleyici istatistik analizi sayesinde sık gerçekleşen hataların giderilmesi için içeriği gözden geçirebilirler. Araştırmacılar birlikte meydana gelen hataları bulmak ve ilişkilendirmek amacıyla birliktelik analizi uygulamışlardır. Birlikte meydana gelen hataların belirlenmesindeki amaç, öğreticilerin öğrencilere kavramları açıklarken hassas noktaları vurgulamalarında yardımcı olmaktadır. Birliktelik analizinde bir öğrencinin yapmış olduğu tüm hatalar kümesi ve özel bir hata türüne ilişkin tüm öğrenci hata kümesi incelenerek birliktelik kuralları elde edilmiştir. Bunun nedeni her öğrencinin her alıştırmayı çözmemesi olarak belirtilmiştir. Analiz sonucu hatalara neden olan iki temel nokta bulunmuş ve öğreticiler tarafından doğrulanmıştır.

3.5.3. Metin Madenciliği Uygulamaları

Metin madenciliği, AIWBES'de öğrenci profiline uygun içeriğin hazırlanmasında ve analiz edilmesinde kullanılabilmektedir. Tang ve diğerleri, web madenciliğine dayalı kişiselleştirilmiş ders yazılımının nasıl yapılandırılacağı hakkında bir çalışma yapmışlardır[131]. Yazarlara göre uzaktan eğitimde yer alan bir öğreticinin sahip olması gereken araçlar aşağıda verilmiştir.

---

[131] Changjie Tang ve diğerleri. "Personalized Courseware Construction Based on Web Data Mining," **The First International Conference on Web Information Systems Engineering'da sunulan bildiri** (Hong Kong. 19 – 20 Haziran 2000), s.2204.

1. *Öğrenci profil kümesi:* Öğrenciye ait isim, yaş, sınıf, ilgi alanı, akademik geçmişi ve benzer verilerdir.
2. *Öğrencilerin kümelenmesi:* Öğrencilerin akademik geçmişlerine göre kümelenmesidir.
3. *Eğitim ağacı:* Her öğrenci kümesi için tasarlanmış öğrenme şemasıdır. Eğitim ağacının her düğümü iki nesneden oluşur. Bunlar "WebObj" olarak simgelenen eğitim nesnesi (bir makale veya alt eğitim ağacı) ve diğeri "weight" olarak adlandırılan bir tamsayı dizinidir. Bu dizin küme sayısı, ders önemi, öğretme süresi gibi değerleri içerir.
4. *Değerlendirme ve güncelleme:* Öğrenme sonuçlarının değerlendirilmesi ve öğrenci profillerinin güncellenmesi faaliyetlerini içerir.

Çalışmada uzaktan eğitimde etkin ve kişiselleştirilmiş web'e dayalı eğitim sistemi için önerilen mimaride eğitim nesnesinin nasıl yapılandırılması gerektiği konusunda önerilerde bulunulmuştur. Örnek bir kişiselleştirilmiş eğitim ağacının yapılandırılmasında kullanılacak algoritmalar ve özellikleri verilmiştir. Bir derse ilişkin eğitim ağacını oluşturmak amacıyla web'de mevcut "16. Çin Ulusal Veritabanı Konferans" yayınlarını çeşitli özelliklerine göre eğitim ağacı dallarına konumlayan algoritmalar önerilmiştir.

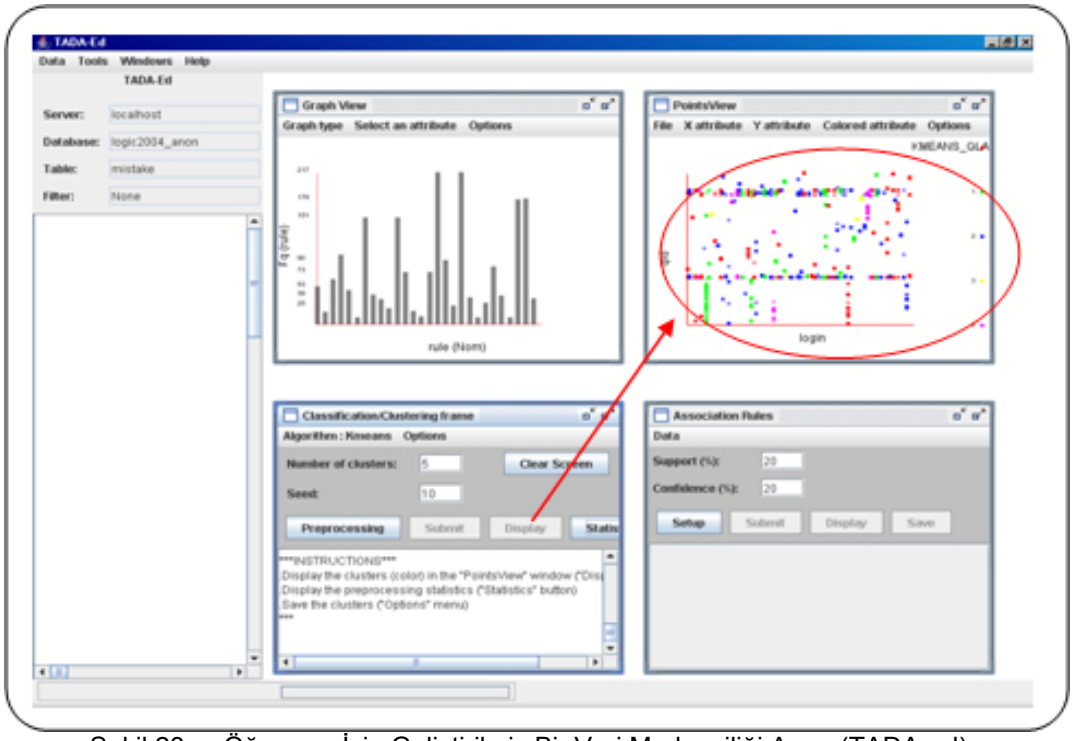
4. Web'e Dayalı Eğitim Sistemlerinin Geleceğinde Veri Madenciliğinin Rolü

Veri madenciliği geleneksel eğitim sistemleri, web'e dayalı dersler, öğrenme yönetim sistemleri ve uyarlanabilir ve zeki web'e dayalı eğitim sistemlerinde uygulama alanı bulmuştur. Farklı veri kaynaklarına sahip bu eğitim sistemlerinde, mevcut veri ön hazırlık sürecinden geçirildikten sonra istatistik, görselleştirme, kümeleme, sınıflama, aykırı değer analizi, birliktelik kuralı ve örüntü madenciliği, metin madenciliği gibi veri madenciliği teknikleri uygulanabilmektedir.

Veri madenciliğinin eğitim sistemlerinde uygulanmasının diğer alanlarda gerçekleştirilen veri madenciliği uygulamalarından farkı öğrenci ve web'e dayalı eğitim sisteminin pedagojik görünüşünün hesaba katılmasıdır. Eğitimde veri madenciliğinin uygulanması yeni bir araştırma alanı olmasına rağmen bu alan yapılmış birçok önemli çalışma bulunmaktadır. Web'e dayalı eğitim sistemlerinin yönetilmesinde karşılaşılan en önemli sorun öğrencinin sistemle olan etkileşiminin izlenememesidir. Bu nedenle öğrencinin sistem içindeki davranışlarını anlamlandırmak için izleme sistemlerinden faydalanılmıştır. GISMO gibi sistemler gizli davranış örüntülerini belirlemese de sistemdeki öğrenci hareketlerini izlemeye yardımcı olmaktadırlar.

Araştırmacılar farklı web'e dayalı eğitim sistemlerinde gerçekleştirdikleri veri madenciliği çalışmalarıyla eğitim sistemlerine katkıda bulunmuşlardır fakat eğitimcilerin kendi başlarına yönetebilecekleri ve kullanılabilecekleri standartlaştırılmış araçlar yaygınlaşmamıştır. Sınırlı sayıda uygulamalardan biri olan "TADA-ed" (Tool for Advanced Data Analysis for Education) aracı, pedagojik olarak ilgili örüntülerin keşfedilmesi amacıyla çevrimiçi alıştırmaları çözen öğrencilere ilişkin verilerin görselleştirilmesi ve madenciliğinin yapılması için geliştirilmiştir[132]. TADA-ed, öğreticilere anlamlı sonuçları türetmek için gerekli veri madenciliği algoritmaları sağlamakta ve algoritmaların ihtiyaç duyduğu verinin ön hazırlık sürecini de yazılımın içinde gerçekleştirmektedir. Bu sayede veri madenciliği uzmanı olmayan eğitimciler sistemi rahatlıkla kullanabilmektedir. Şekil 23'de TADA-ed aracının "K-means" kümeleme tekniğini kullanarak öğrencilerin çevrimiçi ortamla etkileşimlerini gruplandırdığı bir ekran görüntüsü verilmiştir.

[132] Agathe Merceron, A. ve K. Yacef, "Tada-ed for Educational Data Mining," **Interactive Multimedia Electronic Journal of Computer-Enhanced Learning**. Cilt No 7, Sayı No 1: 267–287, (2005), s.267.



Şekil 23. e-Öğrenme İçin Geliştirilmiş Bir Veri Madenciliği Aracı (TADA-ed).

Agathe Merceron, A. ve K. Yacef, "Tada-ed for Educational Data Mining," **Interactive Multimedia Electronic Journal of Computer-Enhanced Learning**. Cilt No 7, Sayı No 1: 267–287, (2005)den uyarlandı.

Eğitim için veri madenciliği araçlarının gelecekte öğrenme yönetim sistemlerinin bir parçası haline geleceği beklenmektedir. Eğitim alanındaki veri madenciliği çalışmalarının yönelmesi gereken konular aşağıda belirtilmiştir[133].

- *Veri madenciliği uzmanı olmayan eğitimcilerin daha kolay kullanabilecekleri madencilik araçlarının geliştirilmesi:* Veri madenciliği araçları analizlerin kolay olarak gerçekleşmesi için değil daha güçlü ve esnek olmaları için tasarlanır. Günümüzde kullanılan araçların çoğu eğitimcilerin kullanamayacağı kadar karmaşıktır ve bu araçlarla istedikleri hedefe ulaşmaları zordur. Bu nedenle bu araçlar öncelikle e-öğrenme tasarımcılarının ve eğitimcilerin elde edilen sonuçları anlamalarında daha iyi görselleştirme araçları sunmalıdırlar. Ayrıca bu araçlar analizin daha basit olarak yürütülmesi için bilinçli ve kolay kullanıma sahip bir ara-yüzü kullanıcıya sunmalıdırlar.

[133] Romero, **Ön.ver.**, s.144.

- *Tekniklerin ve verinin standartlaştırılması:* Web'e dayalı eğitim sistemlerinde depolanan verilerin ve yapıların farklı olması nedeniyle gerçekleştirilen veri madenciliği uygulamasının sisteme özel olmasına ve yeniden başka bir sisteme uygulanmasına engel teşkil etmektedir. Bu nedenle veri madenciliği tekniklerinin ve veri yapılarının standart hale getirilmesi, veri madenciliği uygulamalarının yaygınlaşmasını sağlayacaktır.
- *e-öğrenme sistemiyle bütünleşme:* Veri madenciliği uygulama adımlarının tümünün tek bir e-öğrenme uygulamasında gerçekleştirilmesi ile elde edilen geribildirim ve sonuçlar doğrudan kullanılabilir.
- *Özel veri madenciliği teknikleri:* Eğitim ilgi alanına özel madencilik tekniklerinin geliştirilmesi içerik tasarımı ve pedagojik kararların geliştirilmesine fayda sağlayacaktır.

ÜÇÜNCÜ BÖLÜM

ANADOLU ÜNİVERSİTESİ UZAKTAN EĞİTİM SİSTEMİ ÖĞRENCİ VERİLERİNDE VERİ MADENCİLİĞİ UYGULAMALARI

1. ANADOLU ÜNİVERSİTESİ UZAKTAN EĞİTİM SİSTEMİ

Anadolu Üniversitesi 1958 yılında Eskişehir İktisadi ve Ticari İlimler Akademisi olarak eğitime başlamış ve 1982 yılında Anadolu Üniversitesi adını almıştır. Anadolu Üniversitesi 9 fakülte, 6 yüksekokul, Devlet Konservatuarı, 3 meslek yüksekokulu, 9 enstitü ve 27 araştırma merkezine ek olarak uzaktan eğitim sistemi uygulayan 3 fakültesiyle evrensel üniversite değerlerine sahip bir üniversitedir.

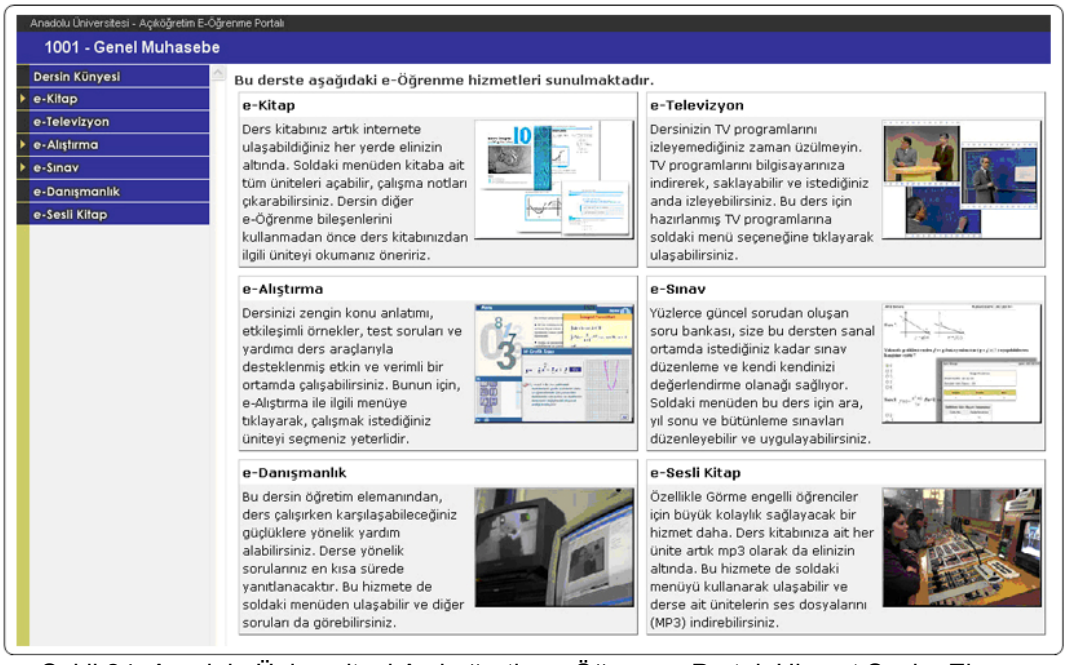
Anadolu Üniversitesi Uzaktan Eğitim Sistemi 1982 yılında 29.500 öğrenciyle başladığı eğitim faaliyetlerine Açıköğretim, İktisat ve İşletme Fakültelerine kayıtlı yaklaşık 1.050.000 öğrenci ile sürdürmektedir. Uzaktan Eğitim Sistemi bünyesinde dört yıllık lisans eğitimi veren İşletme ve İktisat Fakültesi ve 2 yıllık önlisans eğitimi veren Açıköğretim Fakültesi yer almaktadır. İşletme, İktisat, Kamu Yönetimi, Maliye, Çalışma Ekonomisi ve Endüstri İlişkileri bölümleri, İşletme ve İktisat Fakültelerine bağlı olarak hizmet vermektedirler. Açıköğretim Fakültesinde; İktisadi ve İdari Programlar Bölümü, Sağlık Programları Bölümü, Uzaktan Eğitim ve Yaygın Eğitim Programları Bölümü iki yıllık önlisans eğitimi veren programlardır. 2000-2001 öğretim yılında Milli Eğitim Bakanlığı ile işbirliği yapılarak Okul Öncesi ve İngilizce Öğretmenliği Lisans Programları ilk uzaktan eğitim sistemiyle öğretmen yetiştiren programlar olmuştur. 2001-2002 öğretim yılında ise internet üzerinden eğitim veren İnternet'e Dayalı Bilgi Yönetimi Önlisans Programı hizmete başlamıştır. Ayrıca İktisat, İşletme, Turizm ve Otelcilik, Dış Ticaret, Halkla İlişkiler, Bilgi Yönetimi programları ile Açık ilköğretim ve Açık Lise programları, Batı Avrupa Programları kapsamında 6 Batı Avrupa ülkesinde yaşayan Türk vatandaşlarına hizmet vermektedir.

Anadolu Üniversitesi Uzaktan Eğitim Sisteminde Meslek Eğitimi (Kara, Hava ve Deniz Komutanlıkları, Jandarma Genel komutanlığı, Emniyet Genel Müdürlüğü Meslek Eğitimi ve Adalet Meslek Eğitimi Önlisans), Dikey Geçiş, İkinci Üniversite ve Lisans Tamamlama projeleri yaşam boyu eğitim çerçevesinde yürütülmektedir. Ayrıca Uzaktan Eğitim Sisteminde e-MBA, e-konaklama ve e-Gelişimsel Yetersizlikleri Olan Çocukların Öğretmenliği internete dayalı uzaktan eğitim modellerini uygulayan programlardır.

Anadolu Üniversitesinde uygulanan Uzaktan Eğitim Sisteminin temel malzemesi ders kitaplarıdır. Öğrenci merkezli olarak yürütülen uzaktan eğitim sürecinde hazırlanan tüm öğretim materyalleri biçim ve içerik açısından öğrencinin kendi kendine öğrenmesini sağlayacak şekilde tasarlanmaktadır. Televizyon programları, akademik danışmanlık hizmetleri, videokonferans ve bilgisayar/internet destekli eğitim uygulamaları diğer öğretim materyallerini oluşturmaktadır.

Anadolu Üniversitesi Uzaktan Eğitim Sistemi'nde öğrenci işlerini yürütme görevini üstlenen bürolar, İşletme, İktisat ve Açıköğretim Fakültesine kayıtlı öğrencilerin her türlü öğrenci hizmetlerini merkeze gelmeden yürütmelerini sağlamak amacı ile hizmet vermektedirler. Sistemin sınav organizasyonunda Test Araştırma Birimi sınav sorularının hazırlanmasından ve Bilgisayar Araştırma ve Uygulama Merkezi (BAUM) ise 82 ilde 89 merkezde, çeşitli Avrupa ülkeleri ve Kuzey Kıbrıs Türk Cumhuriyeti'nde sınavların uygulanmasından sorumludur.

Anadolu Üniversitesi Uzaktan Eğitim Sistemi, öğrencilerine internet yoluyla zaman ve mekan bakımından bağımsız olarak ders çalışmalarına olanak sağlayan e-öğrenme hizmetlerini bilişim teknolojisindeki gelişimlere paralel olarak yaygınlaştırmaktadır. Öğrencilere sunulan e-öğrenme hizmetleri 2005 yılının mayıs ayında e-öğrenme portalı adı altında birleştirilerek öğrencilerin kullanımına sunulmuştur. e-öğrenme portalı hizmet seçim ekranı Şekil 24'de verilmiştir. e-öğrenme portalında öğrencilere sunulan hizmetler ve özellikleri aşağıda özetlenmiştir.



Şekil 24. Anadolu Üniversitesi Açıköğretim e-Öğrenme Portalı Hizmet Seçim Ekranı

- *e-Kitap:* Ders kitaplarının internet üzerinden elektronik olarak okunmasına olanak sağlayan e-kitap hizmeti kapsamında 224 derse ait 2877 üniteye ulaşım sağlanmaktadır. e-kitap hizmeti 2003-2004 öğretim yılında hizmete başlamıştır.
- *e-Televizyon:* Uzaktan Eğitim dersleri için hazırlanmış TV programlarına öğrencilerin internet üzerinden ulaşımlarına olanak sağlayan e-öğrenme hizmetidir. 2003-2004 öğretim yılında hizmete giren e-televizyon, öğrencilerin 163 ders için hazırlanmış 1190 TV programını izleme ve kendi bilgisayarlarına kaydetme olanağı sunar.
- *e-Alıştırma:* Sınırlı sayıda ders ile 2002-2003 öğretim yılında başlayan e-alıştırma hizmeti 50 derse ait 738 bileşen ile öğrencilere sunulmaktadır. Çoklu-ortam etkileşimi yaratan e-alıştırma öğrencilerin ders içeriklerini izleyerek, soru örneklerini çözerek, kendilerini sınayarak öğrenmelerine yardımcı olmaktadır.
- *e-Sınav:* Öğrencilerin en fazla ilgi gösterdikleri hizmetlerden biri olan e-sınav, içeriğinde 132 derse ait 11.500 soru bulundurmaktadır. Öğrencilerin sınavlar öncesi kendilerini denemelerine olanak sağlayan hizmet 1999-2000 öğretim yılında başlamıştır.

- *e-Danışmanlık:* Öğrencilerin akademik danışmanlara soru sormalarına izin veren e-danışmanlık hizmetinde 2006-2007 öğretim yılında 74 ders uygulamaya dahil edilmiştir.
- *e-Sesli kitap:* Görme engelli öğrenciler için hazırlanan e-sesli kitap hizmeti 2005 yılında hizmete girmiştir. 2006-2007 öğretim yılında 24 derse ait kitap için verilen hizmet görme engellilerin dışında diğer öğrencilerin de talep ettikleri bir hizmet olmuştur.

Anadolu Üniversitesi Uzaktan Eğitim Sisteminde e-öğrenme portalı dışında bazı özel programlar için farklı e-öğrenme hizmetleri de sunulmaktadır. Bilgi Yönetimi Önlisans Programı, İngilizce Öğretmenliği Lisans Programı ve Okul Öncesi Öğretmenliği Lisans Programları için farklı öğrenme bileşenlerini içeren e-öğrenme hizmetleri sunulmaktadır.

2. ARAŞTIRMANIN AMACI VE ÖNEMİ

Anadolu Üniversitesi Açıköğretim ve Uzaktan Eğitimde 25 yıla dayanan tecrübesi ve birikimiyle büyük bir organizasyondur. Sistem planlanan eğitim faaliyetlerinin yürütülmesi amacıyla birbiriyle koordineli çalışan birçok birim ve organizasyondan oluşmaktadır. Sınav organizasyonu, büro organizasyonu, akademik danışmanlık organizasyonu, bilgisayar destekli eğitim birimi, test araştırma birimi gibi birçok organizasyon öğrencilerin kaliteli bir eğitim sürecinden geçmesi için ilgili faaliyetleri yürütürler. Sistemin başarıyla yönetilmesinde akademik ve idari konularda alınan kararların önemi büyüktür.

Öğrenci sayısı göz önüne alındığında mega üniversite olarak değerlendirilen Anadolu Üniversitesinin büyük hacimli veritabanlarına sahip olduğu söylenebilir. Öğrenci verileri üzerinde gerçekleştirilecek veri madenciliği çalışmalarının sistem hakkında farkına varılmamış bilgilerin elde edilmesine katkı sağlayabileceği düşünülmektedir.

Araştırmanın amacı veri madenciliği analizleri sonucu elde edilecek bilgilerle sürekli gelişen e-öğrenme sisteminin yapılanmasına yardımcı

olabilmek ve Uzaktan Eğitim Sistemi planlama faaliyetlerine katkı sağlamaktır. Araştırma kapsamında Uzaktan Eğitim Sisteminde eğitim gören öğrencilere ilişkin farklı kaynaklardaki veriler bir araya getirilerek aşağıdaki sorulara cevap aranmıştır.

- Öğrenci özellikleri ve e-öğrenme faaliyetlerine bakılarak öğrenci performansı tahmin edilebilir mi?
- Öğrencilerin mezuniyet süreleri ve özellikleri arasındaki örüntüler belirlenebilir mi?

Çalışmada sistemle ilgili tüm verileri içeren bir veritabanı oluşturularak veri madenciliği hazırlık aşamaları ve model geliştirme adımları sırasıyla uygulanmıştır.

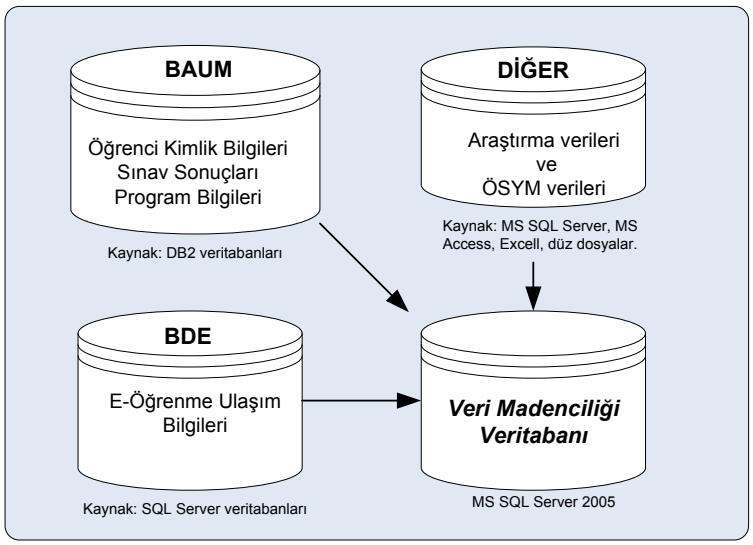
3. ARAŞTIRMA YÖNTEMİ

Araştırmanın gerçekleştirilmesinde önceki bölümlerde ayrıntılı olarak yer verilen veri madenciliği uygulama adımları takip edilmiştir. Cevabı aranan sorulara yanıt olabilecek veri kümesini oluşturabilmek amacıyla veri madenciliği veritabanı oluşturulmuştur. Oldukça uzun zaman alan bu süreçte farklı kaynaklardan toplanan veriler birleştirilerek veri temizleme, tamamlama ve özetleme işlemleri gerçekleştirilmiştir. Veri madenciliği veritabanı hazırlama süreci, modelleme süreci ve elde edilen bulgular izleyen bölümlerde anlatılacaktır. Veri madenciliği veritabanı MS SQL Server 2005 veritabanı yönetim yazılımında, veri madenciliği modelleri ise SPSS Clementine veri madenciliği yazılımında gerçekleştirilmiştir. SPSS firmasının veri madenciliği çözümü olan Clementine yazılımına ilişkin bilgiye EK 2'de yer verilmiştir.

3.1. Veri Madenciliği Veritabanının Hazırlanması

Anadolu Üniversitesi Uzaktan Eğitim Sisteminde öğrencilere ait veriler farklı kaynaklardan elde edilmiştir. Bunlardan ilki Anadolu Üniversitesinin tüm öğrenci bilgi siteminin yer aldığı Bilgisayar Araştırma ve Uygulama Merkezi

(BAUM) veritabanlarıdır. Bu veritabanlarında öğrencilere ait kimlik bilgileri ve sınav sonuç bilgileri yer almaktadır. Diğer bir veri kaynağı ise Uzaktan Eğitim Sistemine kayıtlı öğrencilerin faydalandığı e-öğrenme faaliyetlerinin sunulduğu Bilgisayar Destekli Eğitim Birimi (BDE) veritabanlarıdır. e-öğrenme sisteminde öğrenci verileri BAUM'dan alınarak sistemde tanımlanır ve öğrencilerin e-öğrenme ortamındaki faaliyetleri kısıtlı da olsa güncelere kaydedilir. Öğrencilere ait diğer bir veri kaynağı da öğrenciler üzerinde gerçekleştirilen araştırmalardır. Bu araştırmalar öğrenci profillerini ortaya çıkarmak veya özel bir araştırma konusunda öğrenci fikirlerini almak amacıyla düzenlenmiş anket çalışmalarıdır. Ancak bu araştırmaların çoğu anketi dolduran öğrencinin kimliğinin gizlenmesi veya oldukça az sayıdaki örneklem nedeniyle veri madenciliği analizlerinde kullanılamamaktadır. ÖSYM tarafından sisteme yerleştirilen öğrencilere ilişkin ÖSYM verileri öğrencilere ait bir diğer veri kaynağıdır. Şekil 25'te Anadolu Üniversitesi Uzaktan Eğitim Sistemi öğrencilerine ait veri kaynakları gösterilmiştir.

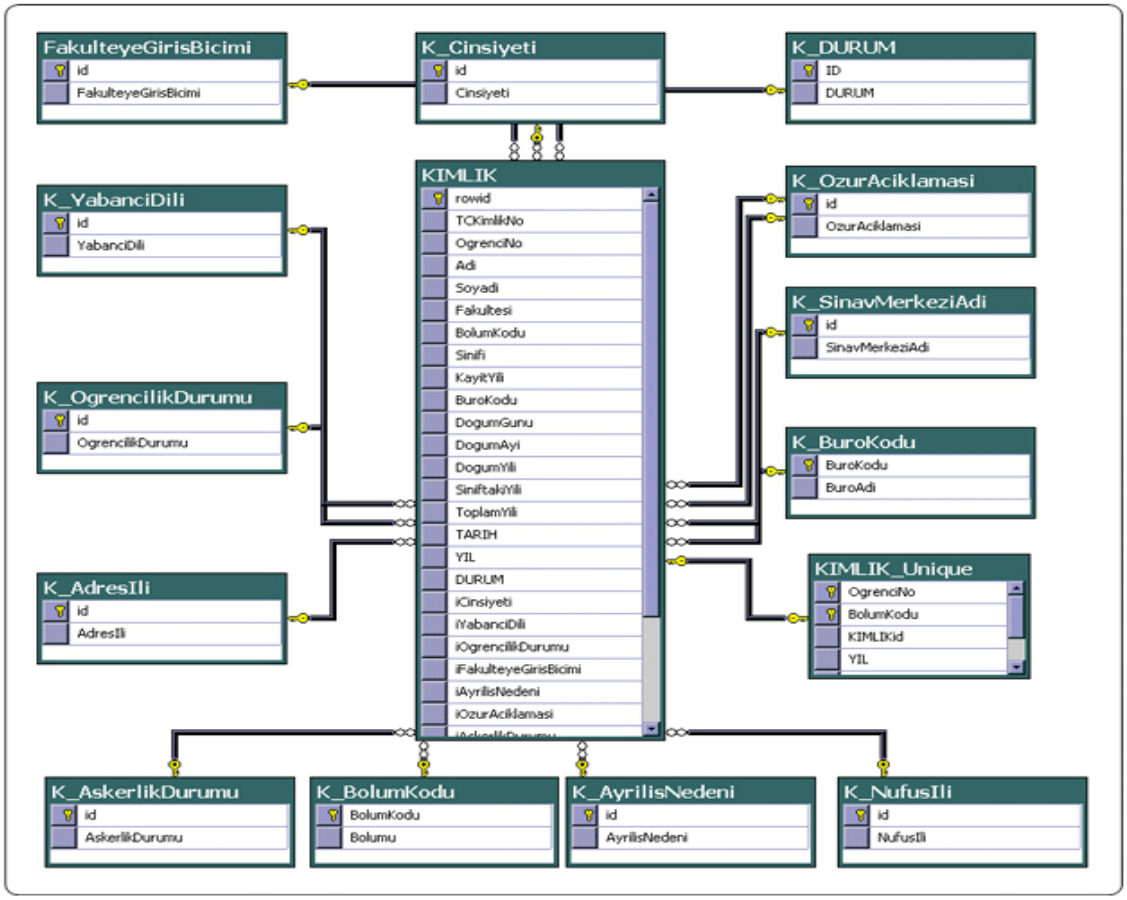


Şekil 25. Uzaktan Eğitim Sistemi Öğrenci Verileri Organizasyonu

Veri madenciliği veritabanının hazırlanması için sanal olarak kurulan MS 2003 Server işletim sisteminde MS SQL Server 2005 veritabanı yönetim yazılımı kurulmuştur. İşlemci performansının yetersiz olması nedeniyle oluşturulan sanal sunucu üç PC'ye kopyalanarak hazırlık süreçleri paralel olarak sürdürülmüştür.

3.1.1. BAUM Veritabanları

BAUM veritabanlarında yer alan uzaktan eğitim öğrencilerine ait veriler, gerekli dönüşümler yapılarak SQL server ortamına aktarılmıştır. Öğrencilere ait kimlik tablosu 3.404.235 satır veriden, öğrencilere ait notların yer aldığı tablo ise 66.255.456 satır veriden oluşmaktadır. Uzaktan Eğitim Sistemine katılan tüm öğrencilere ait veriler uygun şekilde biçimlendirilmiş ve daha hızlı işlenebilmesi amacıyla normallik dereceleri arttırılmış ve indekslenmiştir. BAUM'dan sağlanan verilerde öğrencilerin takip edilebilmesi için en uygun alan olan "Öğrenci Numarası" indeks olarak belirlenmiştir. 2000'li yıllardan sonra öğrenci numarası olarak kullanılan TC kimlik numarası da veri tablolarında yer almakta ancak hatalı TC kimlik numaralarının düzeltilmesi sonucu öğrencinin geçmiş bilgilerine erişimde sorunlar çıkabilmektedir. Öğrencilerin kimlik verilerinin ilişki şeması Şekil 26'da gösterilmiştir.



Şekil 26. BAUM'dan Sağlanan Kimlik Verisi Tablolarının Yapılandırılmış Görünümü

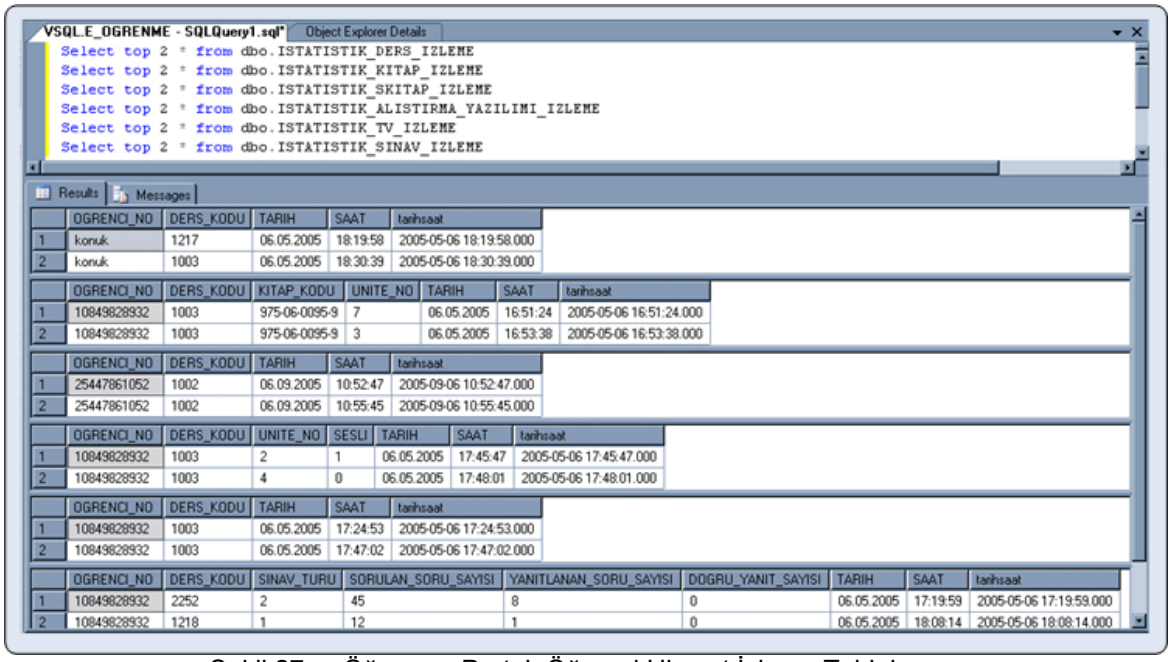
Öğrencilerin sistemde birden fazla kayıt oluşturması ve öğrenci aflarıyla geri dönen öğrenci kayıtlarının oluşturduğu karmaşanın ayıklanabilmesi için öğrencilerin kayıt oldukları bölüm ve yıllar dikkate alınarak benzersiz kayıtlardan oluşan bir tablo oluşturulmuştur. Bu sayede bir öğrenci kaydını diğer kaynaklardan gelen verilerle eşleştirme mümkün kılınmıştır. Öğrenci not verileri ders ve öğrenci bazında gruplanarak daha küçük boyuttaki veri tablolarına dönüştürülmüş ve indekslenmiştir.

3.1.2. BDE Veritabanları

Öğrencilerin e-öğrenme faaliyetlerine ilişkin veriler iki kaynakta yer almaktadır. Bunlardan ilki web sunucu günceleri diğeri ise e-öğrenme uygulamasının oluşturduğu veritabanlarıdır. Genellikle e-öğrenme uygulamaları öğrencinin öğrenme faaliyetlerine odaklandığından oluşturdukları veriler web güncelerine göre oldukça zengin olmaktadır. Ancak Anadolu Üniversitesi Uzaktan Eğitim Sisteminde e-öğrenme sisteminden faydalanan öğrenci sayısının yüksek olması donanımların yetersiz kalmasına neden olmaktadır. Özellikle sınav dönemlerinde e-öğrenmenin sunulduğu sistemlerde yoğunluk yaşanabilmektedir. Bu teknolojik kısıtlar nedeniyle öğrencinin öğrenme faaliyetlerine ilişkin davranışları sınırlı olarak depolanabilmektedir. Örneğin öğrencinin e-alıştırma hizmetinde gerçekleştirdiği faaliyetlerin ayrıntısı bilenememektedir. Diğer yandan web sunucu günceleri incelendiğinde en temel sorun günce hareketini hangi öğrencinin gerçekleştirdiğinin bilinmemesidir. Web sunucu güncelerinde öğrenci kimliğinin kaydedilmesi, öğrenci oturumunu başlatan web uygulamasının yetkilendirme sonrası ilgili web sunucu değişkenlerine öğrenci kimlik numarasının atanması ile mümkündür. Ancak mevcut e-öğrenme sisteminde bu ilişki kurulmamıştır.

Araştırmada BDE birimi e-öğrenme sistemi web uygulaması tarafından oluşturulan izleme tabloları veri madenciliği veritabanına aktarılmıştır. Şekil 27'de gösterilen tablolarda öğrencilerin sisteme girişi ve aldığı hizmete ilişkin veriler yer almaktadır. Bu veri tabloları öğrencinin herhangi bir hizmeti talep ettiğinde tıklama sonrası oluşturulan verileri içerir. Ders izleme, kitap izleme,

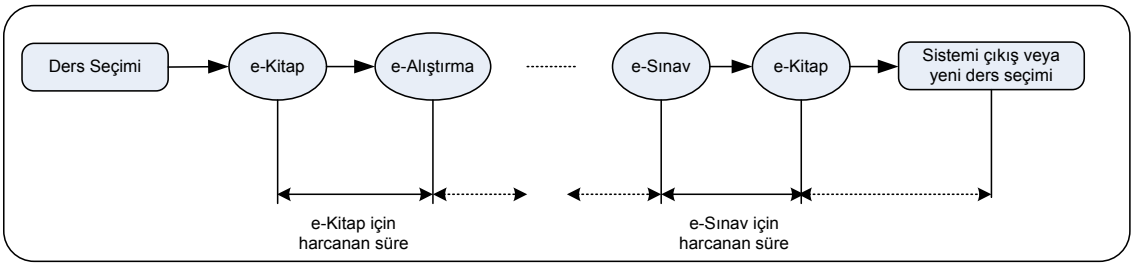
sesli kitap izleme, alıştırma yazılımı izleme, TV izleme ve sınav izleme tablolarında depolanan öğrenci izleme verileri hizmete göre farklı yapıda olabilmektedir. Ancak her tabloda öğrenci TC kimlik numarası, tarih ve saat verileri yer almaktadır. Veri madenciliği veritabanında öncelikle web uygulaması tarafından metin formatında kaydı oluşturulmuş tarih ve saat alanları SQL server “datetime” formatına dönüştürülmüş ve tüm hizmetlere ait günceler tek tabloda birleştirilmiştir. İkinci aşamada ise diğer veri kaynakları ile eşleşmenin sağlanabilmesi amacıyla mevcut TC kimlik numaralarının BAUM veritabanlarındaki karşılığı bulunarak öğrenci numaraları bu tabloya eklenmiştir. Veri içersinde öğrencilere ait olmayan hareketler tablolardan temizlenmiştir.



Şekil 27. e-Öğrenme Portalı Öğrenci Hizmet İzleme Tabloları.

Öğrencilerin hangi hizmetten ne kadar süre faydalandıklarını belirlemek amacıyla birleştirilmiş öğrenci hizmet tabloları üzerinde SQL sorguları çalıştırılmıştır. Öğrencilerin sistem içersindeki hareketleri Şekil 28’de gösterildiği gibi temsil edilebilir. Öğrenci, TC kimlik numarasını girdikten sonra e-öğrenme sayfasına alınır ve hizmet almak istediği dersi seçtiği anda bir izleme kaydı oluşturulur. Bu çalışmada ders oturumu olarak adlandırılan bu öğrenci hareketlerinin sayıları EK 3’de verilmiştir. Her hizmet seçiminde ilgili hizmet izleme tablosuna gerekli veriler kaydedilir. Birleştirilmiş öğrenci hizmet

tablosunda öğrencinin bir sonraki hareket ve zamanına bakılarak öğrencinin ilgili hizmette harcadığı süre hesaplanabilir. Ancak son aldıkları hizmetin süresini hesaplamak mümkün olmamaktadır. Bunun nedeni öğrencinin sistemi terk ettiği anda herhangi bir kaydın alınmamasıdır. Öğrencinin ilgili ders oturumunda aldığı son hizmetin süresi eksik veri olarak nitelendirilebilir. Alınan son hizmetin süresine ilişkin eksik veri, öğrencinin süresi belli olan aynı derste aynı hizmette harcadığı sürelerin ortalaması ile tamamlanmıştır. Ayrıca ders seçimi yaptıktan sonra hiçbir hizmet almadan sistemi terk etmiş ya da yeni bir ders seçmiş öğrenci hareketleri geçersiz olarak işaretlenmiştir.



Şekil 28. Öğrencilerin e-Öğrenme Sistemi İçersindeki Hareketleri Ve Süre Hesaplaması.

Öğrencilerin hizmetlerden faydalandıkları sürelerin hesaplanmasında sayfa içi gezinme hareketlerinin temizlenmesi de oldukça önemlidir. Bu nedenle öğrencilerin e-hizmet sayfasında bir dakikadan daha az süre harcadıkları hareketler hesaba katılmamıştır. Örneğin öğrencinin ilgili bir konuda bilgi arama aşamasında birçok ünite veya hizmet arasında dolaşması bu tür hareket kayıtlarına neden olmaktadır.

Veri madenciliği veritabanında yer alan e-öğrenme öğrenci hareketleri sistemin açıldığı mayıs 2005 tarihinden 2006 yılı bütünleme sınavının yapıldığı tarih aralığındaki verilerden oluşmaktadır. Öğrencinin faydalandığı hizmetlerin yer aldığı tabloda her öğrenci hareketi ilgili sınav yılı ve sınav dönemi alanları ile etiketlenmiştir. Böylece her öğrencinin ilgili sınav döneminde ilgili dersin hangi hizmetten ne kadar süreyle faydalandığı özetlenebilmektedir. Bu sayede öğrencinin e-öğrenme hizmetlerinde harcadıkları süreler ile sınav notları ve başarı notları ilişkilendirilebilmektedir.

Veri madenciliği veritabanında e-öğrenmeye ilişkin veriler cevabı aranan soruya dayalı olarak çeşitlendirilebilir. Çalışmada öğrencilerin deneme sınavlarında gösterdikleri performansları özetleyen bir özet tablo elde edilmiştir.

e-Öğrenme faaliyetlerine ilişkin akla gelen bir diğer soru ise öğrencilerin e-öğrenme sistemine girdikten sonra gerçekten öğrenme faaliyetinde bulunup bulunmadıklarıdır. Bunun aksi bir durum ise öğrencinin ilgili hizmette harcadığı sürenin kısa olmasına rağmen daha sonra çalışılmak üzere e-öğrenme hizmet ekranının çıktısının alınması veya bilgisayara kaydedilmesi olasılığıdır. Bu iki durum, öğrencinin sistemden faydalanma süresinin hatalı hesaplanmasına neden olmaktadır. Gürültülü veri olarak değerlendirilmesi mümkün olmayan bu hataların belirlenmesi veya temizlenmesi mümkün değildir. Ancak bazı durumlarda örneğin aynı öğrenci numarası ile farklı derslere defalarca girilmesi gibi aykırı veriler analiz dışında tutulabilir. Çok kişinin internet bağlantısı kurduğu internet kafe gibi ortamlarda oturumların kapanmaması ya da öğrenci numaralarının hatırlanması nedeniyle aynı öğrenci kimliği ile birçok kişi faydalanabilmektedir. Bu nedenle oturum sayısı uç noktalarda olan veriler ya da öğrencinin sorumlu olmadığı bir derse ilişkin açtığı ders oturumları, gerçeği ifade etmemesi nedeniyle analiz dışı bırakılmıştır.

3.1.3. Anket Araştırmaları Ve ÖSYM Verileri

Çeşitli araştırmalarda Uzaktan Eğitim Sistemi öğrencileri üzerinde birçok anket çalışması uygulanmıştır. Bu anketlerin bir kısmı posta ile öğrenciye ulaştırılmış bir kısmı ise internet üzerinden uygulanmıştır. Ancak bu çalışmaların çoğu veri madenciliği çalışması için uygun değildir. Bunun en temel nedeni anket katılımcısının kimliğinin gizlenmesidir. Dolayısıyla öğrencinin kimlik bilgileri ile ankette belirttiği görüşlerini eşleştirme olanağı yoktur. Uzaktan Eğitim Sistemi tarafından öğrenci profilini belirleme amaçlı yapılan ve kimliği belirli öğrencilerden toplanan büyük hacimli araştırma bilgileri veri madenciliği için uygun olabilmektedir.

2000-2001 öğretim yılında uzaktan eğitim öğrencilerinin “TC kimlik” numaralarını tebliğ etmesi sırasında optik forma kodlanan kısa bir anket gerçekleştirilmiştir. Bu ankette öğrenci profili, Uzaktan Eğitim Sistemi hizmetlerine ilişkin öğrenci tercihleri ve medya sahipliğini belirlemeyi amaçlayan sorular hazırlanarak uygulanmıştır. Ankete 396.394 öğrencinin katılımı gerçekleşmiş ve anket sonuçları BAUM’da optik formlardan manyetik ortama aktarılmıştır. Anket sonuçları çalışmada oluşturulan veri madenciliği veritabanına aktarılarak diğer verilerle eşleştirilmiştir.

Uzaktan Eğitim Sisteminde yer alan öğrencilere ilişkin önemli bir veri kaynağı da sisteme ÖSYM tarafından yerleştirilen öğrencilere ilişkin ÖSYM veritabanlarıdır. ÖSYM tarafından her öğretim yılı başlangıcı Anadolu Üniversitesine gönderilen veriler sistemde biriktirilmediğinden üniversite bünyesinde bu verilere ulaşmak mümkün olamamaktadır. Belirli yıllara ait bir kısım ÖSYM verisini veritabanına aktarılmasına rağmen mevcut analizlerin veri kümesinde oldukça fazla eksik veriye neden olmuştur. ÖSYM verilerinin içeriğinde öğrencinin üniversite öncesi eğitim deneyimlerini gösteren bilgilerin olması bu kaynağın önemini arttırmaktadır.

3.2. Uzaktan Eğitim Sistemi Veri Madenciliği Modelleri

3.2.1. Uzaktan Eğitim Sistemi Veri Madenciliği Veritabanının Yapısı ve Özellikleri

Çalışmada oluşturulan veritabanı verileri başlıca üç veri kaynağından farklı formatlardaki verilerin birleştirilmesiyle yaratılmıştır. Tablolar ve özellikler uygun veri biçimlerine dönüştürüldükten sonra veri madenciliği analizi için gerekli olabilecek yeni özellik ve tablolar oluşturulmuştur.

Ek 4’de önemli tablolara ait görünümleri verilen veritabanında ayrıca birçok “view” olarak adlandırılan kaydedilmiş sorgular ve “Transactional SQL” sorgu dilinde yazılmış prosedürler yer almaktadır. Öğrencilerin e-öğrenme hizmetlerde harcadıkları sürelerin hesaplanması amacıyla yazılmış bir

prosedürün kodları Şekil 29'da verilmiştir. Veritabanında yer alan tablo ve özellikleri aşağıda sıralanmıştır.

- *Kimlik tablosu:* TC kimlik numarası, ögrenci numarası, ad, soyad, fakülte, bolüm, sınıf, kayıt yılı, toplam yılı, tarih (mezuniyet, kayıt silme), yıl (mezuniyet), durum (aktif, aktif-pasif, mezun, kaydı silindi), cinsiyet, yabancı dil, öğrencilik durumu, fakülteye giriş biçmi, ayrılış nedeni, özür durumu, askerlik durumu, adres ili, nüfus ili, doğum tarihi ve yaş özelliklerinden oluşan ve BAUM'dan sağlanan veritabanı tablosudur.

```
ALTER PROCEDURE [dbo].[SureleriHesapla] AS
BEGIN  Declare @Oturum_SıraID bigint Declare @DOturumid int
       Declare @DOSira smallint Declare @OldDOSira smallint
       Declare @TarihSaat datetime Declare @OldTarihSaat datetime
       Declare c Cursor For
        SELECT Oturum_SıraID, DOturumid, DOSira, TarihSaat
        FROM DM_DersOturumAyrıntı       ORDER BY Oturum_SıraID DESC
       OPEN c
       FETCH NEXT FROM c
       INTO @Oturum_SıraID , @DOturumid , @DOSira , @TarihSaat
       Set @OldDOSira=0
       WHILE @@FETCH_STATUS = 0
        BEGIN
         If @OldDOSira<>0 and @DOSira<>0
                Begin  Update DM_DersOturumAyrıntı
                     Set BitisTarihi=@OldTarihSaat,
                     Sure_sn=datediff(ss,@TarihSaat,@OldTarihSaat)
                     where Oturum_SıraID=@Oturum_SıraID end
          Set @OldDOSira=@DOSira          Set @OldTarihSaat=@TarihSaat
          FETCH NEXT FROM c
          INTO @Oturum_SıraID , @DOturumid , @DOSira , @TarihSaat
        END
END
```

Şekil 29. e-öğrenme Hizmet Sürelerinin Hesaplanması İçin Oluşturulan Prosedür.

- *Benzersiz kimlik tablosu:* Bir öğrencinin farklı zamanlarda birden fazla kaydının bulunması ve kimlik tablosunda tekrarlanması kimlik bilgilerinin diğer tablolarla ilişkilendirilmesinde soruna yol açmaktadır. Bu nedenle öğrenci kayıtlarını benzersiz hale getirmek için "ÖğrenciNo", "Bolumkodu", "Yıl" ve "Kayıtsayısı" özelliklerinin yer aldığı bir tablo yaratılmıştır. e-öğrenme verileri ve anket verilerinin kimlik bilgileri ile ilişkilendirilmesinde bu tablo kullanılmıştır.

- *Öğrenci not tabloları:* BAUM veritabanlarında lisans ve önlisans bölümleri için iki ayrı tabloda saklanan öğrenci not bilgileri tek bir tabloda birleştirilerek öğrenci genel not ortalamaları hesaplanmıştır. Öğrencinin başarılı derslerine ilişkin genel not ortalaması "GPA", tüm derslerine ilişkin genel not ortalaması ise "GPA_All" adında özelliklerde depolanmıştır. Öğrencinin çeşitli nedenlerle değerlendirilmeyen sınavları hesaplama dışı bırakılmıştır.
- *e-öğrenme kullanım veri tabloları:* Anadolu Üniversitesi Bilgisayar Destekli Eğitim biriminden sağlanan öğrenci web kullanım günceleri öncelikle gerekli dönüşüm ve eşleştirmeler yapılarak tek tablo haline getirilmiştir. "DM_Tum_Session_3" adlı tablo kullanılarak web kullanım bilgileri ders oturum verilerine dönüştürülmüş ve her oturum ve hizmete ilişkin harcanan süreler hesaplanmıştır. Ayrıca her oturumun ait olduğu sınav dönemleri tarih bilgisi kullanılarak belirlenmiştir. Öğrencilerin her oturumda almış oldukları hizmet sayıları ve süreleri de sınav dönemlerine ve öğrenim yıllarına göre özet tablolar haline getirilmiştir. Ayrıca öğrencilerin deneme sınavlarındaki performanslarını yansıtan ve "yanıtlanan soru sayısı", "doğru yanıtlanan soru sayısı" ve "test soru sayısı" özellikleri kullanılarak yeni özellikler türetilmiştir.
- *Anket veri tablosu:* 2000-2001 öğrenim yılında düzenlenmiş anket verilerinin yer aldığı "AOF_Anket_2000" tablosu kimlik bilgileri ile ilişkilendirilmiştir. Tabloda öğrencinin medeni durumu, bilgisayar kullanım bilgisi ve mekanı, çalışma durumu, mesleği, internet kullanım durumu ve mekanı, gelir düzeyi, medya sahiplik bilgileri ve faydalandıkları uzaktan eğitim hizmetleri bilgileri yer almaktadır.

MS SQL Server 2005 veritabanı yönetim yazılımında oluşturulan veritabanı, çalışmada veri madenciliği yoluyla cevabı aranan soruların yanıtlanmasında kullanılmıştır. Analiz sırasında tekrarlı olarak veritabanına dönülerek ek özellik türetilmesi, eksik değerler sorunlarının giderilmesi ve aykırı değerlerin veri setinden çıkarılması işlemleri gerçekleştirilmiştir.

3.2.2. Öğrenci Performans Tahmin Modelleri

Araştırmada odaklanılan ilk problem öğrenci özellikleri ve e-öğrenme faaliyetlerine ilişkin veriler kullanılarak öğrencinin performansının tahmini yapabilecek bir modelin oluşturulmasıdır. Öğrencinin başarısını tahmin edilmesinde veri madenciliği sınıflama algoritmalarından faydalanılabilmektedir. Bu amaçla ilgili veri bir araya getirilerek veri seti oluşturulmuştur. Oluşturulan veri seti MS SQL veritabanından MS Access veritabanına aktarılarak modelleme ve veri analizinin gerçekleştirildiği SPSS Clementine 9.0 veri madenciliği yazılımına bağlanmıştır. Tahmin modellerini geliştirmede kullanılan veri setinin özellikleri aşağıda tanımlanmıştır.

- Analizde 2004-2005 ve 2005-2006 öğrenim yıllarına ait öğrenci not, kimlik ve e-öğrenme faaliyetlerinden elde edilmiş veriler kullanılmıştır. ÖSYM'den elde edilen verilerin az sayıda olması nedeniyle analize dahil edilmemiştir.
- Herhangi bir nedenle arasınavı veya yılsonu ve bütünleme sınavı değerlendirilmemiş öğrenciler analiz verisinden çıkarılmıştır. Sınava girmemiş öğrenciye ait yapılacak bir başarı tahmini geçeğe uygun olmayacaktır.
- e-öğrenme hizmetlerinde bir dakikadan daha az süre harcanmış öğrenci hareketleri ve öğrencinin sorumlu olmadığı derse veya derslere ilişkin e-öğrenme faaliyetleri filtre edilerek veri setine dahil edilmemiştir.
- e-öğrenme ders oturumlarının son faaliyetlerinde harcanan sürelere ilişkin eksik değerler, öğrencinin aynı dersine ilişkin aynı hizmette harcadığı sürenin ortalaması ile tamamlanmıştır.
- e-öğrenme sürelerine ilişkin aykırı değerler analiz dışı bırakılmıştır. Mevcut veri kümesinde e-öğrenme hizmetlerinde toplam 250 bin saniyeden daha fazla süre harcanmış 13.333 satır veri ve çeşitli nedenlerle ilgili hizmette 1 dakikadan daha az süre harcanmış 234.304 satır (öğrenci-ders hizmet faaliyeti) veri analiz dışı bırakılmıştır.

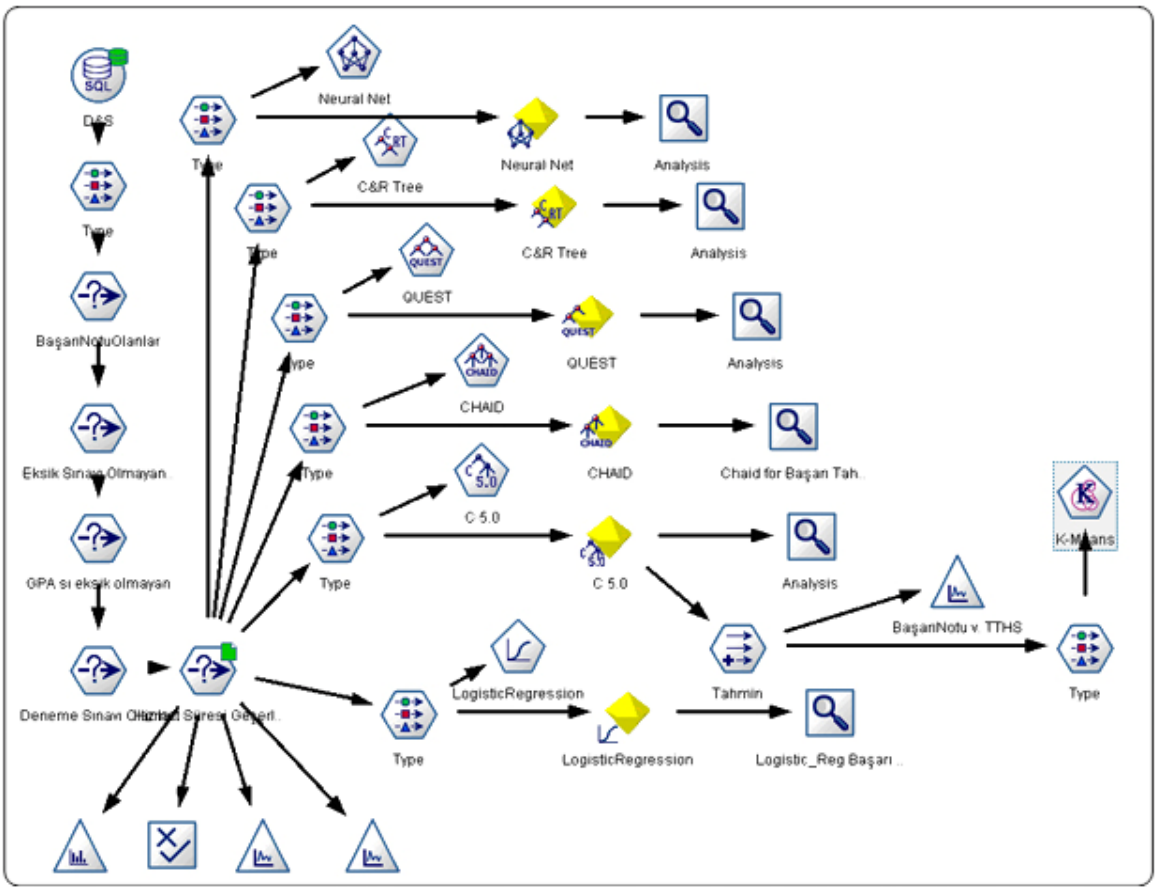
- Deneme sınavları oturumunda hiçbir soru yanıtlamamış olan öğrencilere ilişkin veriler analiz dışı tutulmuştur.
- Analizde kullanılan veri seti 180.554 adet öğrencinin 129 adet dersine ilişkin 429.757 satır kayıttan oluşmaktadır.

Şekil 30'da öğrenci performans tahmini için oluşturulan Clementine analiz görünümü yer almaktadır. Sırasıyla girdi değişkenleri dersin adı, e-hizmet faydalanma süreleri (sırasıyla, e-Kitap, e-Sesli kitap, e-Alıştırma, e-TV ve e-Sınav), öğrencinin dersi kaçıncı kez aldığı, değerlendirilen sınavlarının ortalaması, öğrenci yaşı, deneme sınavlarında doğru cevapladığı soru sayısı ve yanıt verdiği soruların doğruluk oranıdır. Modelde kullanılan veri alanlarının özellik ve dağılım grafikleri EK 5'de verilmiştir.

Clementine veri madenciliği uygulama yazılımında C5.0, Logistic Regression, Neural Net, C&RT, CHAID ve QUEST sınıflama algoritmaları kullanılarak tahmin modelleri oluşturulmuştur. Uygulanan bu algoritmaların özellikleri aşağıda tanımlanmıştır[134].

- *C5.0*: C5.0 algoritması hem karar ağacı hem de kural kümesi üreten ve veri kümesinde en büyük enformasyon kazanımını sağlamak amacıyla bölümleme yapan bir algoritmadır. Hedef alanı kategorik veri türünde olmalıdır. C5.0 her düğümü ikiden fazla alt guruba bölebilir.
- *Logistic Regression:* Nominal regresyon olarak da adlandırılan lojistik regresyon, girdi değişkenlerinin değerine dayalı olarak kayıtları sınıflamak için istatistiksel bir teknik kullanır. Bu modelleme tekniği doğrusal regresyona benzer bir yöntemdir ancak hedef değişken sayısal değil kategorik veri türündedir. Lojistik regresyon hedef değişkenin olası tüm değerleri ile girdi değişkenleri arasındaki ilişkiyi olasılıklarla ifade ederek bir denklemler kümesi oluşturmaya çalışır.

[134] SPSS Inc. **Clementine 9.0 Node Reference.** USA: 2004, s.197.



Şekil 30. Öğrenci Performans Tahmini İçin Oluşturulan Clementine Analiz Görünümü.

- *C5.0*: C5.0 algoritması hem karar ağacı hem de kural kümesi üreten ve veri kümesinde en büyük enformasyon kazanımını sağlamak amacıyla bölümleme yapan bir algoritmadır. Hedef alanı kategorik veri türünde olmalıdır. C5.0 her düğümü ikiden fazla alt guruba bölebilir.
- *Logistic Regression:* Nominal regresyon olarak da adlandırılan lojistik regresyon, girdi değişkenlerinin değerine dayalı olarak kayıtları sınıflamak için istatistiksel bir teknik kullanır. Bu modelleme tekniği doğrusal regresyona benzer bir yöntemdir ancak hedef değişken sayısal değil kategorik veri türündedir. Lojistik regresyon hedef değişkenin olası tüm değerleri ile girdi değişkenleri arasındaki ilişkiyi olasılıklarla ifade ederek bir denklemler kümesi oluşturmaya çalışır.
- *Neural Net:* Yapay sinir ağları olarak adlandırılan hesaplamaya dayalı modelleme tekniğidir. Modelin oluşturulmasında en önemli noktalardan

biri ağ yapısının belirlenmesidir. Neural Net modellerinde girdi ve hedef değişkenler sayısal, sembolik veya ikili veri türünde olabilir.

- *C&RT:* Sınıflama ve regresyon ağacı algoritması katışıklık ölçümünü (impurity measure) enazlamaya dayanan Breiman ve diğerleri tarafından 1984 yılında geliştirilmiş bir karar ağacı algoritmasıdır. Hedef ve girdi alanları aralık veya kategorik veri türünde olabilir. Her düğüm sadece iki alt gruba ayrılabilir.
- *CHAID:* CHAID algoritması eniyi bölümlemeyi yapabilmek için "ki-kare" istatistiğinden yararlanan bir karar ağacı modelidir. Bu algoritma mümkün olan tüm durumları hesapladığından çalıştırılması zaman alıcı bir algoritmadır. Hedef ve girdi değişkenleri kategorik veya aralık değerine sahip olabilir. Her düğüm iki veya daha fazla alt gruba bölünebilir.
- *QUEST:* Adını hızlı, önyargısız, etkin istatistiksel ağaç yönteminin ilk harflerinin kısaltılmasından alan QUEST algoritması 1997 yılında Loh ve Shih tarafından tanımlanmıştır. Girdi değişkenleri sayısal, hedef değişkeni ise kategorik veri türü olmalıdır. Her düğüm ikili olarak bölünebilir.

3.2.3. Uzaktan Eğitim Sisteminden Mezun Olan Öğrenci Verileriyle Gerçekleştirilen Kümeleme Analizi

Anadolu Üniversitesi Uzaktan Eğitim Sisteminden mezun olan öğrencilerin özelliklerine göre gruplandırılması sisteme yeni gelen öğrencilerin özelliklerine bakılarak mezuniyet süresinin tahmin edilmesinde kullanılabilir. Bu tahmin için öncelikle sisteme kayıt olan öğrencinin profil bilgilerinin ve geçmiş öğrenme performans bilgilerinin sistemde toplanması gerekmektedir. Çalışmada Uzaktan Eğitim Sisteminde 2000-2001 öğretim yılında kayıt esnasında gerçekleştirilen büyük ölçekli bir anket ve kimlik verileri kullanılarak veri madenciliği çalışması yapılmıştır. Ankette sorulan soruların bir kısmı bugünkü geçerliliğini yitirmiştir. Örneğin bilgisayar kullanımı, medya sahipliği ve talep edilen Uzaktan Eğitim Sistemi hizmeti gibi öğrenci özellikleri son yıllarda teknolojinin gelişimine bağlı olarak farklılaşabilmektedir. Diğer taraftan

mezunlara ilişkin gruplama yapıldığından analizde değerlendirilecek verilerin, öğrencinin öğrencilik dönemini yansıtması gerektiğinden 2000-2001 öğretim yılında gerçekleştirilen anket çalışmasını kullanmak yerinde olacaktır.

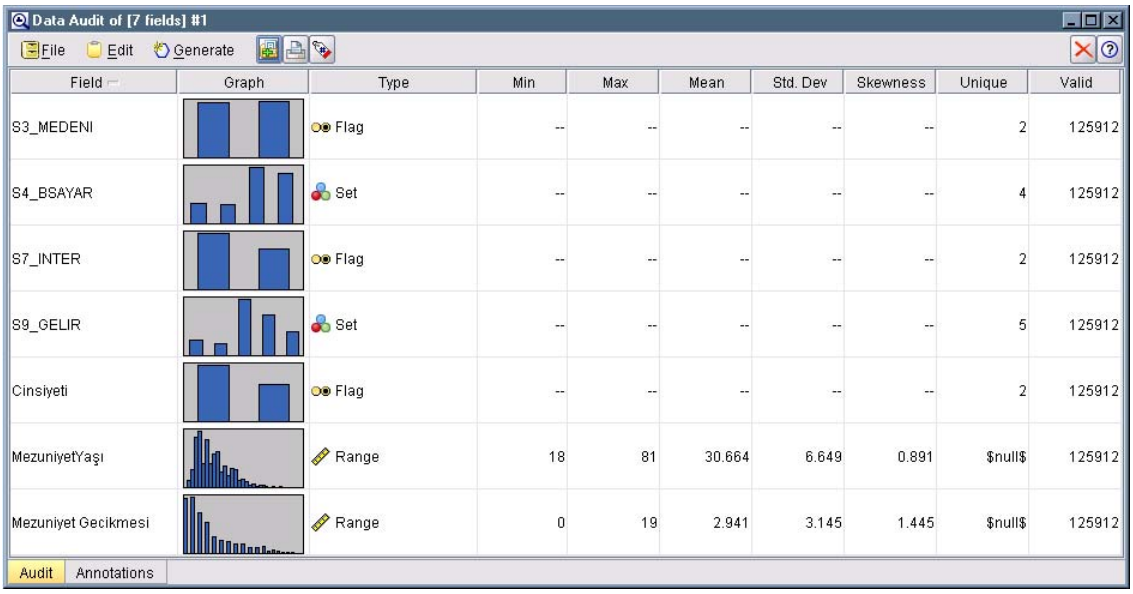
Analizde kullanılan verinin özellikleri aşağıda tanımlanmıştır.

- 2000-2001 yılında 396.394 öğrenciye uygulanan anket ve kimlik verileri birleştirilmiştir.
- Analizde kullanılan alanlara ilişkin anket sorularına cevap vermeyen öğrencilerin eksik verilerinin tamamlanma olanağı bulunmadığından bu öğrencilere ait veriler veri kümesinden çıkarılmıştır.
- Öğrencilerin sistemde geçirdiği süre ve mezun olmaları gereken süre göz önüne alınarak mezuniyet gecikmesi alanı türetilmiştir. Bu hesaplamada İşletme ve İktisat Fakültesinden ön lisans diploması alan öğrenciler ve Açıköğretim Fakültesi lisans bölümleri mezunları veri kümesi dışında bırakılmıştır.
- Öğrencilerin mezun oldukları yıl temel alınarak mezuniyet yaşları hesaplanmıştır.
- Kümeleme analizinde girdi olarak kullanılan 125.912 kayıta ait alanlar, sırasıyla öğrencinin medeni durumu (evli-bekar), bilgisayar kullanımı ve mekanı (Evet/Evde, Evet/Ev ve işte, Evet/İşte, Hayır), internet kullanımı (Evet, Hayır), cinsiyet, mezuniyet yaşı (18-81), mezuniyet gecikmesi (0-19) alanlarından oluşmaktadır. Verilere ilişkin özet Tablo 5'te verilmiştir. Kümeleme analizi için uygun olmayan alanlar veri setinde yer almamaktadır. Örneğin öğrencinin çalışma durumu kümeleme analizi dışında bırakılmıştır. Mevcut veri kümesinde mezun öğrencilerin %79'unun çalışıyor olması nedeniyle yapılacak kümeleme analizinde anlamlı bir ayrım yapılamayacaktır.

Kümeleme modelleri önceki bölümlerde bahsedildiği gibi denetimsiz öğrenme modelleridir. Sonucu bilinen verilerden hareketle hedef değişkeni tahmin etmeye çalışan denetimli öğrenmenin aksine kümeleme modelleri için

doğru ya da yanlış yargısı kullanmak mümkün değildir. Kümeleme yöntemleri kayıtlar arasındaki uzaklık ölçümlerine dayanarak aynı kümede yer alan kayıtlar arasındaki uzaklığı enazlamaya çalışmaktadırlar.

Tablo 5. Kümele Analizi Girdi Verilerine İlişkin Veri Özet Tablosu



Field	Graph	Type	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
S3_MEDENI		Flag	--	--	--	--	--	2	125912
S4_BSAYAR		Set	--	--	--	--	--	4	125912
S7_INTER		Flag	--	--	--	--	--	2	125912
S9_GELIR		Set	--	--	--	--	--	5	125912
Cinsiyeti		Flag	--	--	--	--	--	2	125912
MezuniyetYaşı		Range	18	81	30.664	6.649	0.891	$null$	125912
Mezuniyet Gecikmesi		Range	0	19	2.941	3.145	1.445	$null$	125912

Çalışmada mezun öğrencilere ait veri kümesi yaygın olarak kullanılan “K-means” kümeleme algoritması kullanılarak gruplara ayrılmıştır. “K-Means” kümeleme algoritması veri özelliklerine bakarak küme merkezlerini belirlemekle işe başlar. Daha sonra her kaydı girdi değişkenlerinin özelliklerine bakarak bir kümeye atamaya başlar. Tüm kayıtlar kümelere atandıktan sonra küme merkezlerini, atanan kayıtların merkezlere olan uzaklıklarının ortalamasına göre güncelleyerek kayıtları yeni merkezlere tekrar atar. Bu işlem enyüksek yineleme işlemine ulaşıncaya kadar veya kümelerin uzaklık ölçümlerine bakılarak durdurulur. “K-means” algoritmasının özellikleri aşağıda sıralanmıştır[135].

- Algoritma uzaklık hesaplamalarında öklid uzaklığını kullanır.
- Nümerik değerleri 0-1 arasına ölçekledikten sonra uzaklık hesaplamalarını yapar.
- Kategorik veriler matris halinde sayılara dönüştürülerek mesafe ölçümleri gerçekleştirilir.

[135] SPSS Inc. **Clementine 9.0 Node Reference.** USA: 2004, s.299.

- Kayıtlar küme merkezine olan öklid uzaklığı hesaplanarak en yakın olan kümeye atanır.
- Algoritmada başlangıç parametresi olarak küme sayısı, en büyük yineleme sayısı, tolerans oranı ve kategorik değerlere sahip verilerin sayıya dönüştürülmesinde kullanılan katsayıyı tanımlamak mümkündür.

4. MODELLERİN DEĞERLENDİRİLMESİ VE YORUMLANMASI

Bu bölümde, Anadolu Üniversitesi Uzaktan Eğitim Sistemi veritabanları kullanılarak gerçekleştirilen kümeleme analizi ve geliştirilen tahmin modellerinin değerlendirilmesi yapılacaktır.

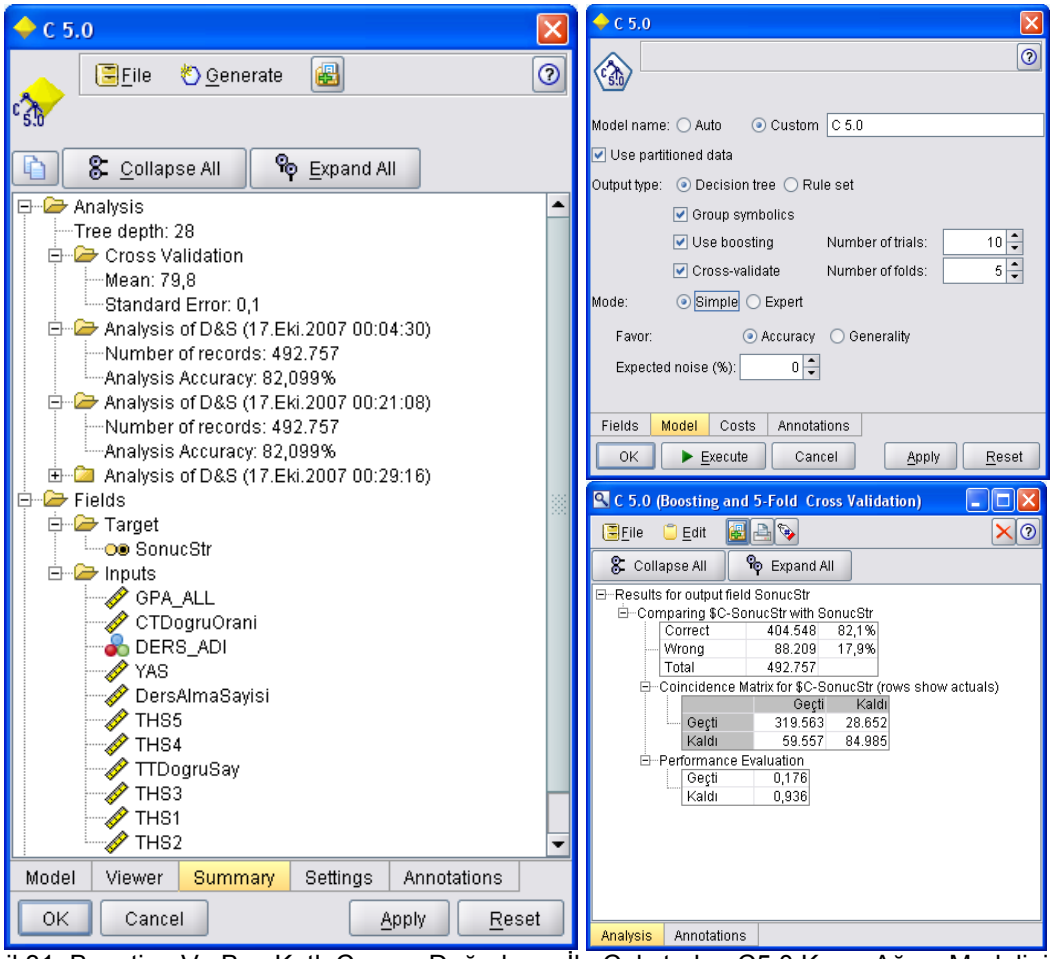
4.1. Öğrenci Performansı Tahmin Modellerinin Değerlendirilmesi

Öğrencilerin bir derse ilişkin başarı durumunu tahmin etmeyi amaçlayan tahmin modellerinin oluşturulmasından önce modelde yer alacak girdi değişkenlerinin belirlenmesi gerekir. Veri madenciliği veritabanında öğrenciye ve ilgili derse ilişkin birçok özellik mevcuttur. Öğrencinin cinsiyeti, yaşı, daha önceki dönemlerde elde ettiği başarı notu ortalaması, deneme sınavlarında elde ettiği skorlar, e-öğrenme hizmetlerinde harcadığı süreler gibi öğrencinin başarısını etkileyecek özelliklerin yanında öğrencinin adı, soyadı, yaşadığı il, bölümü, özür durumu, askerlik durumu gibi onlarca özellik mevcuttur. Öğrencinin başarı durumu ile ilişkili olabilecek değişkenlerin seçilmesi için SPSS Clementine yazılımında “Özellik Seçme” (Feature Selection) fonksiynu kullanılarak öğrencinin başarı durumu ile ilgili olan 11 özellik modellerin girdi değişkeni olarak belirlenmiştir. Bu değişkenler “Dersin adı”, “e-hizmet faydalanma süreleri” (e-Kitap, e-Sesli kitap, e-Alıştırma, e-TV ve e-Sınav), “öğrencinin dersi kaçıncı kez aldığı”, “değerlendirilen sınavlarının ortalaması”, “öğrenci yaşı”, “deneme sınavlarında doğru cevapladığı soru sayısı” ve “yanıt verdiği soruların doğruluk oranı” özellikleridir. Modellerde mevcut verinin %50’si eğitim %50’si test verisi olarak kullanılmış ve modellerin geçerliliği test

edilmiştir. Tahmin modellerinin oluşturulmasında SPSS Clementine yazılımında tanımlanan parametre ve yöntemler aşağıda tanımlanmıştır.

- ***C5.0:*** C5.0 algoritması hem karar ağacı ve kural kümesi (ruleset) oluşturmak için geliştirilmiş bir algoritmadır. C4.0 ve C5.0 algoritmaları ID3 algoritması geliştirilerek oluşturulmuştur. C5.0 algoritması ikili bölme yerine çoklu bölmeleme uygulayan bir algoritmadır. C5.0 algoritmasının bölmeleme kriteri entropi kazanımı ve kazanç oranıdır. Çalışmada C5.0 algoritması, doğruluk oranının arttırılması için "Boosting" olarak adlandırılan özel bir teknikle çalıştırılmıştır. Bu teknik sırasıyla birden çok model üretir. İlk aşamada standart C5.0 algoritması kullanılarak model üretilir. İkinci aşamada ise ilk modelin sınıflayamadığı veriler üzerine odaklanılır. Son aşamada ise ikinci modelin hatalarının giderildiği üçüncü bir model oluşturulur. Sonuçta üç model bir araya getirilerek son model oluşturulur. Modelin geliştirilmesinde ayrıca beş katlı çapraz geçerlilik tekniği kullanılmıştır. Çapraz geçerlilik tekniği sınıflama algoritmalarında karşılaşılabilen "overfitting" sorunlarının belirlenmesinde kullanılabilmektedir. Modelin parametre ve sonuçları Şekil 31'de verilmiştir. Şekilde ilk ekran görüntüsü modelin oluşturulması aşamasında elde edilen değerleri görüntülemektedir. Modelin oluşturulması aşamasında veri beş parçaya bölünerek her bir parça üzerinde model türetilip diğer kümelerde test edilmiştir. Bu aşamada çapraz geçerlilik yöntemi ile belirlenen doğruluk oranı ortalama %79,8 ve standart sapması %0,1 olarak hesaplanmıştır. Model üretildikten sonra tüm veri üzerinde yapılan testte model %82,1 doğruluk oranı sergilemiştir. Modelin karar ağacının derinliği 28 olarak gerçekleşmiştir. Oluşturulan modelin tamamının çalışmada yer alması mümkün olmadığından modelin ilk kuralına ilişkin ayrıntı EK 7'de verilmiştir. Mevcut veri setine uygulanan C5.0 algoritmasının parametreleri aşağıda verilmiştir.
  - Çıktı türü: Karar ağacı
  - Sembolleri gruplama: Evet

- Boosting kullanımı: Evet (deneme sayısı: 10)
- Çapraz doğrulama: Evet (beş katlı çapraz doğrulama)
- Mod: Basit



Şekil 31. Boosting Ve Beş Katlı Çapraz Doğrulama İle Çalıştırılan C5.0 Karar Ağacı Modelinin Parametre Ve Geçerlilik Analiz Sonucu.

- ***Logistic Regression:*** Nominal regresyon olarak da adlandırılan bu istatistiksel sınıflama tekniği doğrusal regresyon tekniğine benzer ancak hedef alan olarak sayısal değil sembolik bir alanı tahmin etmede uygulanır. Modelin girdi alanları sayısal ve sembolik veri türünde olabilmektedir. Çalışmada uygulanan lojistik regresyon modelinde "stepwise" yöntemi kullanılmıştır. Bu yöntem en basit regresyon modelinden başlayarak girdi alanlarını modele ekleyerek tahmin modelini sınar. Clementine yazılımında 11 adet girdi alanı ile öğrenci ders

başarısını tahmin etmeye yönelik oluşturulan model %78,53 doğruluk oranını elde etmiştir. Modelde “e-Sesli kitap hizmet alma süresi” alanı hariç tüm değişkenler modelde yer almıştır.

- ***Neural Net***: Yapay sinir ağı modelinin oluşturulmasında Clementine yazılımında “Neural Net” düğümü kullanılmıştır. “Neural Net” düğümü doğru ağ topolojisinin oluşturulabilmesi için “Quick”, “Dynamic”, “Multiple”, “Prune”, “RBFN” ve “Exhaustive Prune” model türlerini sunmaktadır. Modelin oluşturulmasında “Prune” modeli kullanılmıştır. “Prune” modeli oldukça yavaş çalışmasına rağmen diğer modellere göre daha iyi sonuç vermektedir. “Prune” yöntemi eğitim sürecine oldukça büyük bir ağ yapısıyla başlar ve süreç içerisinde giriş ve gizli katmandaki en zayıf düğümleri belirleyerek eler. “Prune” yöntemi kullanılarak uygulanan “neural net” modelleme tekniği en iyi ağ yapısını 11 girdi düğümü, 26 düğümün bulunduğu bir gizli katman ve 1 çıktı düğümü olarak belirlemiştir. Mevcut algoritmada birden fazla gizli katman tanımlamak mümkündür. Model oluşturma aşamasında 2 ve 3 gizli katman denemeleri %77.80 geçerlilik oranından daha yüksek bir performans gösterememiştir. Prune yapay sinir ağı modelleme tekniğinin varsayılan parametreleri aşağıda verilmiştir.
  - Gizli katman sayısı:1
  - Gizli katmandaki başlangıç düğüm sayısı: Veri setindeki girdi düğüm sayısı k_i, çıktı düğüm sayısı k_0 ve eğitim verisindeki kayıt sayısı n_r olmak üzere enküçük(50,yuvarla(log(n_r) log(k_i+k_0))) ile hesaplanır.
  - Alpha: 0,9
  - Initial Eta:0,4
  - High Eta: 0,15
  - Low Eta: 0,01
  - Persistence: 100
  - Overall persistence: 4

- Hidden persistance: k_h gizli katmandaki düğüm sayısı olmak üzere min(10,max(1,k_i+k_h/10)) olarak hesaplanır.
- Hiddern Rate: 0,15
- Input Persistance: min(10, max(2, k_i-k_0/5))
- Input rate: 0,15

- ***CHAID***: CHAID algoritması en iyi bölmeleri tanımlamak için ki-kare istatistiğini kullanarak karar ağaçlarını oluşturmada kullanılan bir sınıflama yöntemidir. Algoritma karar ağacının ilk dalını oluşturmak için ilk tahmin edici özelliği seçer ve alt daldaki her düğüm seçilen değişkenin homojen değerler grubundan oluşturulur. Bu yineleme ağacın tamamı geliştirilinceye kadar devam eder. Homojen değerler gruplarının oluşturulmasında kullanılan istatistik testi, eğer hedef alan sürekli ise F testi, hedef alan kategorik ise ki-kare testidir. Mevcut veri kümesinde %77,65 geçerlilik oranı sağlayan CHAID algoritmasının uygulama parametreleri aşağıda verilmiştir.
  - En büyük ağaç derinliği: 10
  - Alfa (bölme için) : 0,05
  - Alfa (Birleştirme için) : 0,05
  - Kategorik hedef için ki-kare hesaplama metodu: Pearson
  - Durdurma kriterleri: Kök düğüm için %2, alt düğüm için %1
  - Epsilon: 0,001
  - En büyük yineleme: 100

- ***C&RT:*** C&RT algoritması en iyi bölmelemeyi elde etmek için bölme sonrası oluşacak katışıklık ölçümündeki azalmayı dikkate alır. Bu ağaç oluşturma algoritmasında her düğüm iki alt gruba bölünür. Bölme işlemi durma kriterlerinden biri sağlanıncaya kadar devam eder. Mevcut veri seti üzerinde SPSS Clementine yazılımında uygulanan C&RT algoritmasının parametreleri aşağıda tanımlanmıştır.

- En büyük ağaç derinliği: 10
- En büyük vekil sayısı (maxiumum surrogate): 0 (veri setinde eksik değer yoktur)
- En küçük katışıklık değişimi: 0,0001
- Kategorik hedef alanı için katışıklık ölçümü: Gini
- Durdurma kriterleri: Kök düğüm için %2, alt düğüm için %1

- ***QUEST:*** QUEST algoritmasında bölmeleri belirlemek için kuadratik ayırma analizi (quadratic discriminant analysis) kullanılır. QUEST algoritmasında her düğün iki alt gruba bölünür. Tahmin edici özellikler sayısal, hedef özellik kategorik veri türünde olmalıdır. Bu algoritmanın uygulanmasında "Ders adı" girdi özelliği veri türü olarak algoritmaya uygun olmadığından veri seti dışında bırakılmıştır. Algoritmanın uygulanmasında belirlenen parametreler aşağıda verilmiştir.
  - En büyük ağaç derinliği: 10
  - Alfa (bölme için) : 0,05
  - En büyük vekil sayısı (maxiumum surrogate): 0 (veri setinde eksik değer yoktur)
  - Durdurma kriterleri: Kök düğüm için %2, alt düğüm için %1

Öğrenci performans tahmini için oluşturulan modeller verinin tamamı üzerinde test edilmiştir. Clementine yazılımında "Test" düğümü kullanılarak elde edilen sonuçların ekran görünümü EK 6'da verilmiştir. Modellerin değerlendirilmesi amacıyla uygulanan test sonuçlarından elde edilen doğruluk oranları ve risk matrisleri Tablo 6'da özetlenmiştir. Risk matrislerinde satırlar gerçek durumları sütunlar ise test aşamasında tahmin edilen değerleri ifade etmektedir. C5.0 algoritması 319.563 "Geçti" ve 84.985 "Kaldı" sonucunu doğru, ancak 28.652 "Kaldı" ve 59.557 "Geçti" sonucunu hatalı tahmin ederek en yüksek geçerlilik oranını sağlamıştır.

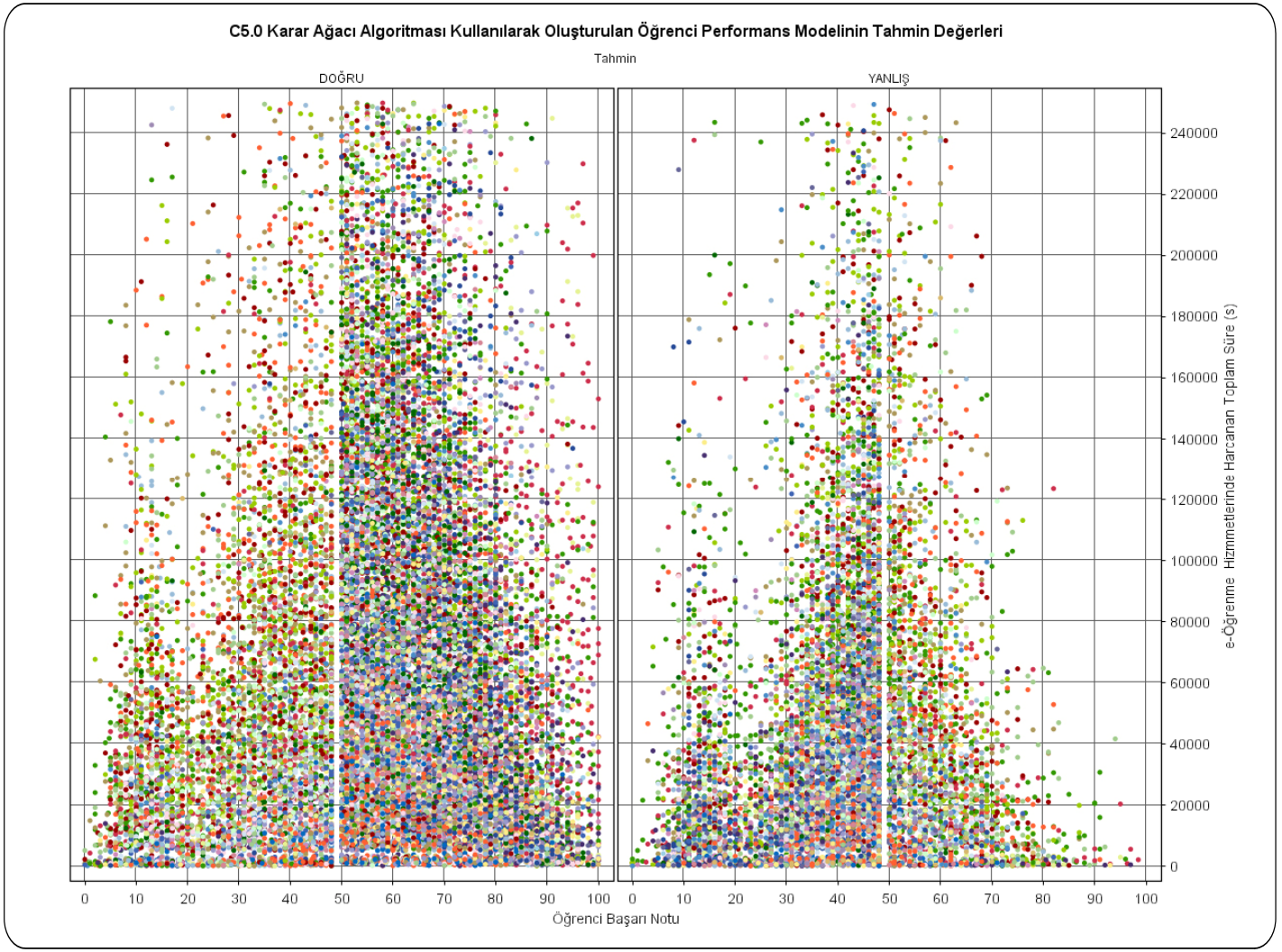
Şekil 32'de doğru ve hatalı tahmin edilmiş değerlerin, öğrencilerin başarı notu ve toplam e-öğrenme hizmet sürelerine göre saçılma grafiği verilmiştir. Bu

grafiğe göre hatalı tahmin edilmiş verilerin daha çok e-öğrenme hizmet sürelerinin nispeten daha az olduğu veriler olduğu görülmektedir. Ayrıca hatalı tahmin edilen verilerin başarı notunun 50 etrafında yoğunlaştığı görülmektedir.

Tablo 6. Öğrenci Performansı Tahmini İçin Farklı Tahmin Modelleme Teknikleri Kullanılarak Yapılan Analizlerin Doğruluk Oranları Ve Risk Matrisleri

Algoritma	Doğruluk Oranı	Risk Matrisi		
C5.0	% 82.14		***Geçti***	***Kaldı***
		Geçti	*319.563*	*28.652*
		Kaldı	*59.557*	*84.985*
Logistic Regression	% 78.53		***Geçti***	***Kaldı***
		Geçti	*314.154*	*34.061*
		Kaldı	*71.733*	*72.809*
Neural Net	%77.80		***Geçti***	***Kaldı***
		Geçti	*310.196*	*38.019*
		Kaldı	*71.376*	*73.166*
CHAID	%77.65		***Geçti***	***Kaldı***
		Geçti	*312.956*	*35.259*
		Kaldı	*74.878*	*69.664*
C&RT	%77.45		***Geçti***	***Kaldı***
		Geçti	*381.949*	*36.266*
		Kaldı	*74.860*	*69.682*
QUEST	%75.00		***Geçti***	***Kaldı***
		Geçti	*315.525*	*32.690*
		Kaldı	*90.505*	*54.037*

Modelin eğitim ve test aşamasında iki öğretim yılına ait verilerin kullanılması modelin tutarlılığını da arttırmıştır. Sadece tek öğretim yılı için oluşturulan modelin doğruluk oranının %85 üzerinde olmasına rağmen modelin diğer öğretim yılı verileriyle çalıştırılması doğruluk oranını %65 değerine kadar düşürmektedir. Bu nedenle tek bir öğretim yılı yerine birden fazla öğretim yılı verisiyle türetilen tahmin modelleri daha tutarlı olmaktadır.



Şekil 32. C5.0 Algoritması Tahmin Değerlerinin Başarı Notu Ve Toplam Hizmet Süresi Saçılma Grafiği

4.2. Uzaktan Eğitim Sisteminden Mezun Olan Öğrenci Verileriyle Gerçekleştirilen Kümeleme Analizinin Değerlendirilmesi

Çalışmada hazırlanan veri kümesi SPSS Clementine veri madenciliği uygulamasında "K-Means" algoritması uygulanarak kümelenmiştir. Elde edilen sonuçlara ilişkin ekran görünümü EK 8'de özeti ise Tablo 7'de verilmiştir. Clementine uygulamasında alanların küme için anlamı da bir katsayıyla ölçülür. Bu katsayının bire yaklaşması o alanın kümeler için farklılığını yani önemini vurgular. Bu katsayı sürekli değerler için t testi, kesikli yani kategorik değerler için ki-kare testi ile hesaplanmaktadır. Örneğin veri kümesi içersindeki öğrencinin çalışma durumunu gösteren alan modele katıldığında bu alanın kümeler için anlamının önemsiz olduğu sonucuna ulaşılmaktadır. Bunun nedeni bu alanın büyük bir miktarının değerinin aynı olmasıdır. Bu sayede kümeleme analizinde yer alacak değişkenlerin belirlenmesi mümkün olabilmektedir. Oluşturulan modelde tüm alanlar, kümeler için önemli olarak hesaplanmıştır.

Tablo 7. "K-Means" Kümeleme Algoritması Sonucu Elde Edilen Kümeler Ve Özellikleri

	Küme1	Küme2	Küme3	Küme4	Küme5	Tüm Veri
Küme Büyüklüğü Adet (%)	20.612 (%16)	20.600 (%16)	24.883 (%19)	27.488 (%21)	32.369 (%25)	125.912 (%100)
Mezuniyet Yaşı *Ortalama (Std. Sapma)*	27,59 (5,30)	34,41 (5,95)	26,26 (3,55)	28,85 (5,96)	35,16 (6,23)	30,66 (6,65)
Mezuniyet Gecikmesi *Ortalama (Std. Sapma)*	2,35 (2,64)	3,49 (3,53)	1,61 (1,97)	2,89 (3,01)	4,03 (3,53)	2,94 (3,14)
Medeni Hali *B:Bekar (%)* *E:Evli (%)*	B	E	B	B (66,40) E (33,60)	E	E (50,61) B (49,39)
Bilgisayar Kullanımı *E:Evde Kullanıyor (%)* *Ş:Hem Ev Hem İş (%)* *İ:İşte (%)* *H:Kullanmıyor (%)*	E (9,18) Ş (2,18) İ (27,75) H (60,90)	E (11,97) Ş (24,43) İ (46,64) H (16,96)	E (23,56) Ş (20,91) İ (33,19) H (22,35)	E (25,19) Ş (18,64) İ (39,77) H (16,41)	E (4,88) Ş (3,06) İ (35,90) H (56,17)	E (14,86) Ş (13,33) İ (36,63) H (35,18)
İnternet Kullanımı *E:Evet (%), H:Hayır (%)*	H	E	E	E	H	E (57,92) H (42,08)
Gelir Durumu* A: Gelir<250 (%) B: 250≤Gelir<500 (%) C: 500≤Gelir<750 (%) D: 750≤Gelir<1000 (%) E: Gelir>1000 (%) *:milyon TL	A (19,78) B (46,62) C (23,21) D (7,69) E (2,69)	A (3,09) B (26,50) C (32,21) D (24,09) E (14,12)	A (13,79) B (37,05) C (27,09) D (13,21) E (8,86)	A (7,72) B (35,10) C (24,39) D (18,37) E (14,42)	A (3,64) B (38,26) C (32,40) D (18,47) E (7,24)	A (9,09) B (36,77) C (28,07) D (16,56) E (9,51)
Cinsiyet E: Erkek (%) K: Kadın (%)	E (38,68) K (61,32)	E	E	K	E (70,56) K (29,44)	E (60,59) K (39,41)

Elde edilen kümelere ilişkin önemli özellikler aşağıda sıralanmıştır.

- Bilgisayar ve internet kullanan bekar erkek öğrencilerin mezuniyet gecikmesi ortalaması (1,61 yıl) en düşük kümenin üçüncü küme olduğu gözlenmektedir.
- Mezuniyet gecikme ortalamasının en büyük (4,03 yıl) olduğu kümenin tümünün evli, %70'inin erkek, internet kullanmayan, çoğunluğu bilgisayar kullanmayan ve en yüksek mezuniyet yaş ortalamasına sahip bireylerden oluşan beşinci küme olduğu görülmektedir.
- Mezuniyet gecikme ortalaması ikinci büyük küme ise 3,49 yıl ortalama ile yine evli erkeklerin oluşturduğu ancak internet kullanan gelir düzeyi daha yüksek bireylerin oluşturduğu iki numaralı kümedir.
- Veri belirlenen kümelere göre etiketlendikten sonra mesleklerine göre kutu grafiğini oluşturulmuştur. Ek 9'da verilen grafikte 4,03 yıl ortalama ile en yüksek mezuniyet gecikmesi olan beşinci kümenin meslek guruplarının sırasıyla çiftçi, işveren, serbest meslek, emekli olduğu gözlenmektedir. Mezuniyet gecikme ortalaması 3,49 yıl olan ikinci kümeyi oluşturan mezunların mesleklerinin sırasıyla işveren, emekli ve serbest meslek olduğu görülmektedir. Mezuniyet gecikme ortalaması 2,35 yıl olan bekar öğrencilerin oluşturduğu birinci kümede mezuniyet gecikme ortalamasını yükselten meslek gurubunun emekliler olduğu görülmektedir.

5. UZAKTAN EĞİTİM SİSTEMLERİ VERİ ORGANİZASYONU VE VERİ MADENCİLİĞİ

Uzaktan eğitim alanında gerçekleştirilen veri madenciliği çalışmalarında elde edilen deneyimler e-öğrenme ortamlarının kişiselleştirilmesinde, tasarlanmasında ve iyileştirilmesinde anahtar olabilecek bilgilerin elde edilebileceğini göstermiştir. Bu çerçeveden bakıldığında e-öğrenme sistemlerinin ve öğrenci bilgi sistemlerinde depolanan verilerin niteliği önem kazanmaktadır. Öğrencilerin sistem içindeki davranışları ve öğrenme faaliyetlerinin ayrıntılı olarak saklanması çevrim-içi akıllı web'e dayalı öğretim ortamlarının geliştirilmesinin yanı sıra çevrim-dışı yapılacak analizlerin hammaddesini oluşturmaktadır. Bir e-öğrenme sisteminde kullanıcının öğrenme davranışlarını ilişkin saklanması gereken veriler aşağıdaki gibi sıralanabilir.

Öğrencinin internet üzerinden ulaştığı öğrenme hizmetinin özelliğine göre saklanması gereken veriler farklılık gösterebilir.

- *Yazılı materyal:* Ulaşılan belgenin içerdiği konu ve bölümleri, öğrencinin ilgili kısım için harcadığı süre, materyali okuma sırası gibi bilgiler okuma materyali ile öğrencinin etkileşimini ölçebilecek verilerdir. Ayrıca içeriğinin indekslenerek sunulması materyalin diğer öğrenme hizmetleri ile ilişkilendirilebilmesine olanak sağlayacaktır.
- *Problem ve sınavlar:* Web üzerinden gerçekleştirilen ölçme ve değerlendirme faaliyetleri öğrencinin kendini değerlendirmesinin yanı sıra aslında e-öğrenme ortamının da değerlendirilebilmesine olanak sağlamalıdır. e-öğrenme ortamında öğrenciye yöneltilen soru veya problemin zorluk derecesi, problemin çözülebilmesi için gerekli bilgilerin saptanması, ilgili bilgilerin e-öğreneme materyallerindeki konumları gibi bilgiler e-öğrenme sistemi veri organizasyonu içinde saklanmalıdır. Öğrencinin problemi ya da soruyu kaçıncı denemede doğru cevapladığı, cevaplama süresi, cevaplama denemeleri arasında diğer öğrenme faaliyetleri bilgileri e-öğrenme güncelerinde saklanması yerinde olacaktır.

Saklanacak veriler öğrenme materyalindeki ve problemlerdeki olası hataların fark edilmesinde kullanılabilecektir.

- *Etkileşimli öğrenme ortamları*: Forum, e-posta, mesaj gibi yazışma ile yürütülen iletişim araçlarında gerçekleştirilen öğrenme faaliyetlerinin güncelerde saklanması gerekmektedir. Büyük hacimli bu günceler metin madenciliği ile analiz edilebilmektedir. Çeşitli yazılım teknolojileri kullanılarak gerçekleştirilen etkileşimli öğrenme ortamları da son yıllarda yaygın olarak kullanılmaktadır. Özellikle teknik alanlarda faydalanılan bu ortamlarda öğrencinin öğrenme faaliyetlerinin güncelere kaydedilmesi bu araçların gelişmesine katkı sağlayacağı söylenebilir.

SONUÇ

Birçok kurum; operasyonel sistemler, internet ve enformasyon teknolojilerinin sağladığı olanaklar sayesinde depolanan verilerden anlamlı bilgiler türetme çabasındadır. Son yıllarda bazı eğitim kurumları da kendi sistemleri için faydalı olabilecek bilgileri veritabanlarından elde etmeyi fark etmeye başlamışlardır. Anadolu Üniversitesi Uzaktan Eğitim Sistemi, 1982 yılından günümüze sistemin çeşitli birimlerinde kimlik, not, sınav, e-öğrenme verileri, öğrenci ve uzaktan eğitim araştırmalarından sağlanan verileri manyetik ortamlarda saklamaktadır.

Veri madenciliği konusundaki kavramsal çerçeve ve uygulama yaklaşımlarına ilişkin literatür incelenip değerlendirildikten sonra Anadolu Üniversitesi Uzaktan Eğitim Sisteminde yer alan veritabanları kullanılarak anlamlı bilgilerin türetilmeye çalışıldığı veri madenciliği uygulaması gerçekleştirilmiştir.

Eğitim alanında gerçekleştirilmiş veri madenciliği çalışmalarını iki kategoride özetlemek mümkündür. Bunlardan ilki yüz-yüze eğitimin yürütüldüğü geleneksel eğitim sistemleri diğeri ise internet üzerinden yürütülen uzaktan eğitim sistemleridir. Geleneksel eğitim sistemlerindeki veri madenciliği çalışmaları genellikle öğrencilerin benzer özelliklerine göre gruplanması, öğrenci memnuniyetinin tahmin edilmesi gibi daha çok yönetsel kararların alınmasında katkı sağlayacak çalışmalardır. Eğitim alanındaki veri madenciliği çalışmalarının yoğun olarak sürdürüldüğü uzaktan eğitim sistemlerinde özellikle internet üzerinden yürütülen eğitim faaliyetlerinin analiz edilmesinde veri madenciliğinden faydalanılmaktadır. e-öğrenme sisteminde öğrenci davranış örüntülerinin keşfedilmesi, öğrenme materyallerinin değerlendirilmesi ve olası hataların ortaya çıkarılması, otomatik öneri sistemlerinin geliştirilmesi vb. görevlerin yerine getirilmesinde web madenciliği ve metin madenciliğinden faydalanan birçok çalışma gerçekleştirilmiştir.

Bu çalışmada gerçekleştirilen veri madenciliği uygulamalarından ilkinde öğrencilerin performanslarına ilişkin bir tahmin modelinin oluşturulması hedeflenmiştir. Tahmin modelinin oluşturulabilmesi için öncelikle sistemde yer alan veri kaynakları bir araya getirilerek veritabanı oluşturulmuştur. Öğrenci bilgi sistemi ve e-öğrenme sisteminden sağlanan veriler, veri madenciliği uygulama adımları takip edilerek modelleme aşamasına hazır hale getirilmiştir. Uygulamada SPSS Clementine veri madenciliği yazılımı kullanılmıştır. Öğrencinin kimlik, geçmiş başarı ve e-öğrenme kullanım güncelerini girdi parametresi olarak kullanan tahmin modelleri C5.0, Logistic Regression, Neural Net, C&RT, CHAID ve QUEST algoritmaları çalıştırılarak elde edilmiştir. Bu modellere geçerlilik testi uygulanarak C5.0 ile elde edilen karar ağacı modeli en iyi tahmin modeli olarak seçilmiştir. C5.0 algoritması ile elde edilen tahmin modeli %82,1 doğruluk oranı sağlamıştır.

Araştırmanın ikinci aşaması mezun öğrencilerin kümeleme analizidir. Bu analizde veri kaynağı 2000-2001 öğretim yılında öğrencilere uygulanan anket çalışması verileri ve mezun öğrenci kimlik verileridir. K-means kümeleme algoritması kullanılarak mezun öğrenciler; medeni durum, bilgisayar ve internet kullanım verileri, cinsiyet, mezuniyet yaşı ve mezuniyet gecikmesi özelliklerine göre beş küme halinde gruplandırılmıştır. Kümeleme sonucu bilgisayar ve internet kullanan bekar erkek öğrencilerin diğer öğrencilere göre daha kısa sürede mezun oldukları gözlenmiştir. Mezuniyet gecikmesi en fazla olan öğrencilerin ise yaşı büyük, internet kullanmayan ve bilgisayarı sınırlı kullanan öğrenciler olduğu belirlenmiştir. Kümeleme analizi sonucu elde edilen bilgilerin bilgisayar kullanımı ve başarı arasındaki doğru orantıyı doğrular nitelikte olduğu görülmektedir. Araştırma sonucu elde edilen bir diğer bilgi ise mezuniyeti geciken öğrencilerin çiftçi, işveren, serbest meslek ve emekli meslek gruplarında yoğunlaştığı bilgisidir.

Veri madenciliği analizleri ile elde edilen modellerin uzaktan eğitim sisteminin planlama faaliyetlerine katkıda bulunacağı düşünülmektedir. Mevcut sistemde gelecek dönem öğrenci sayısı tahminleri geçmiş dönemde elde edilen başarı oranları kullanılarak gerçekleştirilmektedir. Veri madenciliği analizi ile

geliştirilen model ise öğrenci performansını; e-öğrenme faaliyetlerini, geçmiş başarı ve yaş bilgilerini kullanarak öğrenci bazlı tahmin gerçekleştirmektedir. Tahmin modelinde kullanılan verilerin çevrimiçi hale getirilmesi ve modelin karar kurallarına dönüştürülerek uygulamaya konması etkin planlama faaliyetlerinin gerçekleştirilmesine katkı sağlayacaktır. Tahmin modelinin e-öğrenme sistemi için de katkı sağlayabileceği düşünülmektedir. e-öğrenme sistemi içersinde performansı düşük tahmin edilen öğrencilere ilişkin önlemler alınabilir ya da bu öğrencilerin başarısını arttıracak çözümler geliştirilebilir.

Kümeleme çalışmasından elde edilen bilgiler uzaktan eğitim sisteminde öğrenim gören öğrencilerin profilleri ve mezuniyet süreleri hakkında bilgi sağlamıştır. Bu bilgiler uzaktan eğitim sistemi bünyesinde gerçekleştirilecek araştırma ve projelerde alan uzmanlarına yol gösterebilecektir.

Bu çalışmada farklı veri kaynaklarından oluşturulan veritabanı, sistemdeki diğer problemlerin cevabını araştırmada veri kaynağı olarak kullanılabilir. Çalışmada oluşturulan veritabanına öğrencilerin geçmiş eğitim kurumlarındaki başarıları ve Uzaktan Eğitim Sistemi Test Araştırma Birimi sınav değerlendirme bilgilerinin eklenmesi ile sınav sisteminde karşılaşılan sorunlar veri madenciliği analizleri ile incelenebilecektir.

Anadolu Üniversitesi e-öğrenme sisteminde öğrencilerin sistem içersindeki davranışlarına ilişkin ayrıntılı verinin toplanmadığı gözlenmiştir. Yeni e-öğrenme sistemleri tasarımı gerçekleştirilirken öğrencilerin sistem içindeki davranışları hakkında ayrıntılı bilgilerin toplanması; öğrenci, yönetici, sistem tasarımcısı ve eğitimcilere büyük katkılar sağlayacaktır. Öğrencilerin öğrenme özelliklerini yansıtabilecek verilerin toplanması ve öğrenme örüntülerinin keşfedilmesi, gelecekte kişiselleştirilmiş ya da zeki e-öğrenme ortamlarının tasarımına katkı sağlayacaktır.

Günümüzde veri madenciliğinin iş dünyasında olduğu gibi geleneksel eğitim sistemlerinde önemli faydalar sağlayacağı görülmüştür. Bu nedenle veri madenciliği uygulamaları geleneksel eğitim sistemlerinde yönetsel ve akademik

faaliyetler için bir karar destek aracı olarak kullanımı yaygınlaşacaktır. Aynı zamanda uzaktan eğitim sistemlerinde içerik tasarımı ve pedagojik kararların alınmasında da önemli katkılar sağlayabilecektir. Eğitim alanı için özelleştirilmiş çevrimiçi veri madenciliği araçlarının geliştirilmesiyle, web'e dayalı eğitim ortamlarının öğrenciyi tanıyan, yönlendiren ve başarılı olmasında bir öğretici gibi davranabilen sistemler haline dönüşebileceği öngörülmektedir. Bu çerçeveden bakıldığında web'e dayalı eğitim ortamlarına öğrencinin öğrenme stilinin algılanmasında gerekli olabilecek verileri saklayabilme yeteneği kazandırılmalıdır. Birçok enformasyon sisteminde olduğu gibi web'e dayalı eğitim sistemlerinde de veri saklama standartları oluşturulmalıdır.

EKLER

EK 1. Veri Depolama Ve Yönetim Sistemleri

Kategori	Yazılım	Tanım
Metin Editörleri	Not Defteri	Basit belgelerin oluşturulması için kullanılabilen bir metin editörüdür. Not defterinin en yaygın kullanımı metin (.txt) dosyalarını düzeltmek ve okumaktır. Not defteri dosyaları, kullanıcılara farklı karakter setlerini kullanan belgelerle çalışma esnekliği sunan Unicode, ANSI veya UTF-8 olarak kaydedebilir.
	Ultra Edit-32	Metin ve 16'lık sayı sistemi editörüdür. Bu yazılım sütun modlu düzenlemeyi sağlar ve DOS tan Unix'e dosya dönüşümünü gerçekleştirebilir. Ayrıca html dosyaları da düzenleyebilir.
Hesap Tablosu	Microsoft Excel	Hesap tablosu pazarının %90 tarafından kullanılan hesap tablosu yazılımıdır.
	Lotus 1-2-3	IBM DB2 ve Oracle gibi veritabanlarına öncülük eden, Excel ve Lotus Notes 'la uyumlu hesap tablosu yazılımıdır.
	Quatro Pro	İlk versiyonu Borland tarafından geliştirildi ardından Novell tarafından daha sonra da Corel tarafından satın alındı. Quatro Pro Lotus 1-2-3 model alınarak geliştirilmiştir. Son versiyonu ise Microsoft Excel model alınarak tasarlandı.
Veritabanı	Microsoft SQL	Microsoft SQL Server ileri düzey veritabanı programlaması sunan bir veritabanı yazılımıdır. Zengin XML ve internet standartlarını destekler ve kullanıcılara bünyesindeki "stored procedureler" sayesinde XML formatındaki dosyaları kolayca depolama ve okuma olanağı tanır.
	Oracle 9i	Bir istemci/sunucu veritabanı yönetim yazılımıdır. Tam XML veritabanı işlevi sağlar. Bünyesinde OLAP işlevlerini barındırır ve Windows ve Linux işletim sistemleri için oluşturulmuştur.
	IBM DB2	IBM DB2 veritabanı sektörünün ilk çoklu ortam platformu sağlayan yazılımdır. Web'e hazır ilişkisel veritabanı yönetim sistemidir.
OLAP	Microsoft OLAP	Microsoft SQL Server tarafından sağlanan bir servistir.
	Oracle Discover	Oracle'ın OLAP çözümüdür. Sorgu, rapor, arama ve web yayını işlevlerini sağlar.
Veri Ambarı	SAP	Birçok ön tanımlı analiz modelini içerir. Raporlama aracı olarak Excel ve web sayfalarını kullanabilir. Yeni sorgular oluşturmada sürükle ve bırak teknolojisini kullanır.
	SAS	SAS veri ambarı ileri yükleme, çıkarım ve dönüşüm tekniklerine sahip yazılımdır.

EK 2. SPSS Veri Madenciliği Çözümü Clementine

Merkezi ABD Chicago'da bulunan SPSS 1967 yılından bu yana verideki gizli bilgileri keşfetme ve stratejik karar desteği sağlama yönünde ileri analitik çözümler sunmaktadır. Clementine veri madenciliği uygulamaları için SPSS tarafından geliştirilmiş görsel modelleme aracıdır. Clementine farklı veri kaynaklarına ulaşıma olanak sağlayan, modelleri hızlı bir şekilde oluşturma ve karşılaştırma olanaklarıyla ön plana çıkan ve pazarın %50'sini elinde bulunduran güçlü bir yazılımdır.

Clementine her türlü veri madenciliği projesinde kullanılması olası bütün modelleme yöntemlerini içeren ve modelleme yöntemlerinin daha nitelikli sonuçlar elde edilebilmesi için birbiri ile ardışık olarak kullanılmasına olanak sağlayan bir çözümdür. Clementine veri madenciliği metodu olarak CRISP-DM standardını kullanmaktadır. Birinci bölümde "Veri Madenciliği Uygulama Adımları" başlığı altında hakkında bilgi verilen CRISP-DM kullanıcılarına teknolojiden ziyade çözümü aranan probleme odaklanma imkanı sunmaktadır. Çalışmada kullanılan Clementine yazılımı fonksiyon ve modelleme türleri aşağıdaki tabloda özetlenmiştir.

Düğüm	İşlev
Veri Kaynakları	
	ODBC (Open Database Connectivity) veri kaynaklarından veri getirilmesini sağlayan düğümdür.
Kayıt Seçenekleri	
	Veri kayıtlarını bir alt kümesini seçmek veya analiz dışı bırakmak için kullanılır.
	Verileri özetlemek amacıyla gruplama fonksiyonlarını içeren bir düğümdür. İstenilen nümerik alanların toplam, ortalama vb. özet istatistiğini hesaplamada kullanılır.
Alan Seçenekleri	
	Veri alanlarının türünü belirlemek için kullanılan düğümdür. Ayrıca model düğümlerinden önce kullanılarak alanların model için girdi, hedef seçimlerinin yapılabildiği veri özelliklerinin tanımlandığı düğümüdür.

Düğüm	İşlev
	Kullanılmayacak alanların veri akışından bertaraf edilmesi için kullanılan filtre düğümüdür.
	Mevcut veri alanlarından yeni bir alanın türetilmesi işlevini yerine getirir. Clementine fonksiyonları kullanılarak veri içeriğine göre farklı değerlerin türetilmesi sağlanabilir.
	Eksik değerlerin boş değerlere dönüştürülmesini sağlayan düğümdür.
	Sayısal verilerin bölmelenerek kategorik verilere dönüştürülmesinde kullanılan veri işlem noktasıdır. Bölmeleme seçenekleri arasında ortalama, sabit değer, eşit sayı, standart sapma gibi seçenekleri kullanmak mümkündür.
	Model öncesi veri kümesini eğitim, test veya değerlendirme gibi farklı kısımlara bölerek model geliştirme aşamasında ilgili faaliyetler için kullanmasını sağlayan düğümdür.
Grafikler	
	Nümerik değerlere sahip alanların saçılma grafiğinin çizimi için kullanılan düğümdür. 3 boyutlu çizimlerin yapılabildiği saçılma grafiği düğümü farklı alanların grafik üzerine yansıtılması için renk, animasyon, panel, çizim şekli, ve saydamlık seçeneklerini kullanabilir.
	Kategorik değerlere sahip alanların dağılımını görselleştirmek için kullanılan grafiklerdir. Renk özelliği kullanılarak farklı bir özellik değeri grafikte gösterilebilir.
	Nümerik değerlerin histogramını oluşturmak için kullanılan grafik aracıdır.
Modelleme	
	C&RT algoritması kullanılarak bir karar ağacı üreten model düğümüdür. Her model düğümünde olduğu gibi parametreleri belirlenerek çalıştırıldığında yeni bir model çıktı düğümü üretir.
	CHAID algoritmasının uygulandığı karar ağacı modelini simgeler. Eniyi bölümlemeyi yapabilmek için ki-kare istatistiğinden yararlanan ve mümkün olan tüm durumları hesapladığından çalıştırılması zaman alıcı bir algoritmadır.
	Etkin istatistiksel ağaç yöntemini uygulayan QUEST algoritmasını uygulayan düğümdür.
	C5.0 algoritması hem karar ağacı hem de kural kümesi üreten ve veri kümesinde en büyük enformasyon kazanımını sağlamak amacıyla bölümleme yapan bir algoritmadır.
	Yapay sinir ağı kullanılarak tahmin modelinin geliştirilmesini sağlayan düğümdür. Ağ topolojisinin oluşturulması için gerekli parametrelerin tanımlanabildiği model seçeneğidir.
	Nominal regresyon olarak da adlandırılan lojistik regresyon girdi değişkenlerinin değerine dayalı olarak kayıtları sınıflamak için istatistiksel tekniklerin kullanıldığı tahmin modeli seçeneğidir.
	Kümeleme analizinin gerçekleştirildiği “K-means” algoritmasını simgeler. Algoritma küme elemanlarını belirlemede kayıtların küme merkezlerine olan uzaklıklarının ortalamasını enazlamayı hedefler.
Türetilen Modeller	
	Türetilmiş C&RT model düğümüdür. Diğer üretilmiş model düğümleri gibi üretildikten sonra veri akış diyagramına sürüklenerek veri akışına bağlanır. Tahmin değerini içeren bir alanı veri akışına ekler.

Düğüm	İşlev
	Üretilmiş Logistic Regresyon modelini temsil eder.
	Üretilmiş CHAID karar ağacını temsil eder.
	Yapay sinir ağı tekniği ile üretilen modelini temsil eder ve üretilen modele ilişkin performans değerlerini görüntüler.
	QUEST karar ağacı modelini temsil eder.
	K-means kümeleme algoritması modelini temsil eder. Kayıtları atadığı küme adları ile etiketler ve istenirse uzaklık ölçümlerini de veri setine dahil edebilir.
	C5.0 karar ağacı modelini temsil eder.
Çıktı	
	Bağlandığı noktadan veri akışını tabloya dönüştürerek görüntüleyen bir düğümdür.
	Analiz düğümü tahmin edici modellerin geçerliliğini test etmek için kullanılır. Çapraz geçerlilik ve n katlı çapraz doğrulama geçerlilik ölçümlerini destekler.
	Verinin incelenmesinde kullanılabilen veri inceleme düğümüdür. Mevcut verideki alanlar hakkında betimleyici istatistikler ve dağılım grafikleri üreterek veri hakkında kısa ve öz bilgi görüntüler.
	İstatistik düğümü olarak adlandırılan bu düğüm sayısal değer içeren alanlar için betimleyici istatistikler hesaplar. Ayrıca alan arasındaki korelasyon hesaplayabilir.
	Kalite düğümü veri içersindeki eksik veya boş değerleri raporlar.
	Veri hazırlama veya modelleme süreci içersinde elde edilen veri kümelerini ODBC veri kaynağı kullanılarak veritabanı yönetim yazılımlarına tablo olarak kaydeden bir çıktı düğümüdür.

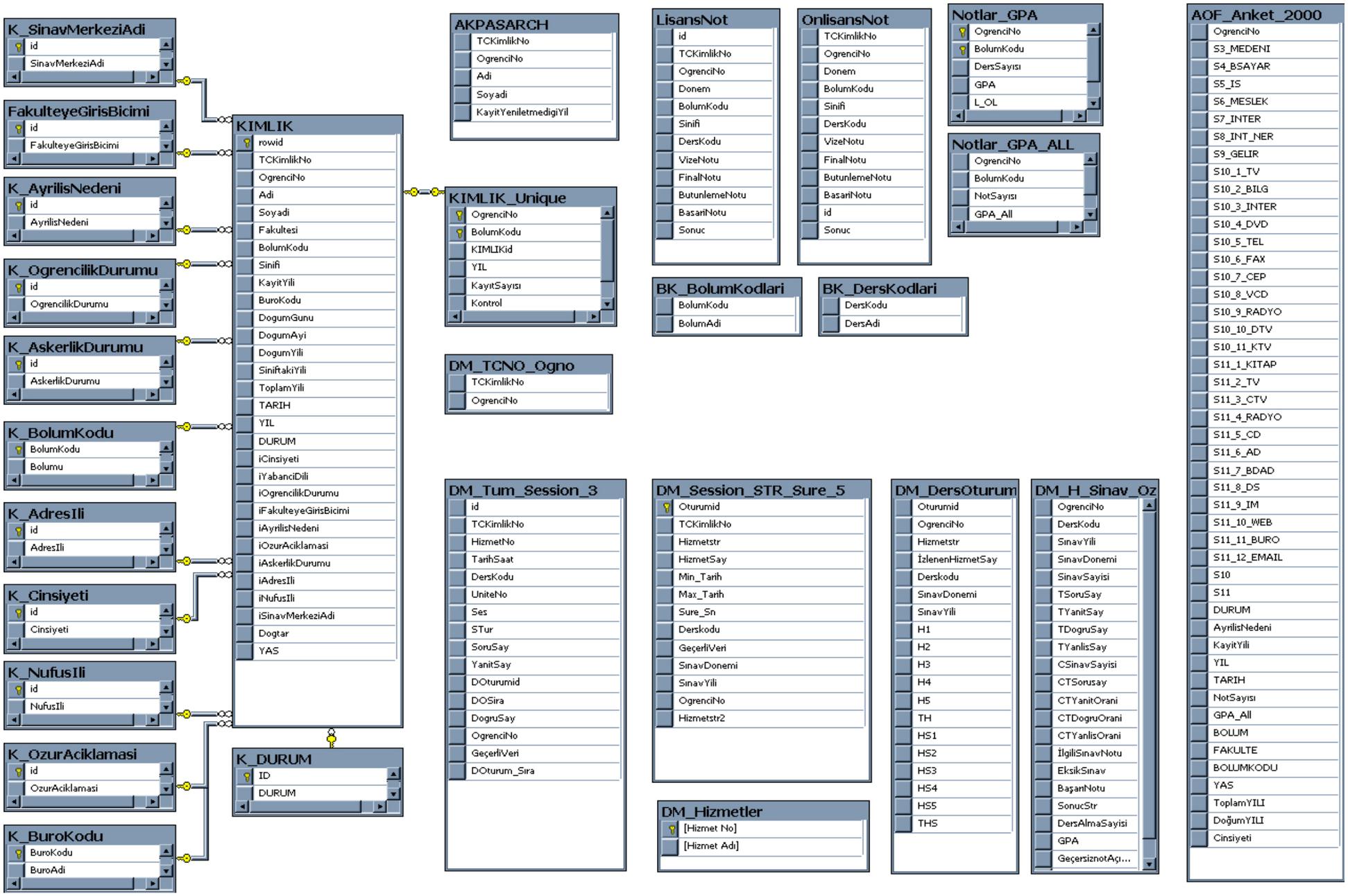
EK 3. Sınavlara Göre e-Öğrenme Ders Oturum Sayıları

Ders Oturum Sayısı	2005		2006		
	2005YS	2005BS	2006AS	2006YS	2006BS
1	18693	19770	41494	33928	30475
2	12441	11090	24386	20201	17712
3	8508	6992	16456	13633	11851
4	6047	5099	12510	10064	8809
5	4724	3728	10008	8064	6808
6	3820	3017	8425	6563	5529
7	3115	2389	7269	5476	4667
8	2550	2056	6265	4731	3838
9	2141	1707	5495	4044	3368
10	1848	1495	4923	3376	2872
11	1617	1215	4256	3148	2492
12	1435	1080	3935	2773	2243
13	1121	932	3589	2465	1915
14	1124	847	3251	2206	1731
15	935	758	2906	1978	1534
16	854	637	2623	1757	1373
17	789	557	2571	1574	1212
18	658	496	2302	1415	1128
19	585	445	2169	1357	992
20	549	444	1925	1155	951
21	508	348	1917	1148	768
22	405	315	1671	990	746
23	412	283	1557	905	705
24	369	280	1477	889	665
25	332	261	1437	791	545
26	300	246	1356	780	546
27	255	204	1301	725	496
28	251	207	1148	619	463
29	227	169	1122	613	397
30	231	144	992	574	402
31	210	151	968	512	377
32	180	123	930	476	331
33	155	145	810	457	264
34	152	148	798	397	287
35	131	124	756	397	271
36	120	108	701	366	234
37	102	88	681	363	258
38	102	96	657	344	207
39	99	75	579	325	231
40	94	87	591	270	201
41	82	68	547	258	182
42	73	79	507	267	163
43	57	64	483	258	149
44	80	48	475	209	148
45	60	74	431	194	154

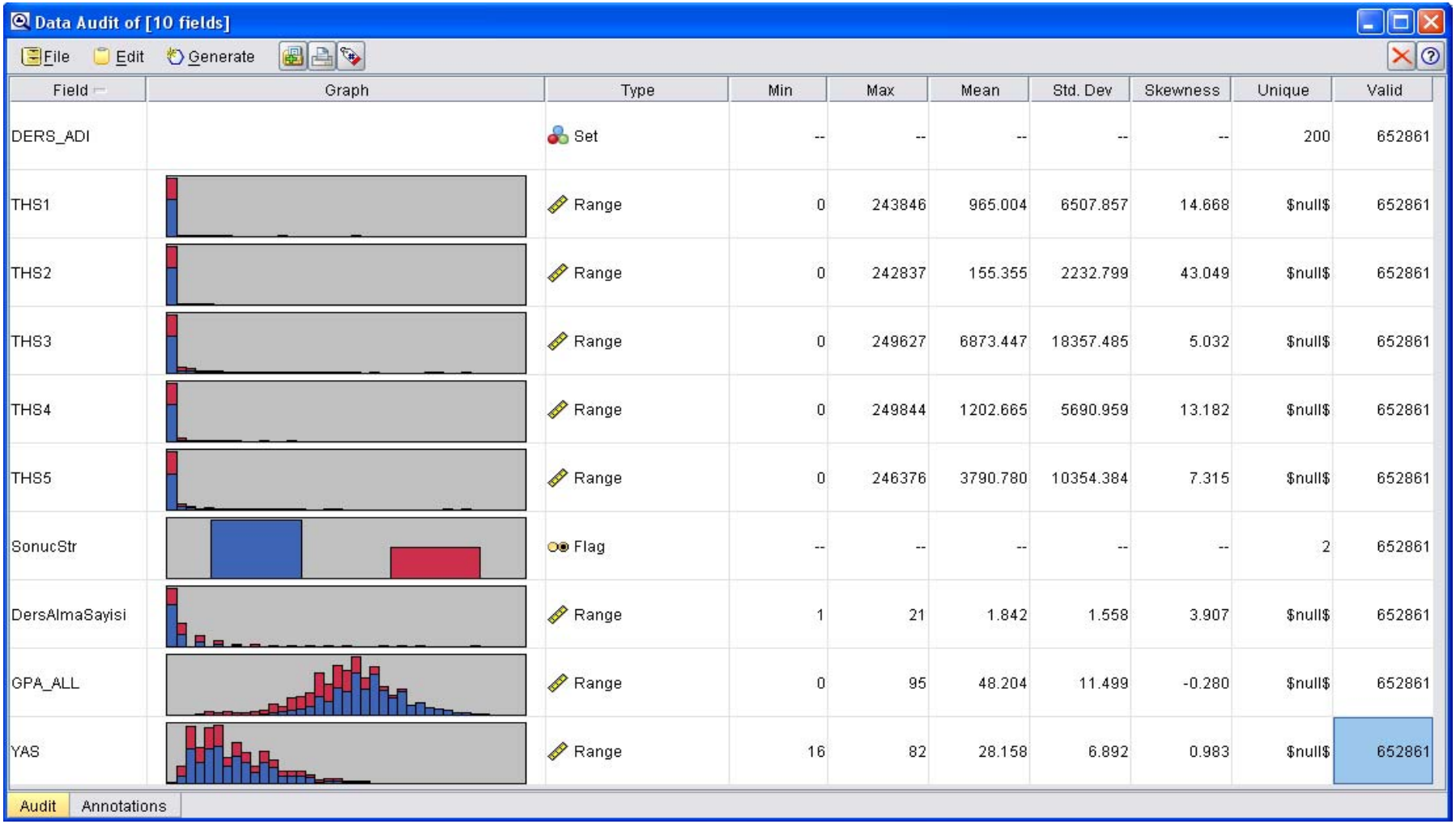
Ders Oturum Sayısı	2005		2006		
	2005YS	2005BS	2006AS	2006YS	2006BS
46	62	54	430	216	88
47	46	51	424	183	120
48	38	55	410	176	121
49	50	36	400	154	110
50	45	28	340	147	98
51	38	27	331	146	96
52	32	35	307	151	88
53	42	33	316	127	104
54	37	42	288	109	96
55	41	31	296	132	79
56	17	25	287	103	67
57	22	22	244	120	76
58	28	24	258	94	49
59	31	30	241	106	65
60	21	19	209	101	77
61	22	20	221	84	72
62	12	28	227	88	52
63	24	28	185	87	50
64	16	23	204	88	40
65	18	23	199	80	38
66	22	11	183	62	35
67	12	14	160	77	50
68	11	11	174	62	33
69	15	11	177	51	40
70	12	11	161	55	31
71	16	11	144	60	30
72	9	8	134	60	32
73	11	11	118	57	29
74	6	10	143	53	39
75	11	8	128	47	28
76	11	6	132	47	21
77	8	5	118	42	27
78	12	5	118	37	19
79	7	9	106	36	23
80	4	11	110	31	21
81	3	9	92	19	18
82	6	3	107	34	14
83	3	6	86	30	20
84	4	8	90	29	13
85	2	11	91	23	16
86	2	9	87	23	21
87	4	10	85	34	22
88	10	5	87	26	14
89	3	7	71	26	20
90	1	4	53	28	17
91	1	4	73	20	16
92	2	5	68	22	11
93	1	6	72	17	17

Ders Oturum Sayısı	2005		2006		
	2005YS	2005BS	2006AS	2006YS	2006BS
94	3	6	74	22	15
95	3	2	55	17	19
96	2	6	68	14	8
97	2	3	69	18	12
98	1	2	58	12	15
99	2	1	60	17	10
100 ve Üzeri	60	84	2000	465	303
Genel Toplam	79465	69656	202019	147200	123335

EK 4. Uzaktan Eğitim Sistemi Veri Madenciliği Veritabanı Tabloları

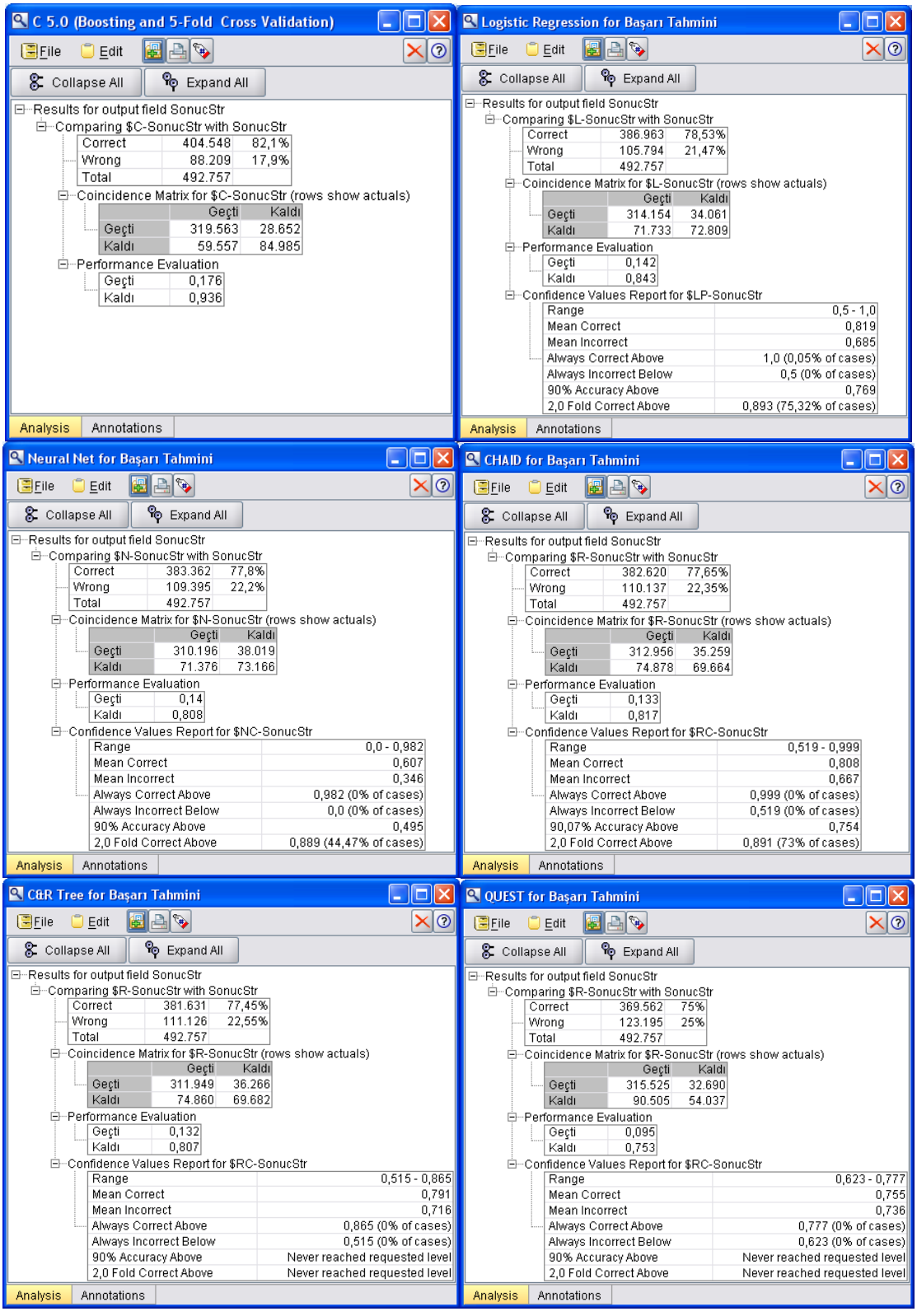


EK 5. Öğrenci Performans Tahmin Modeli Veri Özellik Ve Dağılım Grafikleri



Field	Graph	Type	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
DERS_ADI		Set	--	--	--	--	--	200	652861
THS1		Range	0	243846	965.004	6507.857	14.668	$null$	652861
THS2		Range	0	242837	155.355	2232.799	43.049	$null$	652861
THS3		Range	0	249627	6873.447	18357.485	5.032	$null$	652861
THS4		Range	0	249844	1202.665	5690.959	13.182	$null$	652861
THS5		Range	0	246376	3790.780	10354.384	7.315	$null$	652861
SonucStr		Flag	--	--	--	--	--	2	652861
DersAlmaSayisi		Range	1	21	1.842	1.558	3.907	$null$	652861
GPA_ALL		Range	0	95	48.204	11.499	-0.280	$null$	652861
YAS		Range	16	82	28.158	6.892	0.983	$null$	652861

EK 6. Öğrenci Performans Tahmini İçin Oluşturulan Modeller



C 5.0 (Boosting and 5-Fold Cross Validation)

File Edit

Collapse All Expand All

Results for output field SonucStr

Comparing $C-SonucStr with SonucStr

Correct	404.548	82,1%
Wrong	88.209	17,9%
Total	492.757	

Coincidence Matrix for $C-SonucStr (rows show actuals)

	Geçti	Kaldı
Geçti	319.563	28.652
Kaldı	59.557	84.985

Performance Evaluation

Geçti	0,176
Kaldı	0,936

Analysis Annotations

Logistic Regression for Başarı Tahmini

File Edit

Collapse All Expand All

Results for output field SonucStr

Comparing $L-SonucStr with SonucStr

Correct	386.963	78,53%
Wrong	105.794	21,47%
Total	492.757	

Coincidence Matrix for $L-SonucStr (rows show actuals)

	Geçti	Kaldı
Geçti	314.154	34.061
Kaldı	71.733	72.809

Performance Evaluation

Geçti	0,142
Kaldı	0,843

Confidence Values Report for $LP-SonucStr

Range	0,5 - 1,0
Mean Correct	0,819
Mean Incorrect	0,685
Always Correct Above	1,0 (0,05% of cases)
Always Incorrect Below	0,5 (0% of cases)
90% Accuracy Above	0,769
2,0 Fold Correct Above	0,893 (75,32% of cases)

Analysis Annotations

Neural Net for Başarı Tahmini

File Edit

Collapse All Expand All

Results for output field SonucStr

Comparing $N-SonucStr with SonucStr

Correct	383.362	77,8%
Wrong	109.395	22,2%
Total	492.757	

Coincidence Matrix for $N-SonucStr (rows show actuals)

	Geçti	Kaldı
Geçti	310.196	38.019
Kaldı	71.376	73.166

Performance Evaluation

Geçti	0,14
Kaldı	0,808

Confidence Values Report for $NC-SonucStr

Range	0,0 - 0,982
Mean Correct	0,607
Mean Incorrect	0,346
Always Correct Above	0,982 (0% of cases)
Always Incorrect Below	0,0 (0% of cases)
90% Accuracy Above	0,495
2,0 Fold Correct Above	0,889 (44,47% of cases)

Analysis Annotations

CHAID for Başarı Tahmini

File Edit

Collapse All Expand All

Results for output field SonucStr

Comparing $R-SonucStr with SonucStr

Correct	382.620	77,65%
Wrong	110.137	22,35%
Total	492.757	

Coincidence Matrix for $R-SonucStr (rows show actuals)

	Geçti	Kaldı
Geçti	312.956	35.259
Kaldı	74.878	69.664

Performance Evaluation

Geçti	0,133
Kaldı	0,817

Confidence Values Report for $RC-SonucStr

Range	0,519 - 0,999
Mean Correct	0,808
Mean Incorrect	0,667
Always Correct Above	0,999 (0% of cases)
Always Incorrect Below	0,519 (0% of cases)
90,07% Accuracy Above	0,754
2,0 Fold Correct Above	0,891 (73% of cases)

Analysis Annotations

C&R Tree for Başarı Tahmini

File Edit

Collapse All Expand All

Results for output field SonucStr

Comparing $R-SonucStr with SonucStr

Correct	381.631	77,45%
Wrong	111.126	22,55%
Total	492.757	

Coincidence Matrix for $R-SonucStr (rows show actuals)

	Geçti	Kaldı
Geçti	311.949	36.266
Kaldı	74.860	69.682

Performance Evaluation

Geçti	0,132
Kaldı	0,807

Confidence Values Report for $RC-SonucStr

Range	0,515 - 0,865
Mean Correct	0,791
Mean Incorrect	0,716
Always Correct Above	0,865 (0% of cases)
Always Incorrect Below	0,515 (0% of cases)
90% Accuracy Above	Never reached requested level
2,0 Fold Correct Above	Never reached requested level

Analysis Annotations

QUEST for Başarı Tahmini

File Edit

Collapse All Expand All

Results for output field SonucStr

Comparing $R-SonucStr with SonucStr

Correct	369.562	75%
Wrong	123.195	25%
Total	492.757	

Coincidence Matrix for $R-SonucStr (rows show actuals)

	Geçti	Kaldı
Geçti	315.525	32.690
Kaldı	90.505	54.037

Performance Evaluation

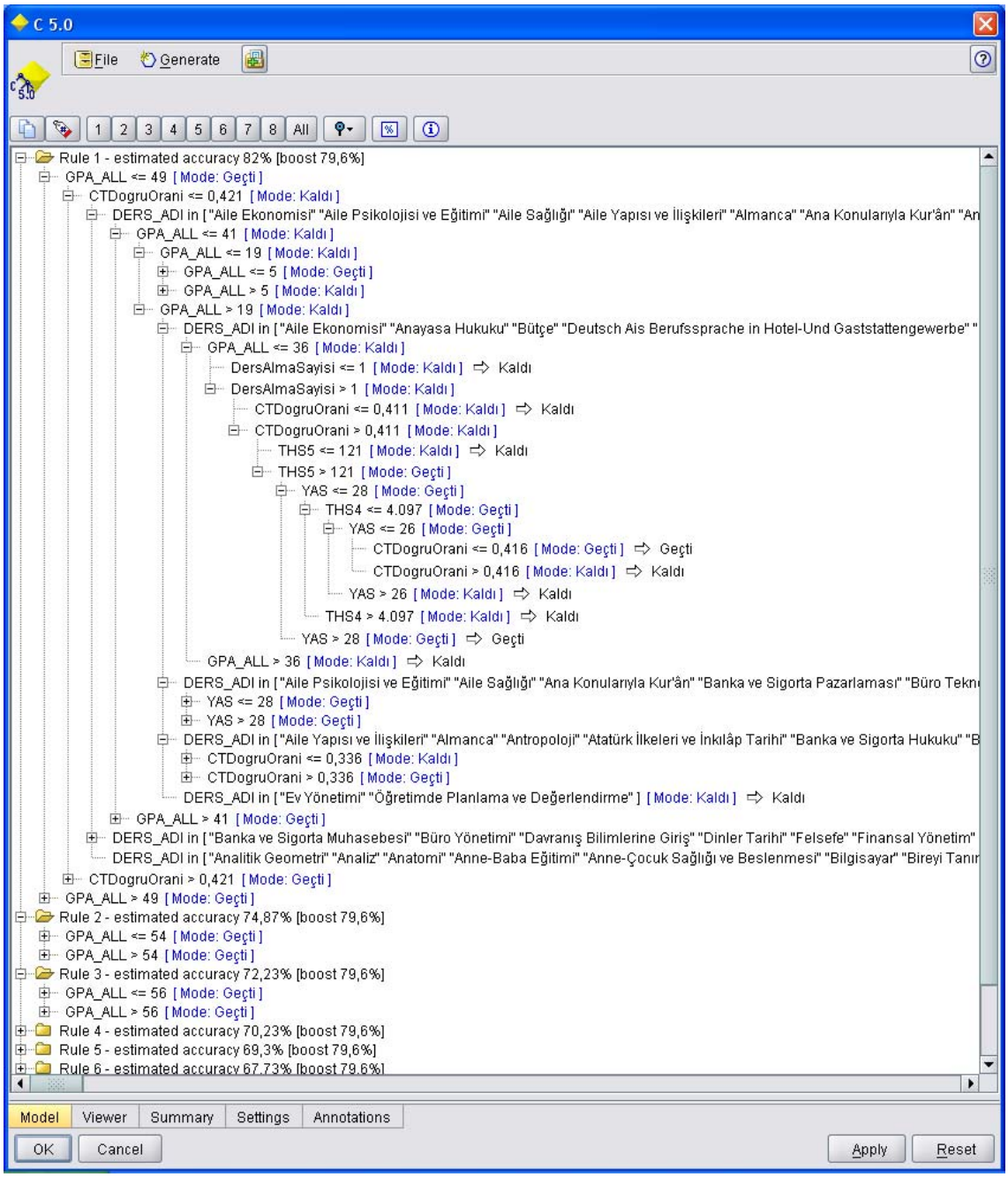
Geçti	0,095
Kaldı	0,753

Confidence Values Report for $RC-SonucStr

Range	0,623 - 0,777
Mean Correct	0,755
Mean Incorrect	0,736
Always Correct Above	0,777 (0% of cases)
Always Incorrect Below	0,623 (0% of cases)
90% Accuracy Above	Never reached requested level
2,0 Fold Correct Above	Never reached requested level

Analysis Annotations

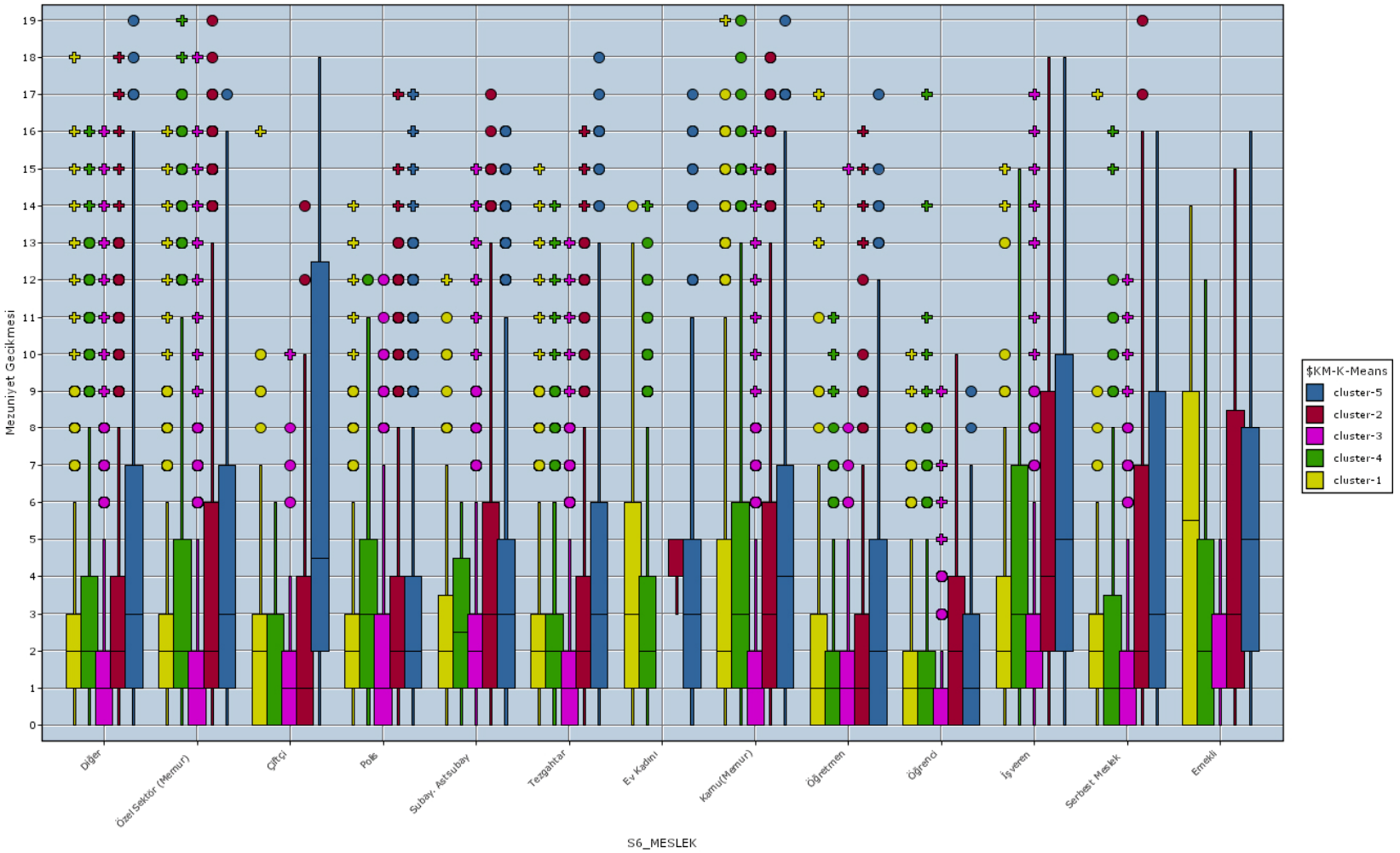
EK 7. C5.0 Karar Ağacı Algoritması Model Görünümü



EK 8. Mezun Öğrencilere İlişkin K-Means Kümeleme Analizi Sonuç Ekran Görünümü

	cluster-1	cluster-2	cluster-3	cluster-4	cluster-5	Overall	Importance
							>=0.95 >=0.90 <0.90 Unknown
MezuniyetYaşı	27.59 (5.30)	34.41 (5.95)	26.26 (3.55)	28.85 (5.96)	35.16 (6.23)	30.66 (6.65)	Important 1.00
Mezuniyet Gecikmesi	2.35 (2.64)	3.49 (3.53)	1.61 (1.97)	2.89 (3.01)	4.03 (3.53)	2.94 (3.14)	Important 1.00
S3_MEDENI	**B** 20612 100.00% **E** 0 0.00%	**B** 0 0.00% **E** 20600 100.00%	**B** 24883 100.00% **E** 0 0.00%	**B** 18226 66.40% **E** 9222 33.60%	**B** 0 0.00% **E** 32369 100.00%	**B** 63721 50.61% **E** 62191 49.39%	Important 1.00
S4_BSAYAR	**E-Ev** 1892 9.18% **E-Ev İş** 449 2.18% **E-İş** 5719 27.75% **H** 12552 60.90%	**E-Ev** 2466 11.97% **E-Ev İş** 5032 24.43% **E-İş** 9608 46.64% **H** 3494 16.96%	**E-Ev** 5862 23.56% **E-Ev İş** 5202 20.91% **E-İş** 8258 33.19% **H** 5560 22.35%	**E-Ev** 6913 25.19% **E-Ev İş** 5116 18.64% **E-İş** 10915 39.77% **H** 4504 16.41%	**E-Ev** 1579 4.88% **E-Ev İş** 989 3.06% **E-İş** 11620 35.90% **H** 18181 56.17%	**E-Ev** 18712 14.86% **E-Ev İş** 16788 13.33% **E-İş** 46120 36.63% **H** 44291 35.18%	Important 1.00
S7_INTER	**E** 0 0.00% **H** 20612 100.00%	**E** 20600 100.00% **H** 0 0.00%	**E** 24883 100.00% **H** 0 0.00%	**E** 27448 100.00% **H** 0 0.00%	**E** 0 0.00% **H** 32369 100.00%	**E** 72931 57.92% **H** 52981 42.08%	Important 1.00
S9_GELIR	**1000 de...** 555 2.69% **250 den az** 4077 19.78% **250 ile...** 9609 46.62% **500 ile...** 4785 23.21% **750 ile...** 1586 7.69%	**1000 de...** 2908 14.12% **250 den az** 636 3.09% **250 ile...** 5458 26.50% **500 ile...** 6635 32.21% **750 ile...** 4963 24.09%	**1000 de...** 2205 8.86% **250 den az** 3431 13.79% **250 ile...** 9219 37.05% **500 ile...** 6741 27.09% **750 ile...** 3286 13.21%	**1000 de...** 3957 14.42% **250 den az** 2119 7.72% **250 ile...** 9634 35.10% **500 ile...** 6695 24.39% **750 ile...** 5043 18.37%	**1000 de...** 2343 7.24% **250 den az** 1177 3.64% **250 ile...** 12383 38.26% **500 ile...** 10488 32.40% **750 ile...** 5978 18.47%	**1000 de...** 11968 9.51% **250 den az** 11440 9.09% **250 ile...** 46303 36.77% **500 ile...** 35344 28.07% **750 ile...** 20856 16.56%	Important 1.00
Cinsiyeti	**ERKEK** 7973 38.68% **KADIN** 12639 61.32%	**ERKEK** 20600 100.00% **KADIN** 0 0.00%	**ERKEK** 24883 100.00% **KADIN** 0 0.00%	**ERKEK** 0 0.00% **KADIN** 27448 100.00%	**ERKEK** 22840 70.56% **KADIN** 9529 29.44%	**ERKEK** 76296 60.59% **KADIN** 49616 39.41%	Important 1.00

EK 9. Mezun Öğrencilerin Özelliklerine Göre Elde Edilen Kümelerin Meslek Dağılımları



KAYNAKÇA

Abdi, Hervé ve D. Valantin. "Cilt I: Distance", Neil J. Salkind (Ed.), **Encyclopedia of Measurement and Statistics**, USA: Sage Publications, Inc., 2007.

Akpınar, Haldun. "Veritabanlarında Bilgi Keşfi ve Veri Madenciliği", **İşletme Fakültesi Dergisi**, Cilt No: 29, Sayı No:1, 1-22, 2000.

Arroyo, I. ve diğerleri. "Inferring Unobservable Learning Variables from Students' Help Seeking Behavior", **ITS2004 Workshops - Analyzing Student-Tutor Interaction Logs to Improve Educational Outcomes'da sunulan bildiri**. 782-784, Maceió, Alagoas, Brazil: 30 Ağustos 2004.

Becker, K., C. Ghedini ve E.L. Terra. "Using KDD to analyze the impact of curriculum revisions in a Brazilian university", **SPIE 14th Annual International Conference on Aerospace/Defesense, Sensing, Simulation and Controls'da sunulan bildiri**. 412-419, Orlando: Nisan 2000.

Berry, Michael J.A.ve Gordon Linoff. **Data Mining Techniques for Marketing, Sales, and Customer Support.** USA: John Wiley & Sons, Inc., 1997.

Berthold, Michael ve David Hand. **Intelligent Data Analysis.** Second revised and extended edition. Berlin: Springster, 2003.

Brusilovsky, P. ve C. Peylo. "Adaptive and Intelligent Web-based Educational Systems", **International Journal of Artificial Intelligence in Education**, Cilt no13: 156–169, 2003.

Chen, G. ve diğerleri. "Discovering Decision Knowledge from Web Log Portfolio for Managing Classroom Processes by Applying Decision Tree and Data Cube Technology", **Journal of Educational Computing Research**, Cilt No 23, Sayı No 3: 305–332, 2000.

CRISP-DM Consortium, **CRISP-DM 1.0 Step-by-Step Data Mining Guide**. www.crisp-dm.org, 2000.

Damez, M. ve diğerleri. "Fuzzy Decision Tree for User Modeling from Human–Computer Interactions",**5th International Conference on Human System Learning ICHSL.5'de sunulan bildiri**. 287 – 302, Marrakech, Morocco: 22-25 Kasım 2005.

Delavari, N., M. R. Beikzadeh ve S. Phon-Amnuaisuk. “Application of Enhanced Analysis Model for Data Mining Process in Higer Educational System”, **ITHET 6th Annual International Conference’da sunulan bildiri**. Juan Dolio, Dominican Republic: 8 Haziran 2005**.**

Dringus, L. ve T. Ellis. “Using Data Mining as a Strategy for Assessing Asynchronous Dicussion Forums”, **Computer & Education Journal**, Cilt No 45: 141–160, 2005.

Dunham, Margaret H. **Data Mining Introductory and Advanced Topics.** New Jersey: Pearson Education, Inc., 2003.

Elmasri, Ramez ve Shamkant B. Navathe. **Fundementals of Database Systems**. 2.Baskı. USA: Benjamin/Cummings Publishing Company, Inc., 1994.

Erdoğan Şenol ve Mehpare TİMOR. “A Data Mining Application in a Student Database”, **Havacılık ve Uzay Dergisi**, Cilt No 2,Sayı 2: 57-64, Temmuz 2005.

Fayyad, Usama M., Georgy Piatetsky-Shapiro, Padhraic Smyth ve Ramasamy Uthurusamy. **Advances in Knowledge Discovery and Data Mining.** USA: MIT Press, 1996.

Freyberger, J., N. Heffernan ve C. Ruiz. “Using Association Rules to Guide a Search for Best Fitting Transfer Models of Student Learning”, **ITS2004 Workshops - Analyzing Student-Tutor Interaction Logs to Improve Educational Outcomes’da sunulan bildiri**. Maceió, Alagoas, Brazil: 30 Ağustos 2004.

Georges Grinstein, M. Trutschl ve U. Civek, “High-dimensional Visulations,” **KDD-2001-Visual Data Mining Workshop’da sunulan bildiri**. San Francisco: 26-29 Ağustos 2001.

Giudici, Paolo. Applied Data Mining Statistical Methods for Business and Industry. England: John Wiley and Sons, 2003.

Ha, S., S. Bae ve S. Park. “Web Mining for Distance Education”, **IEEE International Conference on Management of Innovation and Technology’de sunulan bildiri**. 715–719, Singapore:12-15 Kasım 2000.

Hammouda, K. ve M. Kamel. “Chapter 13: Data Mining in e-learning (374-404)”, Samuel Pierre (ed.), **e-learning Networked Environments and Architectures: A Knowledge Processing Perspective**, London: Springer, 2006.

Han, Jiawei ve Micheline Kamber. **Data Mining: Concept and Techniques.** USA: Academic Press, 2001.

Holsheimer, M. and A. Siebes, **Data Mining: The Search for Knowledge in Databases**. CWI Technical Report, Amsterdam: 1994.

Koutri, M., N. Avouris ve S. Daskalaki. "Chapter 7: A Survey on Web Usage Mining Techniques for Web-based Adaptive Hypermedia Systems (125-149)", S. Y. Chen ve G. D. Magoulas (ed.), **Adaptable and Adaptive Hypermedia Systems**, Hershey: IRM Press, 2005.

Larose, Daniel T. **Discovering Knowledge in Data an Introduction to Data Mining**. USA: John Wiley & Sons, Inc., 2005.

Li, J. ve O. Zaïane. "Combining Usage, Content, and Structure Data to Improve Web Site Recommendation", **e-Commerce and Web Technologies, 5th International Conference'da sunulan bildiri**. 305–315, Zaragoza, Spain: 31 Ağustos-3 Eylül 2004.

Luan J. **Data Mining Applications in Higher Education**, SPSS Inc., http://www.spss.com/home_page/wp114.htm, 2004.

Luan, J. "Data Mining, Knowledge Management in Higher Education, Potential Applications", **42nd Associate of Institutional Research International Conference çalıştayında sunulan bildiri**. 1–18, Toronto,Canada: 2002.

Ma, Y.ve diğerleri. "Targeting the Right Students Using Data Mining", **The Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining'de sunulan bildiri**. 457–464, San Francisco, California, USA: 26-29 Ağustos 2001.

Markellou, P. ve diğerleri. "Using Semantic Web Mining Technologies for Personalized e-learning Experiences", **Web-based Education'da sunulan bildiri**. 461–466, Grindelwald, Switzerland: 21-23 Şubat 2005.

Mazza, R. ve C. Milani. "Exploring Usage Analysis in Learning Systems: Gaining Insights from Visualisations", **12th International Conference on Artificial Intelligence in Education'da sunulan bildiri**. 65-72, Amsterdam, The Netherlands: 18 Temmuz 2005.

Merceron, A. ve K. Yacef. “Tada-ed for Educational Data Mining”, **Interactive Multimedia Electronic Journal of Computer-Enhanced Learning**, Cilt No 7, Sayı No 1: 267–287, 2005.

Merceron, A. ve K.Yacef. “Mining Student Data Captured From a Web-based Tutoring Tool: Initial Exploration and Results”, **Journal of Interactive Learning Research**, Cilt No 15, Sayı No 4: 319–346, 2004.

Minaei-Bidgoli, B. ve W. Punch. “Using Genetic Algorithms for Data Mining Optimization in an Educational Web-Based System”, **Genetic and Evolutionary Computation Conference’da sunulan bildiri**. 2252–2263, Chicago, IL, USA: 12-16 Temmuz 2003.

Minaei-Bidgoli, B., P. Tan ve W. Punch. “Mining Interesting Contrast Rules for a Web-Based Educational System”, **2004 International Conference on Machine Learning and Applications (ICMLA ‘ 04)’da sunulan bildiri**. 320- 327, Louisville, Kentucky, USA: 6 - 18 Aralık 2004.

Mor, E.ve J. Minguillon. “E-learning Personalization Based on Itineraries and Long-term Navigational Behavior”, **13th International World Wide Web Conference’da sunulan bildiri**. 264–265, New York, USA: 17 - 22 Mayıs 2004.

Mostow, J. “Some Useful Design Tactics for Mining Its Data”, **ITS2004 Workshops - Analyzing Student-Tutor Interaction Logs to Improve Educational Outcomes’da sunulan bildiri**. Maceió, Alagoas, Brazil: 30 Ağustos 2004.

Muehlenbrock, M. “Automatic Action Analysis in an Interactive Learning Environment”, **12th International Conference on Artificial Intelligence in Education AIED-2005 - workshop on Usage Analysis in Learning Systems’de sunulan bildiri**. 73-80, Amsterdam, The Netherlands: 18-22 Haziran 2005

Nemati, Hamid R. ve Christopher D. Barko. **Organizational Data Mining: Leveraging Enterprise Data Resources for Optimal Performance**. USA: Idea Group Inc., 2004.

Pendharkar, Prag C. **Managing Data Mining Technologies in Organizations: Techniques and Applications**. USA: Idea Group Inc., 2003.

Roiger, Richard J. ve Michael W. Geatz. **Data Mining a Tutorial-Based Primer.** USA: Pearson Education, 1993.

Romero, C. ve S. Ventura."Educational Data Mining: A Survey from 1995 to 2005", **Expert Systems with Applications**, Cilt No 33, Sayı No 1: 135-146, 2007.

Rud, Olivia Parr. **Data Mining Cookbook**. New York: John Wiley & Sons, Inc., 2001.

Sanjeev, P. ve J. M. Zytkow. "Discovering Enrollment Knowledge in University Databases", **1th Conference on Knowledge Discovery and Data Mining'de sunulan bildiri**. 246–251, Montreal, Canada: 20-21 Ağustos 1995.

Shen, R. ve diğerleri. "Data Mining and Case-based Reasoning for Distance Learning", **Journal of Distance Education Technologies**, Cilt No1, Sayı No 3: 46–58, 2003.

Silva, D. R. ve M. T. P Vieira. "Using Data Warehouse and Data Mining Resources for Ongoing Assessment in Distance Learning", **IEEE International Conference on Advanced Learning Technologies'de sunulan bildiri.** 40–45, Kazan,Russia: 9-12 Eylül 2002.

SPSS Inc. **Clementine 9.0 Node Reference.** USA:2004.

Srivastava, J. ve diğerleri. "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data", **SIGKDD Explor. Newsl.**, Cilt No 1, Sayı No 2: 12–23,Ocak 2000.

Standford University, **Cross-Validation and the Bootstrap: Estimating the Error Rate of a Prediction Rule**. Standford:1995.

Talavera, L.ve E. Gaudioso. "Mining Student Data to Characterize Similar Behavior Groups in Unstructured Collaboration Spaces", **16th European Conference on Artificial Intelligence (ECAI 2004) - Workshop on Artificial Intelligence'da sunulan bildiri**, 17–23, Valencia,Spain: 22-27 Ağustos 2004.

Tan, Pang-Ning, Michael Steinbach ve Vipin Kumar. **Introduction to Data Mining.** USA: Pearson Education, Inc., 2006.

Tane, J., C.Schmitz ve G. Stumme. "Semantic Resource Management for the Web: an e-learning Application", **13th International World Wide Web Conference'da sunulan bildiri**. 1–10, New York, USA: 17 - 22 Mayıs 2004.

Tang, C. ve diğerleri. “Personalized Courseware Construction Based on Web Data Mining”, **The First International Conference on Web Information Systems Engineering’da sunulan bildiri**. 2204 – 2211, Hong Kong, China: 19 – 20 Haziran 2000.

Tang, T. ve G. McCalla. “Smart Recommendation for an Evolving e-learning System”, **International Journal on E-Learning**, Cilt No 4, Sayı No 1: 105–129, Haziran 2005.

Thomas, E.H. ve N.Galambos. “What satisfies students?”, **Resaerch in Higher Education**, Cilt No 45, Sayı No 3: 251-269,Mayıs 2004.

Two Crows Corp., **Introduction to Data Mining and Knowledge Discovery** (Versiyon 3: www.trocrows.com, 1999).

Ueno, M. “Data Mining and Text Mining Technologies for Collaborative Learning in an ILMS ‘ Samurai’ ”, **4th IEEE International Conference on Advanced Learning Technologies (ICALT’04)’da sunulan bildiri**. 1052-1053, Joensuu, Finland: 30 Ağustos - 1 Eylül 2004.

Ueno, M. “Online Outlier Detection System for Learning Time Data in e-learning and Its Evaluation”, **7th IASTED International Conference on Computers and Advanced Technology in Education’da sunulan bildiri**. 248–253, Kauai, Hawaii, USA: 16-18 Ağustos 2004.

Vranić M., D. Pintar ve Z. Skoćır, “The Use of Data Mining in Education Environment ”, **9th International Confrtence on Telecommunucations’de sunulan bildiri,** 243-250, Zagrep, Croatia: 13-15 Haziran 2007.

Wang, W. ve diğerleri. “Learning Portfolio Analysis and Mining in SCORM Compliant Environment”, **34th Annual Frontiers in Education’da sunulan bildiri**. 17-24, Savannah, Georgia, USA: 20-23 Ekim 2004.

Witten, Ian H. ve Eibe Frank. **Data Mining: Practical Machine Learning Tools and Techniques**. USA: Elsevier Inc., 2005.

Ye, Nong. **The Handbook of Data Mining.** USA: Lawrence Erlbaum Assoc., Inc., 2003.

Zaïane, O. ve J. Luo. “Web Usage Mining for a Better Web-based Learning Environment”, **Advanced Technology for Education’da sunulan bildiri**. 60-64, Banff, Alberta: 27-28 Haziran 2001.

Zaïane, O., M. Xin ve J. Han. “Discovering Web Access Patterns and Trends by Applying OLAP and Data Mining Technology on Web Logs”, **Proceeding of the Advances in Digital Libraries Conference (ADL’98) de sunulan bildiri**. 19–29, Santa Barbara, California, USA: 22-24 Nisan1998.

Zorrilla, M. E. ve diğerleri. “Web Usage Mining Project for Improving Web-Based Learning Sites”, **EUROCAST 2005 sunulan bildiri**. 205-210, Canary Islands, Spain: 7-11 Şubat 2005.