T.C.
DİCLE ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
İŞLETME ANABİLİM DALI
YÜKSEK LİSANS TEZİ

VERİ MADENCİLİĞİ
(Öğrenci Başarısına Etki Eden Faktörlerin Regresyon Analizi ile Tespiti)

MEHMET TAŞDEMİR

DANIŞMAN

Doç.Dr. SAİD PATIR

DİYARBAKIR
2012

T.C.
DİCLE ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
İŞLETME ANABİLİM DALI
YÜKSEK LİSANS TEZİ

VERİ MADENCİLİĞİ
(Öğrenci Başarısına Etki Eden Faktörlerin Regresyon Analizi ile Tespiti)

MEHMET TAŞDEMİR

DANIŞMAN

Doç. Dr. SAİD PATIR

DİYARBAKIR
2012

ÖZET

Akademik başarı kavramı günümüzde çok önem kazanan konulardan biri olmaya devam ediyor. Akademik başarı öğrencilerin ilerde sahip olacakları mesleki becerilerini öngören bir başarı kriteri olarak görülmektedir. Öğrencilerin üniversiteye yerleştiği yıllar onlar için stresli bir sürecin başlangıcı olmaktadır. Hayatı boyunca yapacağı bir mesleğin bilgisini almaya geldikleri bilinci öğrenciler üzerinde bir baskı unsuru olmaktadır.

Bu tez çalışmasında öğrencilerin başarı kriteri olarak aldığımız akademik başarıya etki eden nedenler incelenmiştir. Başarıya etki eden nedenleri tespit edebilmek için ÖSYM tarafından gönderilen verilerden ve öğrencinin öğrenimi sırasında aldığı ders başarılarından yararlanılmıştır. Bu çalışmada verilerin incelenmesi için veri madenciliği tekniklerinden olan regresyon kullanılmıştır. Bulunan ilişkilerin yönünü ve şiddetini tespit edebilmek içinde korelasyon kullanılmıştır. Bu tez çalışmasında iki bağımlı değişken ve sekiz bağımsız değişken üzerinde araştırma yapılmıştır. Bu değişkenler üzerinde onaltı adet hipotez tespit edilip bunlara cevaplar bulunmaya çalışılmıştır. Bu işlemlerin ardından öğrencinin akademik başarısına etki eden etmenler bulunmuştur.

Anahtar Kelimeler: Veri Madenciliği, Akademik başarı, Akademik Başarısızlık, Regresyon, Korelasyon.

ABSTRACT

The term academic success continues to be one of the most important issues today. Academic success is seen as success criteria for the students' future professional skills. The time that the students enter university becomes the beginning of a stressful process for them. The consciousness of getting the knowledge of a profession that they will do throughout their lives is an element of pressure on students.

In this thesis, we examined the factors that affect academic success which is taken as success criteria. In order to identify the causes that affect the success, we used the data sent by Student Selection and Placement Center and the course success of the students during their education. In this study, in order to examine the data, regression is used which is one of the data mining techniques. In order to detect the direction and intensity of the relations, correlation is used. In this thesis the research is done on two dependent variables and eight independent variables. On these variables, sixteen hypotheses were identified and answers were tried to be found for them. After following these procedures, the factors that affect the student's academic success were found.

Key Words: Data Mining, Academic Success, Academic Failure, Regression, Correlation

TUTANAK

Sosyal Bilimler Enstitüsü Müdürlüğüne

Bu çalışma jürimiz tarafından İktisadi ve İdari Bilimler Fakültesi, İşletme Anabilim Dalında YÜKSEK LİSANS TEZİ olarak kabul edilmiştir.

Başkan : **Doç. Dr. Said PATIR**

Üye : **Prof.Dr. Selim ERDOĞAN**

Üye : **Yrd.Doç.Dr. Abdurrahim EMHAN**

Onay
Yukarıdaki imzaların, adı geçene öğretim üyelerine ait olduğunu onaylarım.

İmza

Enstitü Müdürü
Prof. Dr. Sabri EYİGÜN

ÖNSÖZ

Bu çalışmamızda, veri madenciliği ve veri madenciliği süreçleri detaylı bir şekilde incelenmiştir. Çalışmamızda veri madenciliği içerisinde bulunan tekniklerden bahsedilmiştir. Yaptığımız araştırmada veri madenciliği analiz yöntemlerinden olan regresyon kullanılmıştır. Araştırma içerisinde veri madenciliğini akademik başarıya etki eden etmenlerin bulunması için elde ettiğimiz veri yığınına uygulanmıştır.

Üniversiteye yerleşen öğrencilerin yapısı incelenmiş ve veri madenciliği süreci sonunda akademik başarılarına etki eden etmenler bulunmuştur. Bu sonuçlar ışığında daha önceki yapılan araştırmalarla karşılaştırmalar yapılmış ve daha sonra yapılabilecek araştırmalara yön vermesi amacıyla öneriler sunulmuştur.

Bu tez çalışmasında değerli fikir ve önerileriyle beni yönlendiren, sabır ve destek gösteren, bilgi ve deneyimlerini benden esirgemeyen danışman hocam Sayın Doç.Dr.Said PATIR'a ve bana destek veren, teşvik eden hocalarım Sayın Yrd.Doç.Dr.Abdurrahim EMHAN ve Yrd.Doç.Dr.Mehmet METE'ye teşekkürlerimi sunarım.

Ayrıca tüm çalışmam boyunca manevi olarak bana her zaman destek olan eşim Canan'a, kızlarım Zeynep ve Şevvale teşekkürlerimi sunmayı borç bilirim.

Mehmet TAŞDEMİR

Diyarbakır, 2012

İÇİNDEKİLER

ÖZET ..... **i**
ABSTRACT ..... **ii**
TUTANAK ..... **iii**
ÖNSÖZ ..... **iiv**
ŞEKİLLER DİZİNİ ..... **viii**
TABLOLAR DİZİNİ ..... **ix**
SİMGELER VE KISALTMALAR LİSTESİ ..... **x**
I- VERİ MADENCİLİĞİ **1**
1- Veri Madenciliğine Giriş 1
1.1- Veri Madenciliğinin Tanımı 2
1.2- Veri Madenciliği Sistemlerinin Sınıflandırılması 5
1.3- Veri Tabanı Bilgi Keşfi (VTBK) ile Diğer Disiplinler Arasındaki İlişki 7
1.3.1- VTBK ile Makine Öğrenimi Arasındaki İlişki 7
1.3.1.1- VTBK ile İstatistik Arasındaki İlişki 7
1.3.1.2- VM ile Veri Tabanı Arasındaki İlişki 8
1.4- Veri Madenciliğinin Kulanım Amacı 8
1.5- Veri Madenciliği Uygulamaları ve Kullanım Alanları 9
1.5.1- Pazarlama alanında 9
1.5.2- Bankacılık alanında 9
1.5.3- Sigortacılık alanında 9
1.5.4- Perakendecilik alanında 9
1.5.5- Borsa alanında 10
1.5.6- Telekomünikasyon alanında 10
1.5.7- Sağlık ve ilaç alanında 10
1.5.8- Endüstri alanında 10
1.6- Veri Madenciliğinin Gereksinimleri 12
1.7- Veri Madenciliği Uygulamalarında Karşılaşılan Problemler 12
1.8- Veri Madenciliği Sistemleri Üzerine Yapılan Çalışmalar 13
1.8.1- Analysis Manager 14

1.8.2- Darwin 14
1.8.3- Clementine 15
1.8.4- DBMiner 15
1.8.5- Data Logic/R 17
1.8.6- INLEN 17
1.8.7- KDW (Knowledge Discowery Workbench) 18
1.8.8- SKICAT (Sky Image Classification & Archiving Tool) 18
1.8.9- R-MINI 18
1.8.10- TASA (Telecommunication Network Alarm Sequence Analyzer) 19
1.8.11- GCLUTO (Graphical CLUstering TOolkit) 19
1.8.12- Enterprise Miner 19
1.8.13- Veri Madenciliği Araçlarının Karşılaştırmaları 20
1.9- Veri Tabanlarında Bilgi Keşfi Süreci 24
1.9.1- Problemin tanımlanması 27
1.9.2- Veri madenciliği veritabanın oluşturulması 28
1.9.2.1- Veri Kaynaklarının Belirlenmesi 28
1.9.2.1.1- Metin Dosyaları ve İşlem Tabloları 30
1.9.2.1.2- Veri Tabanı Sistemleri 30
1.9.2.1.3- OLAP ve Veri Ambarları 32
1.9.2.2- Veri Tanımlama 35
1.9.2.3- Seçim 35
1.9.2.4- Veri Kalitesini İyileştirme ve Ön Hazırlık Süreçleri 36
1.9.2.4.1- Veri Temizleme 36
1.9.2.4.2- Eksik Veri 36
1.9.2.4.3- Gürültülü Veri 38
1.9.2.4.4- Tutarsız Veri 38
1.9.2.5- Veri Birleştirme 39
1.9.2.6- Veri Dönüştürme 39
1.9.2.7- Veri Azaltma 40
1.9.3- Model Oluşturma 41
1.9.3.1- Kümeleme 42
1.9.3.2- Birliktelik kuralları ve sıralı örüntüler (Sepet Analizi) 44

1.9.3.3- Sınıflama 45
1.9.3.3.1- Diskriminant analizi 46
1.9.3.3.2- Naive Bayes 46
1.9.3.3.3- Karar ağaçları 48
1.9.3.3.4- Sinir ağları 49
1.9.3.3.5- Kaba kümeler 51
1.9.3.3.6- Genetik algoritma 51
1.9.3.3.7- Regresyon ve Korelasyon Analizi 52
1.9.3.3.7.1- Basit Doğrusal Regresyon Analizi 58
1.9.3.3.7.2- Çoklu Doğrusal Regresyon 59
1.9.3.3.7.3-Aşamalı Regresyon 61
1.9.3.3.7.4- Polinomiyal Regresyon 61
1.9.3.3.7.5-En İyi Regresyon Modeli Bulma 62
1.9.3.3.7.6- Robust (Sağlam) Regresyon 62
1.9.4- Modelin Değerlendirilmesi 63
1.9.5- Modelin Uygulanması 64
II- UYGULAMA 65
2- Dicle Üniversitesinin Kısa Tarihçesi 65
2.1- Literatür Özeti 66
2.2- Araştırmanın Metodolojisi 68
2.2.1- Araştırmanın Amacı 69
2.2.2- Problem Cümlesi 69
2.2.3- Hipotezler 70
2.2.4- Araştırmanın Evreni ve Örneklemi 71
2.2.5- Veri Toplama Araçları 71
2.2.6- Verilerin Analizi 76
2.3- Bulgular 76
III. SONUÇ VE ÖNERİLER 87
KAYNAKLAR 91

ŞEKİLLER DİZİNİ

Şekil 1-1: Veri Tabanlarında Bilgi Keşfi Aşamaları
Şekil 1-2: Veri madenciliğinin uygulandığı alanlar
Şekil 1-3: DBMiner sisteminin yazılım mimarisi
Şekil 1-4: CRISP-DM Veri Madenciliği Uygulama Süreci
Şekil 1-5: Veri Küpü Örneği
Şekil 1-6: Üç Katmanlı Veri Ambarı Mimarisi
Şekil 1-7: Serpilme diyagramı gösterimi

TABLOLAR DİZİNİ

Tablo 1-1: Veri Madenciliğinin gelişim Adımları
Tablo 1-2: Veri madenciliği araçlarının çalışabildiği platformlar
Tablo 1-3: Veri madenciliği araçlarının veri girdisi ve çıktısı açısından karşılaştırmaları
Tablo 1-4: Veri madenciliği araçlarının desteklediği karar ağaçları
Tablo 1-5: Veri madenciliği araçlarının kullanılırlık karşılaştırmaları
Tablo 1-6: Veri madenciliği araçlarının görüntüleme açısından karşılaştırması
Tablo 1-7: Veri madenciliği araçlarının otomasyon açısından karşılaştırması
Tablo 1-8: Veri madenciliği araçlarının güçlü ve zayıf olduğu alanlar
Tablo 1-9: Veri Depolama Ve Yönetim Sistemleri
Tablo 2-1: Giriş Yaşı Veri Dönüşüm Tablosu
Tablo 2-2: Not Otalaması Veri Dönüşüm Tablosu
Tablo 2-3: Kaldığı Ders Sayısı Veri Dönüşüm Tablosu
Tablo 2-4: ÖSS Puanı Veri Dönüşüm Tablosu
Tablo 2-5: Tercih Sırası Veri Dönüşüm Tablosu
Tablo 2-6: Cinsiyet Veri Dönüşüm Tablosu
Tablo 2-7: Anne Öğrenim Durumu ve Baba Öğrenim Durumu Veri Dönüşüm Tablosu
Tablo 2-8: Bekleme Süresi Veri Dönüşüm Tablosu
Tablo 2-9: Fakültelere göre frekans tablosu
Tablo 2-10: Cinsiyete göre frekans tablosu
Tablo 2-11: Üniversiteye Giriş Yaşına göre frekans tablosu
Tablo 2-12: Öss Puanına göre frekans tablosu
Tablo 2-13: Üniversiteyi kaçıncı sırada tercih ettiğine göre frekans tablosu
Tablo 2-14: Anne öğrenim durumu frekans tablosu
Tablo 2-15: Baba öğrenim Durumu frekans tablosu
Tablo 2-16: Bekleme Süresi frekans tablosu
Tablo 2-17: Fakültelere göre başarı durumunun dağılımı (Not Ortalamasına göre)
Tablo 2-18: Fakültelere göre başarı durumunun dağılımı (Kaldığı Ders sayısına göre)
Tablo 2-19: Kaldığı Ders Sayısı ve Not Ortalaması Karşılaştırması
Tablo 2-20: Regresyon Tablosu
Tablo 2-21: Korelasyon Tablosu

SİMGELER VE KISALTMALAR LİSTESİ

VT	Veri Tabanı
VTBK	Veri Tabanlarında Bilgi Keşfi
VM	Veri madenciliği
OLAP	OnLine Transaction Processing (Çevrim İçi Analitik İşlem)
SQL	Structured Query Language (Yapılandırılmış Sorgu Dili)
SPSS	Statistical Packages for the Social Sciences-Sosyal Bilimler için İstatistik Paketi
OLAM	Online Analytical Mining (Çevrim İçi Analitik Madencilik)
DMQL	Data Mining Query Language
ODBC	Open DataBase Connectivity (Veri bankası bağlantısı)
OLE DB	Object Linking and Embedding Database
KDW	Knowledge Discowery Workbench
SKICAT	Sky Image Classification & Archiving Tool
TASA	Telecommunication Network Alarm Sequence Analyzer
GCLUTO	Graphical CLUstering TOolkit
CRISP-DM	CRoss-Industry Process For Data Mining
CHAID	Chi-Squared Automatic Interaction Detector
C&RT	Classification and Regression Trees
ID3	Induction of Decision Trees
ROC	Receiver Operating Characteristic (Alıcı Çalışma Karakteristik Grafiği)

I- VERİ MADENCİLİĞİ

1- Veri Madenciliğine Giriş

Manyetik ortamda veri saklama süreci ilk bilgisayarların üretilmesiyle başlamıştır. Veri saklama teknolojinin gelişmesi sonucu günümüzde çok büyük miktarda ve çeşitlilikte veri saklama ve depolanmasına imkân vermektedir. Günümüzde veri tabanları artık terabayt'larla ifade edilebilecek büyüklüğe ulaşmışlardır. Bu saklanan veri türleri çok çeşitli ve stratejik bilgiler içermektedir. Bankacılık sektör verileri, market verileri, sağlık sektörü verileri, sigorta verileri, eğitim verileri, hava tahmini için uydu verileri vb. bu sektörlerde veri tabanı yönetim sistemlerinin kullanılması sayesinde manyetik ortamda saklanabilmekte ve yönetilebilmektedir.

Veri tabanlarında (VT) saklanan verilerin çığ gibi büyümesi ve karmaşık hale gelmesiyle verileri analiz etmede kullanılan yöntemler ve basit araçlar yetersiz kalmıştır. Bu veri yığınlarından anlamlı bilgiler çıkarmak için yeni yöntem ve teknolojilerin geliştirilmesi ihtiyacı ortaya çıkmıştır. Bunun sonucu olarak yapılan çalışmalar sonucunda Veri Tabanlarında Bilgi Keşfi-VTBK (Knowledge Discovery in Databases-KDD) adı altında yeni bir kavram ortaya çıkmıştır. [1]

Veri madenciliği (VM) yakın zamanda oldukça dikkat uyandırmıştır. Ticari ve bilimsel keşifler için büyük potansiyeli ile yeni sorunlarla uğraşan, yeni bir teknolojidir.[2]

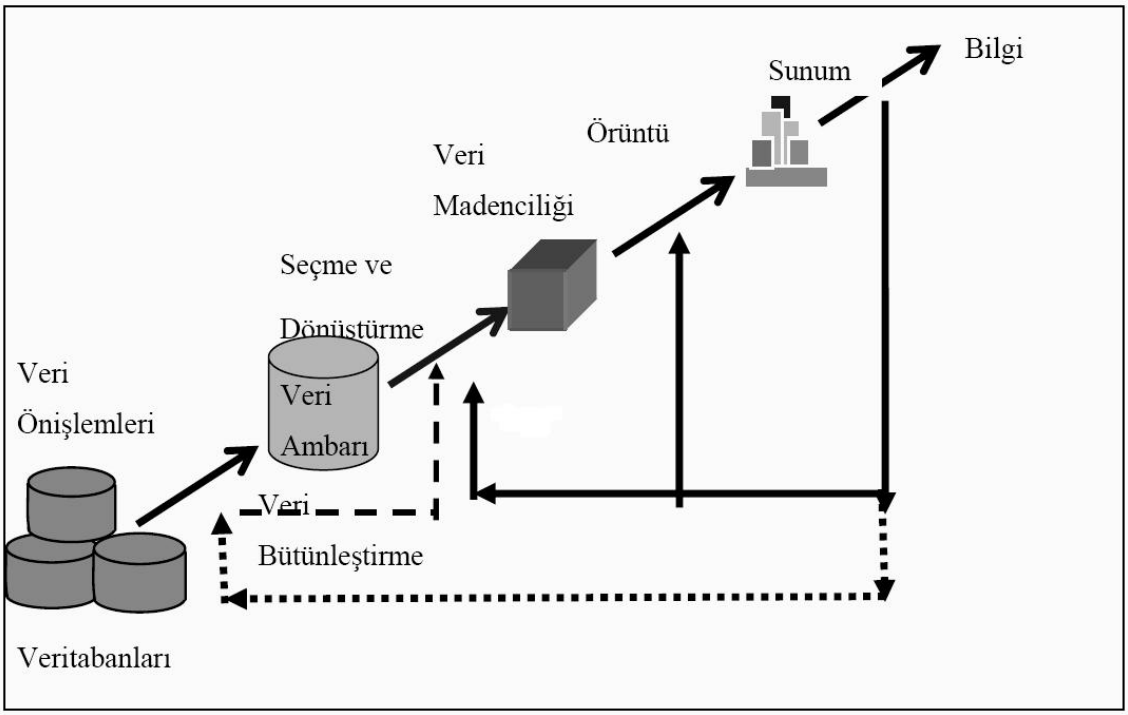
VTBK süreci, veriden yararlı bilgiyi keşfetmedeki tüm faaliyetleri ifade ederken, VM bu süreçteki özel bir adımı ifade etmektedir. VTBK sadece veri madenciliğini içeren bir süreç değildir. Aşağıdaki şekilde görüldüğü gibi 5 aşamadan oluşmaktadır. VM bu sürecin içerisinde yer almaktadır.[3]

---

[1] Akpınar H., "Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği", İstanbul, İ.Ü. İşletme Fakültesi Dergisi, Sayı:1, Nisan 2000, sf:1-22.

[2] Koldere Akın Y., "Veri Madenciliğinde Kümeleme Algoritmaları ve Kümeleme Analizi", Yayınlanmamış Doktora Tezi, İstanbul, 2008, sf:30

[3] Kiremitci B., "Veri Ambarlarında Veri Madenciliği ve Ulaştırma-Lojistik Sektöründe Bir Uygulama", Yayınlanmamış Yüksek Lisans Tezi, İstanbul, 2005. Sf:24



Şekil 1-1: Veri Tabanlarında Bilgi Keşfi Aşamaları

Kaynakça: Kiremitçi B., "Veri Ambarlarında Veri Madenciliği ve Ulaştırma-Lojistik Sektöründe Bir Uygulama", Yayınlanmamış Yüksek Lisans Tezi, İstanbul, 2005, Sf:25

Bununla beraber endüstride, medya ve veritabanı araştırmalarında "veri madenciliği" terimi "veri tabanlarında bilgi keşfi" teriminden daha yaygın olarak kullanılmaktadır. Bu nedenle sürecin tamamı genellikle veri madenciliği olarak anılmaktadır.[4]

1.1- Veri Madenciliğinin Tanımı

Literatürde "Veri Madenciliği" ile ilgili yapılan birçok tanım bulunmaktadır. Bunlardan bazılarını aşağıda sıralarsak:

[4] Aydın S., "Veri Madenciliği ve Anadolu Üniversitesi Uzaktan Eğitim Sisteminde Bir Uygulama", Yayınlanmamış Doktora Tezi, Eskişehir, 2007, sf:4

Veri Madenciliği;

- Kullanıcılara veri madenciliği yöntemleriyle anlaşılabilir ve faydalı olan verileri özetlemek ve veriler arasındaki beklenmeyen kayda değer ilişkileri bulmak için büyük ölçekli gözlemsel veri kümelerinin analiz edilmesidir,[5]
- Veri Madenciliği, Geçerli tahminler yapmak için kullanılan verilerdeki örüntüleri ve ilişkiyi açığa çıkarmak için çeşitli veri analiz araçlarını kullanan süreçtir,[6]
- Büyük veritabanlarında bulunan veri yığınlarından yaralanarak gizli ilişkilerin ve genel örüntülerin araştırılmasıdır,[7]
- Büyük veri yığınlarında bulunan verilerden anlamlı örüntülerin otomatik veya yarı otomatik olarak keşfedilme sürecidir,[8]
- Veritabanında yer alan verilerden bilginin otomatik olarak çıkarılması ve analiz edilmesinde bir veya daha fazla bilgisayar öğrenme tekniklerinin uygulanması sürecidir.[9]

Veri madenciliğinin zaman içindeki gelişim adımları aşağıdaki Tablo 1'de gösterilmektedir.

[5] Hand D. ve diğerleri, Principles of Data Mining, The MIT Press, London, 2001, sf:1

[6] Two Crows Corp., Introduction to Data Mining and Knowledge Discovery (Versiyon 3: http://www.twocrows.com/intro-dm.pdf, 1999), sf:1.

[7] M. Holsheimer ve A. Siebes, "Data Mining: The Search for Knowledge in Databases" (CWI Technical Report, Amsterdam: 1994), sf:2.

[8] Ian H. Witten ve E. Frank, Data Mining (USA: Elsevier Inc., 2005), sf:5.

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

Tablo 1-1: Veri Madenciliğinin gelişim Adımları

Gelişim Adımları	**Ticari Sorular**	**Geçerli Teknolojiler**	**Üretici Firmalar**
Veri Toplanması (1960)	Geçmiş beş yıldaki toplam gelirim nedir?	Bilgisayarlar, Teypler, Diskler	IBM, CDC
Veri Erişim (1980)	Geçen mart ayında İngiltere'de ki şube ne sattı?	İlişkisel veritabanları, Yapısal sorgulama dili, ODBC	Oracle, Sysbase, Informix, IBM, Microsoft
Veri Ambarı ve Karar Destek Sistemleri (1990)	Geçen mart ayında New England'daki şube satışları ne kadardı?(Boston bölgesi dahil)	OLAP, Çok boyutlu veri tabanları, veri ambarı	Pilot, Comshare, Arbor, Cognos, Microstrategy
Veri Madenciliği (Bugün)	Boston'da satışların gelecek ay ne kadar olmasını bekliyorsunuz? Neden?	İleri algoritmalar, Çok işlemcili bilgisayarlar, büyük veritabanları	Pilot, Lockheed, IBM, SGI

Kaynakça: Erdoğan Ş.Z., "Veri Madenciliği ve Veri Madenciliğinde Kullanılan K-means Algoritmasının Öğrenci Veri Tabanında Uygulanması", Yayınlanmamış Yüksek Lisans Tezi, İstanbul, 2004, sf:5

Dünyada 1960'larda veri toplama sistemleri, 1970'lerde ise ilişkisel veri tabanları kullanılmaya başlanmış, 1980'lerde ise ilişkisel veri tabanları popüler olmaya başlamış, 1990 ve 2000'lerde ise bilgisayar sistemlerindeki teknolojik gelişmelere paralel ilişkisel veri tabanlarında tutulan veri depoları kullanılmaya başlanmıştır. Bugün, dünya gündeminde de veri madenciliğinin, veri ambarlarının, multimedya ve web veri tabanlarının yaygınlaşmaya başladığı görülür. VM, son 10 yılda dünyada hızla yaygınlaşmaya başlayan bir disiplinler arası disiplin olarak göze çarpmaktadır. Günümüzde artan veri sayısı, bilgisayar kullanımının yaygınlaşması ve bilgi toplumu

olma yolundaki adımlar bu disiplinin daha fazla gündeme gelmesine neden olmaktadır. Yurt dışında yaygın bir şekilde kullanılan veri madenciliği, ülkemizde daha yeni yeni tanınmaya ve kullanılmaya başlanmıştır.[10]

Veri hacminin hangi boyutlara ulaşabileceği ve bunların işlenmesinin ne kadar güç olduğu kolayca anlaşılabilmektedir. Süper market örneği verirsek, önceleri market kasalarında sadece alınan malların hesap bilgileri yapılırdı ancak günümüz teknolojilerinde artık çok daha fazla veri kayıt altına alınmaktadır. Süper market örneği incelendiğinde, veri analizi yaparak her mal için bir sonraki ayın satış tahminleri çıkarılabilir; müşteriler satın aldıkları mallara bağlı olarak gruplanabilir; yeni bir ürün için potansiyel müşteriler belirlenebilir; müşterilerin zaman içindeki hareketleri incelenerek onların davranışları ile ilgili tahminler yapılabilir. Binlerce malın ve müşterinin olabileceği düşünülürse bu analizin gözle ve elle yapılamayacağı, otomatik olarak yapılmasının gerektiği ortaya çıkar. Veri madenciliği burada devreye girer; veri madenciliği büyük miktarda veri içinden gelecekle ilgili tahmin yapmamızı sağlayacak bağıntı ve kuralların aranmasıdır.[11]

Veri madenciliği uygulamalarının farklı disiplinlerde kullanılmasından dolayı veri madenciliği sistemlerinde sınıflandırma yapılması gerekmektedir.

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.[12]

- **Veritabanına göre:** Veritabanı yönetim sistemleri; veri modelleri, veri tipleri veya uygulama alanları gibi farklı özelliklere göre kendi içlerinde sınıflandırılırlar ve

[10] Altıntaş T., Veri Madenciliği Metotlarından Olan Kümeleme Algoritmalarının Uygulamalı Etkinlik Analizi, Sakarya, 2006, sf: 4

[11] Kayaalp K., Asenkron Motorlarda Veri Madenciliği ile Hata Tespiti, Isparta, 2007, sf:13

[12] Aydın A.g.e., sf:7

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 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 Tabanı Bilgi Keşfi (VTBK) ile Diğer Disiplinler Arasındaki İlişki

1.3.1- VTBK ile Makine Öğrenimi Arasındaki İlişki

Makine öğrenimi gözlem ve deneye dayalı ampirik kuralların otomatik biçimde bulunması olan VTBK sistemleri ile yakından ilgilidir. Genel olarak makine öğrenimi ve örüntü tanıma alanlarında yapılan çalışmaların sonuçları VTBK'de veri modelleme ve örüntü çıkarmak için kullanılmaktadır. Bu çalışmalardan bazıları örneklerden öğrenme, düzenli örüntülerin keşfi, gürültülü ve eksik veri ve eksik belirsizlik yönetimi olarak sayılabilir.[13]

VTBK'nın makine öğreniminden en büyük farkı aşağıda sıralanmıştır:

• VTBK büyük veri kümeleriyle çalışabilir,

• VTBK gerçek dünya verileriyle uğraşır.

Veri görselleştirmede kullanılan yöntemler, VTBK sistemi ile elde edilen örüntülerin, kullanıcıya grafikler aracılıyla sunumunu sağlar.

1.3.1.1- VTBK ile İstatistik Arasındaki İlişki

İstatistik ile VTBK arasındaki ilişkinin ana sebebi veri modelleme ve verideki gürültüyü azaltmadan kaynaklanmaktadır. İstatistiğin VTBK'de kullanılan tekniklerinden bazıları aşağıda sıralanmıştır:[14]

• Özellik seçimi,

• Veri bağımlılığı,

• Tanıma dayalı nesnelerin sınıflandırılması,

• Veri özeti,

• Eksik değerlerin tahmini,

• Sürekli değerlerin ayrımı

[13] Tiryaki S., "Lojistik Alanında Bir Veri Madenciliği Uygulaması", İstanbul, 2006, sf:4

[14] Tiryaki, A.g.e., sf:5

1.3.1.2- VM ile Veri Tabanı Arasındaki İlişki

VM sorgularına girdi sağlamak amacıyla VT kullanılmaktadır. VT'deki sorgu cümlecikleri VM'nin istediği örneklem kümesini elde etmek amacıyla kullanılmaktadır. Özellikle ilişkilendirme sorgusunda fazla miktarda VT sorgusu yapmak gerekmektedir. VM, VT'den farklıdır, çünkü VT'de var olan örüntüler için sorgular çalıştırılırken, VM'deki sorgular genelde keşfe dayalı ve ortada olmayan örüntüleri keşfetmeye dayalıdır.[15]

1.4- Veri Madenciliğinin Kulanım Amacı

İstatistiğin amacı nasıl ana kütle hakkında anlamlı bilgiler elde etmek ve yorum yapmaksa veri madenciliğinin amacı da anlamlı bilgiler elde etmek ve bunu eyleme dönüştürecek kararlar için kullanmaktır. Buradaki temel amaç, değişkenler arasındaki ilişkilerden çok, geleceğe yönelik sağlıklı öngörülerin üretilmesidir. Bu anlamda VM, özbilginin keşfedilmesi anlamında bir "kara kutu" bulma yaklaşımı olarak kabul edilmektedir ve bu doğrultuda yalnızca keşifsel veri analizi tekniklerini değil, sinir ağı tekniklerinden hareketle geçerli öngörüler yapmak ve öngörülen değişkenler arasındaki ilişkilerin belirlenmesi mümkün olduğu için aynı zamanda sinir ağı tekniklerini de kullanmaktadır.[16]

Veri madenciliği, analitik bir teknik değil, veri analizine bir yaklaşımdır. Diğer veri analiz yaklaşımlarından farklılıklarından biri, araştırmacının çoğu zaman çok genel bir araştırma sorusu çevresinde, keşfe yönelik bir tarzda işlem yapmasıdır. Veri madenciliğinin bu keşfe yönelik doğası sebebiyle, elimizdeki veri grubu için bir teknik seçerken diğer analiz yaklaşımlarına oranla daha az yol gösterici vardır. Veri madenciliği uygulamasının içerdiği istatistiksel teknikler, çoğunlukla genel kullanıma uygun olarak modifiye edilmiş ve böylece uygulama farklı veri grupları ve birçok değişken tipi üzerinde kullanılabilir hale gelmiştir.[17]

---

[15] Tiryaki, A.g.e., sf:5

[16] Altıntaş, A.g.e., sf: 5

[17] Kayaalp, A.g.e., sf:15

1.5- Veri Madenciliği Uygulamaları ve Kullanım Alanları

Günümüzde veri madenciliğinin başlıca uygulama alanları aşağıdaki gibi sayılabilir;[18]

1.5.1- Pazarlama alanında

- Müşteri segmentasyonu
- Müşterilerin demografik özellikleri arasındaki bağlantıların kurulması
- Çeşitli pazarlama kampanyaları
- Mevcut müşterilerin elde tutulması
- Yeni müşterilerin kazanılması
- Pazar sepeti analizi
- Çapraz satış analizleri ve satış tahminleri
- Müşteri değerlendirme ve müşteri ilişkileri yönetimi

1.5.2- Bankacılık alanında

- Farklı finansal göstergeler arasındaki gizli ilişkilerin bulunması
- Kredi kartı dolandırıcılıklarının tespiti
- Kredi taleplerinin değerlendirilmesi
- Usulsüzlük tespiti
- Risk analizleri

1.5.3- Sigortacılık alanında

- Yeni poliçe talep edecek müşterilerin tahmin edilmesi
- Sigorta dolandırıcılıklarının tespiti
- Riskli müşteri tipinin belirlenmesi

1.5.4- Perakendecilik alanında

- Satış noktası veri analizleri

[18] Eker H., "Veri Madenciliği veya Bilgi Keşfi", http://www.ikademi.com/insan-kaynaklari-bilgi-sistemleri/621-veri-madenciligi-veya-bilgi-kesfi.html, Erişim Tarihi:01.07.2011

- Alış-veriş sepeti analizleri
- Tedarik ve mağaza yerleşim optimizasyonu

1.5.5- Borsa alanında

- Hisse senedi fiyat tahmini
- Genel piyasa analizleri
- Alım-satım stratejilerinin optimizasyonu

1.5.6- Telekomünikasyon alanında

- Kalite ve iyileştirme analizleri
- Abonelik tespitleri
- Hatların yoğunluk tahminleri

1.5.7- Sağlık ve ilaç alanında

- Test sonuçlarının tahmini
- Ürün geliştirme
- Tıbbi teşhis
- Tedavi sürecinin belirlenmesi

1.5.8- Endüstri alanında

- Kalite kontrol analizleri
- Lojistik
- Üretim süreçlerinin optimizasyonu

Şekil 1-2'de 2003 yılında veri madenciliğinin sektörler bazında kullanımına ilişkin bir araştırmanın sonuçları yer almaktadır. Bu çizelgede araştırmaya katılan toplam 421 şirketin 51 adedinin bankacılık alanında veri madenciliğinin kullandığı görülmektedir.[19]

[19] Akbulut S., "Veri Madenciliği Teknikleri ile Bir Kozmetik Markanın Ayrılan Müşteri Analizi ve Müşteri Segmentasyonu", Yayınlanmamış Yüksek Lisans Tezi, Ankara, 2006, sf: 11

Son 3 yıl içinde veri madenciliğinin uygulandığı alanlar	
Bankacılık (51)	12%
Bioteknoloji / Genetik (11)	3%
Kredi skorlama (35)	8%
CRM (52)	12%
Doğrudan pazarlama (34)	8%
e-Ticaret (11)	3%
Eğlence/ Müzik (4)	1%
Sahtekarlık tespiti (31)	7%
Şans oyunu (2)	0,01 %
Kamu uygulamaları (12)	3%
Sigortacılık (24)	6%
Yatırım / Hisse senedi (5)	1%
Junk email / Anti-spam (5)	1%
Sağlık/ İK (15)	4%
İmalat (19)	5%
Tıp/ Farmakoloji (12)	3%
Perakende (25)	6%
Bilim (17)	4%
Güvenlik / Anti-terörizm(5)	1%
Telekomünikasyon (23)	5%
Seyahat (8)	2%
Web (9)	2%
Diğer (11)	3%

Şekil 1-2: Veri madenciliğinin uygulandığı alanlar

Kaynakça: Akbulut S., "Veri Madenciliği Teknikleri ile Bir Kozmetik Markanın Ayrılan Müşteri Analizi ve Müşteri Segmentasyonu", Yayınlanmamış Yüksek Lisans Tezi, Ankara, 2006, sf: 11

1.6- Veri Madenciliğinin Gereksinimleri

Veri madenciliğinin gereksinimleri aşağıdadır:[20]

- Erişilebilir veri,
- Etkin erişim yöntemleri,
- Açık problem tanımı,
- Etkin algoritmalar,
- Yüksek performanslı uygulama sunucusu,
- Sonuç oluşturmada esneklik.

Veri madenciliğinin diğer bir gereksinimi temizlenmiş veridir. Veri madenciliğinde kullanılacak veri yanlış sonuçlar üretmeye yol açabilecek aykırı değerler veriden temizlenmelidir. Doğru veri mevcut değilse ve verinin limitleri bilinmiyorsa; kullanılan yazılımın yanlış sonuçlar üretmesi kaçınılmazdır.[21]

1.7- Veri Madenciliği Uygulamalarında Karşılaşılan Problemler

VM büyük hacimli gerçek dünya verileriyle uğraştığı için, bu büyük hacimli veriler VM'de büyük sorunlar oluşturur. Bundan dolayı mesela küçük veri setleriyle ve yapay hazırlanmış verilerle doğru çalışan sistemler büyük hacimli, eksik, gürültülü, NULL değerli, artık, dinamik verilerle yanlış çalışabilir. Bundan dolayı bu sorunların aşılması gerekmektedir.[22]

Veri madenciliği girdi olarak kullanılacak ham veriyi veritabanlarından alır. Bu da veritabanlarının dinamik, eksiksiz, geniş ve net veri içermemesi durumunda sorunlar doğurur.[23]

Diğer sorunlar da verinin konu ile uyumsuzluğundan doğabilir. Sınıflandırmak gerekirse başlıca sorunlar aşağıdaki gibidir:[24]

---

[20] Akbulut, A.g.e., sf: 6

[21] Akbulut, A.g.e., sf: 6

[22] Kayaalp, A.g.e., sf:18

[23] Kayaalp, A.g.e., sf:19

[24] Kayaalp, A.g.e., sf:19

- **Sınırlı bilgi**: Veritabanları genel olarak veri madenciliği dışındaki amaçlar için tasarlanmışlardır. Bu yüzden, öğrenme görevini kolaylattıracak bazı özellikler bulunmayabilir.
- **Gürültü ve kayıp değerler**: Veri girişi veya veri toplanması esnasında oluşan sistem dışı hatalara gürültü denir. Veri toplanması esnasında oluşan hatalara ölçümden kaynaklanan hatalar da dâhil olmaktadır. Bu hataların sonucu olarak VM'de birçok niteliğin değeri yanlış olabilir.
- **Belirsizlik:** Yanlışlıkların şiddeti ve verideki gürültünün derecesi ile ilgilidir. Veri tahmini, bir keşif sisteminde önemli bir husustur.
- **Ebat, güncellemeler ve konu dışı sahalar:** Veri tabanlarındaki bilgiler, veri eklendikçe ya da silindikçe değişebilir. Veri madenciliği perspektifinden bakıldığında, kuralların hala aynı kalıp kalmadığı ve istikrarlılığı problemi ortaya çıkar. Öğrenme sistemi, kimi verilerin zamanla değişmesine ve keşif sisteminin verinin zamansızlığına karşın zamana duyarlı olmalıdır.
- **Artık veri:** Artık veri, problemde istenilen sonucu elde etmek için kullanılan örneklem kümesindeki gereksiz niteliklerdir.

Artık nitelikleri elemek için geliştirilmiş algoritmalar, özellik seçimi olarak adlandırılır. Özellik seçimi arama uzayını küçültür ve sınıflama işleminin kalitesini de artırır.[25]

1.8- Veri Madenciliği Sistemleri Üzerine Yapılan Çalışmalar

Veri madenciliği tekniklerinin birçok alanda gerekli olan bilgiye erişmek için uygulanabilir olması veri madenciliği teknikleriyle hem genel hem de özel amaçlı birçok uygulamanın geliştirilmesi sağlanmıştır.[26]

- **Özel Amaçlı Sistemler**: Veri madenciliği algoritmalarının belirli problem çözümleri için kullanılmasıdır. Bu uygulamaların çıkış amacı Veri madenciliği'nin kullanıcıdan bağımsız bir şekilde çalıştırılarak kullanıcının istediği bilgilerin

[25] Kayaalp, A.g.e., sf:19

[26] Doğan Ş., "Veri Madenciliği Kullanarak Biyokimya Verilerinden Hastalık Teşhisi", Yayınlanmamış Yüksek Lisans Tezi, Elazığ, 2007, sf:26-36

keşfedilmesi ve/veya keşfedilen bilgilerin gömülü bir uygulama içinde doğrudan karar alınmasında faydalanılmasını sağlamaktır. Veri madenciliği algoritmalarının özel amaçlı uygulandığı yerlerden ilk göze çarpanlar: astronomi, işletmelerdeki satış analizleri, pazarlama, borsa, sigorta vb. alanlardır.

- **Genel Amaçlı Sistemler**: Bu tür sistemlerde amaçlanan veri madenciliği sorgularının problemden bağımsız olarak tanımlanması ve bu özelliğinden dolayı istenen problemde bu sorguların kullanılabilmesidir.

1.8.1- Analysis Manager

Analysis Manager Microsoft firmasının veri madenciliği için üretmiş olduğu ürünüdür. Kümeleme analizi ve karar ağaçları için hazırlanmıştır. Analysis Manager OLAP (çevrim içi analitik işlem) küp desteği sunmaktadır. Analysis Manager'ın güçlü olduğu taraf kullanıcı-dostu bir ara yüze sahip olması ve uygulama kolaylığıdır. Aracın SQL sonucu ile bütünleşik çalışabilmesi bu aracı etkin hale getirmektedir. Analysis Manager'ın bir veri madenciliği sorgusu için farklı algoritmaları desteklememesi en büyük eksikliğidir. Kaynak kodun açık olmaması uygulama geliştiriciler için büyük zorluklar oluşturmaktadır. Kaynak kod yerine, Microsoft kümeleme ve karar ağacı için COM desteği sunsa da bu destek birçok gömülü sistem uygulamalarında geliştiriciler için eksik bir hizmet olarak görülmektedir. Analysis Manager, üretilen sonuçları farklı birçok gösterim şekliyle kullanıcıya sunulabilmektedir. Mesela karar ağaçları için, karar ağacını gösterebildiği gibi sonuçları kural tabloları şeklinde yorumlama imkânı vermektedir.[27]

1.8.2- Darwin

Darwin Oracle firmasının veri madenciliği aracıdır. Darwin regresyon ağaçları, karar ağaçları, kümeleme, yapay sinir ağları, Bayesian öğrenme, k-yakınlığında komşuluk gibi birçok algoritmayı destekleyen bir veri madenciliği aracıdır. Paralel sunucular için geliştirilmiş bir veri madenciliği sistemidir. Darwin kullanımı kolay bir

[27] Doğan, A.g.e., sf:26-36

ara yüze sahiptir. Darwin veri madenciliği algoritmalarından CART, StarTree, StarNet ve StarMatch'i kullanır.[28]

1.8.3- Clementine

Clementine SPSS firmasının veri madenciliği için geliştirmiş olduğu bir modüldür. SPSS istatistiksel bir araçtır. Clementine'nin SPSS içinde bir modül olarak kullanılması kullanıcıların SPSS'in istatistiksel fonksiyonlarından faydalanmasına imkan verir. Yapay sinir ağları ve kural tümevarım yöntemlerini kullanır. Clementine müşteri hizmetleri yönetimi, kimya sektöründe maddelerin aşındırıcılık tahmininde ve bankacılık alanında kredi kartı dolandırıcılıkları gibi konularda kendine uygulama alanı bulmuştur.[29]

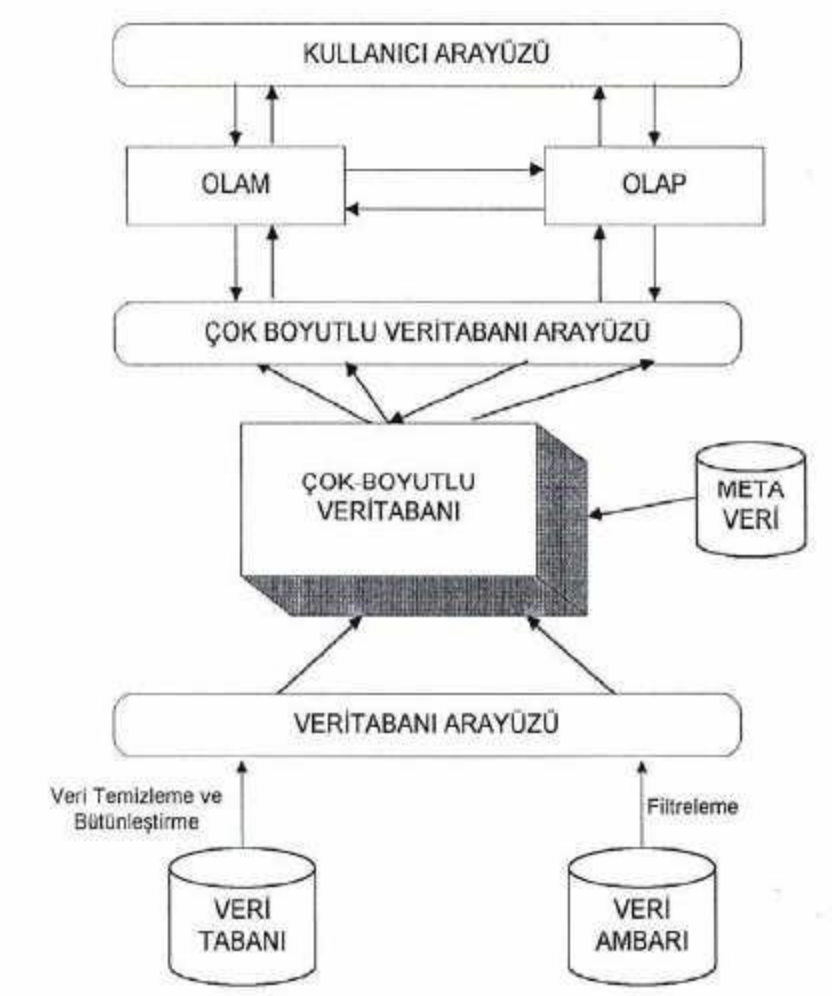
1.8.4- DBMiner

Kanada Simon Fraser Üniversitesi tarafından geliştirilen bir sistemdir. DBMiner sınıflama, kümeleme, eşleştirme ve sıra örüntüleri sorgularını yapabilecek veri madenciliği algoritmalarını kullanır. DBMiner çevrimiçi analitik işleme özelliğiyle veri madenciliği algoritmalarının bütünleşik çalışabilme özelliği sayesinde ön plana çıkmaktadır. Bu özellik OLAM (Online Analytical Mining) olarak anılır. DBMiner OLAP ve veri madenciliği yöntemlerini dinamik bir şekilde seçebilme imkânına sahiptir. Kullanıcının kolay kullanabileceği bir ara yüze sahiptir. Bu ara yüz sayesinde elde edilen sonuçlar çok yönlü bir soyutlama kullanılarak gösterilebilmektedir. DBMiner sisteminin mimarisi Şekil 2,8'de verilmiştir.[30]

[28] Doğan, A.g.e., sf:26-36

[29] Doğan, A.g.e., sf:26-36

[30] Doğan, A.g.e., sf:26-36



Şekil 1-3: DBMiner sisteminin yazılım mimarisi

Kaynakça: Doğan Ş., "Veri Madenciliği Kullanarak Biyokimya Verilerinden Hastalık Teşhisi", Yayınlanmamış Yüksek Lisans Tezi, Elazığ, 2007, sf:26-36

Şekil 1-3'de de görüldüğü üzere DBMiner verilerini ilişkisel veri tabanından ve/veya veri ambarından alarak veri küpleriyle bütünleştirerek çok boyutlu veri tabanına aktarır. Bu aktarım kaynaktan, ya verilerin bir bütün olarak çekilmesiyle ya da belli bir bölümünün çekilmesiyle gerçekleşir. DBMiner'ın diğer sistemlere göre en büyük avantajı geliştirilen DMQL'i (data mining query language) kullanmasıdır. DMQL, SQL benzeri bir veri madenciliği sorgu dilidir. DMQL sayesinde çevrimiçi sorgular OLAM veya OLAP modülüne yönlendirilerek işlenir. DBMiner'ın Veri tabanı ara yüzü çok boyutlu veri tabanına temizlenmiş, filtrelenmiş ve bütünleştirilmiş verileri aktarmaya yarar. Veri aktarımı için ODBC ve OLE DB(Object Linking and Embedding Database)

gibi bağlantılar da kullanılabilmektedir. OLAP ve OLAM modülleri arasındaki ilişkinin varlığı iki modülün birbirlerinin sonuçlarını kullanılabilmesine imkân tanır.[31]

DBMiner ürettiği sonuçları farklı birçok şekilde gösterebilme imkânına sahiptir. Mesela karar ağaçları için, karar ağacı şeklinde, kural tabloları şeklinde eşleştirme sorgusu için kural tablosu ve grafikleri üretebilmektedir. DBMiner ne kadar genel amaçlı bir sistemse de DBMiner'ı kullanarak ortaya çıkarılan özel amaçlı sistemler de mevcuttur. Bunlar arasında MultiMediaMiner, GeoMiner ve WeblogMiner sayılabilir.[32]

1.8.5- Data Logic/R

DataLogic/R kümeleme ve sınıflama analizi için kullanılan ticari bir veri madenciliği aracıdır. DataLogic/R artık nitelik ve verilerin temizlenmesi işlemlerini yapabilmektedir. Sistemin en güçlü olduğu taraf, üretilen kuralların öğrenme-test geçerliliği ve güvenlik gibi kriterlerde değerler üretmesidir. Bu değerler üretilen kuralların kalitesini belirlemek için kullanılabilmektedir. Bu araç, kimya ve ticaret sektöründeki çeşitli uygulamalarda kullanılmaktadır.[33]

1.8.6- INLEN

İlişkisel veri tabanından aldığı verileri makine öğrenimi teknikleriyle işledikten sonra ortaya çıkan sonuçları Veri tabanına yazmaktadır. Üretilen bilgi kesimi, basit ya da bileşik olabilmektedir.[34]

INLEN aracında dört işleç vardır:[35]

- **Veri tabanı yönetim işleci:** Veri tabanı sorgularını yazmak için geliştirilen bir işleçtir.
- **Bilgi yönetim işleci:** Üretilen bilgiyi yönetmek için kullanılır.

---

[31] Doğan, A.g.e., sf:26-36

[32] Doğan, A.g.e., sf:26-36

[33] Doğan, A.g.e., sf:26-36

[34] Doğan, A.g.e., sf:26-36

[35] Doğan, A.g.e., sf:26-36

- **Bilgi üretim işleci:** Veri tabanından bilgi almak ve makine öğrenimi algoritmalarını çağırmak için kullanılır.
- **Makrolar:** INLEN işleçlerini bir sırada tanımlamayı ve tek bir işleç gibi kullanabilmeyi sağlar.

1.8.7- KDW (Knowledge Discowery Workbench)

Knowledge discowery workbench; kümeleme, sınıflama, bağımlılık analizi algoritmalarını kullanan bir araçtır. Etkileşimli veri analizine imkân vermektedir. INLEN sistemiyle birçok ortak özelliği bulunmaktadır.[36]

1.8.8- SKICAT (Sky Image Classification & Archiving Tool)

Sky image classification & archiving tool, özel amaçlı bir veri madenciliği sistemidir. Özelleştiği konu astronomidir. Bu araç astronomik verileri indirgemek ve karar ağacı analizi için ID3, GID3, O-Btree algoritmalarını kullanmaktadır. Görüntü işleme, veri sınıflama ve VTYS metotlarını kullanır. SKICAT adından da anlaşılabileceği gibi gökyüzü fotoğraflarındaki gök cisimlerini tanımlamak, bunları sınıflandırmak, kataloglamak için kullanılan bir araçtır. Sayısal gökyüzü fotoğraflarındaki gök cisimlerinin parlaklık, alan, çekirdek büyüklüğü gibi özelliklerini kullanarak sınıflandırma sorgusunu gerçekleştirmektedir. SKICAT'ın deneysel testlerle fotoğraftan cisimleri tanıma ve sınıflandırma performansı %94 olarak saptanmıştır.[37]

1.8.9- R-MINI

R-MINI, SKICAT gibi özel amaçlı bir veri madenciliği sistemidir. Finansal konularda özelleşen R-MINI sınıflama ve sapma tespiti yapmak için kullanılır. R-MINI Veri tabanından çektiği gürültü içerikli verileri kullanarak tamlık ve tutarlılık kriterlerini sağlayan en küçük kural kümesini bulur.[38]

---

[36] Doğan, A.g.e., sf:26-36

[37] Doğan, A.g.e., sf:26-36

[38] Doğan, A.g.e., sf:26-36

1.8.10- TASA (Telecommunication Network Alarm Sequence Analyzer)

Telecommunication network alarm sequence analyzer, telekomünikasyonda kullanılan özel amaçlı bir veri madenciliği sistemidir. Telekomünikasyon hatlarında oluşabilecek bir hatanın önceden tahmini için kullanılır. Zaman serileri arası bağımlılıklarda kullanılan veri madenciliği algoritmaları, hata tahmini için kullanılmaktadır. Hatlarda olağandışı bir olay meydana geldiğinde bu sistem tetiklenir. Tetikleme sayısının, kontrol edilebilecek sayının çok üzerinde olması böyle bir sisteme ihtiyaç doğurur.[39]

1.8.11- GCLUTO (Graphical CLUstering TOolkit)

Graphical CLUstering TOolkit Minnesota Üniversitesi tarafından gerçekleştirilmiş bir araçtır. Bu araç kümeleme algoritmaları için geliştirilmiştir. Girdi kütüğünden aldığı verileri istenen kümeleme algoritmasına göre işleyip sonuçları çıktı kütüğüne yazmaktadır. Kolay kullanılabilir arayüze sahip olması ve görüntüleme problemlerinin iyi çözülmüş olması, üretilen sonuçların farklı gösterimleri ile GCLUTO kümeleme analizi için güçlü bir araçtır.[40]

1.8.12- Enterprise Miner

SAS firmasının veri madenciliği aracıdır. SAS'ın Veri ambarı ve ÇAİ (çevrimiçi analitik işleme) araçlarıyla bütünleşik çalışabilmektedir. Enterprise Miner karar ağaçları, yapay sinir ağları, regresyon analizi, 2-aşama modelleri (two-stage models), kümeleme, zaman serileri, ilişkilendirme, vb. veri madenciliği sorgularını ele alabilmektedir. Grafiksel ara yüzü sayesinde kullanım kolaylığı sağlar ve kullanıcılar uygulamanın karmaşıklığından habersiz bir şekilde sadece girdi ve çıktılara yoğunlaşabilirler. 2 katmanlı mimariyi kullanır. İstemci bilgisayardaki yazılım gereksinimi Windows 98, 2000 ve NT'dir. Sunucu bilgisayardaki yazılım gereksinimi Windows 98, 2000 ve NT ile Linux'dır.[41]

---

[39] Doğan, A.g.e., sf:26-36

[40] Doğan, A.g.e., sf:26-36

[41] Doğan, A.g.e., sf:26-36

1.8.13- Veri Madenciliği Araçlarının Karşılaştırmaları

Bu bölümdeki veri madenciliği araçlarının karşılaştırmaları, Elder ve Abbott (1998)'un "A Comparison of Leading Data Mining Tools" isimli sunum sonuçlarından faydalanılarak oluşturulmuştur. Bu bölümdeki çizelgeler için, aşağıda verilen tablodaki anahtarlar kullanılmıştır.[42]

Tablo (1-2 . 1-8)'de kullanılan anahtarların anlamları

Anahtar Anlamı

- Boşluk Sıfır kapasite
- X Düşük kapasite
- X- Normal kapasite
- X+ Yüksek kapasite

Tablo 1-3'de veri madenciliği araçlarının veri girdisi ve çıktısı açısından karşılaştırmaları, Tablo 1-4'de veri madenciliği araçlarının desteklediği karar ağaçları, Tablo 1-5'de veri madenciliği araçlarının kullanılırlık karşılaştırmaları, Tablo 1-6'da veri madenciliği araçlarının görüntüleme açısından karşılaştırması, Tablo 1-7'de veri madenciliği araçlarının otomasyon açısından karşılaştırması, Tablo 1-8'de veri madenciliği araçlarının güçlü ve zayıf olduğu alanlar verilmiştir.[43]

[42] Doğan, A.g.e., sf:26-36

[43] Doğan, A.g.e., sf:26-36

Tablo 1-2: Veri madenciliği araçlarının çalışabildiği platformlar

Platformlar	PC Standalone (95 / NT)	Unix Standalone	Unix Sunucu /PC İstemci	NT Sunucu PC İstemci	VT Bağlanabilirlik
Clementine	X	X +			X
Darwin			X		X
DataCruncher	X		X		X
Enterprise Miner	X		X +	X	X
GainSmarts	X	X			X
Intelligent Miner			X		X
MineSet		X			X
Model 1	X		X	X	X
ModelQuest	X	X			X
PRW	X				X
CART	X	X +			
Scenario	X				X
NeuroShell	X				
OLPARS	X	X			
See5	X	X +			
S-Plus	X				X -
WizWhy	X				

Kaynakça: Doğan Ş., "Veri Madenciliği Kullanarak Biyokimya Verilerinden Hastalık Teşhisi", Yayınlanmamış Yüksek Lisans Tezi, Elazığ, 2007, sf:26-36

Tablo 1-3: Veri madenciliği araçlarının veri girdisi ve çıktısı açısından karşılaştırmaları

Veri Girdisi & Model Çıktısı	Otomatik Başlık	Veri Formatını Kaydetme	ODBC	Yerel Veri Tabanı Sürücüleri	Özet Raporlar	Çıktı aynak Kodu
Clementine	X		X			X
Darwin		X	X			X
DataCruncher	X	X	X	X	X	
Enterprise Miner	X-		X	X	X-	X
GainSmarts	X	X		X	X	X
Intelligent Miner				X-		X
MineSet		X		X		
Model 1	X	X	X	X	X	X
ModelQuest	X			X	X	X
PRW	X	X	X		X	X
CART						
Scenario	X				X	
NeuroShell	X					
OLPARS		X				
See5	X-					
S-Plus	X		X		X	X
WizWhy	X				X	

Kaynakça: Doğan Ş., "Veri Madenciliği Kullanarak Biyokimya Verilerinden Hastalık Teşhisi", Yayınlanmamış Yüksek Lisans Tezi, Elazığ, 2007, sf:26-36

Tablo 1-4: Veri madenciliği araçlarının desteklediği karar ağaçları

Karar Ağaçları	Sınıflandırılmış Maliyetler	Budama	Görsel Ağaçlar	"Çizelge"	C5 veya C4. 5	CHAID	Diğerleri
Clementine	X	X	X-		X		
Darwin	X			X			
Enterprise Miner	X	X	X	X	X-	X	
GainSmarts			X	X		X	X
Intelligent Miner			X				X
MineSet	X	X	X	X		X	
Model 1				X		X	
ModelQuest		X			X-		
CART	X		X	X+			
Scenario							X
S-Plus		X	X	X			
See5	X	X			X+		

Kaynakça: Doğan Ş., "Veri Madenciliği Kullanarak Biyokimya Verilerinden Hastalık Teşhisi", Yayınlanmamış Yüksek Lisans Tezi, Elazığ, 2007, sf:26-36

Tablo 1-5: Veri madenciliği araçlarının kullanılırlık karşılaştırmaları

Kullanılırlık	Veri Yükleme ve İşleme	Model Kurma	Model Anlama	Teknik Destek	Genel
Clementine	X+	X+	X +	X +	X +
Darwin	X	X	X +	X	X
DataCruncher	X+	X +	X	X	X
Enterprise Miner	X	X	X	X	X
GainSmarts	X+	X	X	X	X
Intelligent Miner	X	X	X	X	X
MineSet	X	X +	X +	X	X +
Model 1	X+	X +	X +	X +	X +
ModelQuest Enterprise	X	X +	X +	X +	X +
PRW	X+	X +	X +	X +	X +
CART	X-	X	X	X	X
Scenario	X	X +	X +	X	X +
NeuroShell	X	X	X	X	X
OLPARS	X-	X	X	X	X
See5	X	X	X	X	X
S-Plus	X	X	X +	X	X
WizWhy	X	X	X +	X	X

Kaynakça: Doğan Ş., "Veri Madenciliği Kullanarak Biyokimya Verilerinden Hastalık Teşhisi", Yayınlanmamış Yüksek Lisans Tezi, Elazığ, 2007, sf:26-36

Tablo 1-6: Veri madenciliği araçlarının görüntüleme açısından karşılaştırması

Görüntüleme	Histogramlar	Pasta Grafiği	Dağılım/ çizgi çizimleri	Dönen dağılım	Şartlı çizimler	Sınıflandırma karar bölgeleri	Korelasyon çizimleri
Clementine	X		X		X	X-	X
Darwin	X-	X-	X-				
DataCruncher	X	X	X		X		
Enterprise Miner	X	X	X	X-	X		X
GainSmarts	X		X-				
Intelligent Miner	X-	X	X		X		
MineSet	X	X	X	X	X		
Model 1	X		X	X			
ModelQuest Enterprise	X		X				
PRW	X		X	X			
CART							
Scenario							X
NeuroShell			X				
OLPARS	X	X	X	X-	X	X	
See5	X						
S-Plus	X	X	X		X		X
WizWhy							

Kaynakça: Doğan Ş., "Veri Madenciliği Kullanarak Biyokimya Verilerinden Hastalık Teşhisi", Yayınlanmamış Yüksek Lisans Tezi, Elazığ, 2007, sf:26-36

Tablo 1-7: Veri madenciliği araçlarının otomasyon açısından karşılaştırması

Otomasyon	Otomasyon Metodları	Özgür Metin Açıklamasının Adımları
Clementine	Görsel programlama , Programlama dili	X
Darwin	Programlama dili	X
DataCruncher	(İş Yöneticisi)	
Enterprise Miner	Görsel programlama, Programlama dili	X
GainSmarts	Makro Dili , Sihirbazlar	X-
Intelligent Miner	(Sihirbazlar)	
MineSet	Veri Geçmişi , Log	
Model 1	Model Sihirbazları	
ModelQuest	Toplu Ajanda	
PRW	Deney Yöneticisi ; Makrolar	X
CART	*Built-in Basic Scripting*	
Scenario		
NeuroShell		
OLPARS		
See5		
S-Plus	*Scripting* (S) ; C / C ++	
WizWhy		

Kaynakça: Doğan Ş., "Veri Madenciliği Kullanarak Biyokimya Verilerinden Hastalık Teşhisi", Yayınlanmamış Yüksek Lisans Tezi, Elazığ, 2007, sf:26-36

Tablo 1-8: Veri madenciliği araçlarının güçlü ve zayıf olduğu alanlar

Otomasyon	Otomasyon Metodları	Özgür Metin Açıklamasının Adımları
Clementine	Görsel programlama , Programlama dili	X
Darwin	Programlama dili	X
DataCruncher	(İş Yöneticisi)	
Enterprise Miner	Görsel programlama, Programlama dili	X
GainSmarts	Makro Dili , Sihirbazlar	X-
Intelligent Miner	(Sihirbazlar)	
MineSet	Veri Geçmişi , Log	
Model 1	Model Sihirbazları	
ModelQuest	Toplu Ajanda	
PRW	Deney Yöneticisi ; Makrolar	X
CART	*Built-in Basic Scripting*	
Scenario		
NeuroShell		
OLPARS		
See5		
S-Plus	*Scripting* (S) ; C / C ++	
WizWhy		

Kaynakça: Doğan Ş., "Veri Madenciliği Kullanarak Biyokimya Verilerinden Hastalık Teşhisi", Yayınlanmamış Yüksek Lisans Tezi, Elazığ, 2007, sf:26-36

1.9- Veri Tabanlarında Bilgi Keşfi Süreci

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.[44]

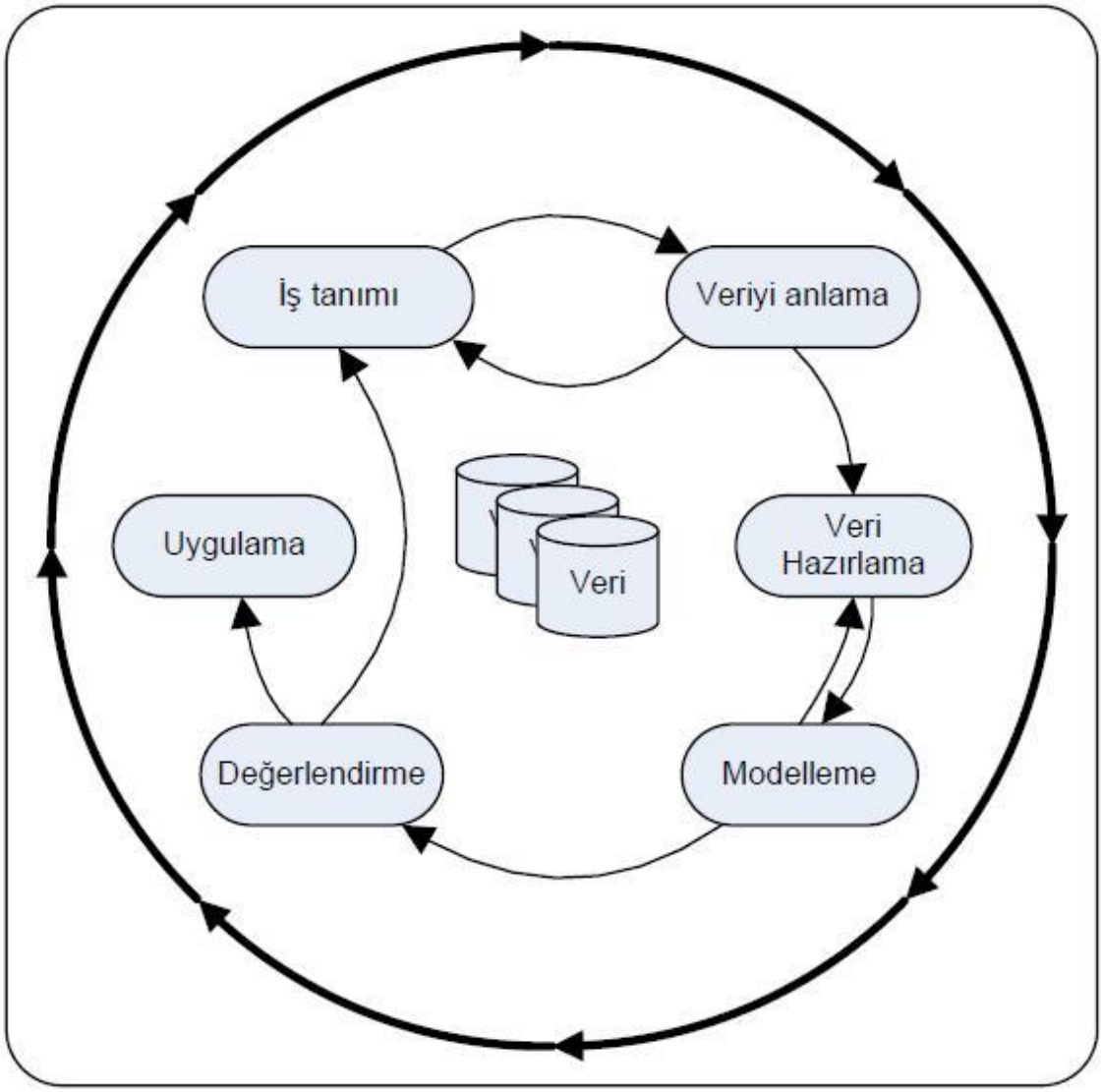
[44] Aydın A.g.e., sf:15

Uygulama süreci, yerine getirilmesi gereken görevler ve bu görevler arasındaki ilişkileri içerir. CRISP-DM tarafından önerilen uygulama süreç adımları Şekil 1-4'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 madenciliği süreci önceki süreçlerde elde edilen tecrübelerden faydalanmaktadır.[45]

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.[46]

[45] Aydın, A.g.e., sf:15

[46] Aydın A.g.e., sf:16



Şekil 1-4: CRISP-DM Veri Madenciliği Uygulama Süreci

Kaynakça: Aydın S., "Veri Madenciliği ve Anadolu Üniversitesi Uzaktan Eğitim Sisteminde Bir Uygulama", Yayınlanmamış Doktora Tezi, Eskişehir, 2007, sf:16

Bir diğer veri madenciliği uygulama süreci Two Crows şirketi tarafından önerilmiştir. 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.[47]

[47] Aydın, A.g.e., sf:16

- Problemin tanımlanması
- Veri madenciliği veritabanın oluşturulması
- Verinin incelemesi
- Model için veri hazırlama
- Modelin oluşturulması
- Modelin değerlendirilmesi
- 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.[48]

1.9.1- Problemin tanımlanması

Veri madenciliği çalışmalarında başarılı olmanın en önemli şartı, projenin hangi işletme amacı için yapılacağının açık bir şekilde tanımlanmasıdır. İlgili işletme amacı işletme problemi üzerine odaklanmış ve açık bir dille ifade edilmiş olmalı, elde edilecek sonuçların başarı düzeylerinin nasıl ölçüleceği tanımlanmalıdır. Ayrıca yanlış tahminlerde katlanılacak olan maliyetlere ve doğru tahminlerde kazanılacak faydalara ilişkin tahminlere de bu aşamada yer verilmelidir.[49]

Bu aşamada mevcut iş probleminin nasıl bir sonuç üretilmesi durumunda çözüleceğinin, üretilecek olan sonucun fayda - maliyet analizinin başka bir değişle üretilen bilginin işletme için değerinin doğru analiz edilmesi gerekmektedir. Analistin işletmede üretilen sayısal verilerin boyutlarını, proje için yeterlilik düzeyinin iyi analiz edilmesi gerekmektedir. Ayrıca analistin işletme konusu hakkındaki iş süreçlerinin de iyi analiz edilmesi gerekmektedir.[50]

---

[48] Aydın A.g.e., sf:17

[49] Tiryaki, A.g.e., sf:24

[50] Tiryaki, A.g.e., sf:24

1.9.2- Veri madenciliği veritabanın oluşturulması

Burada kullanılacak verinin kalitesi sonuçları da etkileyeceğinden kullanılacak verilerin öncelikle ön işlemden geçirilmesi büyük bir önem taşımaktadır. Sonuçta kaliteli verilerden ancak kaliteli çıktılar elde edilebilecektir. Bu nedenle verilerin kalitesini arttırmanın yolu, verilerin ön işlemden geçirilmesidir.[51]

Modelin kurulması aşamasında ortaya çıkacak sorunlar, bu aşamaya sık sık geri dönülmesine ve verilerin yeniden düzenlenmesine neden olacaktır. Bu durum verilerin hazırlanması ve modelin kurulması aşamaları için, bir analistin veri keşfi sürecinin toplamı içerisinde enerji ve zamanının %50 - %85'ini harcamasına neden olmaktadır. Bu aşamada firmanın mevcut bilgi sistemleri üzerinde ürettiği sayısal bilginin iyi analiz edilmesi, veriler ile mevcut iş problemi arasında ilişki olması gerektiği de unutulmamalıdır. Proje kapsamında kullanılacak sayısal verilerin, hangi iş süreçleri ile elde edildiği de bu veriler kullanılmadan analiz edilmelidir. Bu sayede analist veri kalitesi hakkında fikir sahibi olabilir. Verilerin hazırlanması aşaması kendi içerisinde toplama, değer biçme, birleştirme ve temizleme, seçme ve dönüştürme adımlarından oluşmaktadır.[52]

1.9.2.1- 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. Veri depolama ve veri sistemleri aşağıda yer alan bölümlerde açıklanmıştır. Günümüzde veri depolama ve yönetim sistemlerinin uygulandığı yazılımların listesi aşağıdaki tabloda verilmiştir.[53]

---

[51] Altıntaş, A.g.e., sf: 10

[52] Altıntaş, A.g.e., sf:10

[53] Aydın A.g.e., sf:4

Tablo 1-9: 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.

Kaynakça: Aydın S., "Veri Madenciliği ve Anadolu Üniversitesi Uzaktan Eğitim Sisteminde Bir Uygulama", Yayınlanmamış Doktora Tezi, Eskişehir, 2007, sf:143

1.9.2.1.1- Metin Dosyaları ve İşlem Tabloları

Metin dosyaları en eski veri saklama ve depolama yöntemidir. Düz dosya yapısına sahip olduklarından dolayı ilişkisel alanların tespitinin zor olmasından dolayı pek çok veri madenciliği uygulamalarında kullanılamamaktadır. Sadece küçük boyut ve hacimli veri madenciliği uygulamaları için kullanılabilmektedir.[54]

1.9.2.1.2- Veri Tabanı Sistemleri

Veri Tabanı Yazılımı verileri sistematik bir biçimde depolayan yazılımlara verilen isimdir. Birçok yazılım bilgi depolayabilir ama aradaki fark, veri tabanın bu bilgiyi verimli bir şekilde düzenleyebilmesi ve ona hızla ulaşabilmesidir. Bilgiye gerekli olduğu zaman ulaşabilmek esastır. Veri tabanı içinde düzenlenmemiş bilgiler, katalogu olmayan bir kütüphaneye benzetilebilir. İmkânlarının sağlandığı, bilgilerin bütünlük içerisinde tutulabildiği ve birden fazla kullanıcıya aynı anda bilgiye erişim imkânının sağlandığı programlardır.[55]

[54] Aydın, A.g.e., sf:19

[55] http://tr.wikipedia.org/wiki/Veri_taban%C4%B1#Veri_modelleme (Erişim Tarihi: 04.07.2011)

Günümüzde veri tabanı sistemleri bankacılıktan otomotiv sanayisine, sağlık bilgi sistemlerinden şirket yönetimine, telekomünikasyon sistemlerinden hava taşımacılığına, çok geniş alanlarda kullanılan bilgisayar sistemlerinin alt yapısını oluşturmaktadır. Veri tabanı fiziksel olarak bilgileri tutarken mantıksal bir sisteme de sahiptir. Veritabanı sistemlerinin kurulumu, konfigürasyonu, tasarımı, sorgulaması, güvenliği ve denetiminin karmaşık bir hal alması veri tabanı yöneticiliği kavramının oluşmasına neden olmuştur.[56]

Veri tabanı yönetim sistemi yazılımları :[57]

- Filemaker
- MySQL
- PostgreSQL
- Oracle
- Sybase
- MsSQL
- Berkeley
- Firebird
- Ms access
- OpenOffice.org Veritabanı
- 1Ç İşletme - Açık kaynak kodlu, görsel veri tabanı geliştirme ortamı

Veri tabanlarını 9 başlık altında toplayabiliriz.

1. İlişkisel Veri Tabanları
2. Veri Ambarları
3. Transactional (İşlemsel) Veri Tabanları
4. Gelişmiş Veri Tabanı Sistemleri ve Uygulamaları
5. Nesneye Yönelik Veri Tabanları
6. Nesne İlişkisel Veri Tabanları
7. Uzaysal Veri Tabanları

[56] http://tr.wikipedia.org/wiki/Veri_taban%C4%B1#Veri_modelleme (Erişim Tarihi: 04.07.2011)

[57] http://tr.wikipedia.org/wiki/Veri_taban%C4%B1#Veri_modelleme (Erişim Tarihi: 04.07.2011)

8. Time Series-Temporal Veri Tabanları

9. Text ve Multimedya Veri Tabanları

1.9.2.1.3- 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. OLAP uygulamaları genellikle gerçek verilerin analizini içerir. Bu uygulamalar SQL'deki mevcut temel 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.[58]

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.[59]

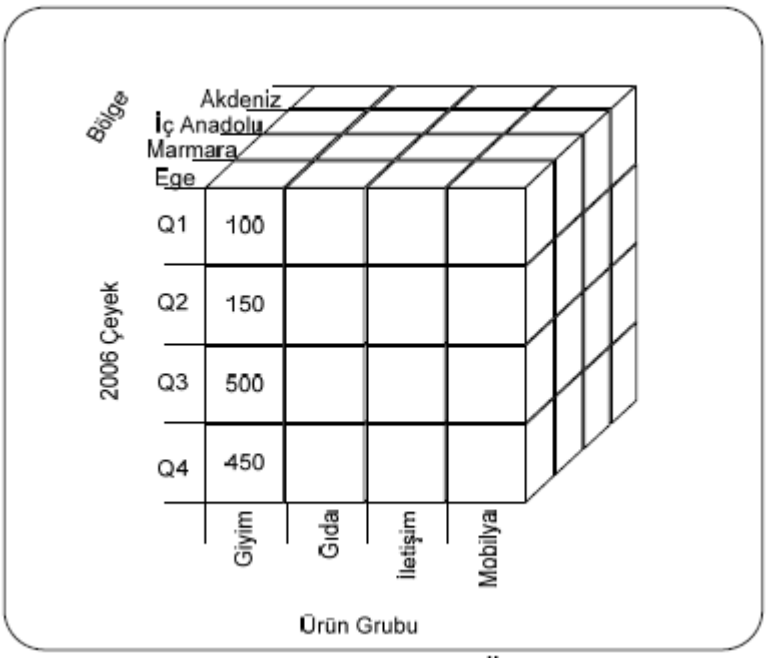
Veri ambarları ve OLAP araçları çok boyutlu veri modeline dayalıdır. Çok boyutlu veri modeli, verileri bir küp şeklinde ele alır. Bir veri küpü verilerin çok boyutlu olarak yapılandırılmasını ve görüntülenmesini sağlar. Örneğin bir şirketin satışlarına ilişkin veri küpü oluşturulduğunu düşünelim. Veri küpünün ana temasının satış tutarları olduğunu, boyutlarının ise satış bölgesi, zaman ve ürün grubu olduğunu varsayalım. Bu durumda Şekil 1-5'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.[60]

---

[58] Aydın, A.g.e. sf:22

[59] Aydın , A.g.e. sf:23

[60] Aydın , A.g.e. sf:23



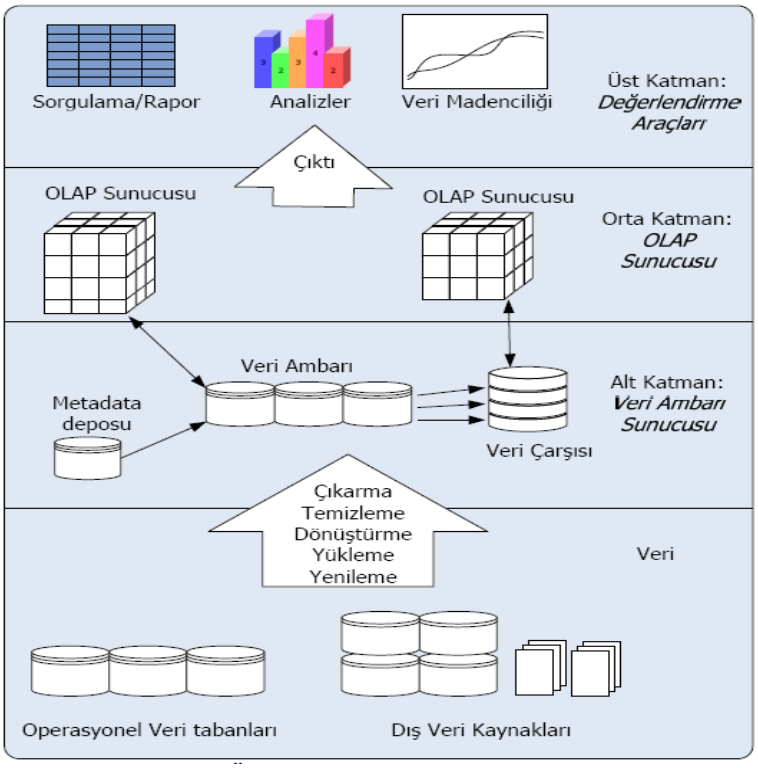
Şekil 1-5: Veri Küpü Örneği

Kaynakça: Aydın S., "Veri Madenciliği ve Anadolu Üniversitesi Uzaktan Eğitim Sisteminde Bir Uygulama", Yayınlanmamış Doktora Tezi, Eskişehir, 2007, sf:23

OLAP ve Veri Küpünü tanımladıktan sonra veri ambarı mimarisi ve kullanım şeklini tanımlamak faydalı olacaktır. Şekil 1-6'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.[61]

[61] Aydın, A.g.e., sf:24

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.[62]



Şekil 1-6: Üç Katmanlı Veri Ambarı Mimarisi

Kaynakça: Aydın S., "Veri Madenciliği ve Anadolu Üniversitesi Uzaktan Eğitim Sisteminde Bir Uygulama", Yayınlanmamış Doktora Tezi, Eskişehir, 2007, sf:25

[62] Aydın, A.g.e., sf:24

1.9.2.2- 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.

- Tabloda yer alan alanların sayısı
- Eksik değerler yer alan kayıtların sayısı ve yüzdesi
- Alan isimleri
- Veri türü
- Açıklaması
- 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"[63]

1.9.2.3- 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.[64]

[63] Aydın A.g.e., sf:25

[64] Aydın A.g.e., sf:26

1.9.2.4- Veri Kalitesini İyileştirme ve Ön Hazırlık Süreçleri

Günümüzde veritabanları büyük boyutlarından ve birçok farklı kaynaktan gelmelerinden dolayı gürültülü, eksik, tutarsız veriler ile doludur. Verilerin kalitesiz olması veri madenciliğinden elde edilen sonuçların da kalitesiz olmasına yol açabilir. Veri Madenciliğinde veri kümesinin büyüklüğünden kaynaklanan en fazla zaman alıcı aşama, verilerin ön işlemden geçirilmesi aşamasıdır. Veri Madenciliği uygulamalarında kaynakların %80'i verilerin ön işlemden geçirilmesi ve temizlenmesi süreçleri için harcanmaktadır.[65]

1.9.2.4.1- 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. Veri temizleme yoluyla eksik değerler tamamlanarak, gürültülü veri düzeltilerek, aykırı değerler tanımlanarak veya çıkarılarak ve tutarsızlıklar giderilerek veri kalitesi arttırılmaya çalışılır.[66]

Veri temizleme için verinin özelliklerini bilmek gerekir. Buna üst veri (metadata) denir. Bir başka ifade ile verinin içeriği hakkındaki veriye üst veri denir. Her özelliğin alabilecek değerleri ve uzunlukları bilinmelidir.[67]

1.9.2.4.2- Eksik Veri

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

[65] Bilekdemir G., "Veri Madenciliği Tekniklerini Kullanarak Üretim Süresi Tahmini ve Bir Uygulama", Yayınlanmamış Yüksek Lisans Tezi, İzmir, 2010, sf:10

[66] Aydın A.g.e., sf:27

[67] Bilekdemir, A.g.e., sf:10

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.[68]

Örneğin, satış ve müşteri verilerinin analizi yapıldığını düşünelim. Veri setinde müşteri geliri kaydının tutulmadığı durumda bu özellik için eksik veri nasıl tamamlanır? Herhangi bir değişkene ilişkin eksik değerlerin doldurulması için farklı yollar vardır. Bunlardan bazıları aşağıda kısaca açıklanmaktadır:[69]

- Eksik değer içeren kayıt veya kayıtlar atılabilir. Bu metot genellikle sınıf etiketi eksik olduğu durumda yapılır. Bu metot satır birden fazla özellik eksik veri içermediği sürece verimli değildir.
- Eksik veri manüel olarak tamamlanabilir. Bu metot zaman alıcı bir yöntemdir ve büyük veri setlerinde uygulanabilir değildir.
- Eksik veri genel bir sabit ile doldurulur. Bütün eksik veriler "Bilinmiyor", "∞" gibi aynı sabitle doldurulur. Bu yöntemde Veri Madenciliği yazılımı verilerin hepsinin ortak "Bilinmiyor" verisini içerdiği sonucunu çıkarabilir.
- Aynı sınıfa ait tüm örneklemler için değişkenin ortalaması kullanılabilir. Örneğin aynı kredi risk kategorisine giren müşteriler için ortalama gelir değeri eksik değerler yerine kullanılabilir.
- Eksik değer ortalama değer ile doldurulur. Örneğin müşteri geliri eksik ise tüm müşterileri gelirlerinin ortalaması eksik değere yazılır.
- Var olan verilere dayalı olarak en uygun değer kullanılabilir. Burada sözü edilen en uygun değerin belirlenmesi için regresyon veya karar ağacı gibi teknikler kullanılabilir. Örneğin yaşı x, eğitim düzeyi y olan bir kişi için ücret durumu, mevcut verilerden yukarıdaki tekniklerden birinin kullanılmasıyla tahmin edilebilir.

[68] Aydın, A.g.e., sf:28

[69] Bilekdemir, A.g.e., sf:11

1.9.2.4.3- 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.[70]

Veri düzeltme işlemi için çeşitli teknikler bulunmaktadır. Bunlardan bazıları:[71]

- **Binning:** Binning yöntemleri, küçükten büyüğe veya büyükten küçüğe sıralanmış verileri düzeltmek için kullanılır. Binning yönteminde öncelikle sıralanmış veriler eşit büyüklükteki bin'lere ayrılır. Daha sonra bin'ler, bin ortalamaları, bin medyanları veya bin sınırları yardımıyla düzeltilir.
- **Kümeleme (Clustering):** Aykırı değerler kümeler ile belirlenebilir. Benzer değerler aynı grup veya küme içinde yer alırken, aykırı değerler kümelerin dışında yer alacaktır.
- **Regresyon (Regression):** Veriler regresyon ile verilere bir fonksiyon uydurularak düzeltilebilir. Uydurulan fonksiyona uymayan noktalar aykırı değerlerdir.

1.9.2.4.4- Tutarsız Veri

Bazen kaydedilen veriler arasında veya veri birleştirme esnasında tutarsızlıklar olabilmektedir. Bir değişken ismi farklı veritabanlarında farklı şekilde olabilir. Bu veri tutarsızlıkları dış referanslar kullanılarak düzeltilebilir. Örneğin kodların kullanımındaki tutarsızlıklar düzeltilebilir.[72]

---

[70] Aydın, A.g.e., sf:30

[71] Bilekdemir, A.g.e., sf:12

[72] Koldere Akın, A.g.e., sf:61

1.9.2.5- Veri Birleştirme

Veri madenciliğinde bazen farklı veri tabanlarındaki verilerin birleştirilmesi gerekebilir. Farklı veri tabanlarındaki verilerin tek bir veri tabanında birleştirilmesiyle şema birleştirme hataları oluşur. Örneğin, bir veri tabanında girişler "tüketici-ID" seklinde yapılmışken, bir diğerinde "tüketici-numarası" şeklinde olabilir. Bu tip şema birleştirme hatalarından kaçınmak için meta veriler kullanılır.[73]

1.9.2.6- 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. Veri dönüşümünde verilerin veri madenciliği için uygun formlara dönüştürülmesi düzeltme, bir araya getirme, genelleme, normalleştirme ve özellik oluşturma işlemleriyle gerçekleştirilir. Bunları kısaca açıklarsak:[74]

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

- **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

[73] Kayaalp, A.g.e., sf:31

[74] Aydın, A.g.e., sf:32

madenciliği terminolojisinde her iki terim birbiri yerine kullanılmaktadır. Ancak buradaki normalleştirme terimi, istatistikte kullanılan bir değişkenin normal dağılmış bir değişkene dönüştürülmesi ile karıştırılmamalıdır. 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.

1.9.2.7- Veri Azaltma

Veri seti çok büyük olduğunda veri madenciliği uygulamaları çok zaman almakta hatta bazı algoritmaların uygulanması oldukça zorlaşmaktadır. Orijinal veri setinin özelliklerini koruyan ancak hacim olarak daha küçük veri seti elde etmek için veri indirgeme teknikleri kullanılmaktadır. Böylece indirgenmiş veri seti üzerinde veri madenciliği daha etkili olmakta ve hemen hemen orijinal veri seti ile aynı analitik sonuçları vermektedir. Veri indirgeme tekniklerini şu şekilde sıralamak mümkündür.[75]

- **Veri Küpü Birleştirme:** Veri küplerinde çeyrekler bazında ya da aylar bazında satışların 6 aylık veya yıllık satışlar olarak gösterilmesidir.

- **Boyut İndirgeme:** Burada ilgisiz, az ilgili veya gereksiz olan değişkenlerin kaldırılmasıdır. Karar ağaçları boyut indirgeme amacıyla kullanılmaktadır.

[75] Koldere Akın, A.g.e., sf:64-65

- **Veri Sıkıştırma:** Burada veri seti büyüklüğünü azaltmak amacı ile veri şifreleme veya veri dönüşümü kullanılmaktadır. Temel Bileşenler analizi sıkıştırılmış veri seti elde edilmesinde kullanılmaktadır.

- **Kesikleştirme:** Bazı veri madenciliği algoritmaları sadece kategorik verileri dikkate aldığından, sürekli, verilerin kesikli değerlere dönüştürülmesi gerekir. Böylece elde edilen kategorik değerler, orijinal veri değerlerinin yerine kullanılırlar.

1.9.3- 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 zekâ, 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.[76]

Veri madenciliği modelleri işlevlerine göre 3 temel grupta toplanır:[77]

- Kümeleme (Clustering),
- Birliktelik kuralları ve sıralı örüntüler (Association rules and sequential patterns).
- Sınıflama (Classification),

[76] Aydın, A.g.e., sf:49

[77] Akbulut, A.g.e., sf: 20

1.9.3.1- Kümeleme

Kümeleme analizi, nesnelerin altdizinlere gruplanmasını yapan bir işlemdir. Böylece nesneler, örneklenen kitle özelliklerini iyi yansıtan etkili bir temsil gücüne sahip olmuş olur. Sınıflamanın aksine, yeniden tanımlanmış sınıflara dayalı değildir. Kümeleme, denetimsiz bir öğrenme (unsupervised learning) yöntemidir.[78]

Hiyerarşik olmayan kümelemede, kümeler arasında ilişki bulunmamaktadır. Örneğin, 'k-ortalamalar', hiyerarşik olmayan bir kümeleme algoritmasıdır. Hiyerarşik kümelemede, her kümede veri nesnelerini içerecek bir bağlantı kurulur. Hangi yöntem olursa olsun kümeler birbirine benzer özellik gösteren nesnelerden oluşturulur. Böylece kümeler kendi içinde aynı özelliği taşıyan nesneleri içermiş olur. Manhattan ve Euclid uzaklık fonksiyonları çoğunlukla benzerliklerin bulunmasında kullanılır. Uzaklık fonksiyonunun sonucu yüksek bir değer ise az benzerlik, düşük bir değer ise çok benzerlik olduğunu ifade eder. p-boyutlu veri nesneleri i:(x_{i1},x_{i2},.....,x_{ip}), j:(x_{j1},x_{j2},.....,x_{jp}) için aşağıda verilen uzaklık fonksiyonları tanımlanabilir.[79]

Euclid uzaklık fonksiyonu:

$$d_{ij} = \sqrt{|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \cdots + |x_{ip} - x_{jp}|^2}$$

Manhattan uzaklık fonksiyonu:

$$d_{ij} = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{ip} - x_{jp}|$$

Uzaklıkları karşılaştırmak için Euclid uzaklık fonksiyonu kullanıldığında, denklemin sağ tarafındaki karekökü hesaplamak gereksizdir. Çünkü uzaklıklar her zaman pozitif sayılardır ve bundan dolayı, d_1 ve d_2 gibi iki uzaklık için, $\sqrt{d_1} > \sqrt{d_2}$ iken $d_1 > d_2$'dir. Bir nesnenin bazı özellikleri farklı ölçeklerde ölçülüyorsa, Euclid uzaklık fonksiyonu kullanılarak büyük ölçeklerle ölçülen nitelikler küçük bir ölçekte ölçülen niteliklere baskın gelebilir. Böyle bir sorundan kaçınabilmek için, nitelik

[78] Altıntaş, A.g.e., sf: 18

[79] Altıntaş, A.g.e., sf:19

değerleri çoğunlukla 0 ile 1 arasında normalleştirilir. Veri kümeleri için uygulanacak uzaklık fonksiyonlarının verimleri farklı olabilir, bundan dolayı Euclide ve Manhattan'ın haricindeki uzaklık fonksiyonları bazı veri kümeleri için daha uygun olabilir. VM'de kullanılmakta olan birçok kümeleme algoritması vardır ve bunlar analiz edilecek olan verinin yapısına göre belirlenir. Altıntaş'a göre, "kümeleme metotları" genel olarak şunlardır:[80]

- **Bölümlendirme Metodu:** n tane nesnenin olduğu veritabanında, nesneler mantıksal gruplara ayrılarak analiz edilir. Küçük ve orta boyutlu veritabanlarında birkaç grup olabilirken, veritabanının büyüklüğü arttığında daha çok grup oluşabilir. Gruplandırma yapılırken değişik kriterler değerlendirilebilir. Yapılan gruplandırma analizin kalitesine etki eder. En çok bilinen bölümlendirme algoritmaları

✓ **K-means algoritması**
✓ **K-medoids algoritması**
✓ **EM algoritması**

- **Hiyerarşik Metot:** Hiyerarşik kümeleme nesnelerin yakınlık ilişkisine göre oluşturulan kümelerden bir ağaç inşa eder. Analiz etmeden önce nesneler, hiyerarşik bir yapıya göre düzenlenir. Veriyi hiyerarşik bir yapıya çevirmek için değişik yöntemler kullanılır. Bunların arasında BIRCH ve CURE yöntemleri bulunur. Hiyerarşik kümeleme algoritmaları

✓ **Toplayıcı (Agglomerative) hiyerarşik algoritması**
✓ **Bölücü (Divisive) hiyerarşik algoritması**

- **Yoğunluk Bazlı Metot:** Birçok kümeleme yöntemi nesnelerin birbirleri arasındaki farklılıklarına göre kümeleme yaparken, bu metot nesnelerin yoğunluğuna göre gruplama yapar. Yoğunluktan kasıt, analiz edilen nesnelerin sayısıdır. Yoğunluk bazlı metotlara örnek olarak DBscan verilebilir.

[80] Altıntaş, A.g.e., sf: 20

- **Grid Bazlı Metot:** Grid temelli kümeleme yaklaşımı çok çözümlü grid veri yapısını kullanır. Kümeleme yapılacak alanın sonlu sayıda hücrelere bölünmesiyle oluşur. Ana avantajı genelde birbirinden bağımsız sayıda veri nesnelerinde hızlı işlem zamanıdır. Grid temelli yaklaşımın bazı genel örnekleri: STING, grid hücrelerindeki istatiksel bilgiyi araştırır; Wavecluster, wavelet dönüşüm metodunu kullanan nesneleri kümeler; CLIQUE, yüksek boyutlu veri alanlarını kümelemek için grid ve yoğunluk temelli yaklaşımı temsi eder.

- **Model Bazlı Metot:** Her küme için bir model belirlenir ve bu modele uyan veriler uygun kümeye yerleştirilir. Model tabanlı kümeleme metotları, verilen veri ile bazı matematiksel modeller arasında uygunluğu optimize etmeye çalışır. Bu metotlar genelde olasılık dağılımlarına göre varsayımlar üretir. Model tabanlı kümeleme metotlarının 2 ana yaklaşımı vardır:

✓ **İstatiksel Yaklaşım ve COBWEB**

✓ **Nöral Network Yaklaşımı**

1.9.3.2- Birliktelik kuralları ve sıralı örüntüler (Sepet Analizi)

Birliktelik kuralları madenciliği, büyük veri yığınları arasındaki ilginç ilişkileri ya da ilişki bağıntılarını yakalar. Amaç, veritabanındaki ilgili veri gruplarından, "$A_1 \Lambda A_2 \Lambda ... \Lambda A_m \Rightarrow B_1 \Lambda B_2 \Lambda ... \Lambda B_n$" ($A_m$ ve B_n değer kümelerini ifade eder) şeklinde birliktelik kuralları çıkarmaktır. İlişki kuralları analizinin uygulandığı tipik bir alan sepet analizi alanıdır. Bu tür uygulamalar, müşterilerin sepetlerinde yer alan farklı ürünlerden yola çıkarak; müşterilerin alışveriş alışkanlıklarının yakalanabilmesini hedefler. Birliktelik kuralları, iki olayın aynı anda ortaya çıkma durumunu ölçen kümeleme analizinde kullanılan eşleştirme yöntemine benzer. Bu ilişkileri değerlendirirken genellikle kullanılan iki ölçüt vardır. İlki olan "güven", B olayı olduğu zaman A olayının da olma olasılığı olarak hesaplanır. "Destek" terimi ise, kesişen olayların ortaya çıkma sayısının toplam olay sayısı içindeki yüzdesini ifade eder. Bir ilişkiyi kural olarak belirlemek için eldeki olay kümesinin önemli bir kısmı tarafından desteklenmesi gerekir. Bu nedenlerden dolayı, $X \Rightarrow Y$ ilişki kuralı kullanıcı tarafından minimum değeri belirlenmiş güvenirlik ve destek eşik değerlerini sağlayacak biçimde

üretilir. İlişki kurallarını tespit etmek için büyük veri yığınları üzerinde, tekrar eden sorgular yapılması gerektiğinden, işlem miktarı çok fazla olur ve performans konusunda iyileştirme yapmak gerekliliğinin ortaya çıkması muhtemeldir.[81]

Birliktelik kuralına ilişkin olarak geliştirilen bazı algoritmalardan en çok bilineni Apriori algoritmasıdır. Birliktelik kurallarını bulmak için genellikle problem iki parçaya bölünür. Önce sık tekrarlanan öğeler bulunur. Bu öğelerin her biri en az, önceden belirlenen minimum destek sayısı kadar sık tekrarlanırlar. Sonra sık tekrarlanan öğelerden güçlü birliktelik kuralları oluşturulur. Bu kurallar minimum destek ve minimum güven değerlerini karşılamalıdır.[82]

1.9.3.3- Sınıflama

Sınıflandırma veri madenciliğinin en çok kullanıldığı alandır. Amacı yeni bir nesnenin özelliklerini açıklamak ve bu yeni nesnenin daha önceden tanımlanmış sınıf setlerinden birine atamasını yapmaktır. Sınıflandırmanın yapılabilmesi için sınıflandırılmamış veriye uygun bir model uygulanabilir. Veri madenciliğinin sınıflandırma grubu içerisinde en sık kullandığı teknik karar ağaçlarıdır. Aynı zamanda lojistik regresyon, diskriminant analizi, sinir ağları ve fuzzy setleri de kullanılmaktadır. İnsanlar verileri daima sınıflandırdıkları, kategorize ettikleri ve derecelendirdikleri için sınıflandırma, hem veri madenciliğinin temeli olarak hem de veri hazırlama aracı olarak da kullanılabilmektedir.[83]

Temel sınıflama algoritmaları aşağıdadır:

- Diskriminant analizi,
- Naive Bayes,
- Karar ağaçları,
- Sinir ağları,
- Kaba kümeler,

[81] Kolay G., "İşletmelerde Bilgi Sistemleri Verimliliğini Arttırmada Veri Madenciliği Yöntemi: Bir Simülasyon Çalışması", Yayınlanmamış Yüksek Lisans Tezi, Zonguldak, 2006, sf:40

[82] Bilekdemir, A.g.e., sf:22

[83] Koldere Akın, A.g.e., sf:55

- Genetik algoritma,
- Regresyon analizi.

1.9.3.3.1- Diskriminant analizi

Diskriminant fonksiyonları aracılığı ile gruplar arası ayırıma en fazla etki eden ayırıcı değişkenleri belirlemede ve hangi gruptan geldiği bilinmeyen bir bireyin hangi gruba dahil edileceğini belirlemede kullanılır. Genel anlamda ayırma olup, bireylere ait p tane özellikten yararlanarak ait oldukları grupları (kütle) belirlemede veya mevcut grupları birbirinden ayıracak en iyi fonksiyonu bulmada kullanılan çok değişkenli istatistik tekniklerinden birisidir. Genel olarak birimlerin gruplanmasında bazı matematiksel eşitliklerden faydalanılır. Diskriminant fonksiyonu olarak adlandırılan bu eşitlikler birbirine en çok benzeyen grupları belirlemeye olanak sağlayacak şekilde grupların ortak özelliklerini belirlemek amacıyla kullanılmaktadır. Grupları ayırmak amacıyla kullanılan karakteristikler ise diskriminant değişkenleri olarak adlandırılmaktadır. Kısaca, iki veya daha fazla sayıdaki grubun farklılıklarının diskriminant değişkenleri vasıtasıyla ortaya konması işlemidir. Araştırıcı, hatalı sınıflandırma olasılığını en aza indirgeyerek gözlemleri ait oldukları gruplara ayırmak veya bu gözlemlerin çekilmiş oldukları grupları belirlemek isteyecektir. Belirlenecek grupların ortalamaları arasındaki farklılığın maksimum olması amaçlanmaktadır.[84]

İşlevi, grup farklılıklarını anlamak ve bir varlığın (birey, nesne) belirli bir sınıfa veya birkaç metrik bağımsız değişkene dayalı gruba ait olması olasılığını tahmin etmektir. Gruplar arası farklılığa etki eden tahmin değişkenlerinin hangileri olduğunu ortaya çıkarır.[85]

1.9.3.3.2- Naive Bayes

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 edebilirler. Veri

[84] Cangül O., "Diskriminant Analizi ve Bir Uygulama Denemesi", Yayınlanmamış Yüksek Lisans Tezi, Uludağ Üniversitesi, Bursa, 2006, sf:5

[85] Cangül, A.g.e., Sf:5

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.[86]

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.[87]

Naive Bayes, sürekli veri ile çalışmaz. Bu nedenle sürekli değerleri içeren bağımlı ya da bağımsız değişkenler kategorik hale getirilmelidir. Örneğin; bağımsız değişkenlerden biri yaş ise, sürekli değerler "<20" "21-30", "31-40" gibi yaş aralıklarına dönüştürülmelidir. Naive Bayes, modelin öğrenilmesi esnasında, her çıktının öğrenme kümesinde kaç kere meydana geldiğini hesaplar. Bulunan bu değer, öncelikli olasılık olarak adlandırılır. Örneğin; bir banka kredi kartı başvurularını "iyi" ve "kötü" risk sınıflarında gruplandırmak istemektedir. İyi risk çıktısı toplam 5 vaka içinde 2 kere meydana geldiyse iyi risk için öncelikli olasılık 0,4'tür. Bu durum, "Kredi kartı için başvuran biri hakkında hiçbir şey bilinmiyorsa, bu kişi 0,4 olasılıkla iyi risk grubundadır" olarak yorumlanır Naive Bayes aynı zamanda her bağımsız değişken / bağımlı değişken kombinasyonunun meydana gelme sıklığını bulur. Bu sıklıklar öncelikli olasılıklarla birleştirilmek suretiyle tahminde kullanılır.[88]

[86] Aydın, A.g.e., sf:54

[87] Aydın, A.g.e., Sf:54

[88] Akbulut, A.g.e., sf: 22

1.9.3.3.3- Karar ağaçları

Karar ağaçları, belirli bir veri setindeki kümeleri hiyerarşik olarak alt kümelere ayırmaktadır. Karar ağaçları sınıflama amaçlı uygulamalarda da oldukça çok kullanılan tekniklerden birisidir. Tahmin edici ve tanımlayıcı özelliklere sahip olan karar ağaçları VM projelerinde

- Meydana getirilmelerinin ucuz olması,
- Yorumlamalarının daha kolay, basit olması,
- Veri tabanı sistemleri ile kolayca entegre olabilmeleri,
- Güvenilirliklerin daha yüksek olması

gibi nedenlerden dolayı sınıflama modelleri arasında en yaygın kullanıma sahip olan tekniktir.[89]

Çok basit bir alt yapıya sahip olmalarına rağmen önemli bilgileri uygulayıcılara sunma imkânına sahip olan karar ağaçları genellikle kümeleme uygulamalarında kullanılmaktadır. Diğer tekniklerle karşılaştırıldığında karar ağaçlarının yorumlanması, anlaşılması ve yapılandırılması daha kolaydır. Bu teknikte sınıflama yapılırken ilk önce veri setinden bir ağaç meydana getirilir. Bu ağaç meydana getirildikten sonra veri setindeki her bir kayıt bu ağaca uygulanarak bu kayıt sınıflandırılır.[90]

Karar ağacı meydana getirilirken, mevcut veri setindeki verilerin bir kısmı öğrenme amaçlı olarak kullanılacak ve bunlardan yola çıkarak karar ağacı oluşturulacaktır. Veri setinin bir kısmı ise meydana getirilen bu karar ağacını test etmek üzere kullanılmaktadır. Ağaç meydana getirilirken, kurulan sistemin doğru şekilde çalışıp çalışmadığı belirlenerek istenen seviyede olduğu anlaşılırsa ağacın dallanmasına son verilir. Durdurma olarak isimlendirilen ağacın dallanmasının sonlandırılması ağacın hassasiyetini gösterir. Geç durdurulan karar ağacı daha fazla dallanacak bu ise istenmeyen sonuçların ortaya çıkmasına ya da önemli ilişkilerin, sınıfların, örüntülerin gözden kaçmasına sebep olabilecektir. Erken durdurulan ağaç ise tam öğrenmenin gerçekleşmemiş olma ihtimalini taşıyabilecektir. Karar ağacı meydana getirilirken ya da ağaç kurulumu tamamlandıkta sonra budama olarak adlandırılan, ağaçta oluşmuş, karar

[89] Akpınar, A.g.e., S: 1-22.

[90] Kiremitçi B., "Veri Ambarlarında Veri Madenciliği ve Ulaştırma-Lojistik Sektöründe Bir Uygulama", Yayınlanmamış Yüksek Lisans Tezi, İstanbul, 2005. Sf:62

verirken sonucu etkilemeyen, sınıflama için bir yararı bulunmayan dalların ağaçtan alınması işlemi yapılabilmektedir.[91]

En yaygın kullanılan karar ağacı algoritmaları;

- CHAID (Chi-Squared Automatic Interaction Detector , Kass 1980),
- C&RT (Classification and Regression Trees, Breiman ve Friedman, 1984),
- ID3 (Induction of Decision Trees, Quinlan, 1986),
- C4.5 (Quinlan, 1993).

1.9.3.3.4- Sinir ağları

İlk kez 1943'te ortaya çıktı ama bilgisayarlarda kullanımı 1980'lerde başladı. **Yapay sinir ağları** (artificial neural networks), beynin yapısından esinlenilmiş bir bilgi işleme sistemidir. Nöronlara benzeştirilmiş işlem öğeleri arasındaki ilişkilerle yapılandırılmıştır. İnsan beyni gibi yapay sinir ağı da birbirine bağlı birçok işlem biriminden oluşmuştur. Sonra, birçok düğüm (işlem birimi) ve arkla (iç bağlantılar) yönetilen bir grafik olarak yapılandırılır. Bu işlem birimleri birbirlerinden bağımsız işlev görürler ve yalnızca yerel veriyi (düğüme gelen girdi ve düğümden çıkan çıktı) kullanırlar. Bu özellik, sinir ağlarının dağıtık ya da paralel ortamlarda kullanımını kolaylaştırır. Sinir ağları, kaynak (girdi), çıktı ve iç (gizli) düğümlerle yönetilen bir grafik olarak görülebilir. Girdi düğümü girdi katmanında, çıktı düğümü ise çıktı katmanında bulunur. Gizli düğümler, bir ya da daha çok gizli katmanda bulunur. Veri madenciliğinde, çıktı düğümü tahmini belirler. Tek bir girdi düğümünün olduğu (ağacın kökü) karar ağaçlarından farklı olarak sinir ağlarında, her öznitelik değeri için bir girdi düğümü vardır. Sinir ağları karmaşık sorunları çözebilir, ayrıca temel uygulamalardan "öğrenebilir". Yani, soruna kötü bir çözüm bulunduysa, ağ bu soruna bir dahaki sefer daha iyi bir çözüm bulacak biçimde değiştirilir. Sinir ağları üç bölümden oluşur:[92]

- Sinir ağının veri yapısını tanımlayan sinir ağı grafiği.
- Öğrenmenin nasıl gerçekleşeceğini belirten öğrenme algoritması.
- Bilginin ağdan nasıl elde edileceğini belirleyen teknikler.

---

[91] Kiremitçi, A.g.e., sf:62

[92] Baykal A., "Veri madenciliği: Öğrenci verileri üzerinde uygulamaları", Yayınlanmamış Doktora Tezi, Diyarbakır, 2003

Yapay Sinir Ağları

- Örüntü tanımada
- Ses tanıma ve çözümlemede
- Tıbbi uygulamalarda (tanı, ilaç)
- Hata algılamada
- Sorun tanılamada
- Robot denetiminde
- Herhangi bir işlevi hesaplamada

kullanılabilir.[93]

Yapay Sinir Ağları, bağlantı ve öğrenme türlerine göre sınıflandırılabilir.

- İleri beslemeli bağlantıda bağlantılar yalnızca yapıda daha sonraki katmanlaradır.
- Geri beslemeli bağlantıda ise bazı bağlantılar daha önceki katmanlardır.[94]

Yapay Sinir Ağları öğrenme türleri ise;[95]

- Denetimli (supervised) öğrenme
- Denetimsiz (unsupervised) öğrenme

Denetimli öğrenme, temel olarak iki aşamalı bir işlemdir:[96]

- Sinir ağını, örnek dizileri göstererek verideki farklı sınıfları tanıyacak biçimde eğitmek.
- Önceden görmediği bir veri grubu sağlayarak sinir ağının bu örneklerden ne kadar öğrendiğini denemektir.

Denetimsiz öğrenmede ise sinir ağına, sunulan verinin doğru olarak sınıflandırılmasına ilişkin hiçbir ön bilgi verilmez. Sinir ağı, denetimsiz öğrenmeyi, o veride doğal olarak var olan kümeleri ve altkümeleri bulmak amacıyla çok boyutlu bir

[93] Sıramkaya E., "Veri Madenciliğinde Bulanık Mantık Uygulaması", Yayınlanmamış Yüksek Lisans Tezi, Konya, 2005, sf:21

[94] Sıramkaya, A.g.e., sf:21

[95] Sıramkaya, A.g.e., sf:21

[96] Sıramkaya, A.g.e., sf:21

veri grubunu çözümlemek için kullanır. Sinir ağları denetimsiz öğrenme tekniği, sağlanan verinin yapısını temel alarak kendi sınıflandırma şemalarını tanımlamak için kullanır. Denetimsiz örüntü tanımaya bazen küme çözümlemesi (cluster analysis) de denir.[97]

1.9.3.3.5- Kaba kümeler

Kaba küme teorisi 1970'li yıllarda Pawlak tarafından geliştirilmiştir. Kaba küme teorisinde bir yaklaştırma uzayı ve bir kümenin alt ve üst yaklaştırmaları vardır. Yaklaştırma uzayı, ilgilenilen alanı ayrı kategorilerde sınıflandırır. Alt yakınlaştırma belirli bir altkümeye ait olduğu kesin olarak bilinen nesnelerin tanımıdır. Üst yakınlaştırma ise alt kümeye ait olması olası nesnelerin tanımıdır. Alt ve üst sınırlar arasında tanımlanan herhangi bir nesne ise "kaba küme" olarak adlandırılır.[98]

1.9.3.3.6- Genetik algoritma

Genetik algoritmalar tam olarak 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. 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.[99]

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. Problemlerin çözümünü parametre değerleriyle

[97] Baykal, A.g.e.

[98] Akbulut, A.g.e., sf: 23

[99] Aydın, A.g.e., sf:54

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.[100]

1.9.3.3.7- Regresyon ve Korelasyon Analizi

İstatistik biliminin en önemli konularından birisini regresyon analizi oluşturmaktadır. Regresyon analizi, araştırma, matematik, finans, ekonomi, tıp gibi bilim alanlarında yoğun olarak kullanılmaktadır. Özellikle birçok istatistiksel analizde bir değişkendeki değişikliğin diğer bir değişkeni nasıl etkilediği araştırılmak istenir. Örneğin iktisatta bir ürün için fiyat ve talep, maliyede gelir ve vergi, işletmede üretim ve kâr, tıpta yaş ve hastalığın iyileşmesi, kimyada bir asidin dozajı ve etkisi vs… Regresyon analizinin temelinde; gözlenen bir olay değerlendirilirken, hangi olayların etkisi içinde olduğunun araştırılması yatmaktadır. Bu olaylar bir veya birden çok olacağı gibi dolaylı veya direkt etkileniyor da olabilirler. Regresyon analizi incelenirken, genellikle konusunu oluşturan, etkilendiği olaylara değişkenler adı verilir bu değişkenlerin yer alacağı matematik model incelenir. Değişken, belirli bir zaman aralığı göz önüne alınıp, o zaman aralığında bir kütleyi oluşturan belli birimdeki olayları içeren örneklerdir. Sayılabilir veya ölçülebilir nitelikte olmalıdır. Regresyon analizinde, herhangi bir bağımlı değişkenin bir veya birden fazla bağımsız değişken ile arasındaki ilişki matematiksel bir fonksiyon şeklinde ifade edilir. Bu fonksiyona regresyon denklemi adı verilir. Regresyon denklemi yardımıyla bağımsız değişkenlerin çeşitli değerlerine karşılık bağımlı değişkenin alacağı değer tahmin edilir. Regresyon analizi, bağımsız değişken sayısına göre basit regresyon analizi ve çoklu regresyon analizi, fonksiyon tipine göre lineer regresyon analizi ve nonlineer regresyon analizi olmak üzere sınıflandırılabilir.[101]

---

[100] Aydın, A.g.e., sf:55

[101] Ertan E., "Regresyon Analizi ve Matematik Programlama Arasındaki İlişki", Yayınlanmamış Yüksek Lisans Tezi, İstanbul Üniversitesi, 2008

Y bağımlı ve X_i=1,2,...,k bağımsız değişken(ler) olmak üzere, Y ile X_i arasındaki nedenselliği (neden-sonuç, faktör-cevap) matematiksel model (matematiksel eşitlik) olarak ortaya koyan ve Y ile X_i arasındaki bağıntının önemliliğini belirleyen yönteme **regresyon analizi** adı verilir.[102]

Y ile X_i arasındaki ikili (bivariate) ilişki düzeyini, yönünü ve ilişkinin önemini belirleyen yönteme **korelasyon analizi** adı verilir.[103]

Regresyon ve korelasyon kavramları birbirleri içine girmiş iki kavramdır. Regresyon analizi ve korelasyon analizi, değişkenlerin ölçeklerine, değişken sayılarına ve regresyon modellerine göre parametrik ve parametrik olmayan yaklaşımlar olarak iki temel sınıfta ele alınarak uygulanır. Verilere parametrik regresyon analizi uygulanırken korelasyon analizi de birlikte uygulanır. Korelasyon Analizi tek başına da uygulanabilir.[104]

Korelasyon analizini iki başlık altında toplarsak;[105]

- **Basit Korelasyon Katsayısı:** İki değişken arasında mevcut bulunan ilişkinin derecesine basit korelasyon denir. İki değişken arasında pozitif korelasyon veya negatif korelasyon olabileceği gibi bunlar arasında hiç korelasyon olmayabilir.

    - **Pozitif Korelasyon:** İki değişken eğer birlikte ve aynı yönde değişme eğiliminde İse yani eğer bunlar birlikte artma veya azalma eğiliminde ise, bunlar arasında pozitif korelasyon olduğu söylenir. Bir malın arz edilen miktarı ve onun fiyatı arasında pozitif korelasyon vardır. Fiyatlar arttığında arz edilen miktar artmakta ve fiyatlar düştüğünde arz edilen miktar azalmaktadır.

[102] Özdamar K., Paket Programlar ile İstatistiksel Veri Analizi-1, Eskişehir, 2011

[103] Özdamar A.g.e.

[104] Özdamar A.g.e.

[105] Tosun İ., "Türkiye'nin Kağıt-Karton Üretim ve Tüketimi Üzerine Zaman Serileri ve Regresyon Analizi", Yayınlanmamış Yüksek Lisans Tezi, İstanbul Üniversitesi, 1996

- **Negatif Korelasyon:** Eğer iki değişken ters yönde değişme eğiliminde ise, yani X artarken Y azalırsa veya bunun tersi durumda, iki değişken arasında negatif korelasyon olduğu söylenir. Örneğin bir maldan talep edilen miktar ve onun fiyatı arasında negatif korelasyon vardır. Fiyatı arttığında bir malın talebi azalmakta fiyat azaldığında bir malın talebi artmaktadır.

- **Korelasyon olmaması veya sıfır korelasyon:** İki değişken birbiri ile hiçbir bağıntı olmaksızın değişme eğilimi olduğunda bunlar arasında korelasyon olmadığı söylenir. Örneğin: Bir ülkede spor yapanların boyları ve kağıt üretimi arasında veya öğrencilerin başarıları ile gözlerinin rengi arasında sıfır korelasyon olduğunu tahmin edebiliriz.

❖ **Çoklu Korelasyon Katsayısı:** Üç veya daha fazla değişken arasında mevcut bulunan ilişkinin derecesine çoklu korelasyon adı verilir.

Bir tek bağımsız değişkenin kullanıldığı regresyon tek değişkenli regresyon analizi, birden fazla bağımsız değişkenin kullanıldığı regresyon analizi de çok değişkenli regresyon analizi olarak adlandırılır. Tek değişkenli regresyon analizi bir bağımlı değişken ve bir bağımsız değişken arasındaki ilişkiyi inceler. Tek değişkenli regresyon analizi ile bağımlı ve bağımsız değişkenler arasındaki doğrusal ilişkiyi temsil eden bir doğrunun denklemi formüle edilir. Çok Değişkenli Regresyon Analizi içinde bir adet bağımlı değişken ve birden fazla bağımsız değişkenin bulunduğu regresyon modelleri çok değişkenli regresyon analizi olarak bilinir.[106]

Regresyon analizinin uygulanması için değişkenlerin bağımlı değişken (cevap değişken, dependent/response variable) ve bağımsız değişken (Açıklayıcı değişken, Faktör, Independent/Predictor Variable) olarak tanımlanması gerekir. Bağımlı değişken ile bağımsız değişken(ler) arasındaki matematiksel bağıntının (doğrusal, eğrisel) regresyon modelinin belirlenmesi ve matematiksel formun oluşturulması gerekir.

[106] Bilekdemir, A.g.e.,

Bağımlı ve bağımsız değişken arasındaki matematiksel eşitliğe **regresyon modeli** ya da **Regresyon denklemi** adı verilir.[107]

Bağımlı değişken (Dependent/Response variable): Değeri başka değişkenler tarafından belirlenen ve diğer değişkenlerin değeri değiştiğinde bu değişimden etkilenen değişkene **bağımlı değişken** denir. İncelenen bir olayda; etkilenen, sonuç değişken bağımlı değişkendir. Bağımlı değişken genelde Y ile gösterilir.[108]

Bağımsız değişken (Independent/Predictor variable): Değeri rastgele koşullara göre oluşan, bağımsız olarak değişim gösteren ve başka değişkenlerin değişimi üzerine etkide bulunan değişkenlere bağımsız değişken denir. Bağımsız değişken, bir olayda sonucun ortaya çıkmasında etkili olduğu düşünülen nedensel faktörler, risk faktörleri vb. olarak alınır. Bağımsız değişken genelde X ile gösterilir. Bir çalışmada birden fazla bağımsız değişken varsa X, S, M... gibi farklı harflerle, X_1, X_2,...X_k ya da X_1, X_2,..., X_k gibi (X_i, i=1,2,...,k) alt indisler biçiminde gösterilir.[109]

Y ve X_i değişkenleri arasındaki regresyon tipleri (regresyon yöntemleri) bağıntı modelindeki bağımlı, bağımsız değişken sayısına, bağımlı ve bağımsız değişken arasındaki matematiksel bağıntının tipine yani regresyon modeline, bağımlı değişken/değişkenlerin Normal dağılım gösterip göstermemesine ve değişkenlerin ölçeklerine göre farklılık gösterir. Aşağıda regresyon yöntemleri verilmiştir.[110]

Parametrik/Parametrik Olmayan Regresyon yöntemleri:

1. **Parametrik Regresyon Yöntemleri**, Bağımlı değişkenin/ değişkenlerin Normal dağılım/Çok değişkenli Normal dağılım göstermesini ön koşul olarak kabul eden Regresyon yöntemlerini içerir.
2. **Parametrik olmayan Regresyon Yöntemleri**, Bağımlı değişkenin/değişkenlerin Normal/çok değişkenli Normal dağılım göstermesini ön koşul olarak ileri sürmeyen regresyon yöntemlerini içerir.

[107] Özdamar A.g.e.

[108] Özdamar A.g.e.

[109] Özdamar A.g.e.

[110] Özdamar A.g.e.

Doğrusal/Doğrusal olmayan (Linear/Nonlinear) Regresyon Yöntemleri:

1. **Doğrusal Regresyon Yöntemleri,** Regresyon modelindeki yer alan bağımsız değişken/değişkenlerin Y_i bağımlı değişkene/değişkenlere etkilerini doğrusal ve eklenebilir formda ele alan regresyon yöntemlerini içerir.
2. **Doğrusal olmayan (Eğrisel, Nonlineer) Regresyon Yöntemleri,** regresyon modelinde yer alan bağımsız değişken/ değişkenlerin Y_i bağımlı değişken/değişkenler üzerine olan etkilerinin toplanabilir olmadığını (eğrisel, üssel, çarpımsal vb.) varsayan regresyon yöntemlerini içerir.

Basit/Çoklu/Çok değişkenli Regresyon yöntemleri:

1. **Basit (Simple) Regresyon Yöntemleri,** regresyon modelinde (denklemde) değişken sayısı p=q+k olmak üzere, q=1 ve k=1 olan doğrusal ve eğrisel regresyon modellerini içerir.
2. **Çoklu (Multiple) Regresyon Yöntemleri,** regresyon modelindeki değişken sayıları q=1 ve k≥2 olan doğrusal ve eğrisel regresyon yöntemlerini içerir.
3. **Çok değişkenli (Multivariate) Regresyon Yöntemleri**, regresyon modelindeki değişken sayıları q≥2 ve k≥1 olan doğrusal ve eğrisel regresyon yöntemlerini içerir.

Uygulamada, veri tipine, değişkenler arası bağıntıya ve modellemeye bağlı olarak; Parametrik/Parametrik olmayan, Doğrusal/Eğrisel ve Basit/Çoklu/Çok değişkenli olmak üzere çok sayıda Regresyon Yöntemi bulunmaktadır. Aşağıda yaygın kullanımı olan regresyon yöntemleri verilmiştir.[111]

1. Modelde, Y ve X olmak üzere iki sürekli değişken varsa, en yaygın kullanılan regresyon modelleri;

- Basit Doğrusal Regresyon (Simple Linear regression)
- Polinomiyal regresyon (Polinomial regression)
- Geometrik Regresyon (Geometric Regression)
- Üssel Regresyon (Exponential Regression)
- Basit Eğrisel Regresyon (Simple Nonlinear regresyon) olarak sayılabilir.

[111] Özdamar A.g.e.

2. Modelde, bir Y bağımlı değişken, X_i (i=1,2,3,..,k) k>2 bağımsız değişken olmak üzere, değişkenler sürekli (interval/orantılı) ölçekli ise en yaygın kullanılan regresyon yöntemleri,

- Çoklu Doğrusal Regresyon (Multiple Linear Regression)
- Çoklu Doğrusal Olmayan Regresyon (Multiple Nonlinear Regresyon)
- Çoklu Cox Regresyon (Multiple Cox's Proportional Hazard Regression)

olarak sayılabilir.

3. Modelde, bir bağımlı değişken Y ve bir ya da daha fazla bağımsız değişken X_i (i=1,2,..,k) varsa ve Y bağımlı değişken en az iki kategorili olmak üzere nominal/ordinal/nominalize interval ölçekli, bağımsız değişkenler ise orantılı/interval/ordinal/nominal ölçekli ve en az iki kategorili iseler en yaygın kullanılan regresyon yöntemleri,

- Lojistik Regresyon (Simple, Multiple)
- Ordinal Regresyon
- Robust Regresyon
- Poisson Regresyon
- Cox Regresyon (Cox's Poroportioanl Hazard regression)
- Parametrik olmayan regresyon
- Ağırlıklı regresyon
- Spline Regresyon olarak sayılabilir.

4. Modelde, en az iki bağımlı değişken Y_i (i=2,3,...,q) ve en az bir bağımsız değişken X_i (i=1,2,..,k) varsa ve Y bağımlı değişkeni çok değişkenli normal dağılım gösterip göstermemesine göre en yaygın kullanılan regresyon yöntemleri,

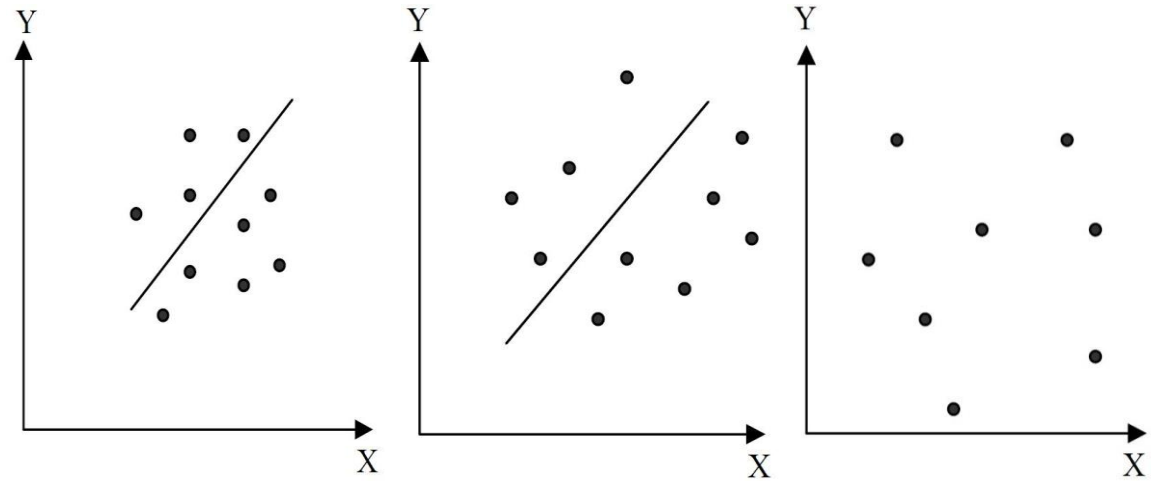
- Çok değişkenli regresyon (Multivariate Regression)
- Çok değişkenli Lojistik Regresyon (Multiple Event data Logistic Regression)
- Çok değişkenli Eğrisel Regresyon (Multivariate Nonlinear Regression)
- Spline Regresyon
- Yerel (Local) Regresyon
- Gauss Karma regresyon olarak sayılabilir.

1.9.3.3.7.1- Basit Doğrusal Regresyon Analizi

Birbiriyle ilişkili olan iki değişkenin olduğu basit doğrusal regresyonda, değişkenlerden birisi bağımlı diğeri ise bağımsız değişken olarak kabul edilir. Y bağımlı değişkeni, X bağımsız (tasarım) değişkeni, β_0 ile β_1 bu değişkenin bilinmeyen parametrelerini ve ε hata terimlerini göstermek üzere kitle için basit doğrusal regresyon denklemi aşağıdaki şekilde yazılır.[112]

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \,, \quad i=1,2,3,\ldots,n$$

n adet (X_i, Y_i) , i=1,2,...,n gözlem çiftinin olduğu varsayılsın. Bu noktalar (X, Y) koordinat düzlemi üzerinde gösterilebilir. Bu gösteriliş serpilme ya da nokta diyagramı olarak adlandırılır. Lineer regresyonda amaç, serpilme diyagramındaki noktalardan geçebilecek en uygun doğru denkleminin belirlenmesidir. Aşağıda serpilme diyagramının üç farklı gösterimi verilmiştir.[113]



(a) (b) (c)

Şekil 1-7: (a)İki değişken arasında kuvvetli bir ilişki vardır. (b) İki değişken arasında daha zayıf bir ilişki vardır (c) İki değişken arasında çok zayıf bir ilişki vardır.

Kaynakça: Ertan E., "Regresyon Analizi ve Matematik Programlama Arasındaki İlişki", Yayınlanmamış Yüksek Lisans Tezi, İstanbul Üniversitesi, 2008

[112] Okutan D., "Bootstrap Yönteminin Regresyon Analizinde Kullanımı ve Diğer Yöntemlerle Karşılaştırılması", Yayınlanmamış Yüksek Lisans Tezi, Ege Üniversitesi, İzmir, 2009

[113] Ertan, A.g.e.

Doğrusal regresyon yöntemini kullanmak için bazı varsayımların bulunduğu kabul edilmektedir. Varsayımlardan herhangi birinin gerçekleşmemesi durumunda geçerli sonuçlara ulaşmak mümkün olmaz. Bu varsayımlar şöyle sıralanmaktadır:[114]

1. Bağımsız değişkenlerin hepsi nicel veya nitel olarak ölçülmüş olması, bağımlı değişken Y'nin ise nicel ve sürekli olması gerekmektedir. X ve Y değişkenleri doğru olarak ölçülmelidir.

2. Bütün bağımsız değişkenlerin varyansının sıfırdan farklı olması gerekmektedir.

3. Bağımsız değişkenler arasında doğrusal bir ilişkinin olmaması gerekmektedir.

4. Hata terimleri ortalaması sıfırdır: $E(\varepsilon_i)=0$

5. Bağımsız değişkenler ve hata terimi arasında korelasyon olmamalıdır.

6. $E(\varepsilon_i X_i)=0$

7. Hata terimlerinin varyansı sabit olmalıdır, $E(\varepsilon_i^2)=\sigma^2$

8. Hata terimleri arasında korelasyon olmamalıdır. $E(\varepsilon_j \varepsilon_i)=0$, $(i\neq j)$

9. Hata terimleri ε_i, normal dağılmalıdır.

1.9.3.3.7.2- Çoklu Doğrusal Regresyon

Bağımlı değişkenin tek, bağımsız değişkenin birden fazla olduğu regresyon modeline çoklu regresyon (multiple regression) modeli denilmektedir. Bazı iktisadi olayların modellerinde hem bağımlı hem de bağımsız değişken sayısı birden fazla olabilmekte ve

$$(Y_1, Y_2,\ldots, Y_q)=f(X_1, X_2,\ldots, X_p)$$

biçiminde ifade edilmektedir. Bu şekilde birden fazla bağımlı ve bir veya birden fazla bağımsız değişkeni bulunan modellere ise çok değişkenli regresyon (multivariate regression) modeli denmektedir. Genel doğrusal çoklu regresyon modellerinin analizini basitleştirmek maksadıyla matris yaklaşımı uygulanmaktadır. Bu suretle genel doğrusal

[114] Okutan, A.g.e.

modele ilişkin varsayımlar, parametre tahminleri, tahminlerin özellikleri, tahminlerin varyansları ve tahmin edilen regresyon denklemi ile ilgili tüm analizler daha kolay anlaşılabilir ve ispatlanabilir olmaktadır. Y bağımlı değişkeni ile p sayıda bağımsız değişken arasında doğrusal bir ilişki varsa ve Y ve X 'lere ait n tane gözlem değerine sahipsek, yığın için bağıntıyı

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \varepsilon_i, \quad (i=1,2,\ldots,n)$$

biçiminde yazabiliriz. Burada ε_i'ler hataları göstermektedir.[115]

Verilere uyan modelin açıklayıcılık yüzdesi olan belirtme katsayısı R^2 hesaplanır. Modelde yer alan her bir açıklayıcı değişkenin Y üzerindeki etkisinin önemliliği b_i'lerin önemliliğinin test edilmesi ile belirlenir. Bunun için; her bir değişken için hesaplanan regresyon katsayısının önemliliği t testi ile test edilir. Regresyon analizi sonuçlarına göre, modelin tutarlılığı/tahmin gücü, tahminin varyansı (s^2) ve belirtme katsayısı ile belirlenir. Modelin belirleyicilik gücünü ifade eden R^2, aşağıdaki gibi hesaplanır.[116]

$$R^2 = \frac{RKT}{KT_Y} = \frac{\text{Regresyon Kareler Toplamı}}{\text{Genel Kareler Toplamı}} \qquad (4)$$

Çoklu belirtme katsayısı (R^2) modele yeni bir değişken eklendiğinde artış gösterir. Modele yeni bir değişken eklenmesine rağmen paydanın değeri değişmezken payın değeri artar. Bu nedenle R^2 hesaplanırken değişken sayısına göre düzeltme yapılması gerekir. Düzeltilmiş R^2 değeri ($R^2_{düz}$),

$R^2_{duz} = 1 - \left[\frac{\sum e^2/(N-k)}{\sum y^2/(N-1)}\right]$ biçiminde yada,

$R^2_{duz} = 1 - \frac{(1-R^2)(N-1)}{N-k-1}$ biçimine hesaplanır.[117]

[115] Çerçi İ., "Çok Değişkenli Regresyon Analizi (Gsm Sektöründe Bir Uygulama)", Yayınlanmamış Yüksek Lisans Tezi, Gazi Üniversitesi, Ankara, 2010

[116] Özdamar A.g.e.

[117] Özdamar A.g.e.

Eğer N yeteri kadar büyük ise R^2 ile R^2_{duz} arasındaki fark minimum düzeye iner. Fakat küçük örneklerde bağımsız değişken sayısı fazla ise $R^2_{düz}$ değeri R^2'den küçük, hatta eksi değerler alabilecektir. $R^2_{düz}<0$ durumunda $R^2_{düz}=0$ alınmalı ve buna göre yorumlanmalıdır.[118]

1.9.3.3.7.3-Aşamalı Regresyon

Aşamalı regresyon, k sayıda belirleyici değişken içeren bir veri setinde bağımlı değişkenin değişimini en iyi biçimde açıklayan açıklayıcı değişkenleri (açıklayıcı değişken alt setini) seçmeyi sağlayan bir yöntemdir. Aşamalı regresyon ile Y'yi etkilediği düşünülen k bağımsız değişken setinden p kadar bağımsız değişkeni (p=k-m, m=0,l,..., p≤k) optimum düzeyde seçerek çoklu doğrusal regresyon modeli oluşturulur.[119]

1.9.3.3.7.4- Polinomiyal Regresyon

Bağımlı ve bağımsız değişken arasındaki ilişki her zaman basit bir doğru fonksiyonu ile açıklanamaz. Bu gibi durumlarda doğrusal bir fonksiyon yerine eğrisel bir fonksiyon kullanılması zorunlu hale gelmektedir. İki değişken arasındaki ilişkinin doğrusal olmadığı durumlarda, serpme diyagramının yolunu en iyi şekilde belirten eğri tipi regresyon eğrisi olarak seçilir. Bu şekilde modelin belirleme gücü arttırılır. Modelin belirleme gücünü arttırmak için modele bağımsız değişkenin 1., 2., vd. üslerinin katılması ile oluşan modele polinomiyal regresyon adı verilir ve aşağıdaki gibi gösterilir.[120]

$$\hat{Y}=b_0+b_1X+b_2X^2+b_3X^3+...+b_kX^k \quad k=1,2,\ldots,n$$

Burada,

X: Değişken

$\hat{Y}$: X değişkeninin belli bir değerine karşın Y değişkeninin tahmin edilen değeri

b_0 : X'in değeri sıfır olduğunda Y'nin aldığı değer (sabit sayı)

[118] Özdamar A.g.e.

[119] Özdamar A.g.e.

[120] Yerel A.g.e. sf:2

$b_1,...,b_k$: X' deki bir birimlik değişmenin Y' de yaptığı değişikliği gösteriyor.

Eğrisel regresyon eşitliği ile grafik çizildiğinde eğrisel bir fonksiyon elde edilir. Bu eğrinin, birden fazla eğimi söz konusu olduğundan, eğrisel regresyonda, regresyon denkleminin gözlem değerlerini açıklayıp açıklamadığı belirlilik katsayısı (R^2) ile saptanabilir.[121]

Belirlilik katsayısının 0,5 olması regresyon denkleminin gözlenen değerlerin %50'sine uyduğunu açıklar. Belirlilik katsayısının değeri 1'e yaklaştıkça değişkenler arasındaki ilişki artar, 0'a yaklaştıkça ise değişkenler arası ilişki zayıflar.[122]

1.9.3.3.7.5-En İyi Regresyon Modeli Bulma

Veri setindeki bağımsız değişkenleri içeren tüm olası basit ve çoklu regresyon modellerini oluşturduktan sonra bu modeller içinden en büyük R^2 değerine sahip modeli uygun çoklu regresyon modeli olarak almayı amaçlayan bir yöntemdir.[123]

En iyi regresyon yaklaşımında veri setinde yer alan bağımsız değişkenler ile bağımlı değişken arasında kurulabilecek tüm basit ve çoklu doğrusal kombinasyonlara ilişkin regresyon denklemleri belirlenir. Bu denklemler içinden en yüksek belirleyicilik katsayısına sahip olan model uygun regresyon modeli olarak seçilir.[124]

1.9.3.3.7.6- Robust (Sağlam) Regresyon

Robust regresyon teknikleri, aykırı değerlerden etkilenmeyen tekniklerdir. Varsayımlara bağlı olmayan, özellikle normallik varsayımına duyarsız yaklaşımlar "robust (sağlam)" olarak adlandırılmıştır.[125]

Gözlem değerlerinin normallik varsayımını bozan en önemli etken büyük hata terimi olarak nitelendirilen aykırı değerlerdir. Aykırı değerler, istatistiksel tekniğe ve

[121] Yerel A.g.e. sf:2

[122] Yerel A.g.e. sf:2

[123] Özdamar A.g.e.

[124] Özdamar A.g.e.

[125] Ergül B., "Robust Regresyon ve Uygulamaları", Yayınlanmamış Yüksek Lisans Tezi, Eskişehir, 2006, sf:2

kurulan regresyon modeline zarar verirler. Bununla birlikte, aykırı değerin modeli açıklayan bir değer olması olasılığı çok düşüktür. Eğer model değiştirilirse, bu aykırı değerin etkisi de büyük ölçüde azalmaktadır. Robust regresyon teknikleri, aykırı değere sahip gözlem değerini güncelleyerek, analizi bu yeni güncellenmiş değer ile yapmaktır.[126]

1.9.4- Modelin Değerlendirilmesi

Modelimizi oluşturulduktan sonra modelden elde ettiğimiz sonuçlarının değerlendirilmesi ve elde edilen sonuçların yorumlanması gerekmektedir. Elimizde mevcut bulunan veriler için elde ettiğimiz modellerin sadece kendi aralarında karşılaştırmasıyla değil de tercih edilen modelin uygulanması ile elde edilecek faydaların karşılaştırılması da gereklidir. Bir sınıflama probleminin değerlendirilmesinde aşağıdaki yöntemler uygulanır:[127]

- **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.

- **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.

- **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

[126] Ergül A.g.e. Sf:2

[127] Aydın, A.g.e., sf:58-61

(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.

1.9.5- Modelin Uygulanması

Bir veri madenciliği modeli oluşturulduktan ve geçerliliği kabul edildikten sonra uygulama aşamasına geçilir. Veri madenciliği sonuçları modelin özelliğine göre iki şekilde uygulanabilir. Bunlardan ilki modelin sonuçlarına göre faaliyetlerin önerilmesidir. 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. Model uygulandıktan sonra sistemin ne kadar iyi çalıştığının ölçülmesi gerekir. Model ne kadar iyi çalışıyor olsa da modelin performansının sürekli olarak izlenmesi gerekir.[128]

[128] Kolay, A.g.e., sf:40

II- UYGULAMA

2- Dicle Üniversitesinin Kısa Tarihçesi

1966 yılında Ankara Üniversitesi Tıp Fakültesi bünyesinde öğrenime açılan Diyarbakır Tıp Fakültesi, bugünkü Dicle Üniversitesinin temelini oluşturmuştur.

Dicle Üniversitesi, tarihi Diyarbakır kentinin doğusunda 27 bin dekarlık arazi üzerine kurulmuş, adını üniversiteyi kentten ayıran Dicle Nehri'nden almıştır. Merkez kampusun yer aldığı Diyarbakır dışında Diyarbakır'ın Ergani, Çermik, Çüngüş, Bismil, Silvan ve Kulp ilçelerinde de Yüksekokulları mevcuttur.

Dicle Üniversitesi'nin bugün 22500 öğrencisi, 3500 civarında kadrolu akademik ve idari elemanı mevcuttur. Üniversite, sunmakta olduğu öğretim hizmetleri dışında 1200 yataklı Araştırma Hastanesiyle, aynı zamanda bölgenin sağlık merkezi konumundadır. Üniversite Araştırma Hastanesi'nde verilen üst düzeyde sağlık hizmeti sayesinde bölgedeki vatandaşlarımız çok önemli sorunlar dışında büyük merkezlere gitmenin maddi ve manevi külfetinden kurtulmuşlardır. Dicle Üniversitesi, bölgede ihtiyaç duyulan hemen her alandaki meslek elemanlarını yetiştiren, ulusal ve uluslar arası düzeyde araştırmalar yapan ve bölge ekonomisine yönelik projeler üreten devletimizin bölgedeki en önemli kurumlarından biridir. GAP'tan en üst düzeyde verim alınarak, bölgede hayat standardının yükseltilebilmesi için üniversitemizde başta tarım ve hayvancılık olmak üzere bölge ekonomisine katkı yapacak projelere öncelik verilmektedir.

Öğrenci kulüplerinin kültür-sanat faaliyetleri ve üniversite takımlarının sportif başarıları, bir taraftan kent yaşamına hareket getirirken diğer taraftan bu faaliyetler, gençleri üniversiteye özendirerek "modern toplum"a geçişte önemli bir işlev görmektedir.

2.1- Literatür Özeti

Akademik başarı hakkında daha önceden yapılan literatür çalışmaları özetlenirse:

- Patır'a göre (2008) İktisadi ve İdari Bilimler Fakültesi bulunan üniversiteler arasında sayısal derslerdeki başarısızlık nedenlerini araştırmıştır. Öğrencilere uyguladıkları 24 soruluk anketin sonuçlarına göre İ.İ.B.F. işletme bölümü öğrencilerinin bulundukları bölümdeki Matematik, İstatistik ve Yöneylem Araştırması gibi sayısal derslerde başarısızlık nedenlerini belirlemeye çalışmışlardır. Buldukları sonuca göre cinsiyet, lise kolu ve Üniversiteler açısından anlamlı bir farkın olduğunu bulmuşlardır.[129]

- Demirok (1990), yaptığı çalışmada Hacettepe Üniversitesi Eğitim Fakültesi bölümünde 1988-1989 öğretim yılında ikinci sınıfa kayıtlı öğrencilerin akademik başarısına ÖSS, ÖYS puanları ile ortaöğretimdeki başarılarının etkisini araştırmıştır. Araştırmacı, öğrencilerin lise ve dengi okullardaki başarıları ile ÖSS ve ÖYS puanları arasında ilişkiyi incelemiş ancak, anlamlı bir ilişki bulamamıştır. Fakat akademik başarı ile lise ve dengi okullarda elde edilen başarı faktörü arasında anlamlı bir ilişki bulmuştur.[130]

- Güngör (1989) ise kendilerini akademik yönden "çok başarılı", "başarılı", "orta başarılı" olarak algılayan öğrencilerin özsaygılarının "başarısız" olarak algılayanlardan daha yüksek olduğunu bulmuştur. Başarılı ve başarısız öğrencileri benlik saygısı bakımından karşılaştırıldığında başarısız öğrencilerin benlik saygısı puanları ortalaması, başarılı öğrencilere kıyasla, anlamlı derecede daha düşük olduğunu bulmuşlardır. Bu bulgular öğrencinin akademik başarısının yüksek olmasının yüksek özsaygı

---

[129] Patır S. ve Yıldız M. S., İktisadi ve İdari Bilimler Fakültesi İşletme Bölümü Öğrencilerinin Sayısal Derslerdeki Başarısızlık Nedenleri ve Çözüm Önerileri, Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi C.13 S.1 s.293-315, Isparta, 2008

[130] (Çil B., Ticaret ve Turizm Eğitim Fakültesinde Akademik Başarıyı Etkileyen Bazı Etkenler, Milli Eğitim Basımevi, Ankara, 1995)'den aktaran Patır S. ve Yıldız M. S., İktisadi ve İdari Bilimler Fakültesi İşletme Bölümü Öğrencilerinin Sayısal Derslerdeki Başarısızlık Nedenleri ve Çözüm Önerileri, Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi C.13 S.1 s.293-315, Isparta, 2008, Sf: 297

geliştirmesine, yüksek özsaygı geliştirmiş olmasının da akademik başarısının artmasına neden olduğu şeklinde yorumlanabilir.[131]

- Akhun'a göre (1980) okullara yerleştirme için yapılan eleme ve giriş sınavlarında alınan puanlar ile öğrencilerin akademik başarılarının bir ölçüsü olarak kabul edilen birinci sınıf not ortalamaları arasındaki korelasyon katsayılarının düşük bulunmuş olması, bu sınav sonuçlarının tek başına akademik başarının kestirilmesinde iyi bir değişken olmadıklarını ortaya koymaktadır.[132]

- Tuzlugöl Dost'a göre (2007) Üniversite öğrencilerinin öğrenim hayatlarının belki de en yoğun ve profesyonel eğitimi aldıkları süreçte akademik başarılarının düşük olması, mezun olamama korkusu yaşamalarına ve zor bir süreçten geçerek elde ettikleri öğrenim olanağını kaybetme endişesi duymalarına neden olabilmektedir. Akademik başarısızlığın kökeninde kimi zaman, bulunduğu alandan memnun olmama, mesleki kararsızlık ve motivasyon düşüklüğü gibi etkenler yatmaktadır. Dolayısıyla akademik başarı düşüklüğü, çok boyutlu ve birden fazla nedene dayanan bir sorun olarak karşımıza çıkabilmektedir. Bireyin hayatının çoğunu kapsayan mesleğe ve kişinin toplumdaki statüsünü belirleyecek profesyonel kimliğe götürecek olan üniversite yıllarında, öğrencilerin akademik olarak başarılı olmaları onların mutluluğundaki en önemli etkenlerden biri olmaktadır.[133]

- Çocuklarda başarıyı etkileyen kişilik özelliklerinin bir kısmı doğuştan gelen yapısal özelliklerle açıklanabilirken, bir kısmı da ailenin ve çevrenin etkisi ile öğrenilmiş davranışlardır. Çocuğun öğrenmeye ve okumaya güdülenmesinde, kendine olan güveni, çevresiyle iletişimi çok önemlidir.[134] Çelenk'in (2003) konuyla ilgili yaptığı bir araştırma ile başarılı öğrencilerde, neşeli, kolay uyum sağlayan, sosyal, hırçın, girişken,

---

[131] (Güngör, A., 1989, "Lise Öğrencilerinin Özsaygı Düzeylerini Etkileyen Etmenler".Yayınlanmamış Doktora Tezi, Hacettepe Üniversitesi, Ankara)'den aktaran Bırtıl, A.g.e., sf:5-6

[132] Akhun, İ.,1980, "Akademik Başarının Kestirilmesi",Ankara Üniversitesi Basımevi- Ankara. 1980.Ankara Üniversitesi Eğitim Fakültesi Yayınları No:88, sf:332

[133] Tuzlugöl Dost M., "Üniversite Öğrencilerinin Yaşam Doyumunun Bazı Değişkenlere Göre İncelenmesi", Pamukkale Üniversitesi Eğitim Fakültesi Dergisi Yıl 2007 (2) 22. Sayı, sf:140

[134] Bırtıl F.S., "Kız Meslek Lisesi Öğrencilerinin Akademik Başarısızlık Nedenlerinin Veri Madenciliği Tekniği ile Analizi", Yayınlanmamış Yüksek Lisans Tezi, Afyon Kocatepe Üniversitesi, 2011, sf:6

mantıklı ve iyimser olma özelliklerinin bulunduğu belirlenmiştir. Buna karşılık, başarısız öğrencilerde daha durgun, içedönük, uysal, uyumsuz, sıkılgan, kötümser veya öfkeli davranış biçimleri tespit edilmiştir.[135]

- Köse'ye (1999) göre açıkça ifade edilsin veya edilmesin, öğrencilerin üniversite giriş sınavlarında göstermiş oldukları başarı düzeyi, sınav öncesi eğitim-öğretim sürecine doğrudan veya dolaylı olarak eklenmiş tüm kişi ve kurumların başarı veya başarısızlıklarının en somut göstergesi olarak algılanmaktadır. Başka bir deyişle, öğrencilerin üniversite giriş sınavlarında göstermiş oldukları başarı düzeyi, mezun oldukları lisenin başarı ve verimlilik düzeyinin de en önemli ölçütü olarak değerlendirilmektedir. Üniversiteye yerleşmede bitirilen lise ve ortamın çok önemli bir etkisinin olduğunu söylemektedir.[136]

- Turan'a göre (1998) öğrencilerin akademik performansı, zihinsel performansın yanında yaş, cinsiyet, benlik saygısı ve yaşam doyum düzeyi gibi kişisel etkenlere bağlı olmakla birlikte ancak anne ve babanın eğitim durumlarına bağlı olmadığını söylemektedir.[137]

2.2- Araştırmanın Metodolojisi

Akademik başarı, öğrencinin bir yıllık çalışmasını yansıtan, bütün derslerden aldığı sınıf geçme notlarının aritmetik ortalamasıdır[138]. Bu notların belirli bir değerin altında olması durumunu akademik başarısızlık olarak tanımlayabiliriz.

Öğrencinin akademik başarısını aldığı dersleri geçme durumuna göre değerlendirdiğimiz zaman kaldığı ders sayısına göre başarıyı değerlendirebiliriz.

---

[135] (Çelenk, S., 2003, "Okul Aile İşbirliği İle Okuduğunu Anlama Başarısı Arasındaki İlişki," Hacettepe Üniversitesi Eğitim Fakültesi, Dergisi, Sayı:24)'den aktaran Bırtıl, A.g.e., sf:6

[136] Köse M.R., Üniversiteye Giriş ve Liselerimiz, Hacettepe Üniversitesi Eğitim Fakültesi Dergisi 15, 1999, sf:51 - 60

[137] Turan M. ve diğerleri, Üniversite Öğrencilerinde Akademik Performansı Etkileyen Faktörler, Selçuk Üniversitesi Eğitim Fakültesi VII. Ulusal Eğitim Bilimleri Kongresi, Konya, 1998, sf: 131-134

[138] Akhun, İ., Akademik Başarının Kestirilmesi,Ankara Üniversitesi, 1980

Araştırmamızda başarı kavramını iki farklı bağımlı değişkene göre değerlendireceğiz. Bunlar not ortalamasına göre ve kaldığı ders sayısına göre olacaktır. Not ortalamasına göre değerlendirdiğimizde 0-59 arası not ortalamasına sahip olan öğrenciler başarısız kabul edilmiştir, 60-100 arası not ortalamasına sahip olan öğrenciler başarılı kabul edilmiştir. Kaldığı ders sayısına göre değerlendirdiğimizde kaldığı ders sayısı 0 (sıfir) olanlar yani hiçbir dersten kalmamış olanlar başarılı, 1 veya daha fazla dersten kalmış olanlar başarısız kabul edilmiştir.

2.2.1- Araştırmanın Amacı

Bu araştırma; üniversiteye yerleşen öğrencilerin üniversiteye yerleştikten sonraki akademik başarısına etki eden faktörlerin neler olduğunu ÖSS verilerine ve aile özelliklerine dayanarak ortaya çıkarmaya çalışılacaktır. Bu araştırmada akademik başarı ölçütü iki farklı şekilde değerlendirilecektir. Bunlar not ortalaması ve öğrencinin kaldığı ders sayısı temel alınacaktır. Bu çerçevede üniversiteye yerleşen öğrencilerin Fakültesi, Üniversiteye Giriş Yaşı, Üniversiteyi kaçıncı sırada tercih ettiği, Öss Puanı, Cinsiyeti, Üniversitede aldığı derslerin Not Ortalaması, Üniversitede aldığı derslerin kaçından kaldığı, Anne Eğitim Durumu, Baba Eğitim Durumu ve liseyi bitirdikten sonra kaçıncı yılında üniversiteye yerleştiği verilerinden yararlanarak bunların arasından akademik başarıya etki eden faktörleri bulmaya çalışılacaktır.

2.2.2- Problem Cümlesi

Öss verileri ve aile eğitim durumu verilerine dayanarak Üniversiteye yerleşen öğrencilerin, üniversite hayatı boyunca akademik başarılarını neler etkilemektedir?

Üniversiteye yerleşen öğrencilerin, üniversite hayatı boyunca akademik başarılarını nelerin etkilediğini Öss verilerine ve aile eğitim durumuna bakarak aralarındaki ilişkiyi regresyon ve korelasyon (Spearman) yöntemleriyle incelenecektir. Bu araştırmada akademik başarı ölçütü iki farklı şekilde değerlendirilecektir. Bunlar not ortalaması ve öğrencinin kaldığı ders sayısı temel alınacaktır.

2.2.3- Hipotezler

Bu araştırma sonucunda aşağıdaki hipotezlere cevap bulmaya çalışılacaktır.

H_{1a}: Üniversiteye giriş yaşının öğrencinin not ortalamasına etkisi vardır.
H_{1b}: Üniversiteye giriş yaşının öğrencinin kaldığı ders sayısına etkisi vardır.

H_{2a}: Öss Puanının öğrencinin not ortalamasına etkisi vardır.
H_{2b}: Öss Puanının öğrencinin kaldığı ders sayısına etkisi vardır.

H_{3a}: Üniversiteyi Kaçıncı sırada tercih ettiği öğrencinin not ortalamasına etkisi vardır.
H_{3b}: Üniversiteyi Kaçıncı sırada tercih ettiği öğrencinin kaldığı ders sayısına etkisi vardır.

H_{4a}: Cinsiyetin öğrencinin not ortalamasına etkisi vardır.
H_{4b}: Cinsiyetin öğrencinin kaldığı ders sayısına etkisi vardır.

H_{5a}: Anne eğitim durumunun öğrencinin not ortalamasına etkisi vardır.
H_{5b}: Anne eğitim durumunun öğrencinin kaldığı ders sayısına etkisi vardır.

H_{6a}: Baba eğitim durumunun öğrencinin not ortalamasına etkisi vardır.
H_{6b}: Baba eğitim durumunun öğrencinin kaldığı ders sayısına etkisi vardır.

H_{7a}: Lisenin bitim tarihi ile üniversiteye yerleşme tarihi arasında geçen sürenin öğrencinin not ortalamasına etkisi vardır.
H_{7b}: Lisenin bitim tarihi ile üniversiteye yerleşme tarihi arasında geçen sürenin öğrencinin kaldığı ders sayısına etkisi vardır.

H_{8a}: Öğrencinin bulunduğu fakültenin öğrencinin not ortalamasına etkisi vardır.
H_{8b}: Öğrencinin bulunduğu fakültenin öğrencinin kaldığı ders sayısına etkisi vardır.

2.2.4- Araştırmanın Evreni ve Örneklemi

Bu uygulamada araştırmanın evreni olarak 2010 yılında Dicle Üniversitesine lisans düzeyinde yerleşen öğrencilerin verileri kullanılmıştır. Bu öğrencilerin verileri 2. Sınıfa geldiklerinde otomasyon ortamından çekilmiştir. 2010 yılında Dicle Üniversitesine 5816 öğrenci yerleşmiştir. Bu öğrencilerin 1875'i önlisans programlarına, 3939'u da lisans programlarına yerleşmişlerdir. Bu araştırmada lisans öğrencileri üzerinde çalışılacağı için önlisans öğrencilerine ait veriler silinmiştir. Lisans öğrencilerinden ilk yılında hazırlık okumuş olanlar, bilgilerinde eksiklik olanlar veya tutarsız bilgi bulunan veriler silinmiştir. Ayrıca Tıp ve Diş Hekimliği Fakültesinin değerlendirme sistemi farklı olduğu için bu fakültelerdeki öğrenciler değerlendirme dışı tutulmuştur.

Bu uygulamada 2010 yılında Dicle Üniversitesine yerleşen lisans öğrencilerinin 2. sınıfları üzerinde durulmuştur. Bu veriler üzerinde yapılan veri temizleme, veri dönüştürme işlemleri sonucunda elimizde 3391 öğrenciye ait öğrenci örneklemi bulunmaktadır.

2.2.5- Veri Toplama Araçları

Verilerin toplanmasında Dicle Üniversitesi öğrencilerinin verilerinin tutulduğu takibinin yapıldığı Öğrenci İşleri Otomasyonundan ve ÖSYM tarafından her sene Dicle Üniversitesine Yerleşen Öğrencilerin verilerinin gönderildiği veri dosyasından yararlanılmıştır. Uygulamada kullanacağımız verileri Dicle Üniversitesi Öğrenci Otomasyonu sisteminden aşağıdaki SQL sorguları kullanılarak elde edilmiştir.

```
SET DATEFORMAT DMY
SELECT DRC,O.OI,O.BOLKOD,O.NUMARA,YEAR(O.GIRTAR1)-
YEAR(O.DOGTAR) G_YASI,
  O.CINSIYET,N.ISIM NUFUS_ILI,M.ISIM MEMLEKET_ILI,Y.PUANI
OSS_PUANI,Y.TERCIH,
```

```
  ADS=(SELECT COUNT(DISTINCT OPTIKKOD) ADS FROM NOTLAR
WHERE NUMARA=O.NUMARA AND YIL>='2010'  AND ( NOT( YIL='2010'
AND DNM = '2')) ),
  GDS=(SELECT COUNT(DISTINCT OPTIKKOD) GDS FROM NOTLAR
WHERE NUMARA=O.NUMARA AND SONUC=1 AND YIL>='2010' AND ( NOT(
YIL='2010' AND DNM = '2')) ),
  ANAEGT, BABAEGT,D.L_OKULTURU,L_DERECE
  FROM OZLUK O
  INNER JOIN OSYM Y ON Y.NUMARA=O.NUMARA
  INNER JOIN OZLUK_DETAY D ON D.NUMARA=O.NUMARA
  INNER JOIN KOD_ILKODLARI N ON N.PLAKANO=O.NUFTRKOD
  INNER JOIN KOD_ILKODLARI M ON M.PLAKANO=O.MEMTRKODU
  WHERE 1=1
  AND YEAR(O.GIRTAR1)='2010'
ORDER BY O.NUMARA
```

Bu sorgu sonucunda elde ettiğimiz verileri, veri temizleme yöntemi çerçevesine göre analiz ederek, “fakülte”, “giriş yaşı”, “not ortalaması”, “kaldığı ders sayısı”, “öss puanı”, “tercih sırası”, “cinsiyet”, “anne öğrenim durumu”, “baba öğrenim durumu” ve “bekleme süresi” başlıklı değişkenler seçildi. Bunların dışında verilerinde eksiklik olanlar ve araştırma sonucunda işimize yaramayacak veriler silindi. Daha sonra veri dönüştürme yöntemi uygulanarak veriler kullanıma uygun hale getirildi.

Veri dönüştürme yöntemi yapılacak istatistikî değerlendirmeyi sadeleştirmek ve sonuçları daha anlaşılır hale getirmek için yapılmıştır.

Not ortalaması ve kaldığı ders sayısı değişkenleri bağımlı değişken olarak diğerleri ise bağımsız değişken olarak alınmıştır.

Üniversiteye yerleşen öğrencilerin üniversiteye giriş yaşını istatistiksel anlamda değerlendirmesini kolaylaştırmak için veri dönüştürme işlemi uygulanmıştır. Aşağıdaki Tablo 2-1'de görüleceği gibi 17 ve öncesi yaşa sahip olanlar 1, 18-21 yaş arası 2, 22-25 yaş arası 3, 26-29 yaş arası 4, 30-33 yaş arası 5, 34 ve daha büyük yaşa sahip olanlar 6 olarak adlandırılmıştır.

Tablo 2-1: Giriş Yaşı Veri Dönüşüm Tablosu

Giriş Yaşı	
17 ve öncesi	1
18-21	2
22-25	3
26-29	4
30-33	5
34+	6

Yapılan veri dönüşümlerinde bağımlı değişkenlerden olan Not ortalaması küçükten büyüğe sıralandı ve Tablo 2-2'görüleceği gibi not ortalaması 0-59 arası olanlar "0" başarısız, 60-100 arası olanlar "1" başarılı olarak belirlendi.

Tablo 2-2: Not Otalaması Veri Dönüşüm Tablosu

Not Ortalaması	
0-59	0(başarısız)
60-100	1 (başarılı)

Yapılan veri dönüşümlerinde bağımlı değişkenlerden olan Kaldığı ders sayısı değişkeni 0 olanlar "hiçbir dersten kalmayanlar" başarılı, 1 veya daha fazla dersten kalmış olanlar başarısız olarak belirlenmiştir.

Tablo 2-3: Kaldığı Ders Sayısı Veri Dönüşüm Tablosu

Kaldığı Ders Sayısı	
0	1 (Başarılı)
1+	0 (Başarısız)

Yapılan veri dönüşümlerinde Öss Puanı küçükten büyüğe sıralanmıştır. Tablo 2-4'te görüleceği gibi 200-249 arası puanlar 1, 250-299 arası puanlar 2, 300-349 arası puanlar 3, 350-399 arası puanlar 4, 400-449 arası puanlar 5, 450-449 arası puanlar 6, 500 ve daha fazla puan alanlar 7 olarak adlandırılmışlardır.

Tablo 2-4: ÖSS Puanı Veri Dönüşüm Tablosu

Öss Puanı	
200-249	1
250-299	2
300-349	3
350-399	4
400-449	5
450-499	6
500+	7

Yapılan veri dönüşümlerinde Tercih Sırası küçükten büyüğe sıralanmıştır. Tablo 2-5'de görüleceği gibi 1-5 tercih arasındakiler 1, 6-10 tercih arasındakiler 2, 11-15 tercih arasındakiler 3, 16-20 tercih arasındakiler 4, 21-25 tercih arasındakiler 5, 26-30 tercih arasındakiler 6 olarak sınıflandırılmıştır.

Tablo 2-5: Tercih Sırası Veri Dönüşüm Tablosu

Tercih Sırası	
1-5	1
6-10	2
11-15	3
16-20	4
21-25	5
26-30	6

Cinsiyete göre veri dönüşümünde cinsiyeti Erkek olanlar 1, Bayan olanlar 2 olarak adlandırılmıştır.

Tablo 2-6: Cinsiyet Veri Dönüşüm Tablosu

Cinsiyet	
1	Erkek
2	Bayan

Anne öğrenim durumu ve Baba öğrenim durumu değişkenlerinin veri dönüşüm işlemleri aşağıdaki tabloya göre yapılmıştır. Tablo 2-7'ye göre tahsil almamış olanlar 0, ilköğretim mezunu olanlar 1, lise mezunu olanlar 2, önlisans programından mezun olanlar 3, lisans mezunları 4, lisansüstü mezunları 5 olarak adlandırılmıştır.

Tablo 2-7: Anne Öğrenim Durumu ve Baba Öğrenim Durumu Veri Dönüşüm Tablosu

Anne Öğrenim Durumu ve Baba Öğrenim Durumu	
0	Tahsil Yok
1	İlköğretim (İlkokul ve ortaokul)
2	Lise
3	Önlisans (MYO)
4	Lisans
5	Lisansüstü

Lise bitimi ile Üniversiteye yerleşmesi arasında geçen süre (Bekleme Süresi) değişkeninin veri dönüşüm işlemleri aşağıdaki tabloya göre yapılmıştır. Tablo 2-8'e göre 0-1 yıl arasında üniversiteye yerleşenler 1, 2-3 yıl arasında yerleşenler 2, 4-5 yıl arasında yerleşenler 3, 6-7 yıl arasında yerleşenler 4, 8-9 yıl arasında yerleşenler 5, 10 ve daha fazla süre sonra yerleşenler 6 olarak dönüştürülmüştür.

Tablo 2-8: Bekleme Süresi Veri Dönüşüm Tablosu

Lise bitimi ile Üniversiteye yerleşmesi arasında geçen süre (Bekleme Süresi)	
0-1	1
2-3	2
4-5	3
6-7	4
8-9	5
10+	6

2.2.6- Verilerin Analizi

Verilerin değerlendirilmesinde SPSS 17.0 istatistik programı kullanılmıştır. Elde edilen verilerin analizinde regresyon ve korelasyon testleri kullanılmıştır. İstatistikî anlamlılık düzeyi $p<0.05$ olarak kabul edilmiştir.

2.3- Bulgular

Araştırma için elde ettiğimiz veriler doğrultusunda yaptığımız analizlerin sonuçlarını demografik değişkenler, korelasyon testi ve regresyon analizi sonuçlarını aşağıdaki tablolarda gösterilmektedir.

Yaptığımız sorgular sonucunda elde ettiğimiz veri sayısı 3391 olup bu verilerle ilgili frekans tabloları aşağıda gösterilmiştir.

Tablo 2-9: Fakültelere göre frekans tablosu

	Frekans	%
Fakülteler		
DİYARBAKIR ATATÜRK SAĞLIK YÜKSEKOKULU	82	2,4
EDEBİYAT FAKÜLTESİ	558	16,5
FEN FAKÜLTESİ	435	12,8
HUKUK FAKÜLTESİ	200	5,9
İKTİSADİ VE İDARİ BİLİMLER FAKÜLTESİ	337	9,9
İLAHİYAT FAKÜLTESİ	30	0,9
MİMARLIK FAKÜLTESİ	93	2,7
MÜHENDİSLİK FAKÜLTESİ	315	9,3
VETERİNER FAKÜLTESİ	51	1,5
ZİRAAT FAKÜLTESİ	121	3,6
ZİYA GÖKALP EĞİTİM FAKÜLTESİ	1169	34,5
TOPLAM	3391	100

Yukarıdaki tabloya göre alınan verilerin büyük çoğunluğunu eğitim fakültesine ait olduğu görülmektedir. Bunun nedeni eğitim fakültesinin bölüm sayısının diğer fakültelere göre daha fazla olmasıdır.

Tablo 2-10: Cinsiyete göre frekans tablosu

	Frekans	%
Cinsiyet		
Erkek	1900	56,0
Bayan	1491	44,0
TOPLAM	3391	100

Cinsiyete göre frekans tablosuna baktığımızda erkek öğrencilerin sayısının kız öğrencilerin sayısından göre daha fazla olduğu görülmektedir.

Tablo 2-11: Üniversiteye Giriş Yaşına göre frekans tablosu

	Frekans	%
Üniversiteye giriş yaşı		
17 ve öncesi	120	3,5
18-21	2338	68,9
22-25	686	20,2
26-29	165	4,9
30-33	64	1,9
34+	18	0,5
TOPLAM	3391	100

Üniversiteye giriş yaşına göre değerlendirdiğimizde 18 ile 21 yaş arasındaki öğrencilerin oranı %69 gibi çok yüksek bir oranda olduğu görülmektedir. 21 yaşından önce üniversiteye yerleşen öğrencilerin yüzdesi %72,4, Tablo 2-9'a göre liseyi bitiminden ilk 3 yıl içinde üniversiteye yerleşen öğrencilerin oranı %73,5 oranı ile uyuştuğu görülmektedir.

Tablo 2-12: Öss Puanına göre frekans tablosu

	Frekans	%
Öss puanı		
200-249	138	4,1
250-299	379	11,2
300-349	505	14,9
350-399	730	21,5
400-449	1070	31,6
450-499	555	16,4
500+	14	0,4
TOPLAM	3391	100

ÖSS puanına göre frekans tablosuna baktığımızda üniversiteye yerleşen öğrencilerin %32'lik kısmı 400-449 puan aralığında olduğu görülmektedir. 400 ve üzeri puan alan öğrenciler %48,4'lük bir orandadırlar.

Tablo 2-13: Üniversiteyi kaçıncı sırada tercih ettiğine göre frekans tablosu

	Frekans	%
Üniversiteyi Tercih Sırası		
1-5	1271	37,5
6-10	674	19,9
11-15	488	14,4
16-20	374	11,0
21-25	293	8,6
26-30	291	8,6
TOPLAM	3391	100

Öğrencilerin Dicle Üniversitesini tercih sırasına baktığımızda ilk 5 tercihi arasında gösterenlerin oranının %37,5 gibi yüksek bir oranda oluşu öğrencilerin Dicle Üniversitesini isteyerek tercih ettikleri söylenebilir.

Tablo 2-14: Anne öğrenim durumu frekans tablosu

	Frekans	%	Kümülâtif %
Anne Öğrenim Durumu			
Tahsil Yok	2509	74,0	74,0
İlköğretim (İlkokul ve ortaokul)	718	21,2	95,2
Lise	122	3,6	98,8
Önlisans (MYO)	31	0,9	99,7
Lisans	11	0,3	100,0
Lisansüstü	0	0	100,0
TOPLAM	3391	100,0	100,0

Anne öğrenim durumu tablosuna bakıldığında ilk dikkat çeken %74 gibi çok büyük bir oranın hiç okula gitmemiş olduğudur. Lisans öğrenimi gören Anne sayısının sadece 11 kişi olduğu görülmüştür.

Tablo 2-15: Baba öğrenim Durumu frekans tablosu

	Frekans	%	Kümülâtif %
Baba Öğrenim Durumu			
Tahsil Yok	1173	34,6	34,6
İlköğretim (İlkokul ve ortaokul)	1367	40,3	74,9
Lise	519	15,3	90,2
Önlisans (MYO)	177	5,2	95,4
Lisans	144	4,2	99,7
Lisansüstü	11	0,3	100,0
TOPLAM	3391	100,0	100,0

Baba öğrenim durumu tablosunda %35 gibi bir oranın hiç okula gitmediği görülmektedir. Bu anne öğrenim durumunun yaklaşık olarak yarısına tekabül etmektedir. Ancak baba öğrenim durumu frekans tablosuna baktığımızda 155 kişi oransal olarak %4,5'lik bir oran sadece lisans ve lisansüstü eğitim görmüştür, bu oranda çok düşük düzeyde kalmıştır.

Tablo 2-16: Bekleme Süresi frekans tablosu

	Frekans	%
Bekleme Süresi		
0-1	1885	55,6
2-3	609	18,0
4-5	457	13,5
6-7	193	5,7
8-9	107	3,2
10+	140	4,1
TOPLAM	3391	%100

Öğrencilerin liseyi bitirdikten sonra kaç yıl içinde üniversiteyi kazandığı durumuna baktığımızda (bundan sonra bekleme süresi olarak ifade edilecektir) sınava

ilk girişinde veya ikinci girişinde üniversiteyi kazanma oranı %56 gibi yüksek bir orana sahiptir.

Tablo 2-17: Fakültelere göre başarı durumunun dağılımı (Not Ortalamasına göre)

		Not Ortalaması		Toplam
		0-59 (Başarısız)	60-100 (Başarılı)	
Fakülteler	Diyarbakır Atatürk Sağlık Yüksekokulu	13 (%15,9)	69 (%84,1)	82
	Edebiyat Fakültesi	127 (%22,8)	431 (%77,2)	558
	Fen Fakültesi	319 (%73,3)	116 (%26,7)	435
	Hukuk Fakültesi	36 (%18,0)	164 (%82,0)	200
	İktisadi ve İdari Bilimler Fakültesi	132 (%39,2)	205 (%60,8)	337
	İlahiyat Fakültesi	0 (%0,0)	30 (%100,0)	30
	Mimarlık Fakültesi	15 (%16,1)	78 (%83,9)	93
	Mühendislik Fakültesi	155 (%49,2)	160 (%50,8)	315
	Veteriner Fakültesi	31 (%60,8)	20 (%39,2)	51
	Ziraat Fakültesi	52 (%43,0)	69 (%57,0)	121
	Ziya Gökalp Eğitim Fakültesi	181 (%15,5)	988 (%84,5)	1169
Toplam		1061 (%31,3)	2330 (%68,7)	3391 (%100)

Not ortalamasına göre akademik başarıyı fakülte bazında incelersek; İlahiyat Fakültesi'nin %100 ile en başarılı fakülte, Ziya Gökalp Eğitim Fakültesi'nin %84,5 ile başarı oranı en yüksek ikinci fakülte, Fen Fakültesinin de %26,7 ile başarı oranı en düşük fakülte oldukları ortaya çıkmaktadır. Genel olarak baktığımızda %69 oranında bir başarıdan söz edilebilir. Not ortalaması ve fakülte arasında ki-kare testini uyguladığımızda anlamlılık değeri $p<0,001$ olduğundan bu iki değer arasında anlamlı bağ olduğu söylenebilir.

Tablo 2-18: Fakültelere göre başarı durumunun dağılımı (Kaldığı Ders sayısına göre)

		Kaldığı ders sayısı		Toplam
		1+ (Başarısız)	0 (Başarılı)	
Fakülteler	Diyarbakır Atatürk Sağlık Yüksekokulu	59 (%72,0)	23 (%28,0)	82
	Edebiyat Fakültesi	438 (%78,5)	120 (%21,5)	558
	Fen Fakültesi	415 (%95,4)	20 (%4,6)	435
	Hukuk Fakültesi	104 (%52,0)	96 (%48,0)	200
	İktisadi ve İdari Bilimler Fakültesi	282 (%83,7)	55 (%16,3)	337
	İlahiyat Fakültesi	6 (%20,0)	24 (%80,0)	30
	Mimarlık Fakültesi	73 (%78,5)	20 (%21,5)	93
	Mühendislik Fakültesi	289 (%91,7)	26 (%8,3)	315
	Veteriner Fakültesi	47 (%92,2)	4 (%7,8)	51
	Ziraat Fakültesi	109 (%90,1)	12 (%9,9)	121
	Ziya Gökalp Eğitim Fakültesi	797 (%68,2)	372 (%31,8)	1169
Toplam		1061 (%77,2)	2330 (%22,8)	3391 (%100)

Kaldığı ders sayısına göre akademik başarıyı fakülte bazında incelersek; İlahiyat Fakültesi'nin %80,0 ile başarı oranı en yüksek fakülte, Hukuk Fakültesi %48 ile ikinci, Fen Fakültesinin de %4,6 ile başarı oranı en düşük fakülte oldukları ortaya çıkmaktadır. Genel olarak baktığımızda %23 oranında bir başarıdan söz edilebilir. Kaldığı ders sayısı ve fakülte arasında ki-kare testini uyguladığımızda anlamlılık değeri $p<0,001$ olduğundan bu iki değer arasında anlamlı bağ olduğu söylenebilir.

Tablo 2-19: Kaldığı Ders Sayısı ve Not Ortalaması Karşılaştırması

		Not Ortalaması		
		0-59 (Başarısız)	**60-100 (Başarılı)**	Toplam
Kaldığı Ders Sayısı	**Başarısız**	1061 (%40,5)	1558 (%59,5)	2619 (%77,2)
	Başarılı	0 (%0)	772 (%100)	772 (%22,8)
Toplam		1061 (%31,3)	2330 (%68,7)	3391(%100)

Kaldığı Ders Sayısı ve Not Ortalaması bağımlı değişkenlerini karşılaştırdığımızda; Kaldığı Ders Sayısına göre başarılı olanların %100'ü Not Ortalamasına göre de başarılı olmuşlardır. Toplamda kaldığı ders sayısına göre başarı oranı %23 dolaylarında olurken not ortalamasına göre başarı yaklaşık 3 katı artarak %69 düzeylerinde olmuştur.

Tablo 2-20: Regresyon Tablosu

Bağımlı Değişken	**R^2**	**Bağımsız değişkenler**	**B**	**Beta**	**p (sig.)**
Kaldığı Ders Sayısı	0,418	Fakülte	,001	,012	,477
		Üniversiteye Giriş Yaşı	,032	,058	,026
		Not Ortalaması	,254	,281	,000
		Öss Puanı	,045	,149	,000
		Tercih Sırası	-,004	-,014	,366
		Cinsiyet	,126	,149	,000
		Anne Öğrenim Durumu	,013	,020	,286
		Baba Öğrenim Durumu	,001	,002	,917
		Bekleme Süresi	-,020	-,064	,014
Not Ortalaması	0,508	Fakülte	,008	,067	,000
		Üniversiteye Giriş Yaşı	-,020	-,033	,187
		Kaldığı Ders Sayısı	,279	,252	,000
		Öss Puanı	,111	,333	,000
		Tercih Sırası	,007	,026	,085
		Cinsiyet	,123	,132	,000
		Anne Öğrenim Durumu	-,011	-,015	,382
		Baba Öğrenim Durumu	-,013	-,031	,076
		Bekleme Süresi	,016	,046	,061

Yukarıdaki regresyon tablosuna baktığımızda $p<0,05$ anlamlılık düzeyinde Kaldığı ders sayısı bağımlı değişkenine göre Üniversiteye Giriş Yaşı, Öss Puanı, Cinsiyet ve Bekleme Süresi arasında anlamlı bir ilişki bulunmaktadır. Not ortalaması bağımlı değişkenini $p<0,05$ anlamlılık düzeyinde incelediğimizde Fakülte, Öss Puanı ve Cinsiyet arasında anlamlı bir ilişki bulunduğu görülmektedir.

Tablo 2-21: Korelasyon Tablosu

	Fakülte	Üniversiteye Giriş Yaşı	Not Ortalaması	Kaldığı Ders Sayısı	ÖSS Puanı	Tercih Sırası	Cinsiyet	Anne Öğrenim Durumu	Baba Öğrenim Durumu	Bekleme Süresi
Fakülte	1,000	-0,041*	0,162**	0,107**	0,270**	-0,187**	-0,010	-0,008	0,036*	-0,053**
Üniversiteye Giriş Yaşı		1,000	-0,067**	-0,070**	-0,137**	-0,067**	-0,109**	-0,100**	-0,056**	0,698**
Not Ortalaması			1,000	0,366**	0,398**	-0,040*	0,152**	-0,029	-0,016	-0,074**
Kaldığı Ders Sayısı				1,000	0,264**	-0,048**	0,188**	0,024	0,018	-0,108**
ÖSS Puanı					1,000	-0,130**	-0,050**	-0,050**	-0,024	-0,160**
Tercih Sırası						1,000	-0,034*	-0,008	-0,036*	-0,049**
Cinsiyet							1,000	0,137**	0,164**	-0,103**
Anne Öğrenim Durumu								1,000	0,501**	-0,067**
Baba Öğrenim Durumu									1,000	-0,012
Bekleme Süresi										1,000
	**. Correlation is significant at the 0.01 level (2-tailed). *. Correlation is significant at the 0.05 level (2-tailed).									

Yukarıdaki Tablo 2-13'de görüldüğü gibi başarı kriteri olarak aldığımız Not ortalaması ve Kaldığı ders sayısı verilerine dayanarak, Kaldığı ders sayısının Fakülte, Öss Puanı ve Cinsiyet ile pozitif bir ilişki içindedir ($p<0.01$). Ayrıca Kaldığı ders sayısının Üniversiteye Giriş Yaşı, Tercih Sırası ve Bekleme Süresi ($p<0,01$) ile negatif yönlü bir ilişki içindedir.

Not ortalamasının Fakülte, Öss Puanı ve Bekleme Süresi ($p<0.01$) ile pozitif bir ilişki içindedir. Ayrıca Not ortalamasının Üniversiteye Giriş Yaşı, Bekleme Süresi ($p<0,01$) ve Tercih Sırası ($p<0,05$) ile negatif yönlü bir ilişki içindedir.

Ayrıca Üniversiteye Giriş yaşının; Not Ortalaması, Kaldığı ders sayısı, Öss Puanı, Tercih Sırası, Cinsiyet, Anne Öğrenim Durumu ve Baba Öğrenim Durumu ile negatif, Bekleme Süresi ile pozitif bir ilişki içinde olduğu görülmektedir ($p<0.01$).

III. SONUÇ VE ÖNERİLER

Öğrencilerin Üniversite yıllarındaki akademik başarısı onları daha iyi bir geleceğe, daha iyi bir iş ortamına sahip olmaları için çok önemlidir. Öğrenicileri başarma konusunda yönlendirme için iyi rehberlere ihtiyaç duyulmaktadır. Öğrencilerin geldikleri aile yapılarına bakıldığında çoğunlukla eğitim düzeyi düşük bir anne ve baba'yla karşılaşılmaktadır, böyle bir durumda öğrencinin yönlendirilmesi büyük önem taşımaktadır. Ailenin eğitim durumunu düşük olması çocukla aile arasında iletişim bozukluğuna sebep olmaktadır. Bu iletişim eksikliği öğrencinin başarısızlığında önemli bir yere sahip olabilir.

Yapılan analizler sonucunda hipotezleri değerlendirirsek;

1. Üniversiteye giriş yaşı değişkenine baktığımızda %68,9 gibi büyük bir oran 18-21 yaş arasında üniversiteye yerleşmektedir. Üniversiteye giriş yaşı ile not ortalaması arasında negatif yönlü kuvvetli bir ilişki olduğundan H_{1a} hipotezi kabul edilir. Üniversiteye giriş yaşı attıkça kaldığı ders arasında negatif yönlü kuvvetli bir ilişki vardır. Yani üniversiteye giriş yaşı arttıkça kaldığı ders sayısında azalma meydana gelmektedir. Bu sonuçtan H_{1b} hipotezi kabul edilir.

2. Öss Puanı değişkenine baktığımızda %32'lik bir oran 400-449 puan aralığında Öss puanı alarak üniversiteye yerleşmiştir. Öss Puanı ile Not ortalaması arasında pozitif yönlü kuvvetli bir ilişki vardır. Öss Puanı ile Kaldığı ders sayısı arasında pozitif yönlü kuvvetli bir ilişki vardır. Bu sonuçlardan hareketle H_{2a} ve H_{2b} hipotezleri kabul edilir.

3. Üniversiteyi kaçıncı sıradan tercih ettiği değişkenine baktığımızda ilk beş tercihinde Dicle Üniversitesini tercih edip yerleşenlerin oranı %37,5 olarak bulunmuştur. Üniversiteyi kaçıncı sıradan tercih ettiği ile Not ortalaması

arasında negatif yönlü zayıf bir ilişki vardır. Üniversiteyi kaçıncı sıradan tercih ettiği ile Kaldığı ders sayısı arasında negatif yönlü kuvvetli bir ilişki vardır. Bu sonuçlardan hareketle H_{3a} ve H_{3b} hipotezleri kabul edilir.

4. Öğrencinin cinsiyeti değişkenine baktığımızda %56'lik bir oranda erkeklerin çoğunlukta olduğu gözükmektedir. Öğrencinin cinsiyeti ile Not ortalaması ve Kaldığı ders sayısı arasında pozitif yönlü kuvvetli bir ilişki bulunmuştur. Bu sonuçlardan hareketle H_{4a} ve H_{4b} hipotezleri kabul edilir.

5. Anne öğrenim durumu değişkenine baktığımızda %74 gibi ciddi bir oranın öğretim almadığı gözükmektedir. Anne öğrenim durumu ile öğrencinin not ortalaması ve kaldığı ders sayısı arasında bir ilişki bulunmadığını görmekteyiz. Bu sonuçlardan hareketle H_{5a} ve H_{5b} hipotezleri reddedilir.

6. Baba öğrenim durumu değişkenine baktığımızda %35 lük bir oranın öğretim almadığı görülmektedir. Baba öğrenim durumu ile öğrencinin not ortalaması ve kaldığı ders sayısı arasında bir ilişki bulunmadığını görmekteyiz. Bu sonuçlardan hareketle H_{6a} ve H_{6b} hipotezleri reddedilir.

7. Lisenin bitim tarihi ile üniversiteye yerleşme süresi arasında geçen süreye bekleme süresi dediğimizde sınava ilk ve ikinci girişinde üniversiteye yerleşenlerin oranı %55,6 ile yüksek bir orana sahiptir. Sınava girme sayısı arttıkça kazanma oranı düşmektedir. Bekleme süresinin öğrencinin not ortalaması ve kaldığı ders sayısı ile aralarında negatif yönlü kuvvetli bir ilişki vardır. Bu sonuçlardan hareketle H_{7a} ve H_{7b} hipotezleri kabul edilir.

8. Öğrencilerin yerleştiği fakültelere göre başarı değerlendirmesi yapılırsa not ortalaması ve kaldığı ders sayısı ile aralarında pozitif yönlü kuvvetli bir ilişki vardır. Bu sonuçlardan hareketle H_{8a} ve H_{8b} hipotezleri kabul edilir.

Patır'ın 2008 yılında yayınlanan araştırmasına göre öğrencilerin sayısal derslerde ki başarı durumunun cinsiyet, lise kolu ve araştırmasını yaptığı Üniversitelerle aralarında anlamlı bir farklılığın olduğunu bulmuşlardı. Yaptığımız araştırma sonucuna göre bağımlı değişken olan Kaldığı ders sayısı değişkeni ve not ortalaması değişkeninin cinsiyet değişkeni ile aralarında pozitif yönlü ($p<0,001$) bir ilişki içerisinde olduğu bulunmuştur.

Akhun'a göre (1980) üniversiteye yerleşmek için yapılan giriş sınavlarında alınan puanlar ile öğrencilerin başarı kriteri olarak alınan birinci sınıf not ortalamaları arasındaki korelasyon katsayısını düşük bulmuştu, ancak yaptığımız araştırma sonucuna göre ise Üniversiteye yerleşmede kullanılan ÖSS puanının başarı kriteri olarak aldığımız not ortalaması ve kaldığı ders sayısı değişkenleri ile aralarında pozitif yönlü kuvvetli bir ilişki içinde oldukları tespit edilmiştir.

Turan'a göre (1998) öğrencilerin akademik performansı yaş ve cinsiyete bağlı olduğunu ancak anne ve babanın eğitim durumlarına bağlı olmadığını söylemekteydi. Yaptığımız araştırmaya göre ise Kaldığı ders sayısı değişkeni Cinsiyet ile pozitif bir ilişki içindedir($p<0,01$); Üniversiteye Giriş Yaşı ($p<0,01$) ile negatif yönlü bir ilişki içindedir. Not ortalaması bağımlı değişkeni Üniversiteye Giriş Yaşı ($p<0,01$) ile negatif yönlü bir ilişki içindedir. Yaptığımız araştırmanın sonucuna göre Kaldığı ders sayısı ve Not ortalaması bağımlı değişkeninin anne ve babanın eğitim durumları ile aralarında bir ilişki bulunamamıştır.

Yapılan bu tez çalışmasında elde edilen veriler göz önüne alınarak araştırmacıların daha sonraki çalışmalarına ışık tutması için bu araştırmanın uygulandığı üniversitede bu çalışmaya paralel olarak yapılacak bir anket çalışması ile elde edilen

sonuçlar arasındaki ilişkiler incelenebilir. Yeni çıkan sonuçlar ile bu tez çalışmasındaki sonuçlar karşılaştırılarak öğrenci başarısına etki eden nedenler karşılaştırmalı olarak belirlenebilir. Ayrıca bu çalışmanın bir benzeri başka bir üniversitede yapılırsa çalışmanın sonuçlarını karşılaştırmak, daha iyi bir değerlendirme imkânı verebilir.

Araştırmacılar için bir diğer öneri aynı çalışmayı 2011 ve 2012 yıllarında üniversiteye yerleşen öğrencilerin 2. sınıfa geldiklerinde oluşan verileri üzerinde tekrarlanarak yıllar bazında akademik başarıya etki eden etmenlerin değişimi karşılaştırmalı olarak bulunup aralarındaki ilişkilerin daha detaylı inceleme fırsatı bulunabilir.

KAYNAKLAR

Akbulut S., "Veri Madenciliği Teknikleri ile Bir Kozmetik Markanın Ayrılan Müşteri Analizi ve Müşteri Segmentasyonu", Ankara, 2006

Akhun İ., "Akademik Başarının Kestirilmesi", Ankara Üniversitesi, 1980

Akpınar H., "Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği", İstanbul, İ.Ü**.** İşletme Fakültesi Dergisi, Sayı:1, Nisan 2000

Altıntaş T., Veri Madenciliği Metotlarından Olan Kümeleme Algoritmalarının Uygulamalı Etkinlik Analizi, Sakarya, 2006

Aydın S., "Veri Madenciliği ve Anadolu Üniversitesi Uzaktan Eğitim Sisteminde Bir Uygulama", Eskişehir, 2007, sf:4

Baykal A., "Veri madenciliği: Öğrenci verileri üzerinde uygulamaları", Diyarbakır, 2003

Bırtıl F.S., "Kız Meslek Lisesi Öğrencilerinin Akademik Başarısızlık Nedenlerinin Veri Madenciliği Tekniği ile Analizi", Afyon Kocatepe Üniversitesi, 2011

Bilekdemir G., Veri Madenciliği Tekniklerini Kullanarak Üretim Süresi Tahmini ve Bir Uygulama, İzmir, 2010

Cangül O., "Diskriminant Analizi ve Bir Uygulama Denemesi", Uludağ Üniversitesi, Bursa, 2006

Çerçi İ., "Çok Değişkenli Regresyon Analizi (Gsm Sektöründe Bir Uygulama)", Gazi Üniversitesi, Ankara, 2010

Dener M. ve diğerleri, "Açık Kaynak Kodlu Veri Madenciliği Programları: WEKA'da Örnek Uygulama", Harran Üniversitesi, Şanlıurfa, 11-13 Şubat 2009

Doğan Ş., "Veri Madenciliği Kullanarak Biyokimya Verilerinden Hastalık Teşhisi", Elazığ, 2007

Eker H., "Veri Madenciliği veya Bilgi Keşfi", http://www.ikademi.com/insan-kaynaklari-bilgi-sistemleri/621-veri-madenciligi-veya-bilgi-kesfi.html (01.07.2011)

Erdoğan Ş.Z. Veri Madenciliği ve Veri Madenciliğinde Kullanılan K-Means Algoritmasının Öğrenci Veri Tabanında Uygulanması, İstanbul, 2004

Ergül B., "Robust Regresyon ve Uygulamaları", Eskişehir, 2006

Ertan E., "Regresyon Analizi ve Matematik Programlama Arasındaki İlişki", İstanbul Üniversitesi, 2008

Gündoğan Y., "Sağlam Regresyonda Kısmi Artık Grafiği", Ondokuz Mayıs Üniversitesi, Samsun, 2005

Hand D. ve diğerleri, "Principles of Data Mining", The MIT Press, London, 2001

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

Ian H. Witten ve **Frank** E., Data Mining (USA: Elsevier Inc., 2005)

Kayaalp K., Asenkron Motorlarda Veri Madenciliği ile Hata Tespiti, Isparta, 2007

Kayhan Atılgan Y., "Orijinden Geçen Çoklu Doğrusal Regresyon Modellerinde Yeni En Küçük Ortanca Kareler Yaklaşımı", Hacettepe Üniversitesi, Ankara, 2011

Kiremitci B., "Veri Ambarlarında Veri Madenciliği ve Ulaştırma-Lojistik Sektöründe Bir Uygulama", İstanbul, 2005

Kolay G., "İşletmelerde Bilgi Sistemleri Verimliliğini Arttırmada Veri Madenciliği Yöntemi: Bir Simülasyon Çalışması", Zonguldak, 2006

Koldere Akın Y., "Veri Madenciliğinde Kümeleme Algoritmaları ve Kümeleme Analizi", İstanbul, 2008

Köse M.R., Üniversiteye Giriş ve Liselerimiz, Hacettepe Üniversitesi Eğitim Fakültesi Dergisi 15, 1999

Okutan D., "Bootstrap Yönteminin Regresyon Analizinde Kullanımı ve Diğer Yöntemlerle Karşılaştırılması", Ege Üniversitesi, İzmir, 2009

Özdamar K., "Paket Programlar ile İstatistiksel Veri Analizi-1", Eskişehir, 2011

Özdamar K., "Paket Programlar ile İstatistiksel Veri Analizi-2", Eskişehir, 2010

Patır S. ve **Yıldız** M. S., İktisadi ve İdari Bilimler Fakültesi İşletme Bölümü Öğrencilerinin Sayısal Derslerdeki Başarısızlık Nedenleri ve Çözüm Önerileri, Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi C.13 S.1 s.293-315, Isparta, 2008

Sıramkaya E., "Veri Madenciliğinde Bulanık Mantık Uygulaması", Konya, 2005

Tiryaki S., "Lojistik Alanında Bir Veri Madenciliği Uygulaması", İstanbul, 2006

Tosun İ., "Türkiye'nin Kağıt-Karton Üretim ve Tüketimi Üzerine Zaman Serileri ve Regresyon Analizi", İstanbul Üniversitesi, 1996

Turan M. ve diğerleri, Üniversite Öğrencilerinde Akademik Performansı Etkileyen Faktörler, Selçuk Üniversitesi Eğitim Fakültesi VII. Ulusal Eğitim Bilimleri Kongresi, Konya, 1998

Tuzlugöl Dost M., "Üniversite Öğrencilerinin Yaşam Doyumunun Bazı Değişkenlere Göre İncelenmesi", Pamukkale Üniversitesi Eğitim Fakültesi Dergisi Yıl 2007 (2) 22. Sayı

Two Crows Corp., Introduction to Data Mining and Knowledge Discovery (Versiyon 3: http://www.twocrows.com/intro-dm.pdf, 1999)

Usama M. Fayyad ve diğerleri. "Advances in Knowledge Discovery and Data Mining", American Association for Articial Intelligence, 1996

Yerel S. ve diğerleri, "Sulu Ortamda Kromit ve Serpantinin Mikroelektroforetik Davranışı Üzerine pH'ın Etkisi", Fırat Üniversitesi Fen ve Mühendislik Bilimleri Dergisi, 17 (2), 435-441, 2005

Yıldırım N., "En Küçük Kareler, Ridge Regresyon ve Robust Regresyon Yöntemlerinde Analiz Sonuçlarına Aykırı Değerlerin Etkilerinin Belirlenmesi", Çukurova Üniversitesi, Adana, 2010

http://tr.wikipedia.org/wiki/Veri_taban%C4%B1#Veri_modelleme (04.07.2011)