

**T.C.
MİMAR SİNAN GÜZEL SANATLAR ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ**

**BİR PERAKENDE FİRMASINA AİT VERİLER ÜZERİNDEN VERİ
MADENCİLİĞİ UYGULAMASI**

YÜKSEK LİSANS TEZİ

Fahrettin CEVAHİR

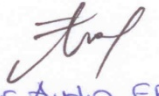
İstatistik Anabilim Dalı

İstatistik Programı

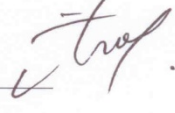
Tez Danışmanı: Prof. Dr. Aydın ERAR

MAYIS 2011

Fahrettin CEVAHİR tarafından hazırlanan BİR PERAKENDE FİRMASINA AİT
VERİLER ÜZERİNDEN VERİ MADENCİLİĞİ UYGULAMASI adlı bu tezin
.....Y.L..... tezi olarak uygun olduğunu onaylarım.


Prof. Dr. Aydın ERAR
Tez Yöneticisi

Bu çalışma, jürimiz tarafından İstatistik..... Anabilim Dalında
Yüksek Lisans tezi olarak kabul edilmiştir.

Başkan: : Prof. Dr. Aydın ERAR 

Üye : Yrd. Doç. Dr. Yalçın ÖZKAN 

Üye : Yrd. Doç. Dr. Elif Özge ÖZDAMAR 

Üye : _____

Üye : _____

Bu tez, Mimar Sinan Güzel Sanatlar Üniversitesi Fen Bilimleri Enstitüsü tez yazım
kurallarına uygundur.

ÖZET

Bu çalışmanın amacı, perakende sektöründe hizmet veren bir kuruma ait farklı sistemlerde tutulan işlemsel bazdaki verileri, veri manipülasyonu yöntemleri ile birleştirerek analiz edilebilir duruma getirmek ve analiz edilebilir hale getirilen bu veri seti üzerinde veri önışleme, EDA teknikleri ve modelleme yöntemleri kullanılarak, veriler arasındaki ilişkilerin ve gizli örüntülerin ortaya çıkarılması amaçlanmıştır.

Birinci bölüm giriş niteliğinde olup, tezi oluşturan konu başlıklarından kısaca bahsedilmiştir.

İkinci bölümde, Veri Madenciliğinin tanım ve sürecinden bahsedilmiştir. Veri Madenciliği uygulamalarında karşılaşılan problemlere değinilmiştir. Karar Destek Sistemlerinden bahsedilmiş, veri tabanı, veri ambarı tanımlamaları yapılmış ve özelliklerine bakılmıştır. Veri madenciliğinde kullanılan programlar incelenmiş ve aralarındaki farklılıklara değinilmiştir. Bu bölümde son olarak, Müşteri İlişkileri Yönetiminin (CRM) Veri Madenciliği ile olan ilişkisinden söz edilmiştir.

Üçüncü bölümde, Veri Tabanında Bilgi Keşfi Sürecinden bahsedilmiş ve bu sürecin aşamalarına değinilmiştir. Büyük veri kümelerinde kullanılan EDA (Exploratory Data Analysis) kavramından bahsedilmiş ve EDA tekniklerinden söz edilmiştir.

Dördüncü bölümde, Veri Madenciliği sonuçlarının elde edildiği ve kararların verildiği modelleme aşamalarından bahsedilmiş ve bu aşamada kullanılan Veri Madenciliği algoritmalarına değinilmiştir.

Beşinci bölümde, müşteri verileri üzerinde SPSS Clementine programı kullanılarak veri manipülasyonu yapılmış ve manipülasyon sonucunda, veri analizi ve veri

madencilięi yöntemleri ile veriler arasındaki ilişkilerin ve gizli örüntülerin ortaya çıkarılması sağlanmıştır.

Son bölüm olan altıncı bölümde, yapılan uygulamanın detayları adımlar bazında ortaya konmuş ve elde edilen sonuçlar değerlendirilmiştir.

DATA MINING APPLICATION ON A RETAIL COMPANY'S DATA

SUMMARY

This study aims to make it possible for analyzing a retail company's transactional data stored in different systems by consolidating the data and using data manipulation methods, followed by data preprocessing, EDA, and modeling methods in order to reveal the relations and unseen patterns among the data.

The first chapter is considered as an introduction and mentions briefly about the topics covered in the thesis.

The second chapter dwells on the definition and process of data mining. Problems encountered in data mining applications are mentioned in this chapter, besides definition of decision support systems, database and data warehouses. Data mining tools are evaluated and comparison between these tools is acquired. The relationship between Customer Relationship Management (CRM) and Data Mining is also covered in this chapter.

In the third chapter, Information Discovery Process in Database and phases of this process are covered. EDA techniques are defined, and special EDA techniques for huge data sets are mentioned in this chapter.

The fourth chapter covers the modeling phase, phase where the decisions are made and results are received, of data mining and besides, algorithms used in this phase are defined.

In the fifth chapter, data manipulation is acquired by using SPSS Clementine software on customer data and, after manipulation, the relations and unseen patterns among data are revealed by using data analysis and data mining methods.

The sixth chapter, which is also the last chapter, details the application and the results obtained are evaluated.

ÖNSÖZ

Tez çalışmam süresince, gece gündüz demeden bana destek olan, her zaman çalışmak için bana vakit ayıran ve yaptığı yapıcı eleştirilerle çalışmanın şekillenmesinde önemli bir rol oynayan sevgili danışmanım Prof.Dr. Aydın ERAR'a,

Bu çalışma süresince, kısıtlı zamanını bana ayırarak desteğini hiç eksiltmeyen ve motivasyonumu her zaman yüksek tutmamı sağlayan sevgili hocam Yrd.Doç.Dr Elif Özge ÖZDAMAR'a,

Mesai saatleri dışındaki dinlenme zamanlarından fedakârlık ederek, tezim için bana yardımcı olan sevgili Selim DELİLOĞLUNA' na

Kendisini tanıdığım günden beri hayranlığımı gizleyemediğim ve tanıştığımız günü sevgiyle hatırladığım, beni her konuda destekleyen ve cesaretlendiren sevgili nişanım Gökçem ÇELEBİ' ye

Üzerimde emeği olan ve adlarını buraya sığdıramadığım herkese teşekkürü bir borç bilirim.

Mayıs, 2011

Fahrettin CEVAHİR

İÇİNDEKİLER DİZİNİ

ÖZET.....	ii
SUMMARY.....	iv
ÖNSÖZ.....	v
ÇİZELGELER DİZİNİ.....	vi
ŞEKİLLER DİZİNİ.....	ix
KISALTMALAR DİZİNİ.....	xii
1.GİRİŞ.....	1
2.VERİ MADENCİLİĞİ TANIMI VE SÜRECİ.....	3
2.1 Veri Madenciliği Nedir?.....	3
2.1.1. Veri Madenciliği Süreci.....	4
2.1.1.1.Araştırma Probleminin Tanımlanması.....	4
2.1.1.2.Verileri Tanıma Aşaması.....	4
2.1.1.3.Verii Hazırlama Aşaması.....	5
2.1.1.4.Modelleme Aşaması.....	5
2.1.1.5.Değerlendirme Aşaması.....	5
2.1.1.6.Uygulama Aşaması.....	5
2.2. Veri Madenciliği Uygulamalarında Karşılaşılan Problemler.....	6
2.3.Karar Destek Sistemleri.....	8
2.3.1.Verii Tabanı Kavramı.....	8
2.3.1.1.Verinin Tekrarlanması Önler.....	9
2.3.1.2.Verinin Tutarlı Olmasını Sağlar.....	9
2.3.1.3.Aynı Andaki Erişimlerde Tutarlılıkların Ortaya Çıkmasını Önler.....	9
2.3.1.4.Verinin Güvenliğini Sağlar.....	9
2.3.2.Verii Ambarları.....	9
2.3.2.1.OLTP.....	10
2.3.2.2.OLAP.....	10
2.4.Verii Madenciliğinde Kullanılan Programlar.....	11
2.4.1. Verii Madenciliği Yazılımları.....	11
2.4.1.1. Intelligent Miner.....	12
2.4.1.2. Darwin.....	12
2.4.1.3. Enterprise Miner.....	12
2.4.1.4.SPSS Clementine.....	13
2.4.1.5.WEKA.....	13
2.5.Müşteri İlişkileri Yönetimi Ve Verii Madenciliği.....	13
3.VERİ ANALİZİ.....	16
3.1.Verii Tabanında Bilgi Keşfi Süreci.....	16

3.1.1 Veri Tabanında Bilgi Keşfi Sürecinin Evreleri.....	17
3.1.1.1 Problemin Tanımlanması.....	19
3.1.1.2. Verilerin Hazırlanması.....	19
3.1.1.3 Toplama (Collection).....	20
3.1.1.4 Değer Biçme (Assessment).....	20
3.1.1.5. Veri Dönüştürme.....	21
3.1.1.6. Veri İndirgeme.....	21
3.1.1.7. Birleştirme ve Temizleme (Consolidation and Cleaning).....	23
3.1.1.8. Seçim (Selection).....	23
3.1.1.9. Dönüştürme (Transformation).....	24
3.1.1.10. Modelin Kurulması ve Değerlendirilmesi.....	24
3.1.1.11. Modelin Kullanılması.....	28
3.1.1.12. Modelin İzlenmesi.....	28
3.2. Büyük Veri Kümelerinde EDA	29
3.3. Veri Çözümlemede Eksik Değerler.....	32
3.3.1. Mevcut Durum Çözümlemesi (Pairwise).....	33
3.3.2. Tam Durum Çözümlemesi (Listwise).....	33
3.4. Veri Çözümlemede Dönüşümler.....	34
3.4.1. Dönüşümler İçin Uyarılar.....	35
3.5. Düzleştirme / Düzleme (Smoothing).....	35
4. VERİ MADENCİLİĞİNDE MODELLEME.....	37
4.2. Sınıflama Ve Regresyon Modelleri.....	38
4.2.1. Karar Ağaçları.....	39
4.2.1.1 Karar Ağaçları'nda Ayırma Kriterleri.....	40
4.2.1.2. Karar Ağaçları'nda Durma Kriterleri.....	43
4.2.1.3. Karar Ağaçları'nda Budama.....	44
4.2.1.4. Kümeleme Yöntemi.....	47
5. UYGULAMA.....	52
5.1. Veri Manipülasyonu.....	55
5.1.1. Manipülasyon Çalışmalarına Ait Akışlar ve Değerlendirmeler.....	56
5.2. Modelleme Öncesi Veri Analizi.....	67
5.3. Modelleme.....	75
5.3.1. Karar Ağaçları Yöntemleri İle Sınıflandırma Teknikleri.....	75
5.3.1.1. EDA Kullanılarak C&R Tree Yöntemi İle Modelleme.....	75
5.3.1.2. EDA Kullanılmadan C&R Tree Yöntemi İle Modelleme.....	77
5.3.1.3. EDA Kullanılarak CHAID Yöntemi İle Modelleme.....	79
5.3.1.4. EDA Kullanılmadan CHAID Yöntemi İle Modelleme.....	81
5.3.2. Müşteri Segmentasyonu.....	83
5.3.3. Birliktelik Kuralları ve Sepet Analizi ile Müşteri Satın Alma Davranışı Modellemesi.....	89
6. SONUÇ VE ÖNERİLER.....	96
KAYNAKLAR.....	99
ÖZGEÇMİŞ.....	101

ÇİZELGELER DİZİNİ

	Sayfa No
Çizelge 2.1. Veri madenciliğinin uygulandığı alanlar	6
Çizelge 2.2. Veri Madenciliğinde kullanılan bazı yazılımlar ve özellikleri.....	12
Çizelge 3.1. Risk Matrisi.....	27
Çizelge 5.1.Ürün tablosuna ait alanlar	53
Çizelge 5.2. Satış tablosuna ait alanlar	54
Çizelge 5.3.Müşteri tablosuna ait alanlar	55

ŞEKİLLER DİZİNİ

	Sayfa No
Şekil 2.1. Veri madenciliğinin veri işleme süreci içindeki	4
Şekil 2.2.Müşteri verilerinin kullanım amaçları.....	15
Şekil 3.1. Veri Madenciliği süreci.....	19
Şekil 5.1.Ürün tablosuna ait veri seti.....	57
Şekil 5.2. Ürün dosyası için gerçekleştirilen manipülasyona ait akış.....	57
Şekil 5.3.Ürün kodunun oluşturulması	58
Şekil 5.4.Kategori kodlarının oluşturulması	58
Şekil 5.5. Kategorilere ait tanımlamalar	59
Şekil 5.6. Manipülasyon sonucu elde edilen ürün tablosu	59
Şekil 5.7.Müşteri tablosuna ait veri seti.....	60
Şekil 5.8.Müşteri dosyası için gerçekleştirilen manipülasyona ait akışı	60
Şekil 5.9. Müşterilerin firmaya kayıt tarihlerinin manipülasyon sonucu elde edilmesi	61
Şekil 5.10. Satış dosyası için gerçekleştirilen manipülasyona ait akış	62
Şekil 5.11.Satış tablosuna ait veri seti	62
Şekil 5.12.Ürün, müşteri ve satış dosyalarına ait manipülasyonların en son halini gösteren akış	63
Şekil 5.13.Birim maliyet dosyasına ait veri seti.....	64
Şekil 5.14.Birim maliyet dosyası için gerçekleştirilen manipülasyona ait akış	64
Şekil 5.15.Ürünlere ait birim maliyetlerin aylara göre elde edilmesi.....	65
Şekil 5.16.Manipülasyon çalışmaları sonucunda oluşturulan yeni iki dosyanın birleştirilmesi.....	66
Şekil 5.17.Analiz edilebilir hale getirilen veriden çeşitli nicel bilgilerin elde edilmesi	66
Şekil 5.18. Nicel değişkenlere ait dağılış histogramı.....	68

Şekil 5.19. Bağımlı ve bağımsız değişkenler arasındaki ilişkilerin incelenmesi (500 örneklem için).....	69
Şekil 5.20. Bağımlı ve bağımsız değişkenler arasındaki ilişkilerin incelenmesi (500 örneklem için).....	70
Şekil 5.21. Bağımlı ve bağımsız değişkenler arasındaki ilişkilerin incelenmesi (2000 örneklem için).....	71
Şekil 5.22.Bağımlı ve bağımsız değişkenler arasındaki ilişkilerin incelenmesi (2000 örneklem için).....	72
Şekil 5.23.Ürün kategorilerine ait boxplot grafiği.....	73
Şekil 5.24. Bölgelere ait boxplot grafiği.....	74
Şekil 5.25. C&R Tree yöntemi ile modellemeye ilişkin akış.....	75
Şekil 5.26. EDA kullanılarak oluşturulan C&R Tree algoritmasına ait karar ağacı...76	
Şekil 5.27.EDA kullanılarak oluşturulan C&R Tree algoritmasına ait Evaluation grafiği.....	77
Şekil 5.28. EDA kullanılmadan oluşturulan C&R Tree algoritmasına ait karar ağacı	78
Şekil 5.29. EDA kullanılmadan oluşturulan C&R Tree algoritmasına ait Evaluation grafiği.....	79
Şekil 5.30.CHAID yöntemi ile modellemeye ilişkin akış.....	79
Şekil 5.31.EDA kullanılarak oluşturulan CHAID Algoritmasına ait karar ağacı	80
Şekil 5.32.EDA kullanılarak oluşturulan CHAID algoritmasına ait Evaluation grafiği	81
Şekil 5.33.EDA kullanılmadan oluşturulan CHAID Algoritmasına ait karar ağacı...82	
Şekil 5.34. EDA kullanılmadan oluşturulan CHAID algoritmasına ait Evaluation grafiği.....	82
Şekil 5.35.EDA kullanılmadan oluşturulan, Özdeğer İstatistiğine bağlı faktör sayısı ve açıklanan varyans yüzdesi	84
Şekil 5.36.EDA kullanılarak oluşturulan, Özdeğer İstatistiğine bağlı faktör sayısı ve açıklanan varyans yüzdesi.....	84
Şekil 5.37.Faktör skorlarının elde edildiği akış.....	85
Şekil 5.38.K-Means yöntemi sonucunda oluşan kümeler.....	86
Şekil 5.39.Kümelerin Skor değerlerine göre yeniden sınıflandırılması.....	86
Şekil 5.40.Kümeler için oluşturulan C&R Tree algoritmasına ait ağaç yapısı	87
Şekil 5.41.Hedef kategorileri için kural kümeleri	88

Şekil 5.42.Müşteri segmentasyon çalışmasına ait akış	88
Şekil 5.43.Birliktelik kuralları analizi için verinin kodlanması	90
Şekil 5.44.Birliktelik analizine ait akış	90
Şekil 5.45.Ürün kategorilerine ait frekanslar	91
Şekil 5.46.Sıklığı yüksek olan bağıntılar	92
Şekil 5.47.Birliktelik kuralları analizi sonucu.....	93
Şekil 5.48.Sıklığı en yüksek olan kategorilere ilişkin C&R Tree algoritmasına ait ağaç yapısı	95

KISALTMALAR DİZİNİ

EDA	: Exploratory Data Analysis
CHAID	: Chi-squared Automatic Interaction Detector
C&RT	: Classification and Regression Trees
VM	: Veri Madenciliği
VTBK	: Veri Tabanlarında Bilgi Keşfi
OLTP	: Online Transaction Processing
OLAP	: Online Analytical Processing

1.GİRİŞ

Günümüzün ağır rekabet koşulları altında, rekabetçi avantaj elde edebilmek için, işletmelerin potansiyel tüketicilerine ya da müşterilerine ait verilerin düzenli bir şekilde, hassas, doğru ve çok boyutlu olarak analiz edilmesi gerekir. Verilerin büyük boyutlarda olması, istatistik yöntemlerin yanı sıra veri madenciliği yöntemlerinin de kullanılmasını gerekli kılmıştır.

Çok büyük boyutta müşteri verisine sahip şirketler, veri madenciliği uygulamaları ile müşterilerinin profillerini, ihtiyaçlarını, satın alma eğilimleri ve sadakatlerini belirleyerek, pazarlama stratejilerini kurdukları bu modeller üzerinden gerçekleştirmektedirler.

Veri Madenciliği uygulamalarının bu yöndeki amacı, toplanmış olan bilgilerin, bir takım sayısal yöntemlerle incelenmesi ve büyük boyutlardaki veri tabanlarında tutulan veriler arasındaki ilişkilerin, gizli örüntülerin ortaya çıkarılması ve keşfi işlemidir.

Veri madenciliğinin en uzun süreci, uygun tekniklerle gerekli işlemlerin yapılarak veri setinin analize hazır hale getirilmesini kapsayan veri manipülasyonu sürecidir. Bu sürecin uzun olması, veri madenciliğinde kullanılan veri setlerinin, farklı fiziksel kaynaklardan elde edilebilmesi, çok büyük boyutlarda olması ve ham verilerde görünen tutarsızlık gibi nedenlerden kaynaklanmaktadır.

Verilerin farklı sistemlerde ve farklı tip dosya formatlarında tutulmuş olması, analiz öncesinde birtakım işlemlerin yapılmasını gerekli kılmaktadır. Veri madenciliği tekniklerinin uygulanabilmesi için, öncelikle farklı sistemlerde farklı formatlarda tutulan verilerin birleştirilerek tek bir sisteme aktarılması ile veri setleri analiz edilebilir bir yapıya kavuşmaktadır.

Analize hazır hale getirilen veri setlerinden bilgi elde edilebilmesi için kullanılacak modelleme yöntemleri ve bunların algoritmaları belirlenmelidir. Veri madenciliği sonuçlarının elde edildiği ve kararların modelleme aşamalarında elde edilen bilgiler üzerinden verildiği düşünüldüğünde modelleme aşamaları büyük önem taşımaktadırlar.

Bu tez çalışmasının amacı, bir perakende firmasına ait farklı sistemlerde tutulan verileri, veri manipülasyonu yöntemleri ile birleştirerek analiz edilebilir hale getirmek ve analiz edilebilir hale getirilen bu veri seti üzerinde veri analizi ve veri madenciliği yöntemleri kullanılarak, veriler arasındaki ilişkilerin ve gizli örüntülerin ortaya çıkarılması amaçlanmıştır.

Tezin ikinci bölümünde veri madenciliğinin tanımı ve süreci anlatılmış; üçüncü bölümünde veri tabanında bilgi keşfi sürecinden ve veri analizi tekniklerinden bahsedilmiştir, dördüncü bölümde veri madenciliği modelleri ve teknikleri anlatılmıştır.

Beşinci bölümde, bir perakende işletmenin belirli bir dönem boyunca alışveriş yapmış müşteri verileri üzerinde, veri madenciliği uygulamalarında kullanılan SPSS Clementine programı ile veri manipülasyonu yapılmış ve manipülasyon sonucunda, veri analizi ve veri madenciliği yöntemleri kullanılarak, veriler arasındaki ilişkilerin ve gizli örüntülerin ortaya çıkarılmasına çalışılmıştır.

Tezin altıncı ve son bölümünde ise perakende firmasına ait veriler üzerinden yapılan uygulamanın detayları adımlar bazında ortaya konmuş ve elde edilen sonuçlar değerlendirilmiştir.

2.VERİ MADENCİLİĞİ TANIMI VE SÜRECİ

2.1 VERİ MADENCİLİĞİ NEDİR?

Veri Madenciliği genel anlamda, çok büyük miktarda veri içerisinde, gizli kalmış, değerli, kullanılabilir bilgilerin açığa çıkarılması biçiminde ifade edilmektedir. Veri Madenciliğindeki amaç, toplanmış olan bilgilerin istatistik yöntemlerle incelenip ilgili kurum ve yönetim destek sistemlerinde kullanılmak üzere değerlendirilmesidir. Veri madencisinin geleneksel yöntemlerde olduğunun aksine başlangıçta herhangi bir amacı ya da varmak istediği bir kavram yoktur. Yapılacak analizlerden sonra elde edilen verilerin bir istatistikçi gözü ile incelenip daha önceden düşünülmemiş kavramların ortaya çıkarılması, başarılı bir Veri Madenciliği süreci olarak kabul edilmektedir.

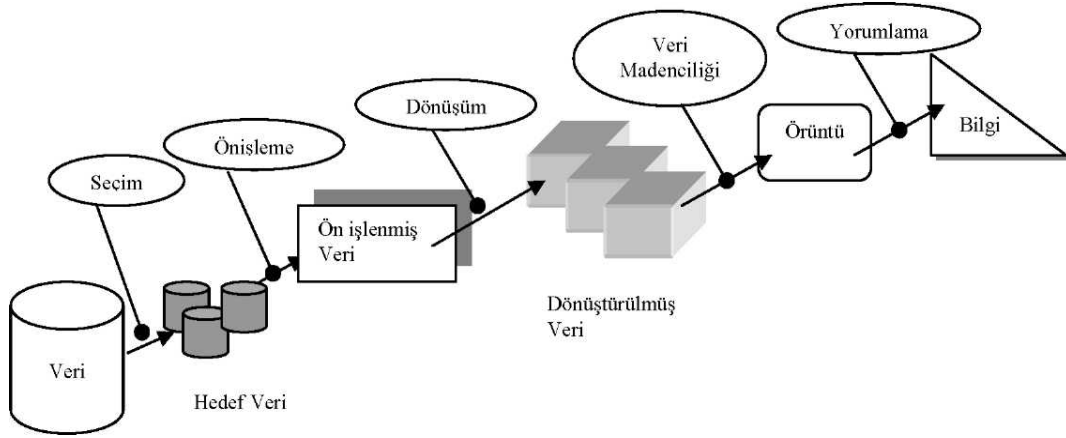
Veri madenciliği, birleşik verilerdeki gizli bilgileri bulmak ve iş uzmanlığını arttırmak amacıyla yapılan yeni bir karar destek analiz işlemidir. Bazı anahtar kelimeler kullanılarak 4 aşamalı ayrıntılı veri madenciliğinin tanımı şöyledir:

- Veri madenciliği, bir süreçtir.
- Veri madenciliği, karar destek araçlarının niteliğini yükseltir.
- Veri madenciliği, gizlenmiş bilgileri bulur.
- Veri madenciliği, iş uzmanları için kavrayış dağıtıcı bir sistemdir (Albayrak ve Yılmaz, 2009).

Veri madenciliği, Pres tarafından İstatistik, Veritabanı teknolojisi, bilgisayar bilimleri ve birkaç başka başka alanın birleştiği çok disiplinli bir alan olarak tanımlanmıştır. Veri madenciliği teknolojisinde kullanılan araçlar genelde, sınıflama, kümeleme, olumsuzluk, temel birleşenler, “Correspondence” analizi, çok boyutlu ölçekleme, regresyon etken çözümlemesi vb. istatistiğin popüler yöntemleridir.

Bunlara ek olarak, karar ağaçları, genetik algoritmalar, market-sepet analizi, sinir ağları gibi yöntemler de sıkça kullanılır (Erar ve Kiroğlu, 2008).

Şekil 2.1' de veri madenciliğinin veri işleme ve veriyi bilgiye dönüştürme süreci içerisindeki yeri gösterilmektedir (Akbulut, 2004).



Şekil 2.1. Veri madenciliğinin veri işleme süreci içindeki yeri

2.1.1. Veri Madenciliği Süreci

Veri madenciliği altı adımlı bir süreç olarak incelenebilir:

2.1.1.1. Araştırma Probleminin Tanımlanması (Business Understanding)

Bu aşama veri madenciliği sürecinin en önemli aşamasıdır. Araştırma probleminin (konusunun) tanımlanması aşaması araştırmanın amacını, mevcut durumun değerlendirilmesini, veri madenciliğinin amaçlarını ve proje planlama sürecinin belirlenmesini kapsamaktadır.

2.1.1.2. Verileri Tanıma Aşaması (Data Understanding)

Veri anlama aşaması veri toplamakla başlamaktadır. Daha sonra benzer verileri bir araya getirme, veri niteliklerini tanımlama, verileri keşfetme, gizli bilgileri sınıflandırma ile sürece devam etmektedir.

2.1.1.3. Veri Hazırlama Aşaması (Data Preperation)

Veri hazırlama aşaması, ham veriden başlayarak son veriye kadar yapılması gereken bütün düzenlemeleri içermektedir. Veri hazırlama, tablo, kayıt, veri dönüşümü ve modelleme araçları için veri temizleme gibi özellikleri içermektedir.

2.1.1.4. Modelleme Aşaması (Modelling)

Bu aşamada, verilerden bilgi çekmek için ileri istatistiksel çözümleme yöntemleri kullanıldığından veri madenciliği sürecinin en gösterişli aşamasıdır. Bu aşama uygun modelleme tekniğinin seçimi, test tasarımının üretimi, model geliştirme ve tahmin işlemlerini içermektedir. Uygun modellerin seçilip uygulanmasıyla birlikte parametreler en uygun değişkenlere dönüştürülmektedir. Veri madenciliği, her problem tipi için farklı yöntemler içermektedir. Bazı yöntemler, veri tipi için uygun değildir ya da özel tanımlamalar gerektirmektedir. Bu nedenle gerekli olduğunda 3. aşama olan veri hazırlama aşamasına geri dönülür.

2.1.1.5. Değerlendirme Aşaması (Evaluation)

Değerlendirme aşamasında, uygun model ya da modeller kurulduktan sonra, veri madenciliği sonuçlarının araştırma probleminin amaçlarını gerçekleştirip gerçekleştirmediği değerlendirilir. Bu aşama sonuçların değerlendirilmesi, veri madenciliği sürecinin gözden geçirilmesi ve sonraki adımların ne olacağı hususlarını içermektedir. Bu aşamanın sonunda veri madenciliği sonuçlarının kullanımı üzerinde ki karara varılmaktadır.

2.1.1.6. Uygulama Aşaması (Deployment)

Son aşama olan uygulama aşaması, araştırmacının tüm emeklerinin karşılığını aldığı bir aşamadır. Bu aşamada veri madenciliği süreciyle üretilen bilgiler, pratik işletme problemlerinin çözümünde kullanılmaktadır. Bu aşamada elde edilen bilgilerin uygulanabilmesine yönelik bir plan hazırlama, gözden geçirme ve bakım faaliyetlerini içerir. Ayrıca bu aşamada nihai araştırma raporunun yazılması ve projenin gözden geçirilmesi işlemleri yer almaktadır (Albayrak ve Yılmaz, 2009).

2.2. VERİ MADENCİLİĞİ UYGULAMALARINDA KARŞILAŞILAN PROBLEMLER

Çizelge 2.1.'de 2003 yılında veri madenciliğinin sektörler bazında kullanımına ilişkin bir araştırmanın sonuçları yer almaktadır. Bu çizelgede araştırmaya katılan toplam 421 şirketin 25 adedinin perakende alanında veri madenciliğinin kullandığı görülmektedir (Akbulut, 2004).

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%
Sahtekârlık 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%

Çizelge 2.1. Veri madenciliğinin uygulandığı alanlar (Akbulut, 2004).

Veri madenciliği girdi olarak kullanılacak ham veriyi veritabanlarından alır. Bu da veritabanlarının dinamik, eksiksiz, geniş ve net veri içermemesi durumunda sorunlar doğurur (Aydoğan, 2003).

Veri tabanları genel olarak basit öğrenme işlerini sağlayan özellik veya nitelikleri sunmak gibi amaçlar için hazırlanmışlardır. Bu yüzden, öğrenme görevini kolaylaştıracak bazı özellikler bulunmayabilir. Örneğin, hasta veri tabanı kırmızı kan hücreli hasta bilgilerini barındırmıyorsa hasta veri tabanından sıtma teşhisi yapılamaz.

Veri özellikleri ya da sınıflarındaki hatalara gürültü adı verilir. Veri tabanlarındaki eksik bilgi ve bu yanlışlardan dolayı veri madenciliği amacına tam olarak ulaşmayabilir. Bu bilgi yanlışlığı, ölçüm hatalarından, ya da öznel yaklaşımdan olabilir (Akbulut, 2004).

Veri tabanı boyutları inanılmaz bir hızla artmaktadır. Veri tabanı algoritması çok sayıda küçük örnekleme ele alabilecek biçimde geliştirilmiştir. Aynı algoritmaların yüzlerce kat büyük örneklerde kullanılabilmesi için çok dikkat gerekmektedir. Örneklemin büyük olması, tahminlerin doğruluğu açısından bir avantaj olsa da dikkatsizlikten kaynaklanacak hatalar göz ardı edilemez.

Veri tabanlarının boyutları sebebiyle, veri madenciliği yöntemlerinden herhangi birinin ham veriyle başarılı olma olasılığı yoktur. Veri madenciliği yöntemleri, bu şekilde elde edilen sonuçların tüm veri tabanını temsil edebileceğini umarak, veri tabanlarından bir örneğin çıkarılmasını gerektirebilir.

Bir veri tabanının boyutlarının küçültülmesi 2 yolla olabilir:

Veri alanında örnekleme: Genellikle rastgele bazı kayıtlar seçilir ve veri madenciliğinin sonraki aşamalarında kullanılır.

Özellik alanında örnekleme: Her veri kaydının bazı özellikleri seçilir. Yine birçok özellik varsa seçim rastgele yapılır (Tezcanlar ve Timor, 2007).

2.3.KARAR DESTEK SİSTEMLERİ

Karar destek sistemleri, değişik kaynaklardan topladığı bilgileri düzenleyerek, kararı modelleyerek, bilgileri analiz ederek ve değerlendirme sonuçlarını sunarak karar vericiye seçim sırasında destek veren bilgisayar tabanlı sistemlerdir. “Bir karar verici için verilen kararın doğruluğu, onun yeteneklerine, deneyimine ve bilgi birikimine olduğu kadar sahip olduğu veri kümesinin yeterliliğine de bağlıdır. Diğer bir deyişle kararın başarısında, verilerin doğru depolanması, doğru sınıflanması, doğru ayıklanıp işlenmesi ve doğru yorumlanması çok önemli bir rol oynar (Yaralıoğlu, 2004).

Günümüzde yöneticiler için karar alma ortamı eskisine oranla daha karmaşık hale gelmiş, yöneticilerin kararları üzerinde etkili olabilecek faktörler artmış, iç ve dış çevredeki değişiklikler ve özellikle kriz durumları karar destek sistemlerine olan eğilimi artırmıştır.

Karar destek sistemleri, yöneticilerin programlanamayan türden karar verme işlemlerine yardımcı olmak üzere geliştirilir. Yöneticinin herhangi bir anda, daha önceden öngörülmemiş bir bilgiye aniden gereksinim olabilir. Böyle bir durumda, hemen yanıt verebilecek bir sistemin varlığı gerekecektir. Karar destek sistemleri bu gibi durumlar için tasarlanır (Özkan, 2008).

2.3.1.Veritabanı Kavramı

Elde edilen verilerin tutulduğu alanlardır. “Bir veri tabanı sistemi, birbiri ile ilişkili verilerin birikimini içeren, veriye erişimi sağlayarak veriyi yönetmeye yardımcı olan yazılım programları kümesidir (Han and Kamber, 2006).

Veritabanı sistemleri yazılımları bilgisayar sistemlerinin önemli bir bileşeni olarak değerlendirilir. Veritabanı yönetim sistemleri, birbirleriyle ilişkili veri ve programlar topluluğundan oluşmaktadır. Veri topluluğu bir “veritabanı” olarak değerlendirilir. Veritabanı bir kuruluşa ilişkin bilgilerin yer aldığı ortamdır. O halde veri tabanı sistemlerini, veri kümesinin düzenli biçimde tutulduğu ve bu verinin çeşitli yazılımlar aracılığıyla yönetildiği bir ortam olarak düşünebiliriz.

Veri tabanı sistemlerinin yararları aşağıda belirtildiği biçimde özetlenebilir:

2.3.1.1.Verinin Tekrarlanması Önler

Veri tabanı sistemleri, uygulamaların bir bütün olarak tasarlanmasını, alt sistemler arasında ilişkinin kurulmasını ve birden çok uygulamada verinin aynı veritabanı içinde ortak kullanılacak biçimde tasarlanmasını öngörür. Tüm uygulamaların gereksinim duyduğu veri birbirleriyle bütünleşik yapıdadır. Dolayısıyla veri kaynağı tek olarak tasarlanır ve böylece veri tekrarları önlenmiş olur.

2.3.1.2.Verinin Tutarlı Olmasını Sağlar

Veri tabanı sistemlerinin en önemli üstünlüklerinden birisi, veri bütünlüğü (data integrity) sağlamasıdır. Veri bütünlüğü, verinin doğruluğu ve tutarlılığını anlatmaktadır. Veritabanındaki verinin doğruluğu ve tutarlı bir bütün oluşturmasını sağlamak için birtakım kısıtlamalara yer verilebilir.

2.3.1.3.Aynı Andaki Erişimlerde Tutarlılıkların Ortaya Çıkmasını Önler

Veri tabanı uygulamalarında veritabanı nesneleri başka başka uygulamalar tarafından paylaşılabilir. Veri aynı anda farklı uygulamalar ve dolayısıyla farklı kullanıcılar tarafından aynı anda paylaşılabilir. Bu şekilde birlikte kullanımın yol açabileceği sorunlar ortaya çıkması durumunda veri tabanı sistemi söz konusu sorunu otomatik olarak çözer.

2.3.1.4.Verinin Güvenliğini Sağlar

Bazı uygulamaların ürettiği verinin güvenliğini sağlamak önemli konudur. Veritabanı kullanıcıların, veritabanının içerdiği tüm bilgilere kolayca erişmesi istenilen bir durum değildir (Oğuzlar, 2003).

2.3.2.Veritabanları

Günümüzde bilişim sistemleri sadece işlemsel düzeydeki personele değil, kurumun üst düzey yöneticilerine de hizmet vermeye odaklanmıştır. Veritabanı sistemlerinin sayılan avantajlarına rağmen, özellikle karar destek uygulamalarında gereksinimleri karşılamakta zorlandığı görülmüştür. Bu gereksinime paralel olarak verinin farklı biçimde saklanması ve veriyi hızlı erişimin sağlanabilmesini sağlayacak başka yollar aranmaya başlanmıştır. Bunun da ötesinde kurumun ürettiği tüm bilgileri, boyutu ne

olursa olsun işlemek söz konusu olacaktır. Söz konusu verinin boyutları bazı kurumlarda o denli fazladır ki, bu verileri geleneksel veritabanı yöntemleriyle işlemek olanaksızdır. Veri ambarı adı verilen kavram bu tür gereksinimlerden dolayı ortaya çıkmıştır (Özkan, 2008).

Veri Madenciliği sık sık veri ambarlarıyla karıştırılmaktadır. En basit anlamda veri madenciliği ve veri ambarları, birbirlerinin tamamlayıcısıdır. Veri ambarları verinin belli bir yapıda saklanması için kurulurken, veri madenciliği bu saklanan verinin bilgiye dönüştürülmesini sağlar. Kısaca veri ambarları, veri madenciliğinin omurgası gibidir (Tezcanlar ve Timor, 2007).

Birçok organizasyon veri ambarlarından çeşitli araçlarla elde ettiği bilgileri, karar alma süreçlerini desteklemek amacıyla kullanır. Bunlara ek olarak veri ambarlarının diğer kullanım amaçları şunlardır:

- Müşterilerin satın alma şekilleri üzerinde analizler yaparak satın alma alışkanlıklarını ortaya çıkarmak.
- Üretim stratejilerinde ince ayarlar yapmak için; satış performanslarını; yıllar, coğrafik bölgeler vb. gibi değişkenlere göre karşılaştırarak ürün portföylerini yönetmek.
- İşlemlerin analizi ve yeni kar alanları yaratmak.
- Müşteri ilişkilerini yönetmek (Han and Kamber, 2006).

2.3.2.1.OLTP

Organizasyonda satın alma, kaydetme, muhasebe, bankacılık gibi günlük işlemlerin yapıldığı işlemsel veritabanı sistemleridir. Detaylı bilgi içerir ve ayrıntılı görüntüye sahiptirler. Veriye erişim sağlanabilir, üzerinde oynama yapılmasına izin verir. Saklanan kayıt sayısı sınırlıdır (Yaralıoğlu, 2004).

2.3.2.2.OLAP

OLAP kavramı ilk olarak ilişkisel veri tabanı teknolojisinin babası sayılan Dr.E.F.Codd tarafından ortaya atılmıştır. OLAP son kullanıcıların her gün ihtiyaç duydukları rapor ve analizleri karşılayan özel bir teknoloji olarak tanımlanabilir. OLAP işlemsel sistemlerin ve ilişkisel veri tabanlarının tamamlanmasını sağlar. Bu teknoloji sayesinde veriler, sadece ihtiyaç duyuldukları zamanlarda değil, sürekli

olarak toplanır. Verilerin bu şekilde elde edilişı istatistiksel değerdendirmelerin çok daha sağılıklı yapılmasına imkân verir. Özellikle son yıllardaki bilgi teknolojilerindeki gelişmeler OLAP uygulamalarını yaygınlaştırmıştır (Şentürk, 2006).

Önceki söylediklerimizi ve işlemsel sistemlerin özelliklerini de göz önüne alarak veri ambarları ile ilgili aşağıda belirtilen hususları sıralayabiliriz:

- a) İşlemsel çevrede yer alan veri bir süzme işlemi sonucunda veri ambarı çevresine aktarılır. İşlemsel ortamdaki verinin tümüyle veri ambarına aktarılması beklenmez. Sadece Karar Destek Sistemlerinin gereksinim duyacağı veri aktarılır.
- b) Zaman yelpazesi her iki sistemde farklılık gösterir. İşlemsel ortamdaki veri çok tazedir. Veri ambarındaki veri ise daha eskidir. Ancak zaman açısından bakıldığında, işlemsel ve veri ambarı ortamındaki veri arasında kısa bir örtüşme vardır.
- c) Veri ambarı özet bilgileri içerebilir. İşlemsel sistemlerde ise özet veriye yer verilmez.

Veri ambarının en önemli özelliklerinden biri olan bütünleştirmeyi sağlamak üzere, verinin önemli bir kısmı belirli bir dönüşümden sonra veri ambarına aktarılır. Böylece işlemsel veri ile veri ambarındaki verinin içeriğı açısından farklılıklar oluşabilecektir (Özkan, 2008).

2.4.VERİ MADENCİLİĞİNDE KULLANILAN PROGRAMLAR

2.4.1. Veri Madenciliğı Yazılımları

Farklı algoritmaları bünyesinde bulunduran ve farklı işletim sistemleri üzerinde çalışabilen birçok veri madenciliğı yazılımı bulunmaktadır. Çizelge 2.2.'de, bu yazılımlardan bazıları, üzerinde çalıştıkları platform ve fonksiyonel özellikleri ile birlikte gösterilmektedir.

Ürün Adı	IBM Intelligent Miner	Oracle Darwin	SAS Enterprise Miner	SPSS Clementine	WEKA
Platform	AIX 4.1. NVS AS/400 Windows	Windows	Windows Macintosh Unix	Windows Macintosh Unix	Windows Macintosh Unix
Karar Ağacı	x	x	x	x	x
Sinir Ağları	x	x	x	x	x
Zaman Serileri	x		x	x	
Tahmin	x	x	x	x	x
Kümeleme	x		x	x	x
Birliktelik	x		x	x	x
Görselleştirme	x	x	x	x	x

Çizelge 2.2. Veri Madenciliğinde kullanılan bazı yazılımlar ve özellikleri

2.4.1.1. Intelligent Miner

Eşleştirme, sınıflama, tahmin modelleme, gruplama, sıralı desen analizi, regresyon analizi gibi fonksiyonları içeren IBM firmasının geliştirdiği bir yazılımdır. Ayrıca yapay sinir ağları algoritmaları, istatistik metotları, veri hazırlama ve görsel gösterim araçları gibi ek özellikler içerir.

2.4.1.2. Darwin

Darwin Oracle firmasının VM aracıdır. Darwin regresyon ağaçları, karar ağaçları, kümeleme, yapay sinir ağları, Bayesian öğrenme, k-yakınlığında komşuluk gibi birçok algoritmayı destekleyen bir VM aracıdır. Darwin kullanımı kolay bir ara yüze sahiptir. Darwin VM algoritmalarından C&RT, StarTree, StarNet ve StarMatch'i kullanır.

2.4.1.3. Enterprise Miner

SAS firmasının VM aracıdır. Enterprise Miner karar ağaçları, yapay sinir ağları, regresyon analizi, kümeleme, zaman serileri, ilişkilendirme, vb. VM sorgularını ele alabilmektedir. Grafiksel ara yüzü sayesinde kullanım kolaylığı sağlar ve kullanıcılar

uygulamanın karmaşıklığından habersiz bir şekilde sadece girdi ve çıktılarına yoğunlaşabilirler.

2.4.1.4.SPSS Clementine

İstatistiksel analize yönelik bir bilgisayar programıdır. Özellikle Sosyal Bilimlerde, pazar araştırmalarında, sağlık araştırmalarında başta anket şirketleri, hükümetler ve eğitim kurumları olmak üzere pek çok kurum tarafından kullanılan bir veri madenciliği yazılımıdır.

Kullanımı grafiksel bir kullanıcı ara yüzüne bağlı olup, açılır menüler yardımıyla kolaylaştırılmıştır. Ayrıca makro dilleri yardımıyla kullanıcı kendi amaçları doğrultusunda programı yönlendirebilmektedir. Programının Windows, Mac OS ve UNIX işletim sistemleri için farklı sürümleri mevcuttur. Windows sürümü daha sık güncellenmekte ve diğer versiyonlara göre daha fazla özellik içermektedir.

2.4.1.5.WEKA

Yeni Zelanda Waikito Üniversitesinde, obje yönelimli programlama dillerinden biri olan Java ile geliştirilmiş ve halen yeni sürümleri geliştirilmeye devam eden açık kaynak kodlu bir VM yazılımıdır. WEKA' nın Java ile geliştirilmiş olması, Linux, Unix, Windows ve Macintosh gibi baslıca işletim sistemi platformlarında kullanılabilirliğini sağlamıştır.

2.5.MÜŞTERİ İLİŞKİLERİ YÖNETİMİ VE VERİ MADENCİLİĞİ

Değişim, günümüz iş dünyasını derinden etkileyen kavramların başında gelmektedir. Değişimin yönlendirdiği ve değişim odaklı bir dünyada, değişime uyum sağlama gücü ve yeteneği olan işletmeler ve organizasyonlar varlıklarını sürdürüp, ayakta kalabileceklerdir.

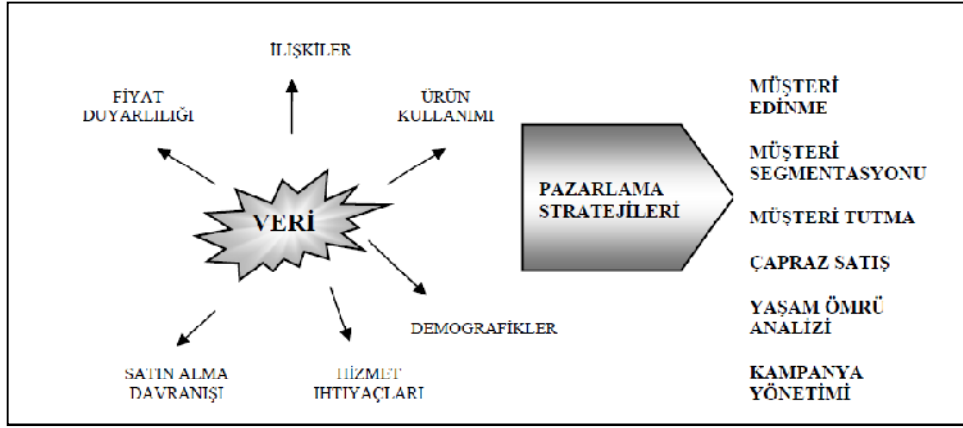
Bu bağlamda değişim, müşteri satın alma davranışlarını ve beklentilerini, büyük ölçüde değiştirmiştir. Değişen müşteri davranışlarını, işletme amaç ve hedeflerine uygun olarak yönetebilmek; değişim yönetimi ve müşteri ilişkileri yönetimi kavramlarını anlayabilmek ve doğru bir şekilde yorumlayabilmek ile mümkün olabilecektir. İşletmelerde başarılı bir müşteri ilişkileri yönetimiyle, küresel bir

dünyada ve giderek artan rekabetçi piyasalarda, işletmeler için yaşamsal önem taşıyan, müşteri için değer yaratmak, müşteri sadakati sağlamak ve bu konularda kurumsallaşmayı gerçekleştirmek mümkün olabilecektir.

İşletmeler açısından uzun dönemli müşteri bağlılığı, yeni rekabet alanında önemli bir değişimi beraberinde getirmiştir. Sanayi devrimi ile ortaya çıkan kitle üretim ve kitle pazarlama, yerini merkezde müşterinin yer aldığı yaklaşımlara bırakmıştır. Eski "tasarla-yap-sat" modelinin (ürün odaklı yaklaşım) yerini "sat-yap-yeniden tasarla" modeli (müşteri odaklı yaklaşım) almıştır. Bu bağlamda müşteri sadakatini gerçekleştirmek ve sürekli kılmak için çeşitli yöntem ve tekniklerin kullanıldığı görülmektedir. Bu yöntemlerden en önemli olanı, müşteriyi merkeze alan ve müşteri memnuniyetinin ötesine geçmeyi amaçlayan müşteri ilişkileri yönetimi yaklaşımıdır. Bu yaklaşım aynı zamanda yeni pazarlama anlayışını da ortaya koymaktadır.

Müşteri ilişkileri yönetimi, geleneksel pazarlama anlayışından çok farklı özellikler taşımaktadır. Geleneksel pazarlama anlayışında kitlesel üretimin etkisiyle kitlesel pazarlama anlayışı geçerli durumda idi. Oysa müşteri ilişkileri yönetiminde pazar payı kavramından, müşteri payı kavramına geçiş söz konusu olmaktadır. Müşteri payı aynı müşteriye birden fazla ürün satabilmeyi ve müşteriyi aktif ve sadık bir müşteriye dönüştürmeyi amaçlamaktadır. Bu durum ise tek yönlü pazarlamadan karşılıklı ilişkiye dayalı pazarlamaya geçişi ifade etmektedir. Günümüzde kitlesel pazarlama yöntemleri, giderek müşteriye hitap eden kişisel pazarlama yöntemlerine yönelmektedir.

Diğer taraftan yazılım ve donanım teknolojilerindeki gelişmeler, işletmelerin çok miktarda veriyi saklayabilmesini ve işleyerek anlamlı bilgilere dönüştürmesini mümkün hale getirmiştir. Bugün, işletmeler satın alma davranışından demografik özelliklere kadar müşterileri ilgili pek çok detayı veritabanlarında tutmaktadır. Veri madenciliği tekniklerini kullanarak bu veriler içerisindeki anlamlı ve gizli örüntülerin ortaya çıkarılması mümkün olmaktadır. Şirketlerin veritabanlarında tuttukları veriler ve bu verilerin hangi amaçlarla kullanıldığı Şekil 2.2' de gösterilmektedir.



Şekil 2.2. Müşteri verilerinin kullanım amaçları

Veri madenciliği sonuçları müşteri edinme, müşteri segmentasyonu, müşteriyi elde tutma, ayrılma eğilimi gösteren müşteri profilini ortaya koyma, müşteri değerlendirme, kredi derecelendirme ve pazar sepeti analizi gibi pek çok müşteri odaklı uygulamaya girdi teşkil etmektedir (Akbulut, 2004).

3.VERİLERİN ÖN İŞLEME SÜRECİ VE EDA

3.1.VERİ TABANINDA BİLGİ KEŞFİ SÜRECİ

Günümüz modern insanının her alışverişinde, her bankacılık işleminde, her telefon edişinde kaydedilen, uzaktan algılayıcılardan, uydulardan toplanan, devlet ve işletme yönetiminde yapılan işlemler sonucunda saklanan veriler her an inanılmaz boyutlarda artmaktadır. Sadece uydu ve diğer uzay araçlarından elde edilen görüntülerin saatte 50 gigabyte düzeyinde olması, bu artışın boyutlarını daha açık bir şekilde göstermektedir.

1995 yılında birincisi düzenlenen *Knowledge Discovery in Databases* konferansı bildiri kitabı sunuşunda, enformasyon teknolojilerinin oluşturduğu veri dağları aşağıdaki cümleler ile vurgulanmaktadır.

“Dünyadaki enformasyon miktarının her 20 ayda bir ikiye katlandığı tahmin edilmektedir. Bu ham veri seli ile ne yapmamız gerekmektedir. İnsan gözleri bunun ancak çok küçük bir kısmını görebilecektir. Bilgisayarlar bilgelik pınarı olmayı vaat etmekte, ancak veri sellerine neden olmaktadır.

Dünyadaki en büyük işletme veri tabanlarının belirlenmesi amacı ile Winter Corporation tarafından yapılan bir araştırmada, Sears, Roebuck and Co.’nun sadece karar destek amaçlı kullanılan veri tabanının 1998 yılında 4630 *gigabyte*’a eriştiği görülmektedir.

Veri tabanı sistemlerinin artan kullanımı ve hacimlerindeki bu olağanüstü artış, organizasyonları elde toplanan bu verilerden nasıl faydalanılabileceği problemi ile karşı karşıya bırakmıştır. Geleneksel sorgu (*Query*) veya raporlama araçlarının veri yığınları karşısında yetersiz kalması, *Veri Tabanlarında Bilgi Keşfi-VTBK*

(*Knowledge Discovery in Databases*) adı altında, sürekli ve yeni arayışlara neden olmaktadır.

3.1.1 Veri Tabanında Bilgi Keşfi Sürecinin Evreleri

Ne kadar etkin olursa olsun hiç bir veri madenciliği algoritmasının üzerinde inceleme yapılan işin ve verilerin özelliklerinin bilinmemesi durumunda fayda sağlaması mümkün değildir. Bu nedenle aşağıda tanımlanan tüm aşamalardan önce, iş ve veri özelliklerinin öğrenilmesi / anlaşılması başarının ilk şartı olacaktır.

- Problemin Tanımlanması,
- Verilerin Hazırlanması,
- Modelin Kurulması ve Değerlendirilmesi,
- Modelin Kullanılması,
- Modelin İzlenmesi

veri tabanlarında bilgi keşfi sürecinde izlenmesi gereken temel aşamalardır (Akpınar, 2000).

Klasik istatistiksel uygulamalar ve veri madenciliği arasındaki en temel farklılık, veri kümesinin büyüklüğüdür. Bir istatistikçi için ‘büyük’ veri kümesi birkaç yüz veya bin veri içerir. Veri madenciliği ile uğraşan birileri için ise milyon veya milyarlık veri beklenmeyen bir sayı değildir. Bu tip büyük veri tabanları gerçek hayatta sıkça ortaya çıkmaktadır.

Çok fazla sayıda değişken olduğunda daha ileri düzeyde zorluklar meydana gelmektedir. Değişken sayısı arttığında uzaydaki birim hücrelerin sayısında üstel oranda artış olacaktır. Örneğin bir tek ikili değişken ele alalım. Bu değişkenin hücrelerinin her biriminde 10 gözlem olduğunu varsayarsak, toplam hücre sayısı 20 olacaktır. İki tane ikili değişken olduğunda gözlem sayısı 40 olurken, 10 tane ikili değişken için gözlem sayısı 10240 ve 20 değişken için sayı 10485760 olacaktır. Ayrıca yüksek boyut sayısına sahip uzayda en yakın noktalar birbirlerinden çok uzakta olabilmektedir. Bazı durumlarda modelin ön seçimiyle ek kısıtlamalar yapılabilir (örneğin doğrusal modellerin varsayılması gibi).

Büyük veri kümelerine erişilmesinin zorlukları açıktır. Pek çok durumda veriler karışık bir haldedir ve farklı bilgisayarlara yüklenilmiş olabilir. Bu durumda verilerden bir rassal örneklem elde edilmesi için çerçevenin tanımlanması ve verilere erişimin ne kadar süreceği önemli konular olacaktır.

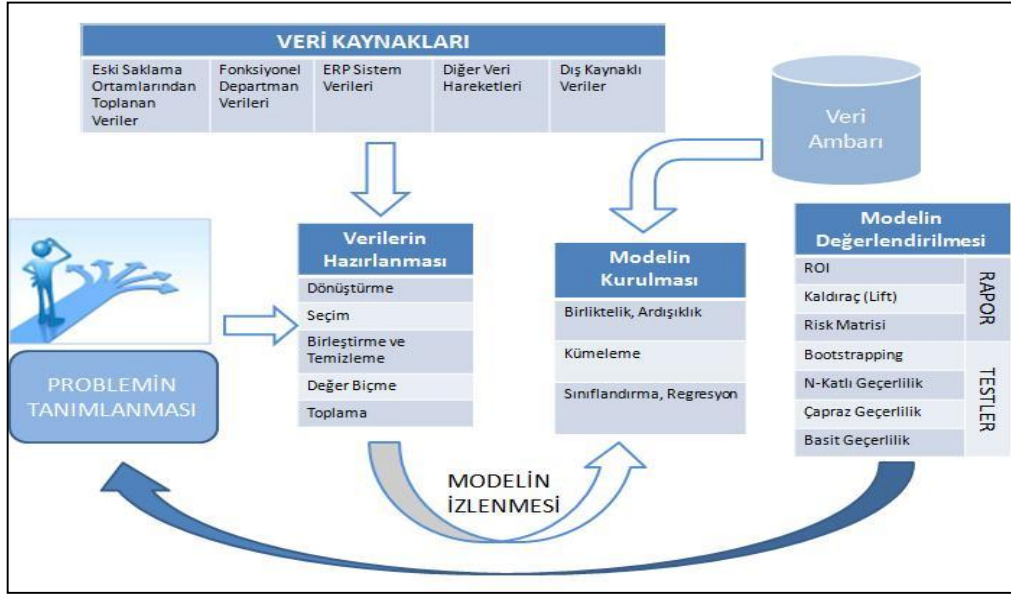
Veri kümesinin büyüklüğü zorluklara yol açarken, standart istatistiksel uygulamalarda sık karşılaşılmayan bir takım özellikler ortaya çıkabilir. Veri madenciliğinde veriler, veri madenciliği uygulamak üzere değil diğer bazı amaçlar için toplanmaktadır. Tersine bir biçimde, pek çok istatistiksel çalışmada veriler akıldaki belirli sorular için toplanır ve bu sorulara yanıt bulmak için analiz edilir.

Verilerin toplanmasında ortaya çıkan problemlerin yanında, büyük veri kümeleri ile çalışılırken başka bir takım problemler de oluşabilir. Büyük veri kümeleri çoğunlukla eksik, kirli ve hatalı veri noktalarını içerecektir. Bu tip hatalara sahip olmayan veri kümeleri az rastlanılan veri kümeleridir. EM algoritması gibi bir tahmin yöntemi veya bir yerine koyma (imputasyon) yöntemi, eksik veriler için kullanılacak aynı genel dağılım özelliklerine sahip yapay veri üretilmesine yardımcı olur. Bu problemlerle standart istatistiksel uygulamalarda da karşılaşılmakla beraber, veri kümesi nispeten küçük olduğundan veri madenciliği uygulamalarındaki gibi büyük problemlere yol açmamaktadır. Özetlemek gerekirse veri madenciliğinde pek çoğu büyüklük ve ele alınan veri kümesinin doğasından kaynaklanan yeni problemler ortaya çıkmaktadır.

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.

Veri ön işleme iki farklı türde işlemi gerektirir. Bu farklı işlemlerin ilki veri kümesinin seçilmesi ve birleştirilmesi, ikincisi ise veri madenciliği için verilerin daha yararlı bir hale getirilmesi amacıyla verilerin işlenmesidir.

Verilerin seçimi, verilerin özellikleri ve veri büyüklüğü ile veri türü gibi teknik sınırlamaların elverdiği ölçüde gerçekleştirilmektedir (Oğuzlar, 2003).



Şekil 3.1. Veri Madenciliği süreci (Akpınar, 2000)

3.1.1.1 Problemin Tanımlanması

Bu adımda ihtiyaç duyulan şeyin tanımlanması için cevaplanması gereken sorular neyin otomatize edilmeye değer olduğu ve neyin insan içeren süreçlere bırakılması gerektiği, amacın ne olduğu ve hangi performans kriterlerinin daha önemli olduğu, sürecin sonucunda elde edilecek çıktının keşif, sınıflandırma, özetleme gibi şeyler için kullanılıp kullanılmayacağı olabilir (Çelik, 2009).

Veri madenciliği çalışmalarında başarılı olmanın ilk şartı, uygulamanın 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 (Akpınar, 2000).

3.1.1.2. Verilerin Hazırlanması

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

ortalama deęerini verme, gürültülü deęerlerde regresyon ile belirli fonksiyonel kalıba sokma gibi yöntemler ile verilerdeki sıkıntı giderilebilir.

Farklı kaynaklardan gelen, aynı deęişkene ait verilerin tiplerinde, alan isimlerinde uyumsuzluk olması halinde gerekli deęişikliklere gidilerek tüm verileri bir arada tutabilecek yapı oluşturulmalıdır (Çelik, 2009).

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 (Akpınar, 2000).

Verilerin hazırlanması aşaması kendi içerisinde toplama, deęer biçme, birleştirme ve temizleme, seçme ve dönüştürme adımlarından meydana gelmektedir.

3.1.1.3 Toplama (Collection)

Tanımlanan problem için gerekli olduęu düşünölen verilerin ve bu verilerin toplanacaęı veri kaynaklarının belirlenmesi adıımıdır. Verilerin toplanmasında kuruluşun kendi veri kaynaklarının dışında, nüfus sayımı, hava durumu, merkez bankası kara listesi gibi veri tabanlarından veya veri pazarlayan kuruluşların veri tabanlarından faydalanılabilir.

3.1.1.4 Deęer Biçme (Assessment)

Veri madencilięinde kullanılacak verilerin farklı kaynaklardan toplanması, doęal olarak veri uyumsuzluklarına neden olacaktır. Bu uyumsuzlukların başlıcaları farklı zamanlara ait olmaları, kodlama farklılıkları (örneęin bir veri tabanında cinsiyet özellięinin e/k, dięer bir veri tabanında 0/1 olarak kodlanması), farklı ölçü birimleridir. Ayrıca verilerin nasıl, nerede ve hangi koşullar altında toplandıęı da önem taşımaktadır.

Bu nedenlerle, iyi sonuç alınacak modeller ancak iyi verilerin üzerine kurulabileceęi için, toplanan verilerin ne ölçüde uyumlu oldukları bu adımda incelenerek deęerlendirilmelidir (Akpınar, 2000).

3.1.1.5. Veri Dönüştürme

Veri dönüştürme ile veriler, veri madenciliği için uygun formlara dönüştürülürler. Veri dönüştürme; düzeltme, birleştirme, genelleştirme ve normalleştirme gibi değişik işlemlerden biri veya bir kaçını içerebilir. Veri normalleştirme en sık kullanılan veri dönüştürme işlemlerinden birisidir. Veri normalleştirme tekniklerinden bazıları aşağıdaki biçimde sıralanabilir.

- Min-Max
- Z Skor
- Ondalık Ölçekleme

Min-max normalleştirme ile orijinal veriler yeni veri aralığına doğrusal dönüşüm ile dönüştürülürler. Bu veri aralığı genellikle 0-1 aralığıdır.

Z Skor normalleştirmede (veya 0 ortalama normalleştirme) ise değişkenin her hangi bir y değeri, değişkenin ortalaması ve standart sapmasına bağlı olarak bilinen Z dönüşümü ile normalleştirilir:

Ondalık ölçekleme ile normalleştirmede ise, ele alınan değişkenin değerlerinin ondalık kısmı hareket ettirilerek normalleştirme gerçekleştirilir. Hareket edecek ondalık nokta sayısı, değişkenin maksimum mutlak değerine bağlıdır. Ondalık ölçeklemenin formülü aşağıdaki şekildedir:

Örneğin 900 maksimum değer ise, $n=3$ olacağından 900 sayısı 0,9 olarak normalleştirilir.

3.1.1.6. Veri İndirgeme

Veri indirgeme teknikleri, daha küçük hacimli olarak ve veri kümesinin indirgenmiş bir örneğinin elde edilmesi amacıyla uygulanır. Bu sayede elde edilen indirgenmiş veri kümesine veri madenciliği teknikleri uygulanarak daha etkin sonuçlar elde edilebilir.

Veri indirgeme yöntemleri aşağıdaki biçimde özetlenebilir:

1. Veri Birleştirme veya Veri Küpü (Data Aggregation or Data Cube)
2. Boyut indirgeme (Dimension Reduction)
3. Veri Sıkıştırma (Data Compression)
4. Kesikli hale getirme (Discretization)

Veri birleştirme veya veri küpü yapılacak 2000-2003 yılları için çeyrek dönemlik satış tutarlarından oluşan bir veri kümesinin bulunduğunu varsayalım. Bu yıllar için yıllık satış tutarları tek bir tabloda toplandığında veri birleştirilmesi yapılmış olur. Sonuç olarak elde edilen veri kümesinin hacmi daha küçüktür fakat yapılacak analiz için bir bilgi kaybı söz konusu değildir. Veri küpleri ise çok değişkenli birleştirilmiş bilginin saklandığı küplerdir. Örneğin bir firmanın satış tutarları yıllar, satışı yapılan ürünler ve firmanın farklı satış yerleri için aynı küp üzerinde gösterilebilir. Veri küpleri özet bilgiye herhangi bir hesaplama yapmadan hızlı bir biçimde erişilmesini sağlarlar.

Veri madenciliği yapılacak veri kümesi bazen gereksiz olarak yüzlerce değişken içerebilir. Örneğin bir ürünün satışına ilişkin olarak düzenlenen bir veri kümesinde, tüketicilerin telefon numaraları gereksiz bir değişken olarak yer alabilir. Bu tür gereksiz değişkenler elde edilecek örüntüleri kalitesizleştirebileceği gibi veri madenciliği sürecinin yavaşlamasına da yol açacaktır. Gereksiz değişkenlerin elenmesi amacıyla ileri veya geri yönlü olarak sezgisel seçimler yapılabilir. İleri yönlü sezgisel seçimde orijinal değişkenleri en iyi temsil edecek değişkenler belirlenir.

Ardından her bir değişken veya değişkenler grubunun, bu kümeye dâhil edilip edilmeyeceği sezgisel olarak belirlenir. Geri yönlü sezgisel seçimde ise öncelikle değişkenlerin tüm kümesi ele alınır. Daha sonra gereksiz bulunan değişkenler kümeden dışlanarak, en iyi değişken kümesi elde edilmeye çalışılır. Boyut indirgeme amacıyla kullanılacak bir diğer yöntem ise karar ağaçlarıdır. Karar ağaçları ele alınacak çıktı değişkenini en iyi temsil edecek değişken kümesini verecektir.

Veri sıkıştırma ise orijinal verileri temsil edebilecek indirgenmiş veya sıkıştırılmış veriler, veri şifreleme veya dönüşümü ile elde edilirler. Bu şekilde indirgenmiş veri kümesi, orijinal veri kümesini bir bilgi kaybı olacak biçimde temsil edebilecektir. Bununla beraber bilgi kaybı olmaksızın indirgenmiş veri kümesi elde edilmesine yarayacak bir takım algoritmalar da mevcuttur. Bu algoritmalar bir takım sınırlamalara sahip olduklarından sıkça kullanılamamaktadır. Bununla beraber temel bileşenler analizi gibi yöntemler, bir bilgi kaybına göz yumularak sıkıştırılmış veri kümesi elde edilmesinde kullanışlıdır.

Kesikleştirme, bazı veri madenciliği algoritmaları yalnızca kategorik değerleri ele aldığından, sürekli verilerin kesikli değerlere dönüştürülmesini içerir. Bu şekilde sürekli verilerin kesikli değer aralıklarına dönüştürülmesiyle elde edilen kategorik değerler, orijinal veri değerlerinin yerine kullanılırlar. Bir kavram hiyerarşisi (concept hierarchy), verilen sürekli değişken için, değişkenin ayrıştırılması olarak tanımlanabilir. Kavram hiyerarşileri, düşük düzeyli kavramların yüksek düzeyli kavramlarla değiştirilmesiyle verilerin indirgenmesinde kullanılır. Örneğin yaş değişkeni 1-15, 16-40, 40+ olacak biçimde daha yüksek kavram düzeyinde ifade edilebilir. Bu şekilde veri indirgemedede detay bilgiler kayboluyorsa da, genelleştirilmiş veriler daha anlamlı olacak, daha kolay yorumlanabilecek ve orijinal verilerden daha düşük hacim kaplayacaktır.

3.1.1.7. Birleştirme ve Temizleme (Consolidation and Cleaning)

Bu adımda farklı kaynaklardan toplanan verilerde bulunan ve bir önceki adımda belirlenen sorunlar mümkün olduğu ölçüde giderilerek veriler tek bir veri tabanında toplanır. Ancak basit yöntemlerle ve baştan savma olarak yapılacak sorun giderme işlemlerinin, ileriki aşamalarda daha büyük sorunların kaynağı olacağı unutulmamalıdır.

3.1.1.8. Seçim (Selection)

Bu adımda kurulacak modele bağlı olarak veri seçimi yapılır. Örneğin tahmin edici bir model için, bu adım bağımlı ve bağımsız değişkenlerin ve modelin eğitiminde kullanılacak veri kümesinin seçilmesi anlamını taşımaktadır.

Sıra numarası, kimlik numarası gibi anlamlı olmayan ve diğer değişkenlerin modeldeki ağırlığının azalmasına da neden olabilecek değişkenlerin modele girmemesi gerekmektedir. Bazı veri madenciliği algoritmaları konu ile ilgisi olmayan bu tip değişkenleri otomatik olarak elese de, pratikte bu işlemin kullanılan yazılıma bırakılmaması daha akılcı olacaktır.

Verilerin görselleştirilmesine olanak sağlayan grafik araçlar ve bunların sunduğu ilişkiler, bağımsız değişkenlerin seçilmesinde önemli yararlar sağlayabilir.

Genellikle yanlış veri girişinden veya bir kereye özgü bir olayın gerçekleşmesinden kaynaklanan verilerin (*Outlier*), önemli bir uyarıcı enformasyon içerip içermediği kontrol edildikten sonra veri kümesinden atılması tercih edilir.

Modelde kullanılan veri tabanının çok büyük olması durumunda tesadüflüğü bozmayacak şekilde örnekleme yapılması uygun olabilir. Günümüzde hesaplama olanakları ne kadar gelişmiş olursa olsun, çok büyük veri tabanları üzerinde çok sayıda modelin denenmesi zaman kısıtı nedeni ile mümkün olamamaktadır. Bu nedenle tüm veri tabanını kullanarak bir kaç model denemek yerine, tesadüfi olarak örneklenmiş bir veri tabanı parçası üzerinde birçok modelin denenmesi ve bunlar arasından en güvenilir ve güçlü modelin seçilmesi daha uygun olacaktır.

3.1.1.9. Dönüştürme (Transformation)

Kredi riskinin tahmini için geliştirilen bir modelde, borç/gelir gibi önceden hesaplanmış bir oran yerine, ayrı ayrı borç ve gelir verilerinin kullanılması tercih edilebilir. Ayrıca modelde kullanılan algoritma, verilerin gösteriminde önemli rol oynayacaktır. Örneğin bir uygulamada bir yapay sinir ağı algoritmasının kullanılması durumunda kategorik değişken değerlerinin evet/hayır olması; bir karar ağacı algoritmasının kullanılması durumunda ise örneğin gelir değişken değerlerinin yüksek/orta/düşük olarak gruplanmış olması modelin etkinliğini artıracaktır (Oğuzlar, 2003).

3.1.1.10. Modelin Kurulması ve Değerlendirilmesi

Modelin kuruluşu, amacımızın ne olduğuna, problemimizi ne şekilde çözmek istediğimize ve sonucun ne kadar işimize yarar olacağına göre değişebilir. Örneğin

görmek istediğimiz gelecek dönemdeki tahmini ciromuz ise, sürekli bir değişkeni tahmin edeceğimiz doğrusal regresyon modelini; müşterilerimizin pasifleşme eğiliminde olup olmadıkları ise kategorik bir değişkeni tahmin edeceğimiz sınıflandırma modelleri olan karar ağaçlarını, yapay sinir ağını veya kategorik değişkenin olasılığını tahmin edeceğimiz lojistik regresyon modelini, hangi ürünlerimizin diğerlerine oranla daha çok beraber alındığı ise birliktelik analizi, beraber alınan bu ürünlerin hangi sırayla alındığı, nedensellikleri ise sıralı örüntü algoritmaları kullanılabilir. Ayrıca müşterilerimizin sahip oldukları alışveriş özelliklerine göre (gelme sıklıkları, uğradıkları mağazalar, satın aldıkları ürünler vb.) belirli gruplara ayırmak için kümeleme algoritmaları kullanılabilir (Çelik, 2009).

Tanımlanan problem için en uygun modelin bulunabilmesi, olabildiğince çok sayıda modelin kurularak denenmesi ile mümkündür. Bu nedenle veri hazırlama ve model kurma aşamaları, en iyi olduğu düşünülen modele varılıncaya kadar yinelenen bir süreçtir.

Model kuruluş süreci denetimli (*Supervised*) ve denetimsiz (*Unsupervised*) öğrenimin kullanıldığı modellere göre farklılık göstermektedir.

Örnekten öğrenme olarak da isimlendirilen denetimli öğrenimde, bir denetçi tarafından ilgili sınıflar önceden belirlenen bir kritere göre ayrılarak, her sınıf için çeşitli örnekler verilir. Sistemin amacı verilen örneklerden hareket ederek her bir sınıfa ilişkin özelliklerin bulunması ve bu özelliklerin kural cümleleri ile ifade edilmesidir.

Öğrenme süreci tamamlandığında, tanımlanan kural cümleleri verilen yeni örneklerle uygulanır ve yeni örneklerin hangi sınıfa ait olduğu kurulan model tarafından belirlenir.

Denetimsiz öğrenimde, kümeleme analizinde olduğu gibi ilgili örneklerin gözlenmesi ve bu örneklerin özellikleri arasındaki benzerliklerden hareket ederek sınıfların tanımlanması amaçlanmaktadır.

Denetimli öğrenimde seçilen algoritmaya uygun olarak ilgili veriler hazırlandıktan sonra, ilk aşamada verinin bir kısmı modelin öğrenimi, diğer kısmı ise modelin geçerliliğinin test edilmesi için ayrılır. Modelin öğrenimi öğrenim kümesi kullanılarak gerçekleştirildikten sonra, test kümesi ile modelin doğruluk derecesi (*Accuracy*) belirlenir.

Bir modelin doğruluğunun test edilmesinde kullanılan en basit yöntem basit geçerlilik (simple validation) testidir (Akpınar, 2000).

Basit geçerlilik yönteminde verilerin bir kısmı test verisi olarak ayrılır, kalan kısım üzerinde modelin öğrenimi gerçekleştirildikten sonra ayrılan kısım üzerinde test işlemi yapılır. “Bir sınıflama modelinde yanlış olarak sınıflanan olay sayısının, tüm olay sayısına bölünmesi ile hata oranı, doğru olarak sınıflanan olay sayısının tüm olay sayısına bölünmesi ile doğruluk oranı hesaplanır (Yarahoğlu, 2004).

Sınırlı miktarda veriye sahip olunması durumunda, kullanılabilecek diğer bir yöntem çapraz geçerlilik (*Cross Validation*) testidir. Bu yöntemde veri kümesi tesadüfi olarak iki eşit parçaya ayrılır. İlk aşamada a parçası üzerinde model eğitimi ve b parçası üzerinde test işlemi; ikinci aşamada ise b parçası üzerinde model eğitimi ve a parçası üzerinde test işlemi yapılarak elde edilen hata oranlarının ortalaması kullanılır.

Bir kaç bin veya daha az satırdan meydana gelen küçük veri tabanlarında, verilerin n gruba ayrıldığı n katlı çapraz geçerlilik (*N-Fold Cross Validation*) testi tercih edilebilir. Verilerin örneğin 10 gruba ayrıldığı bu yöntemde, ilk aşamada birinci grup test, diğer gruplar öğrenim için kullanılır. Bu süreç her defasında bir grubun test, diğer grupların öğrenim amaçlı kullanılması ile sürdürülür. Sonuçta elde edilen on hata oranının ortalaması, kurulan modelin tahmini hata oranı olacaktır.

Bootstrapping küçük veri kümeleri için modelin hata düzeyinin tahmininde kullanılan bir başka tekniktir. Çapraz geçerlilikte olduğu gibi model bütün veri kümesi üzerine kurulur. Daha sonra en az 200, bazen binin üzerinde olmak üzere çok fazla sayıda öğrenim kümesi tekrarlı örneklemelerle veri kümesinden oluşturularak hata oranı hesaplanır.

Model kuruluşu çalışmalarının sonucuna bağlı olarak, aynı teknikle farklı parametrelerin kullanıldığı veya başka algoritma ve araçların denendiği değişik modeller kurulabilir. Model kuruluş çalışmalarına başlamadan önce, imkânsız olmasa da hangi tekniğin en uygun olduğuna karar verebilmek güçtür. Bu nedenle farklı modeller kurarak, doğruluk derecelerine göre en uygun modeli bulmak üzere sayısız deneme yapılmasında yarar bulunmaktadır.

Özellikle sınıflama problemleri için kurulan modellerin doğruluk derecelerinin değerlendirilmesinde basit ancak faydalı bir araç olan risk matrisi kullanılmaktadır. Aşağıda bir örneği görülen bu matriste sütunlarda fiili, satırlarda ise tahmini sınıflama değerleri yer almaktadır. Örneğin fiilen B sınıfına ait olması gereken 46 elemanın, kurulan model tarafından 2'sinin A, 38'inin B, 6'sının ise C olarak sınıflandırıldığı matriste kolayca görülebilmektedir.

	Gerçek sınıf		
Tahmin edilmiş sınıf	A Sınıfı	B Sınıfı	C Sınıfı
A Sınıfı	45	2	3
B Sınıfı	10	38	2
C Sınıfı	4	6	40

Çizelge 3.1. Risk Matrisi

Önemli diğer bir değerlendirme kriteri modelin anlaşılabilirliğidir. Bazı uygulamalarda doğruluk oranlarındaki küçük artışlar çok önemli olsa da, birçok işletme uygulamasında ilgili kararın niçin verildiğinin yorumlanabilmesi çok daha büyük önem taşıyabilir. Çok ender olarak yorumlanamayacak kadar karmaşıklaşsalar da, genel olarak karar ağacı ve kural temelli sistemler model tahmininin altında yatan nedenleri çok iyi ortaya koyabilmektedir.

Kaldıraç (*Lift*) oranı ve grafiği, bir modelin sağladığı faydanın değerlendirilmesinde kullanılan önemli bir yardımcıdır. Örneğin kredi kartını muhtemelen iade edecek müşterilerin belirlenmesi amacını taşıyan bir uygulamada, kullanılan modelin belirlediği 100 kişinin 35'i gerçekten bir süre sonra kredi kartını iade ediyorsa ve

tesadüfî olarak seçilen 100 müşterinin aynı zaman diliminde sadece 5'i kredi kartını iade ediyorsa kaldıraç oranı 7 olarak bulunacaktır.

Kurulan modelin değerinin belirlenmesinde kullanılan diğer bir ölçü, model tarafından önerilen uygulamadan elde edilecek kazancın bu uygulamanın gerçekleştirilmesi için katlanılacak maliyete bölünmesi ile edilecek olan yatırımın geri dönüş (*Return On Investment*) oranıdır.

Kurulan modelin doğruluk derecesi ne denli yüksek olursa olsun, gerçek dünyayı tam anlamı ile modellediğini garanti edebilmek mümkün değildir. Yapılan testler sonucunda geçerli bir modelin doğru olmamasındaki başlıca nedenler, model kuruluşunda kabul edilen varsayımlar ve modelde kullanılan verilerin doğru olmamasıdır. Örneğin modelin kurulması sırasında varsayılan enflasyon oranının zaman içerisinde değişmesi, bireyin satın alma davranışını belirgin olarak etkileyecektir.

3.1.1.11. Modelin Kullanılması

Kurulan ve geçerliliği kabul edilen model doğrudan bir uygulama olabileceği gibi, bir başka uygulamanın alt parçası olarak kullanılabilir. Kurulan modeller risk analizi, kredi değerlendirme, dolandırıcılık tespiti gibi işletme uygulamalarında doğrudan kullanılabilmesi gibi, promosyon planlaması simülasyonuna entegre edilebilir veya tahmin edilen envanter düzeyleri yeniden sipariş noktasının altına düştüğünde, otomatik olarak sipariş verilmesini sağlayacak bir uygulamanın içine gömülebilir.

3.1.1.12. Modelin İzlenmesi

Zaman içerisinde bütün sistemlerin özelliklerinde ve dolayısıyla ürettikleri verilerde ortaya çıkan değişiklikler, kurulan modellerin sürekli olarak izlenmesini ve gerekiyorsa yeniden düzenlenmesini gerektirecektir. Tahmin edilen ve gözlenen değişkenler arasındaki farklılığı gösteren grafikler model sonuçlarının izlenmesinde kullanılan yararlı bir yöntemdir.

3.2.BÜYÜK VERİ KÜMELERİNDE EDA

“Veri analistinin kişisel yargıları ve gösterge açıklamaları, İstatistiğin formal yapısından çok, formal olmayan çıkarımlarına yöneliktir.” Günümüz bilgisayar teknolojisi ve internet ortamı nedeniyle toplanan veriler çok büyük sayı yığınlarına dönüşür. Bunlar üzerinde çözümleme yapmak için yeni kavramlara gereksinim duyulmuştur; veri madenciliği gibi.

Temel EDA kavramları veri genişlediğinden etkilenmez. Dönüşümler, düzleştirme vb. kavramlara büyük veriler için de çok önemli çözümleme araçlarıdır. Ancak istatistiğin alışlagelen kavramları ve teknikleri, zaman zaman, uygulanamaz ya da yanlış yorumlamalara götürür. Büyük veri kümelerine, yeni kavramlar ve açıklamalar ortaya çıkar (Erar ve Kiroğlu, 2008).

- Küçük örneklem genişliği (n) çoğu kez onlarla, yüzlerle belirtilir.
- Küçük veriler genelde “temiz”dir; beklenmeyen değerler olsa bile kolaylıkla ortaya çıkartılabilir. Kaynağa dönüp düzeltme yapılabilir.
- Küçük verilerde el yordamı ile incelemek, elde bazı grafikler üzerinde çalışmak vb. olanaklıdır.
- Veriler birkaç sayfaya sığabilir; gerekirse verilerde sıralama yaparak rank’ lar üzerinde çalışılabilir.
- Küçük genişlikteki veriler, genelde, bilinen bir kitleden seçilmiş bir rastgele örnektir.
- Veriler özel bir problemi çözmek için toplanmıştır.

Büyük verilerde, gözlem (tutanak, kayıt) sayıları 5000-10000’den başlar, milyonlarca satıra ulaşabilir. Bu veriler genelde çok fazla değişken (alan_field,...) içerirler. Zaman zaman değişken sayısı 50’den, 1002den fazla olabilir. Böylece bilgisayar teknolojisi ile ilişkili olarak veri tabanı, veri ambarı, öğrenen makine (machine learning_bağımlı değişkenli modeller), iş zekâsı, daha önce belirtilen örüntü tanımlama (pattern recognition) veri temizleme gibi kavramlar sıkça kullanılır.

EDA’ da uygulanan dal yaprak çizimi gibi bazı grafikler büyük veri kümeleri için uygulanmaz. Bunlar kolay açıklanabilen, kolayca işlenebilen veriler değildir; sağlıklı olmayan rastgele görünümlü bir izlenim verirler. Büyük veriler “ veri çizelgeleri

kümesi” biçiminde veri tabanlarında yer alabilirler. Bu çizelgeler arasındaki bağıntı hiyerarşik ya da ardışık olabilir. Hiyerarşikten kasıt, rank ya da düzey bağımlılığıdır. Ardışıklık ise zaman ya da olay bağımlılığını belirtir.

Büyük veri kümelerinde, veriler, İstatistik derslerinde görüldüğü gibi İstatistikçinin ya da araştırmacının doğrudan önüne gelmez. Oldukça karmaşık biçimlerde ve değişik kaynaklardan geldiklerinden “kirli” olarak ortaya çıkar (farklı kaynaklardan veriler birleştirilirler-merge). Veri kümeleri büyük olduğunda verilerin bazıları geçersizdir. Büyük veri kümelerinde kirlilik (ya da bulaşmışlık) önemli bir sorundur. Küçük bir veri kümesinde %1 lik kirlilik kolayca görülür ve önlem alınabilir. Bir milyon veride %1 lik kirlilik on binlerce kirli veri demektir ki önemli bir oran olabilir. Birleştirme süreci de zaman yoğunluktadır. Özellikle tutanakların birleştirilmesi sırasındaki belli bir anahtar özelliğe göre eşleştirme (matching) işlemi oldukça karmaşıktır, zaman zaman bulanık temizlenmeleri, hazırlanmaları ve belli koşullara göre düzenlenmeleri gerekecektir. Bu da bilgisayar teknolojisini iyi bilmeyi ya da o alandaki uzmanlar ile birlikte çalışmayı gerektirecektir.

“Büyük genişlikteki verilere, bilinmeyen kitlelerin çoklu ve rastgele olmayan örneklemeleri gibi bakılabilir.” Çok kısa bir dönem içinde çok çeşitli kaynaklardan gelen verilerin birleştirilmesi sonucu büyük veriler ortaya çıkar. Ayrıca küçük örneklemdeki gibi belli bir problem için toplanmazlar. Örneğin birçok pazarlama bilgisini, sorgulama sonuçlarını, ürün özellikleri vb. bilgileri kapsayabilir. Farklı zamanlarda farklı amaçlar için bunlar içinden gerekli veriler ayıklanarak amaca uygun veri kümeleri ortaya çıkar.

Gereksiz bilgilerin, kullanılması beklenmeyen verilerin, gereksiz sorgulama sorularının baştan belirlenmesi, ayıklanması vb. hem zaman olarak hem hacim olarak hem de uygulanacak veri analizleri ve istatistik teknikler açısından önem taşır. Bununla birlikte bu belirlemenin de dikkatli yapılması gerekecektir. Burada iki tutum vardır: Her türlü bilgiyi toplamak ya da yalnızca o an için kullanılması gereken bilgileri toplamak. En iyisi, ileriye de düşünerek optimal biçimde bilgi toplamaya karar vermektir.

Çok büyük veri kümelerini işlemek oldukça zaman alıcı ve bilgisayar yoğunluktur. Genellikle büyük veriler için önerilen şey, onların temsil edici nitelikte örneklem seçip veri analizlerini bu örneklem üzerinden uygulamaktır. Önemli bir soru, örneklem genişliği ne olmalıdır? Büyük genişlikte veriler için bu 1000-1000000 arası olabilir. Bu örnekleme veri temizleme ve hazırlama için başlangıçta (Örneğin Clementine akışlarını oluşturmada) seçilerek örneklem ile karıştırılmamalıdır. Bu örneklem için izlenen yol, genelde ilk bir-iki bin veriyi alıp yapılacak işlerin tanımlarını oluşturmaktadır. Bu konuya veri madenciliği açısından da bakılabilir.

Alışlagelen İstatistiksel uygulamalarda serbestlik derecesi ve /ya da birinci tür hata dikkate alınarak çözümler yapılır. Büyük veri kümelerinde bunlar anlamsızlaşabilir. Ancak bazı düzeltimler kullanılarak bunlar uygulanabilir. O nedenle EDA' nın özellikle grafik incelemeleri daha bilgi verici olarak sıklıkla kullanılır. Örneğin ilişki katsayısı $r = 0.05$ büyük veriler için anlamlı çıkar. Test yapılacak ise “ düzeltilmiş ilişki katsayıları” kullanılmalıdır. Quenouille, parametrik olmayan bir grafiğe dayalı birliktelik testi önermiştir. Buna göre düzleştirilmiş saçılım grafiğinde bir ortanca çizgi çizilir ve bu çizginin kaç kez (m) kesildiği sayılır.

Düzleştirilmiş nokta sayısı N iken, $TS = N-1-m$ kritik değeri %95 ya da %99 eşik değeri ile karşılaştırılır. Eğer büyükse birlikteliğin anlamlı olduğu belirtilebilir. Ratner' e göre büyük verilerde istatistiksel anlamlılık terimi yerine “dikkate değer önemde”, ”ön kestirim hatasında dikkate değer küçülme” gibi kavramlar daha sık kullanılır.

Büyük veri kümelerinde, standart istatistik yöntemlerde olan “bağımsız ve aynı dağılımlılık” varsayımı genelde yoktur. Değişken uzayının bazı bölgeleri farklı zamanlarda ötekilerden farklı biçimde örneklenmiş olabilir. Böylece durağan olmama, verilerin dinamikliği gibi durumlar da kestirimlerin doğruluğuna kuşku getirir.

Büyük verilerde, gerçek model bilinmediğinden, örneklem üzerinden model kurmak oldukça akılcıdır. Geçerlilik testlerinde ön kestirimler iyi sonuç veriyorsa gerçek modele iyi bir yaklaşım sağlanmış demektir. Eğer test geçersiz ise örneklem genişliği

arttırılarak çalışma yinelenir. Dikkat edilmelidir ki büyük verilerle çalışıldığında ön kestirim varyansları da artacaktır.

Zaman zaman aykırı değer gibi gözüken ayırık veri noktaları ortaya çıkabilir. Eğer bunların kaynağı belirtilmemiş ve doğru oldukları düşünülüyor ise bu gözlem grupları için 0-1 göstermelik değişkenleri tanımlanarak modele eklenebilir. Bu göstermelik değişken yaratıldığı değişken ile birlikte alınmalı ve incelenmelidir.

Kategorik değişkenlerin modellemeye katılabilmesi için bilindiği gibi göstermelik değişkenlere çevrilmesi gerekir. Ancak büyük veri kümelerinde nitel değişken sayısı ve bunların düzey sayıları genelde fazladır. Örneğin 18 düzeye sahip bir değişken için 17 göstermelik değişken eklenmesi gerekecektir. Bunun için göstermeliğe çevirmeden önce hangi düzeylerin birleştirileceğine, hangilerinin dışarıda bırakılabileceğine yani düzey sayısının nasıl azaltılabileceğine karar verilmelidir. Bunun için alışlagelen istatistik yöntemler (varyans analizi, düzey grafikleri, etkileşim grafikleri, vb.) kullanılabilir (Erar ve Kiroğlu, 2008).

3.3. VERİ ÇÖZÜMLEMEDE EKSİK DEĞERLER

Veri kümesinde “eksik değer” lerin varlığı ön inceleme sırasında giderilmesi gereken bir sorundur. Bazı değerlerin yanlış kaydedilmesi ya da hiç kaydedilmemesi eksik değerlere yol açar. Deney sırasında bir kobayın ölmesi, bir kan örnekleminin bozulması; anket uygulamasında bireylerin bazı sorunlara yanlış yanıt vermesi ya da soruları doldurmaması vb. nedenlerle ortaya çıkabilir.

Eksik değerler, bilgisayar paketlerinin veri girişlerinde “MISSING”, “.”, vb. gösterimlerle tanımlanırlar. Burada önemli bir nokta, “sıfır”ların denetlenmesidir. Eksik değer olarak sıfır girilmemiş olmalıdır. Ayrıca her paket program eksik değerleri farklı algılayabilir. Örneğin SPSS’ de nokta olarak girilen değerler Statgraph’ da sıfır olarak alınırlar; bunlar Statgraph’ da boş göze olurlarsa eksik değer olarak kabul edilirler.

Bir gözlemin deęişken deęerlerinin tümü eksik ise gözlem veri kümesinden çıkartılır. Eđer çok sayıda gözlem içinde birkaç eksik deęer varsa bu deęerlerin bulunduęu gözlemlerin çıkartılması da bir sorun yaratmaz.

Eksik verileri doldurmadaki amaç eksik deęerlerden ötürü ortaya çıkan bilgi kaybını en aza indirmektir. Bu durumda çözümlemeler iki durumda yapılabilir.

3.3.1.Mevcut Durum Çözümlemesi (Pairwise)

İlgilenilen deęişkenin yalnızca mevcut deęerleri kullanılır. İki deęişken birlikte dikkate alındığında, her iki deęişkende de eksik olmayan deęerlerin var olduęu gözlemler kullanılır.

3.3.2.Tam Durum Çözümlemesi (Listwise)

Eksik deęerlere sahip tüm gözlemlerin çıkartıldığı durumdaki analizlerdir. Genelde popüler bir yoldur. Bu durumun üstünlüğü basitliğindendir. Tüm yöntemler uygulanabilir. Ancak bilgi kaybı oldukça yüksektir.

Eksik deęerlerin oranı büyükse, özellikle deęişken sayısına göre küçük örneklemelerde gözlemlerin çıkarılması ciddi sorunlar getirebilir:

1. Ayrı ayrı gözlemlerde bir-iki deęerin eksikliğinden ötürü tüm gözlem deęerlerinin çıkartılmasıyla veri kümesinde az sayıda gözlem kalacak; böylece gözlem deęişken sayılarının oranı büyük ölçüde düşecektir
2. Bu deęerlerden ötürü çok sayıda gözlemin çıkartılması, araştırmacılarca da, veri toplamak için harcanan zaman açısı ve önemli bilgi kaybı olması nedeniyle istenmez.
3. Eksik deęerler veri kümesi içinde rastgele dağılmamış iseler gözlemlerin çıkartılması, örneklerin bozulmasına, deęişimin yeterince açıklanamamasına neden olur; sistematik bir yanılğı getirir.
4. Bağlanım çözümlemesi ve bazı çok deęişkenli istatistik teknikler varyans-kovaryans ya da ilişki matrisleri üzerinden uygulanır. Bu matrisler ve özellikle tersleri, verilerdeki küçük deęişimlere karşı çok duyarlıdır. Eksik deęerli gözlemlerin çıkartılması, bu matrisleri ve böylece çözümlemeyi büyük oranda etkileyebilir.

Eksik deęerlerin varlıęında çözüme gitmeden önce ařaęıdaki deęerlendirmelerin yapılmasında yarar vardır:

1. Eksik deęerlerin veri kümesindeki oranı nasıldır?
2. Bu deęerler nereden kaynaklanmıřtır?
3. Eksik deęerlerin küme içindeki daęılıřı nasıldır?

Bu deęerler belli konumlarda toplanmıř olabilirler; belli gruplarda ya da deęiřkenlerde yer alabilirler; rastgele daęılmıř olabilirler (Erar ve Kiroęlu, 2008).

3.4. VERİ ÇÖZÜMLEMEDE DÖNÜřÜMLER

Ham veriler üzerinden daha bilgi verici ve etkin özetleme yapmak için, daha doęru bir modellemeye gidebilmek için zaman zaman bu verilerin yeniden tanımlanması gerekebilir. Bu tanımlama, veriler üzerinde dönüřüm yapılması, yani veri noktalarının bir fonksiyon yardımıyla kaydedildięi ölçekten farklı bir ölçeęe taşınmasıdır. Veri analistleri için özellikle simetriklik ve doęrusallık, istenilen tekniklerin uygulanabilmeleri ve basitlik aęısından, önemlidir. Dönüřümler, verilerin yalnızca ölçeęi deęiřtirdięi gibi ve/ya da merkezsel eęilimini, biçimini de deęiřtirebilir.

Dönüřümlerin gerekçeleri:

- Verileri basitleřtirmek, yorum kolaylıęı saęlamak,
- Simetriklik ve/ya da normallilik saęlamak,
- Varyansları homojen yapmak, yayılımı kararlı duruma getirmek,
- Doęrusallıęı saęlamak ya da doęru fonksiyonel biçimi bulmak; iki yönlü ya da daha yüksek boyutlu çizelgelerde yapıyı kolaylařtırmak,
- Gözlemler arasındaki baęımlılıęı kaldırmak,

Veri kümesinin yaklaşık normallilięi istatistik aęısından istenilen bir özelliktir. Daęılımlar simetrik olmadıklarında, konum ve yayılım dirençli ölçüleri bile yanlış yola götürebilir. Ayrıca konum ve yayılım kestirimleri büyük oranda daęılımın biçimine baęlıdır. Asimetrik daęılımlarda bu kestirimler zayıftır.

Hartwing'e göre, modellemede sistematik bir yaklaşım: Değişkenler önce simetrik duruma getirilir; ileri çözümlemelerde bu yeni değişkenler kullanılır. Çoğu verileri bu şekilde modellemek olanaklıdır. Sola çarpık dağılımlı iki değişken arasındaki bağıntı doğrusal olabilir. Her iki değişkende de simetriklik dönüşümünün yapılması doğrusal ilişkiyi değiştirmez.

Zaman zaman veri yeni bir ölçekle tanımlandığında, orijinal verilere göre yeni ölçek daha anlamlı gelebilir. Örneğin bir şehrin nüfus büyüme oranları incelendiğinde logaritmik dönüşüm, yorumu kolaylaştırır. Geometrik bir dizinin ortalamasının geometrik ortalama yani logaritmalar üzerinden alınan bir ortalama olduğu unutulmamalıdır. Bazen $1/\sqrt{x}$ gibi bir dönüşümün anlamlı bir yorumu yapılamayabilir.

Çok sayıda değişkenin varlığında, EDA 'da izlenen yol şöyledir: Öncelikle her bir değişken tek tek incelenir; gerekirse simetrikleştirilir. Daha sonra değişkenler ikiye ikiye ve / ya da değişken grupları olarak incelenir ve gerekirse doğrusallık sağlanır. Bu işlemler çok değişkenli model oluşturmada bir zemin oluşturur (Erar ve Kiroğlu, 2008).

3.4.1.Dönüşümler İçin Uyarılar

Değişkenleri dönüştürme ya da yeniden ifade etme veri analizinin önemli bir parçası olmakla birlikte, bu işlemler sırasında bilgi kaybı da ortaya çıkacaktır. Bir dönüşüm ile verilerdeki yayılım azalır; daha dar bir alana sıkışır ve bilgi kaybolur; ya da yayılım genişleyebilir ve anlamsız bilgi kazancı ortaya çıkabilir. Böylece bir dönüşüm yapılacağı zaman,

- Düzgünlük ile doğruluk arasında iyi bir denge kurulmalı,
- Aşırı derecede düzgünleştirmekten kaçınmalı,
- Toplam bilgi kaybına dikkat edilmeli,
- Tam düzgünleştirme yerine yaklaşık düzgünleştirme yeterli sayılmalıdır.

3.5.DÜZLEŞTİRME / DÜZLEME (SMOOTHING)

Düzleme, bir veri kümesinin basitleştirilmesi, verilerdeki varsa düzgünlüğün ya da örüntünün anlaşılması için veri çözümlemede gerekli bir tekniktir. Hartung'a göre bu

düzgünlük, iki değişken arasındaki bağıntıyı anlatan bir doğru ile ya da tek bir değişkenin dağılımını anlatan bir eğriye görülebilir. Bir dağılımın genel biçimi ya da bir bağıntının genel biçimi verilerdeki düzgünlüğü ya da örüntüyü anlatır.

Veriler elde edildiğinde bunların grafiklerinden çoğu kez örüntüyü görmek, çok karmaşık bir şekil ortaya çıktığından, olanaklı değildir. Genel olarak veriler, “Veri = Düzgünlük (Örüntü) değeri + Artık” biçiminde belirtilir. Artık, fark ya da hata anlamındadır.

Unutulmamalıdır ki veri analizinde basitliği yansıtan simetrik ve doğrusallık çok önemli iki kavramdır. İstatistikçiler, özellikle, simetrik ve doğrusal olan nicelikler üzerinde dururlar.

3.5.1.Hangi Amaçlarla Uygulanır

Düzleme teknikleri birkaç amaçla uygulanır. Zamana bağlı verilerde serideki, küçük dalgalanmaları ortadan kaldırarak, eğilimi ya da örüntüyü görmek en sık kullanılan düzleme yöntemidir. Düzleme, tek değişkenli dağılım biçimini ya da iki değişkenli bağıntı biçimini görmede nonparametrik teknikler olarak uygulanabilir. Ayrıca, çok boyutlu grafik tekniklerde, örneğin iki değişkenli sıklık verilerinin üç boyutlu histogram gösterimlerinde, bu üç boyutluyu düzleştirilmiş yüzeye çevirmede, yararlanılır. Düzleme EDA ’nın vazgeçilmez bir parçasıdır (Erar ve Kiroğlu, 2008).

4. VERİ MADENCİLİĞİNDE MODELLEME

Veri madenciliğinde kullanılan modeller, tahmin edici (Predictive) ve tanımlayıcı (Descriptive) olmak üzere iki ana başlık altında incelenmektedir.

Tahmin edici modellerde, sonuçları bilinen verilerden hareket edilerek bir model geliştirilmesi ve kurulan bu modelden yararlanılarak sonuçları bilinmeyen veri kümeleri için sonuç değerlerin tahmin edilmesi amaçlanmaktadır. Örneğin bir banka önceki dönemlerde vermiş olduğu kredilere ilişkin gerekli tüm verilere sahip olabilir. Bu verilerde bağımsız değişkenler kredi alan müşterinin özellikleri, bağımlı değişken değeri ise kredinin geri ödenip ödenmediğidir. Bu verilere uygun olarak kurulan model, daha sonraki kredi taleplerinde müşteri özelliklerine göre verilecek olan kredinin geri ödenip ödenmeyeceğinin tahmininde kullanılmaktadır.

Tanımlayıcı modellerde ise karar vermeye rehberlik etmede kullanılabilecek mevcut verilerdeki örüntülerin tanımlanması sağlanmaktadır. X/Y aralığında geliri ve iki veya daha fazla arabası olan çocuklu aileler ile çocuğu olmayan ve geliri X/Y aralığından düşük olan ailelerin satın alma örüntülerinin birbirlerine benzerlik gösterdiğinin belirlenmesi tanımlayıcı modellere bir örnektir.

Veri madenciliği modellerini gördükleri işlevlere göre,

- Sınıflama (*Classification*) ve Regresyon (*Regression*),
- Kümeleme (*Clustering*),
- Birlikte Kuralları (*Association Rules*) ve Ardışık Zamanlı Örüntüler (*Sequential Patterns*),

olmak üzere üç ana başlık altında incelemek mümkündür. Sınıflama ve regresyon modelleri tahmin edici, kümeleme, birlikte kuralları ve ardışık zamanlı örüntü modelleri tanımlayıcı modellerdir (Akpınar, 2000).

4.2.SINIFLAMA VE REGRESYON MODELLERİ

Sınıflama ve regresyon, önemli veri sınıflarını ortaya koyan veya gelecek veri eğilimlerini tahmin eden modelleri kurabilen iki veri analiz yöntemidir. Sınıflama kategorik değerleri tahmin ederken, regresyon süreklilik gösteren değerlerin tahmin edilmesinde kullanılır (Han and Kamber, 2006).

Sınıflama, verinin önceden belirlenen çıktılara uygun olarak ayrıştırılmasını sağlayan bir tekniktir. Çıktılar, önceden bilindiği için sınıflama, veri kümesini denetimli (supervised) olarak öğrenir (Giudici, 2003).

Sınıflandırmada amaç yeni bir nesnenin belirli sınıflar içinde hangi sınıfa ait olup olmadığını belirleyecek bir sınıflayıcı (classifier) oluşturmaktır. Önceden belirlenmiş sınıflar, veri ambarından veya veri tabanından alınan verinin sınıflandırılması için model geliştirmede kullanılır.

Sınıflandırma bir ürünün özellikleri ile müşterinin eşleşmesi için kullanılır. Böylece bir müşteri için ideal ürün veya bir ürün için ideal müşteri profili çıkartılabilir. Örneğin; genç bayanlar küçük araba satın alır, yaşlı ve zengin erkekler, büyük ve lüks araba satın alır.

Araba satan bir şirket geçmiş müşteri hareketlerinin analizi ile yukarıdaki gibi bir kural bulursa, genç bayanların okuduğu bir dergiye reklam verirken küçük model araba reklamı verecektir (Tezcanlar ve Timor, 2007).

Mevcut verilerden hareket ederek geleceğin tahmin edilmesinde faydalanan ve veri madenciliği teknikleri içerisinde en yaygın kullanıma sahip olan sınıflama ve regresyon modelleri arasındaki temel fark, tahmin edilen bağımlı değişkenin kategorik veya süreklilik gösteren bir değere sahip olmasıdır. Sınıflama ve regresyon modellerinde kullanılan başlıca teknikler,

- Karar Ağaçları (*Decision Trees*),
- Yapay Sinir Ağları (*Artificial Neural Networks*),
- Genetik Algoritmalar (*Genetic Algorithms*),
- K-En Yakın Komşu (*K-Nearest Neighbor*),

- Bellek Temelli Nedenleme (*Memory Based Reasoning*),
- Lojistik Regresyondur (*Logistic Regression*) (Akpınar, 2000).

4.2.1.Karar Ağaçları

Karar ağaçları, sınıfları bilinen örnek veriden tümevarım yöntemiyle öğrenilen ağaç şekilli bir karar yapısı çeşididir. Bir karar ağacı, basit karar verme adımları uygulanarak, büyük miktarlardaki kayıtları, çok küçük kayıt gruplarına bölerek kullanılan bir yapıdır. Her başarılı bölme işlemiyle, sonuç gruplarının üyeleri bir diğeriyle çok daha benzer hale gelmektedir. Büyük veri tabanlarının kullanıldığı pek çok sınıflama probleminde ve karmaşık ya da hata içeren bilgilerde karar ağaçları yararlı bir çözüm olmaktadır. Tahmin edici ve tanımlayıcı özelliklere sahip olan karar ağaçları, veri madenciliğin de kuruluşlarının kolay olması, yorumlanmalarının kolay olması, veri tabanı sistemlerine kolayca entegre edilebilmeleri, güvenilirliklerinin daha iyi olması nedenleri ile sınıflama modelleri içerisinde en yaygın kullanıma sahip olan bir tekniktir (Albayrak ve Yılmaz, 2009).

Karar ağaçları yukarıdan aşağıya, genelden özele doğrultuda eğitilmiş verilerden (training data) türetilmektedir. Bir karar ağacının başlangıç aşaması, kök düğümüdür. Tüm örneklerin aynı sınıfa ait olması durumunda herhangi bir karar alınmasına gerek olmadan çözüm bitirilir. Eğer düğümdeki örnekler bir veya birkaç sınıfa ait ise, düğümde bir test yapılarak bir bölünme oluşur (Tüzüntürk ve Oğuzlar, 2009).

Karar ağaçları, tek bağımlı değişken ve çok sayıda bağımsız değişkene sahip olmaları açısından regresyon modellerine benzerler. Bununla birlikte, ek olarak, veriden regresyon modellerine alternatif olabilecek farklı ve kullanışlı örüntüler keşfederler (Çelik, 2009).

Karar ağacının her düğümünde değişkenler test edilir. “Karar ağacı algoritması, ağacın kökünde hangi değişken ile test edilmesi gerektiği sorusu ile başlayarak yukarıdan aşağıya doğru ağacı oluşturur. Bu işlemi her örnek değişken, eğitim örneklerinin sınıflandırmasına karar vermek için istatistiksel test kullanılarak değerlendirilir. En iyi değişken seçilir ve ağacın kök düğümünde test için kullanılır.” Her bir düğüm için oluşturulacak dalların sayısı, test sonucunda kabul

edilmiş olan değişkenin alabileceği değer sayısına göre farklılaşmaktadır (Kök ve Kuloğlu, 2005).

Karar ağacı temelli analizlerin yaygın olarak kullanıldığı sahalar arasında; belirli bir sınıfın üyesi olacak elemanların belirlenmesi, çeşitli vakaların yüksek, orta, düşük risk grupları biçiminde kategorilere ayrılması, gelecekte gerçekleşebilecek olayların tahmin edilebilmesi için kurallar oluşturulması, parametrik modellerin kurulmasında kullanılacak çok sayıdaki değişken ve veri kümesinden önemli olanlarının seçilmesi, yalnızca belirli alt gruplara özgü ilişkilerin tanımlanması ve sürekli değişkenlerin kesikli değişkenlere dönüştürülmesi ve kategorilerin birleştirilmesi sayılabilir.

Karar ağaçları sınıflandırma ve kestirim için güçlü ve popüler araçlardır. Karar ağaçlarının çekici olan yönü, bir takım kuralları temsil etmesidir. Kurallar, karar ağacından kolaylıkla okunabilir.

Karar ağaçlarının güçlü yönleri aşağıdaki gibi özetlenebilir:

- Karar ağaçları anlaşılabilir kurallar üretirler.
- Karar ağaçları aşırı hesaplamaya gerek kalmadan sınıflandırma yaparlar.
- Karar ağaçları hem sürekli ve hem de kesikli değişkenler için uygundur.
- Karar ağaçları sınıflandırma ve kestirim için hangi alanların en önemli olduğunu açık biçimde gösterir (Tüzüntürk ve Oğuzlar, 2009).

4.2.1.1 Karar Ağaçları'nda Ayırma Kriterleri

Ağacı devam ettirecek olan değişken seçilirken belirli kriterler gözetilmektedir. Bu kriterler sonucu her bir değişkenin aldığı değerlere göre seçim yapılmakta, ağaç dallandırılmaktadır (Çelik, 2009).

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

bakış açısını yakalamaya çalıştıklarından, sonuçta olacak modeller de benzer davranmaya eğilimli olabilirler (Michael and Berry, 2004).

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

Tek Değişkenli Ayırma Kriterleri

Karar ağaçlarında bağımlı değişkenin kategorik olması durumunda karar ağaçlarında dallanma için bilgi kazanımı, gini, X^2 yöntemleri kullanılırken, bağımlı değişken sürekli olduğu durumlarda ise F testi kullanılır.

Bilgi Kazanımı

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

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

X veri kümesindeki herhangi bir kaydı sınıflandırmak için beklenen bilgi 4.1 de gösterilmiştir.

$$\text{Bilgi (X)} = -\sum_{i=1}^m p_i \log_2(p_i) \quad (4.1)$$

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

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

$$\text{Bilgi}_k(x) = - \sum_{j=1}^V \frac{|X_j|}{|X|} * \text{Bilgi}(x) \quad (4.2)$$

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

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

$$\text{Kazanım}(x) = \text{Bilgi}_k(x) - \text{Bilgi}(x) \quad (4.3)$$

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

Gini İndeksi

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

$$\text{Gini}(X) = 1 - \sum_{i=1}^m p_i^2 \quad (4.4)$$

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

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

$$\text{Gini}_A(X) = \frac{|X_1|}{|X|} \text{Gini}(X_1) + \frac{|X_2|}{|X|} \text{Gini}(X_2) \quad (4.5)$$

Burada $\frac{|X_1|}{|X|}$ değeri, o kategori için o sonucun gerçekleşme oranıdır. Değişken seçimi kararı için kullanılan gini indeksi 4.4 ve 4.5 deki iki ifadenin farkını içerir.

$$\Delta \text{Gini}(A) = \text{Gini}(X) - \text{Gini}_A(X) \quad (4.6)$$

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

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

Cok Değişkenli Ayırma Kriterleri

Çok değişkenli ayırma kriterlerinin birçoğu, bağımsız değişkenlerin kombinasyonlarına dayalı olarak oluşturulmaktadır. Burada optimal değişkeni bulma problemi, tek değişkenli parçalama kriterlerine göre daha zordur. En uygun parçalayan değişkeni bulmak için, sonuca deneme yanılma yolu ile giden, doğrusal programlama, doğrusal diskriminant analizi gibi yöntemler kullanılmaktadır.

4.2.1.2. Karar Ağaçları'nda Durma Kriterleri

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

Karar ağacı dallanmayı, belirtilen durma kriterlerinden biri ile karşılaşılan kadar sürdürür. Ortak olarak kullanılan durma kriterlerinden bazıları şunlardır:

- Veri kümesindeki tüm örnekler tek bir sınıfa ait olduğunda veya belirli bir

sayının altına indiği zaman.

- Ağacı oluşturan kişi tarafından belirlenen maksimum ağaç derinliğine ulaşıldığı zaman.
- Son düğümdeki örneklerin sayısı bir önceki düğümdeki minimum örnek sayısından küçük olduğu zaman.
- Kayıtların ayırma için sorgulanabilecek herhangi bir özelliği kalmadığı zaman.
- Ayırma kriteri, belirlenen eşik değerden daha büyük olduğu zaman (Çelik, 2009).

4.2.1.3. Karar Ağaçları'nda Budama

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

Budama ile bazı istatistiksel ölçüler kullanarak fazla güvenilir olmayan dalları ayıklanır. Budanmış ağaçlar daha küçük ve daha az karmaşık olmaya eğilimlidirler ve böylelikle daha kolay yorumlanabilirler. Ayrıca eğitim kümesinden bağımsız olan test kümesini sınıflandırmada budanmamış ağaçlara göre daha iyi ve hızlıdır (Akpınar, 2000).

CHAID

CHAID metodu 1980'de Kaas tarafından en iyi bölmeyi hesaplamak için istatistik olarak anlamlı bir farklılığın olmadığı, hedef değişkene uyan çiftlerde tahmin değişkeninin olası kategori çiftini birleştirmesiyle oluşturulmuştur.

En uygun bölümleri seçmek için kullanılan entropy veya gini metrikleri yerine ki-kare testi kullanılmaktadır.

En iyi bölmeyi hesaplamak için tahmin değişkenleri hedef değişkene uyan bir çiftin içinde istatistik olarak anlamlı bir fark kalmayınca kadar birleştirilmektedir. CHAID ile diğer yöntemler arasındaki en önemli farklılıklardan birisi, ID3, C4.5, ve CRT ikili ağaçlar türetirken, CHAID çoklu ağaçlar türetmektedir (Albayrak ve Yılmaz, 2009).

CHAID, bölümlendirme amaçlı kullanılan etkili bir istatistiksel tekniktir. Bir istatistiksel testin anlamlılığını kriter olarak kullanarak, bir potansiyel ön kestirici değişkenin tüm değerlerini değerlendirir. Hedef değişkene veya aynı anlama gelmek üzere bağlı değişkene göre istatistiksel olarak homojen (benzer) olarak değerlendirilen tüm değerleri birleştirir ve diğer tüm değerleri heterojen (benzer olmayan) olarak değerlendirir. Ardından karar ağacındaki ilk dalın formuna göre en iyi ön kestirici değişkenin seçilmesiyle, her bir düğümün seçilen değişkenin homojen değerlerinin bir grubunu oluşturmasını sağlar. Bu süreç ardıl olarak ağaç tamamıyla büyüyene kadar sürer. Kullanılan istatistiksel test, hedef değişkenin ölçüm düzeyine bağlıdır. Eğer hedef değişken sürekli bir değişken ise, F testi kullanılır. Eğer hedef değişkeni kategorik ise ki-kare testi kullanılmaktadır (Tüzüntürk ve Oğuzlar, 2009).

CHAID algoritması, kategorik değişkenler için gözlenen sıklık değerlerinin X^2 analizini yaparak bu değişkenlerin ne kadar iyi cevap verdiklerine karar verir. CHAID, popülasyondaki istatistiksel önemi olan grupları keşfetmek için kullanılır. CHAID algoritması, dallanan değişken ile bağımlı değişken arasındaki bağımlılığı test eder (Mattison, 1997).

Bağımlı değişken ile ele alınan bağımsız değişken arasındaki ilişkiyi araştıran bu testin sonucu iki değişken arasında bağımlılığı ifade ediyorsa ağacın büyümesine, bağımsızlığı ifade ediyor ise ağacın durmasına sebep olur. Bu, beklenen bir durumdur. Çünkü amaç bağımlı değişkenin dallanarak açıklanmasıdır ve bağımsız olmaları o bağımsız değişkenin, bağımlı değişkeni açıklamadığını ifade etmektedir. Test sonucu olasılık değeri en küçük olan yani önem değeri en yüksek olan değişken dallanma için seçilir (Çelik, 2009).

C&RT (Sınıflandırma ve Regresyon Ağacı)

1984 yılında Breiman, Friedman, Olshen ve Stone tarafından geliştirilen C&RT algoritmasında, her aşamada ilgili grubun, kendinden daha homojen olan iki alt gruba (*Binary Tree*) ayrılması sağlanmaktadır. Ayrım işlemi kategorik bağımlı değişkenler için *gini*, *twoing*, sürekli değişkenler için en küçük kareler sapması (*Least-Squared Deviation*) indeks hesaplamalarına göre yapılmaktadır. Bu hesaplamalarda kar, maliyet değerleri ve değişken kategorileri arasındaki önceliklerin tanımlanabilmesi

gibi sağlanan çeşitli esneklikler, C&RT algoritmasının günümüzde de yoğun olarak tercih edilmesine neden olmaktadır (Akpınar, 2000).

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

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

Modelin doğruluğu, bağımlı değişkenin kategorik olması durumunda doğru tahmin edilen kayıtların oranı, sürekli olması durumunda ise ortalama hata kareler ile ölçülür (Çelik, 2009).

C&RT ağaç modeli, tek değişkenli ikili kararların bir hiyerarşisini içerir. C&RT verileri iki alt kümeye ayırdığı için her bir alt küme içindeki durumlar, bir önceki alt kümeden daha homojen olacaktır. Bu ardışık süreç, homojenlik kriterine ulaşıncaya veya diğer bazı durma kriterleri sağlanıncaya değin kendini tekrar eder. Aynı kestirim değişkeni ağaçta farklı düzeylerde pek çok kez kullanılabilir. Ağacın yapısı önceden belirlenmemekte, verilerden türetilmektedir (Answer Tree 3.0 User's Guide, 2001:189). C&RT, kök düğümünde, verilerin iki gruba bölünmesi için en iyi değişkenin seçilmesini sağlar ve farklı bölümlendirme (splitting) kriterleri kullanır. Bu bölümlendirme kriterlerinin tümü, her bir alt kümedeki sınıf etiketlerini mümkün olduğunca homojen olacak biçimde bölümlendirir (Classification and Regression Trees: An Introduction, 2003: 12). Bölümlendirme prosedürü çocuk düğümlere (child node) veya alt düğümlerin her birine ardışık olarak uygulanır.

C&RT ağaçları, kesin bir heterojenlik (impurity) ölçüsüne bağlı olarak düğümlere ayrılmış iki değerli (binary) ağaçlardır ve bu nedenle de sonuçta homojen dallar oluşmaktadır.

C&RT analizi ağaç yapısına dayalı diğer sınıflama teknikleri ile kıyaslandığında çok sayıda avantaja sahiptir. İlki ve belki de en önemli olan özelliği parametrik olmayışıdır. Diğer bir söyleyişle ön kestirici veya aynı anlama gelmek üzere bağımsız değişken değerlerine ilişkin varsayımlar gerektirmemektedir. Bu nedenle C&RT analizinde kullanılacak değişkenler çok çarpık sayısal değişkenler olabileceği gibi, sınıflayıcı veya sıralayıcı yapıya sahip kategorik değişkenler de olabilir. Bu önemli bir özelliktir ve analizi yapacak araştırmacıya, normallik araştırma ve dönüşüm yapma gibi işlemler gerektirmediğinden zaman kazandırmaktadır. C&RT analizi, ele alınan problem yüzlerce mümkün bağımsız değişken içerse bile, bölümlendirilecek tüm mümkün değişkenleri araştırma gücüne sahiptir (Tüzüntürk ve Oğuzlar, 2009).

4.2.1.4.Kümeleme Yöntemi

Kümeleme, veriyi sınıflara veya kümelere ayırma işlemidir. Aynı kümedeki elemanlar birbirleriyle benzerlik gösterirlerken, başka kümelerin elemanlarından farklıdırlar. Kümeleme modelinde, sınıflama modelinde olan veri sınıfları yoktur. Verilerin herhangi bir sınıfı bulunmamaktadır. Sınıflama modelinde, verilerin sınıfları bilinmekte ve yeni bir veri geldiğinde bu verinin hangi sınıftan olabileceği tahmin edilmektedir. Oysa kümeleme modelinde, sınıfları bulunmayan veriler gruplar halinde kümelere ayrılırlar. Bazı uygulamalarda kümeleme modeli, sınıflama modelinin bir ön işlemi gibi görev alabilmektedir. En iyi bilinen ve en çok kullanılan kümeleme yöntemi, bölme yöntemidir.

Bölme yöntemlerinde, n veri tabanındaki nesne sayısı ve k oluşturulacak küme sayısı olarak kabul edilir. Bölme algoritması n adet nesneyi, k adet kümeye böler ($k \leq n$). Kümeler tarafsız bölme kriteri olarak nitelendirilen bir kritere uygun oluşturulduğu için aynı kümedeki nesneler birbirlerine benzerken, farklı kümedeki nesnelerden farklıdırlar. Bölme yöntemlerinde kullanılan en yaygın metot ise k-means yöntemidir (Akbulut, 2004)

K-means yöntemi, ilk önce n adet nesneden rastgele k adet nesne seçer ve bu nesnelerin her biri, bir kümenin merkezini veya orta noktasını temsil eder. Geriye kalan nesnelerden her biri kendisine en yakın olan küme merkezine göre kümelere dağılırlar. Yani bir nesne hangi kümenin merkezine daha yakın ise o kümeye yerleşir. Ardından her küme için ortalama hesaplanır ve hesaplanan bu değer o kümenin yeni merkezi olur. Bu işlem tüm nesneler kümelere yerleşinceye kadar devam eder (Han and Kamber, 2006).

Birliktelik Kuralı

Birliktelik kuralları, bilgisayar bilimleri alanında geliştirilmiş ve genellikle pazar sepet analizi, belirli müşteriler tarafından alınan ürünlerin birlikteliklerini ölçümlemek veya web sitelerine tıklayan kişilerin ardışık olarak ziyaret ettikleri sayfaları görüntülemek gibi önemli analizlerde kullanılırlar. Amaç, belirli bir işlemler kümesinde tipik olarak birlikte olan kalemlerin altını çizmektir (Guidici, 2003) .

Bir alışveriş sırasında veya birbirini izleyen alışverişlerde müşterinin hangi mal veya hizmetleri satın almaya eğilimli olduğunun belirlenmesi, müşteriye daha fazla ürünün satılmasını sağlama yollarından biridir. Satın alma eğilimlerinin tanımlanmasını sağlayan birliktelik kuralları ve ardışık zamanlı örüntüler, pazarlama amaçlı olarak pazar sepeti analizi (*Market Basket Analysis*) adı altında veri madenciliğinde yaygın olarak kullanılmaktadır. Bununla birlikte bu teknikler, tıp, finans ve farklı olayların birbirleri ile ilişkili olduğunun belirlenmesi sonucunda değerli bilgi kazanımının söz konusu olduğu ortamlarda da önem taşımaktadır (Akpınar, 2000).

Birliktelik Analizi, verilen bir veri kümesinde sıkça tekrarlanan nitelik-değerli koşulları gösteren birliktelik kurallarının araştırılmasıdır. Birliktelik analizi pazar sepet (market basket) analizinde geniş bir kullanıma sahiptir. Pazar sepet analizinde, tüketicinin alışverişi sürecince beraber aldığı mallar belirlenir. Pazar sepet analizinin çıktısı, tüketici satın alma davranışına ilişkin birliktelikler kümesidir. Birliktelikler, birliktelik kuralları (association rules) olarak bilinen kuralların özel bir kümesi formunda verilir. Birliktelik kuralları, uygun ürün pazarlama stratejisinin belirlenmesinde yardımcı olacaktır.

Birliktelik Kuralları $X \Rightarrow Y$ formundadır. Bu birliktelik kuralı, X kuralını sağlayan veritabanındaki kayıtların aynı zamanda Y' nin koşullarını da sağladığı anlamındadır. Birliktelik kurallarında, geleneksel sınıflama kurallarının aksine bir nitelik bir kuralda ön koşul olarak görünürken, diğerinde bağlı koşul olarak görünebilmektedir. Bunun yanında geleneksel sınıflama kuralları bir kuralın bağlı kısmı tek nitelik olacak biçimde sınırlandırmıştır. Birliktelik kuralında ise bir kuralın bağlı kısmı, bir veya birden fazla nitelik değeri içerebilir. Birliktelik kuralları, tek bağlı değişken seçimi sınırlaması olmadan büyük veritabanlarındaki ilişkilerin bulunmasını sağladığından popülerdirler.

Tüketicinin satın alma trendindeki ilginç ilişkiler aşağıdaki dört ürün için belirlenmeye çalışılmış olsun:

- Süt
- Peynir
- Ekmek
- Yumurta

Mümkün birliktelikler ise aşağıdaki biçimde verilmiştir:

- Tüketici süt aldığı anda aynı zamanda ekmek de alır.
- Tüketici ekmek aldığı anda aynı zamanda süt de alır.
- Tüketici süt ve yumurta aldığı anda aynı zamanda peynir ve ekmek de alır.
- Tüketici süt, peynir ve yumurta aldığı anda ekmek de alır.

İlk birliktelik, süt alan tüketicilerin aynı zamanda ekmek de satın aldığını söylemektedir. Bu kural için güven, süt alındığında ekmek almanın koşullu olasılığıdır. Eğer 10000 tüketici süt almışsa ve bunların 5000 aynı zamanda ekmek almışlarsa, süt alındığında ekmek de almanın güveni $5000/10000 = \%50$ olarak bulunur.

İkinci kurala bakıldığında ilk kural ile aynı bilgiyi vermediği anlaşılabilmektedir. Bu birliktelikteki ilk kural ekmek satın alınmış olmasıdır. Bir örnek olarak 20000 tüketicinin ekmek aldığı biliniyorsa ve bunların herhangi 5000 tanesi aynı zamanda süt almış olsalar, ekmek alındığında süt de almanın güven değeri $5000/20000 = \%25$

olacaktır. Üçüncü ve dördüncü birlikteliklerin, birinci ve ikinci ile benzer biçimde güven değerleri hesaplanabilir.

Bir birliktelik kuralından bulunan nitelik değerlerini içeren tüm örneklerin yüzdesi güven değerinden farklı olacaktır. Bu istatistik, kural için destek (support) olarak bilinir. Destek, en basit anlamıyla, belirli bir birliktelik kuralındaki tüm nitelikleri içeren, veri tabanındaki örneklerin minimum yüzdesidir. Destek ve güven değerleri öncül kısım A ile ve bağlı kısım da B ile gösterilmek üzere aşağıdaki biçimde gösterilebilir.

$$\text{Destek}(A \Rightarrow B) = P(A \cup B) \quad (4.7)$$

$$\text{Güven}(A \Rightarrow B) = P(B \setminus A) \quad (4.8)$$

Birliktelik kural madenciliği iki adımlı bir süreçtir:

Tüm sık tekrarlanan kümelerin bulunması: Tüm kümelerin her biri en az önceden belirlenen minimum destek sayısı kadar tekrarlanacaktır.

Sık tekrarlanan kümelerden güçlü birliktelik kurallarının türetilmesi: Bu kurallar minimum destek ve minimum güveni sağlamak zorundadır (Tüzüntürk ve Oğuzlar, 2009).

Apriori Algoritması

Birliktelik kurallarının üretilmesi için birçok yöntem kullanılmaktadır. Bunlardan en yaygın kullanılan “apriori algoritması” adını taşımaktadır. Bu algoritma aşağıda belirtilen aşamalara sahiptir:

- a) Birliktelik çözümlemesinin yapılabilmesi için öncelikle destek ve güven ölçütlerini karşılaştırmak üzere eşik değerler belirlenir. Uygulamadan elde edilen sonuçların bu eşik değerlere eşit ya da büyük olması beklenir.
- b) Veritabanı taranarak çözümlemeye dâhil edilecek her bir ürün için tekrar sayıları, yani destek sayıları hesaplanır. Bu destek sayıları eşik destek sayısı ile karşılaştırılır. Eşik destek sayısından küçük değerlere sahip satırlar çözümlemeden çıkarılır ve koşula uygun kayıtlar göz önüne alınır.

- c) Yukarıdaki adımda seçilen ürünler bu kez ikişerli gruplandırılarak, bu grupların tekrar sayıları, yani destek sayıları elde edilir. Bu sayılar eşik destek sayıları ile karşılaştırılır. Eşik değerden küçük değerlere sahip satırlar çözümlenmeden çıkarılır.
- d) Bu kez üçerli, dörderli vb. gruplandırmalar yapılarak bu grupların destek sayıları elde edilir ve eşik değer ile karşılaştırılır, eşik değere uygun olduğu sürece işlemlere devam edilir.
- e) Ürün grubu belirlendikten sonra kural destek ölçütüne bakılarak birliktelik kuralları türetilir ve bu kuralların her birisiyle ilgili olarak güven ölçütleri hesaplanır (Özkan, 2008).

5. UYGULAMA

Bu uygulamada, bir perakende işletmesinin belirli bir dönem boyunca alışveriş yapmış müşteri verileri üzerinde, veri madenciliği uygulamalarında kullanılan SPSS Clementine programı ile müşteri davranışlarının belirlenebilmesi için veri manipülasyonu ve çeşitli modelleme teknikleri ele alınmıştır.

Bu çalışmada, perakende sektöründe hizmet veren bir kuruma ait işlemsel bazda 2009 yılının ilk 9 ayını kapsayan veriler kullanılmıştır. Firmanın hedef kitlesini, perakende satış yapan işletmeler ve catering işletmeler oluşturmaktadır.

Firma, verilerini tek bir sistem üzerinde değil, kendi aralarında “konuşabilen” farklı sistemler üzerinde tutmaktadır. Bundan dolayı, analiz için gerekli olan bilgileri elde etmek üzere, farklı sistemlerde tutulan veri tablolarının kullanılması bir zorunluluk olarak ortaya çıkmıştır.

Bu çalışmada kullanılacak veriler farklı sistemlerde tutulduğu için firma tarafından farklı dosya formatlarında verilmiştir. Bu dosyalara ait ayrıntılara aşağıda değinilmiştir:

Urun_Master.txt : Firmanın satmış olduğu ürünlere ilişkin bilgilerin tutulduğu dosyadır; 235 adet kayıttan oluşmaktadır. Dosya içerisinde; ürün kodları, ürün ambalaj bilgileri, ürünlere ait marka tanımlamalarına ait kodlar ve ürünlere ait kategori tanımlamalarına ait kodlar mevcuttur.

Urun_Master.txt dosyasındaki alan ve alanlara ait açıklamalar Çizelge 5.1’ de gösterilmektedir.

Urun1	Urun1 ve Urun2 kodlarının birleşiminden bir tek ürüne karşılık gelen Ürün Kodu oluşturulmaktadır.
Urun2	
UrunAmb	Ürüne ait ambalaj bilgilerinin tutulduğu alandır.
AnaMrk	Ürünün sahip olduğu ana markaya ait kodların tutulduğu alandır. Ana markanın bir alt kırılımı olarak marka bulunmaktadır.
Marka	Ürünün sahip olduğu markaya ait kodların tutulduğu alandır.
AnaKat	Ürüne ait ana kategorilerin tutulduğu alandır.
Ktg	Ürüne ait kategorilerin tutulduğu alandır.
AltKtg	Ürüne ait alt kategorilerin tutulduğu alandır.
DetKat	Ürüne ait detay kategorilerin tutulduğu alandır.

Çizelge 5.1.Ürün tablosuna ait alanlar

Çizelge 5.1’ de gösterilen kategorilere ait alanlar birbirlerinin birleşiminden oluşmaktadır. Kategorilere ait kodlar ve oluşturulma kuralları aşağıdaki format yapısında gösterilmektedir:

- Ana Kategori Kod: AnaKat (Örn. Temel Gıda)
- Kategori Kod: AnaKat + Ktg (Örn. Sıvı Yağlar)
- Alt Kategori Kod: AnaKat + Ktg + AltKtg (Örn. Zeytin Yağları)
- Detay Kategori Kod: AnaKat + Ktg + AltKtg + DetKat (Örn. Sızma Zeytin Yağları)

Satış.txt : İşlemsel bazda satış bilgilerinin tutulduğu dosyadır; 765.645 adet kayıttan oluşmaktadır. Dosya içerisinde, alışveriş yapan müşterilere ait müşteri numaraları, satışı yapılan ürünlere ait kod bilgileri, satış tutar bilgileri, satışı yapılan ürünlerin koli veya paket cinsinden satış bilgileri ve satışa ait fatura bilgileri mevcuttur.

Satış.txt dosyasındaki alan ve alanlara ait açıklamalar Çizelge 5.2’ de gösterilmektedir.

Def	Müşterilerin defter kodunu belirten alandır (M : Müşteri, S : Tedarikçi). Verilerde sadece M kodu bulunmaktadır.
Cari	Müşteri numaralarının bulunduğu alandır. Müşteri bazında tekildir.
Sube	Müşterilerin alışveriş yapmış oldukları mağaza kodlarını belirtmektedir.
Srk1	Satış veri dosyasında, Srk1 ve Srk2 kodlarının birleşiminden ürün kodu elde edilmektedir.
Srk2	
BirFia	Satın alınan ürünlerin birim fiyatlarının bulunduğu alandır
Net	Satın alınan ürünlerin tutarlarının bulunduğu alandır.
VerMik	Alınan ürünlerin miktarının tutulduğu alandır.
SatBrm	Alınan ürünün koli cinsinden bilgilerinin tutulduğu alandır (AD: Adet, KL: Koli)
EkbMik	Ürüne ait en küçük birim miktarının tutulduğu alandır. Örneğin Vermik: 8, SatBrm: KL, EkbMik: 48, EkbMik: AD olsun. Bunun anlamı 8 koli ürün alınmış ve bu 48 adet ürüne karşılık gelmektedir.
EkBrm	Ürüne ait koli cinsinden en küçük birimin gösterildiği alandır
FatSer	İlgili alışverişe ait Fatura Seri Numaralarının tutulduğu alandır.
FatNo	İlgili alışverişe ait Fatura Numaralarının tutulduğu alandır.
FatYil	Alışverişin yapıldığı Yıl bilgisinin tutulduğu alandır.
FatAy	Alışverişin yapıldığı Ay bilgisinin tutulduğu alandır.
FatGun	Alışverişin yapıldığı Gün bilgisinin tutulduğu alandır.

Çizelge 5.2. Satış tablosuna ait alanlar

Musteri.txt: Firmanın müşterilerine ait bilgilerin tutulduğu dosyadır; 90.322 adet kayıttan oluşmaktadır. Dosya içerisinde, firmanın müşterilerine ait müşteri numaraları, müşterilere ait grup, tip, sınıf tanımlamaları ve müşterilerin firmaya kayıt oldukları tarih bilgilerini içermektedir.

Musteri.txt dosyasındaki alan ve alanlara ait açıklamalar Çizelge 5.3’ de gösterilmektedir.

DefKD	Müşterilerin defter kodunu belirten alandır (M: Müşteri, S: Tedarikçi). Verilerde sadece M kodu bulunmaktadır.
SicKd	Müşteri numaralarının bulunduğu alandır. Müşteri bazında tekildir.
MGRP	Müşteri Grup tanım bilgilerinin tutulduğu alandır
MTIP	Müşteri Tip tanım bilgilerinin tutulduğu alandır.
MSNF	Müşteri Sınıf tanım bilgilerinin tutulduğu alandır.
KAYSUB	Müşterinin firmaya hangi şubesinden kayıt yaptırdığını gösteren alandır.
KAYYIL	Müşterinin firmaya hangi yıl kayıt yaptırdığını gösteren alandır.
KAYAY	Müşterinin firmaya hangi ay kayıt yaptırdığını gösteren alandır.
KAYGUN	Müşterinin firmaya hangi gün kayıt yaptırdığını gösteren alandır.

Çizelge 5.3.Müşteri tablosuna ait alanlar

Çizelge 5.3’ de gösterilen müşteri grup, tip ve sınıf kodlarına ait alanlar birbirlerinin birleşiminden oluşmaktadır. Bu kodlar ve oluşturulma kuralları aşağıdaki format yapısında gösterilmektedir:

- Müşteri Grup Kod: MGRP (Örn. Perakendeci Müşteri)
- Müşteri Tip Kod: MGRP + MTIP (Örn. Market)
- Müşteri Sınıf Kod: MGRP + MTIP + MSNF (Örn. Büyük Süper Market)

Birim_Maliyet.dbo: Ürünlerin birim maliyetlerini içeren dosyadır.

Müşteri grup, tip, sınıf tanımlamaları ve ürün kategori ve marka tanımlamalarının getirilebilmesi için hazırlanan lookup tablolar excel dosya formatında alınmıştır.

5.1. VERİ MANİPÜLASYONU

Veri madenciliğinin en uzun süreci, verilerin elde edilmesi, düzenlenmesi, temizlenmesi, eksik ve aykırı gözlemlerin belirlenerek uygun tekniklerle gerekli işlemlerin yapılarak veri setinin analize hazır duruma getirilmesini kapsayan manipülasyon sürecidir. Bu sürecin uzun olması, veri madenciliğinde kullanılan veri setlerinin farklı fiziksel kaynaklardan elde edilebilmesi, çok büyük boyutlarda olması ve ham verilerde görünen tutarsızlık gibi nedenlerden kaynaklanmaktadır.

Verilerin farklı sistemlerde ve farklı tip dosya formatlarında tutulmuş olması, analiz öncesinde bazı işlemlerin yapılmasını gerekli kılmaktadır. Veri madenciliği tekniklerinin uygulanabilmesi için, öncelikle farklı sistemlerde farklı formatlarda tutulan verilerin birleştirilerek tek bir sisteme aktarılması gerekmektedir. Bununla beraber, veri tablolarında görülebilecek eksik, aykırı ve/veya tekrarlı gözlemler, analiz için gerekli olmayan değişkenler ve tutarsız veriler uygun EDA ve görselleştirme teknikleriyle belirlenerek var olan problemler giderilmelidir.

Çalışmada uygulanan manipülasyon süreci genel olarak sırasıyla aşağıdaki adımlarda belirtilmiştir:

- 1) Farklı sistemlerden gelen ürün bilgilerine ait *Urun_Master.txt*, satış bilgilerine ait *Satis.txt* ve müşteri bilgilerine ait *Musteri.txt* adlı üç dosyaya çeşitli manipülasyon teknikleri uygulanmış ve içerdikleri ilgili kodlara ait tanımlama bilgileri getirilmiştir.
- 2) Manipülasyon çalışması yapılan bu üç dosya birleştirilerek tek bir dosya olarak kaydedilmiştir.
- 3) Ürünlere ait birim maliyetlerin tutulduğu *Birim_Maliyet.dbo* adlı dosyadan ilgili ürüne ait farklı aylardaki birim maliyetler elde edilmiş ve elde edilen veri *SQL_DB_URUN.sav* adında farklı bir dosya olarak kaydedilmiştir.
- 4) Manipülasyon çalışmaları sonucunda oluşturulan yeni iki dosyanın birleştirilmesi ile bir müşteriye ait farklı yerlerde tutulan bilgiler toplanarak analiz edilebilir hale getirilmiştir.
- 5) Analiz edilebilir hale getirilen veriden, müşterilerin kaç defa mağazalardan alışveriş yaptığı, alışveriş yaptığı ay sayısı, müşterilere ait ortalama ürün kalem sayısı ve ürün çeşit sayısı gibi nicel bilgiler elde edilmiştir.

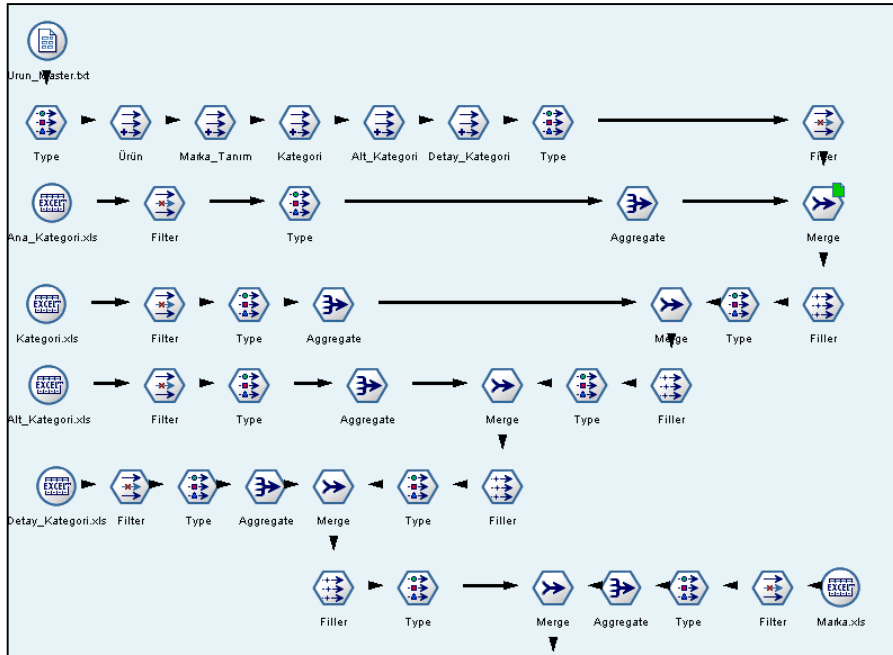
5.1.1. Manipülasyon Çalışmalarına Ait Akışlar ve Değerlendirmeler

Urun_Master.txt dosyası üzerinde yapılan manipülasyon işlemleri adımsal olarak açıklanmadan önce, analiz sürecinde kullanılacak ürün tablosunun ham hali aşağıda Şekil 5.1.'de gösterilmektedir.

	Urun1	Urun2	UrnAmb	AnaMrk	Marka	AnaKat	Ktg	AltKat	DetKat
1	20	26 250GR		68	8	11	2	1	1
2	75	16 10KG		68	3	11	5	1	1
3	75	18 10KG		68	3	11	5	1	3
4	75	22 25KG		68	3	11	5	1	3
5	75	25 10KG		68	3	11	5	4	1
6	75	28 10KG		68	3	11	5	2	1
7	75	31 10KG		68	3	11	5	5	1
8	75	33 10KG		68	3	11	5	3	1
9	75	34 25KG		68	3	11	5	1	4
10	75	36 10KG		68	3	11	5	1	4
11	76	60 25KG		68	3	11	5	2	1
12	76	61 25KG		68	3	11	5	5	1
13	76	62 25KG		68	3	11	5	3	1
14	76	65 25KG		68	3	11	5	1	1
15	76	66 10KG		68	3	11	5	6	1
16	76	68 25KG		68	3	11	5	4	1
17	76	70 10KG		68	3	11	5	2	2
18	76	71 25KG		68	3	11	5	2	2
19	76	96 30KG		68	5	22	2	1	1
20	76	97 25KG		68	5	22	3	1	1
21	76	98 30KG		68	5	22	1	2	1

Şekil 5.1.Ürün tablosuna ait veri seti

Urun_Master.txt dosyasında üzerinde gerçekleştirilen manipölasyon işlemlerinin akışı Şekil 5.2’ de gösterilmektedir.



Şekil 5.2. Ürün dosyası için gerçekleştirilen manipölasyona ait akış

Ürün dosyasında ürünlere ait kodlar ayrı ayrı tutulduğu için öncelikle, Urn1 ve Urn2 kodları birleştirilerek Ürün adında tek bir tekil anahtar alan elde edilmiştir.

	Urun1	Urun2
1	20	26
2	75	16
3	75	18
4	75	22
5	75	25
6	75	28
7	75	31
8	75	33
9	75	34
10	75	36
11	76	60
12	76	61
13	76	62
14	76	65
15	76	66
16	76	68
17	76	70
18	76	71
19	76	96
20	76	97

	Urun1	Urun2	Ürün
1	20	26	2026
2	75	16	7516
3	75	18	7518
4	75	22	7522
5	75	25	7525
6	75	28	7528
7	75	31	7531
8	75	33	7533
9	75	34	7534
10	75	36	7536
11	76	60	7660
12	76	61	7661
13	76	62	7662
14	76	65	7665
15	76	66	7666
16	76	68	7668
17	76	70	7670
18	76	71	7671
19	76	96	7696
20	76	97	7697

Şekil 5.3.Ürün kodunun oluşturulması

Kategorilere ait kodlar ve oluşturulma kuralları aşağıdaki format yapısında birleştirilmiştir:

- Ana Kategori Kod: AnaKat (Örn. Temel Gıda)
- Kategori Kod: AnaKat + Ktg (Örn. Sıvı Yağlar)
- Alt Kategori Kod: AnaKat + Ktg + AltKtg (Örn. Zeytin Yağları)
- Detay Kategori Kod: AnaKat + Ktg + AltKtg + DetKat (Örn. Sızma Zeytin Yağları)

AnaKat	Ktg	AltKat	DetKat
11	2	1	1
11	5	1	1
11	5	1	3
11	5	1	3
11	5	4	1
11	5	2	1
11	5	5	1
11	5	3	1
11	5	1	4
11	5	1	4
11	5	2	1
11	5	5	1
11	5	3	1
11	5	1	1
11	5	6	1
11	5	4	1
11	5	2	2
11	5	2	2
22	2	1	1
22	3	1	1
22	1	2	1

Ana_Kategori	Kategori	Alt_Kategori	Detay_Kategori
11 112	1121	11211	
11 115	1151	11511	
11 115	1151	11513	
11 115	1151	11513	
11 115	1154	11541	
11 115	1152	11521	
11 115	1155	11551	
11 115	1153	11531	
11 115	1151	11514	
11 115	1151	11514	
11 115	1152	11521	
11 115	1155	11551	
11 115	1153	11531	
11 115	1151	11511	
11 115	1156	11561	
11 115	1154	11541	
11 115	1152	11522	
11 115	1152	11522	
22 222	2221	22211	
22 223	2231	22311	
22 221	2212	22121	

Şekil 5.4.Kategori kodlarının oluşturulması

Kategorilere ait tanım bilgileri excel dosya formatı biçiminde tutulmaktadır. Bu dosyalar SPSS Clementine ile okutularak ürünlere ait kategori tanımlarını getirebilmek için akışa eklenmiştir.

Dosyalardaki Ana_Kategori, Kategori, Alt_Kategori ve Detay Kategori alanları anahtar alan olarak kullanılmış ve Ürün dosyası ile bu alanlar üzerinden join yapılarak ilgili ürünlerin yanlarına kategorilere ait tanımlamalar getirilmiştir. Birleştirme yapılmadan önce iki alanda aynı format tipine getirilmiştir. Ürün dosyasında bulunan kategori alanları ile excel dosyalarından gelen kategori alanları aynı format tipine çevrilmiştir. Bu dönüşümlerden sonra birleştirme gerçekleştirilmiştir. Kategorilere ait tanımlama bilgileri aşağıda Şekil 5.5’ de gösterilmektedir.

Ana_Kategori	Ana Kategori Tanım	Kategori	Kategori Tanım	Alt_Kategori	Alt Kategori Tanım
1	10 KRISTAL ÇUVAL ŞEKER	1	101 50KG KRISTAL TOZ ŞEKER	1	1011 50KG KRISTAL TOZ ŞE
2	11 TEMEL GIDA	2	102 25KG KRISTAL TOZ ŞEKER	2	1021 25KG KRISTAL TOZ ŞE
3	12 SÜT ve SÜT ÜRÜNLERİ	3	111 SİM YAĞLAR	3	1111 AYÇİÇEK YAĞLARI
4	15 ATİŞTIRMALIK ÜRÜNLER	4	112 MARGARİN	4	1112 MISIRÖZÜ YAĞLARI
5	16 ŞARKÜTERİ ve KAHVALTIK ÜR	5	113 UN	5	1113 ZEYTİNYAĞLARI
6	17 İÇECEKLER	6	114 MAKARNA	6	1114 BİTKİSEL KARMA YAĞL
7	18 SİGARALAR ve DİĞER TÜTÜN ...	7	115 BAKLIYAT	7	1117 CATERİNG SİM YAĞLA
8	19 EKMEKLER ve DİĞER UNLU M...	8	116 ŞEKERLER	8	1118 ENDÜSTRİYEL SİM YA
9	20 ET ve ET ÜRÜNLERİ	9	117 YARDIMCI ve KATKI MALZEMEL	9	1119 BİTKİSEL KARMA YAĞL
10	21 MEYVE SEBZE	10	118 KONSERVE-TURŞU-HAZIR YE...	10	1121 PAKET MARGARİNLER
11	22 TEMİZLİK ÜRÜNLERİ	11	119 BEBEK MAMALARI-BESİNLERİ	11	1122 KASE MARGARİNLER
12	23 KİŞİSEL BAKIM ve KOZMETİK Ü	12	121 UZUN ÖMÜRLÜ SÜT ve SÜT ...	12	1123 YEMEKLİK MARGARİN...
13	24 KAĞIT ÜRÜNLERİ	13	122 PASTÖRİZE SÜTLER	13	1124 SİM MARGARİNLER
14	26 EV GEREÇLERİ ve BAKIM ÜRÜ...	14	151 BİSKÜVİLER	14	1128 CATERİNG MARGARİ...
15	27 ENERJİ ve AYDINLATMA ÜRÜN...	15	152 KEKLER	15	1131 UNLAR
16	28 DİĞER GIDA DIŞI ÜRÜNLER	16	153 ATİŞTIRMALIK UNLU ÜRÜNLER	16	1132 İRMİKLER
17	31 AMBALAJ MALZEMELERİ	17	154 ÇİKOLATALAR	17	1133 NİŞASTALAR
18	36 KOMÜNİKASYON	18	155 ÇİKOLATA KAPLAMALILAR	18	1134 PİRİNÇ UNLARI
19	50 PL KATEGORİSİ	19	156 ŞEKERLER	19	1139 DİĞER UNLAR
		20	157 SAKIZLAR	20	1141 MAKARNALAR

Şekil 5.5. Kategorilere ait tanımlamalar

Manipülasyon işlemleri sonucunda oluşan yeni ürün tablosu aşağıda Şekil 5.6’ da gösterilmektedir.

Urun Kod	Ana_Kategori	Kategori	Alt_Kategori	Detay_Kategori	Ana Kategori Tanım	Kategori Tanım	Alt Kategori Tanım	Detay Kategori Tanım
1	1061	24	244	2442	24421 KAĞIT ÜRÜNLERİ	KAĞIT PEÇETE ve HAVLULAR	KAĞIT HAVLULAR	KAĞIT HAVLULAR
2	1063	24	245	2451	24512 KAĞIT ÜRÜNLERİ	TUVALET KAĞITLARI	TUVALET KAĞITLARI	ÇİFT KATLI TUVALET
3	10820	22	222	2221	22211 TEMİZLİK ÜRÜNLERİ	BULAŞIK TEMİZLEME ÜRÜNLERİ	ELDE YIKAMA BULAŞIK	SİM BULAŞIK DETRJA
4	1169	23	231	2314	23141 KİŞİSEL BAKIM ve KOZMETİK Ü	SABUNLAR	SİM SABUNLAR	SİM SABUNLAR
5	1176	23	231	2314	23141 KİŞİSEL BAKIM ve KOZMETİK Ü	SABUNLAR	SİM SABUNLAR	SİM SABUNLAR
6	1177	23	231	2314	23141 KİŞİSEL BAKIM ve KOZMETİK Ü	SABUNLAR	SİM SABUNLAR	SİM SABUNLAR
7	1240	15	151	1512	15121 ATİŞTIRMALIK ÜRÜNLER	BİSKÜVİLER	KREMALI BİSKÜVİLER	KREMALI BİSKÜVİL...
8	1241	15	151	1512	15121 ATİŞTIRMALIK ÜRÜNLER	BİSKÜVİLER	KREMALI BİSKÜVİLER	KREMALI BİSKÜVİL...
9	12775	11	115	1151	11511 TEMEL GIDA	BAKLIYAT	PIRİNÇLER	BALDO PİRİNÇLER
10	1313	15	155	1551	15512 ATİŞTIRMALIK ÜRÜNLER	ÇİKOLATA KAPLAMALILAR	ÇİKOLATA KAPLAMALIL	ÇİK.KAP BARLAR
11	13310	11	116	1162	11622 TEMEL GIDA	ŞEKERLER	KESME ŞEKERLER	KESME ŞEKERLER
12	13311	11	116	1162	11622 TEMEL GIDA	ŞEKERLER	KESME ŞEKERLER	KESME ŞEKERLER
13	13312	11	116	1161	11611 TEMEL GIDA	ŞEKERLER	TOZ ŞEKERLER	TOZ ŞEKERLER
14	13313	11	116	1161	11611 TEMEL GIDA	ŞEKERLER	TOZ ŞEKERLER	TOZ ŞEKERLER
15	13314	11	116	1161	11611 TEMEL GIDA	ŞEKERLER	TOZ ŞEKERLER	TOZ ŞEKERLER
16	1339	11	116	1162	11622 TEMEL GIDA	ŞEKERLER	KESME ŞEKERLER	KESME ŞEKERLER
17	13682	16	162	1623	16231 ŞARKÜTERİ ve KAHVALTIK ...	REÇEL- MARMELAT VE BALLAR	BALLAR	BALLAR
18	13683	16	162	1623	16231 ŞARKÜTERİ ve KAHVALTIK ...	REÇEL- MARMELAT VE BALLAR	BALLAR	BALLAR
19	13779	11	113	1131	11311 TEMEL GIDA	UN	UNLAR	UNLAR
20	13780	11	113	1131	11311 TEMEL GIDA	UN	UNLAR	UNLAR
21	13781	11	113	1131	11311 TEMEL GIDA	UN	UNLAR	UNLAR

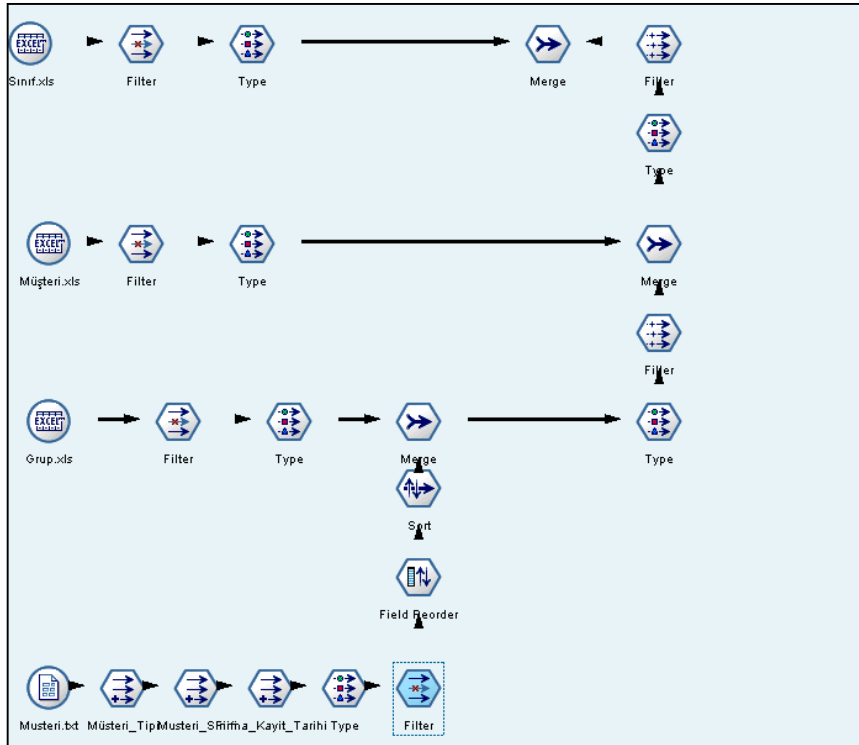
Şekil 5.6. Manipülasyon sonucu elde edilen ürün tablosu

Musteri.txt dosyası üzerinde yapılan manipölasyon işlemleri adımsal olarak açıklanmadan önce, analiz sürecinde kullanılacak ürün tablosunun ham hali aşağıda Şekil 5.7’ de gösterilmektedir.

	DefKD	SicKd	MGRP	MTIP	MSNF	KAYSUB	KAYYIL	KAYAY	KAYGUN
1	M	5	6	1	48	3414	2002	3	23
2	M	52	6	1	48	3418	2002	8	12
3	M	60	6	1	48	3403	2002	1	14
4	M	72	6	1	2	3401	2002	3	17
5	M	213	1	1	2	3504	2002	2	11
6	M	355	1	1	4	3504	2002	1	19
7	M	509	1	1	4	4101	2002	1	12
8	M	529	1	2	1	4101	2002	2	7
9	M	549	1	2	1	4101	2002	1	12
10	M	575	1	2	1	4101	2002	1	16
11	M	601	1	2	1	4101	2002	1	12
12	M	617	1	1	4	4101	2002	3	9
13	M	638	1	2	1	4101	2002	1	28
14	M	652	1	2	1	4101	2002	11	29
15	M	691	2	2	2	4101	2002	1	15
16	M	713	1	1	3	4101	2002	1	17
17	M	734	1	1	3	4101	2003	5	12
18	M	830	1	1	4	4101	2002	1	21
19	M	842	1	2	1	4101	2002	3	10
20	M	847	1	2	1	4101	2002	1	12
21	M	987	1	2	1	4101	2002	2	2

Şekil 5.7.Müşteri tablosuna ait veri seti

Musteri.txt dosyasında üzerinde gerçekleştirilen manipölasyon işlemlerinin akışı Şekil 5.8’ de gösterilmektedir.



Şekil 5.8.Müşteri dosyası için gerçekleştirilen manipölasyona ait akışı

Müşteri gruplarına ait kodlar ve oluşturulma kuralları aşağıdaki yapıda birleştirilerek müşteri tip ve sınıf kodları elde edilmiştir;

- Müşteri Grup Kod: MGRP (Örn. Perakendeci Müşteri)
- Müşteri Tip Kod: MGRP + MTIP (Örn. Market)
- Müşteri Sınıf Kod: MGRP + MTIP + MSNF (Örn. Büyük Süper Market)

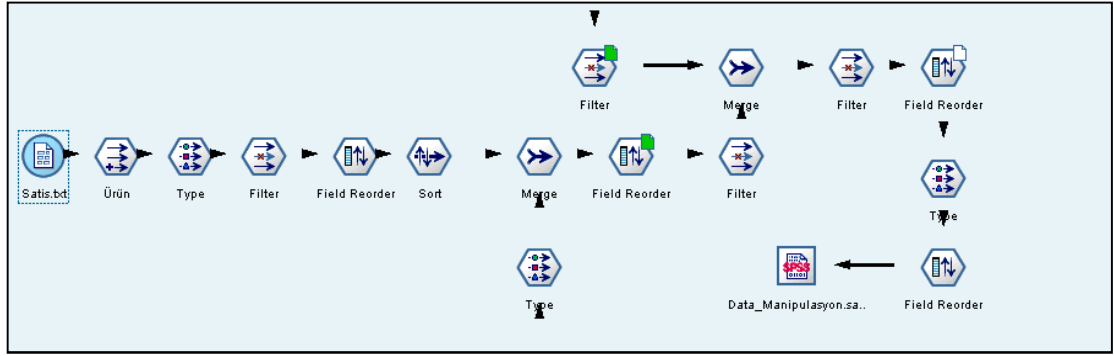
Müşteri guruplarına ait tanım bilgileri excel dosya formatı biçiminde tutulmaktadır. Bu dosyalar SPSS Clementine ile okutularak müşterilere ait grup tanımlarını getirebilmek için akışa eklenmiştir.

Müşterilerin mağazaya kayıt tarihleri Şekil 5.9’ da görüldüğü gibi gün, ay ve yıl olarak ayrı ayrı alanlarda tutulmaktadır. Bu üç alan daha sonra yapılacak analizlerde müşteri yaşının tespit edilip kullanılması amacıyla tarih formatına çevrilmiştir. Buradan müşterinin hangi süre zarfından itibaren firmaya kayıtlı olduğu tespit edilmiştir.

	KAYYIL	KAYAY	KAYGUN	Firma_Kayit_Tarihi
1	2002	3	23	23.03.2002
2	2002	8	12	12.08.2002
3	2002	1	14	14.01.2002
4	2002	3	17	17.03.2002
5	2002	2	11	11.02.2002
6	2002	1	19	19.01.2002
7	2002	1	12	12.01.2002
8	2002	2	7	07.02.2002
9	2002	1	12	12.01.2002
10	2002	1	16	16.01.2002
11	2002	1	12	12.01.2002
12	2002	3	9	09.03.2002
13	2002	1	28	28.01.2002
14	2002	11	29	29.11.2002
15	2002	1	15	15.01.2002
16	2002	1	17	17.01.2002
17	2003	5	12	12.05.2003
18	2002	1	21	21.01.2002
19	2002	3	10	10.03.2002
20	2002	1	12	12.01.2002

Şekil 5.9. Müşterilerin firmaya kayıt tarihlerinin manipölasyon sonucu elde edilmesi

Satis.txt dosyasında üzerinde gerçekleştirilen manipölasyon işlemlerinin akışı Şekil 5.10’ da gösterilmektedir.



Şekil 5.10. Satış dosyası için gerçekleştirilen manipülasyona ait akış

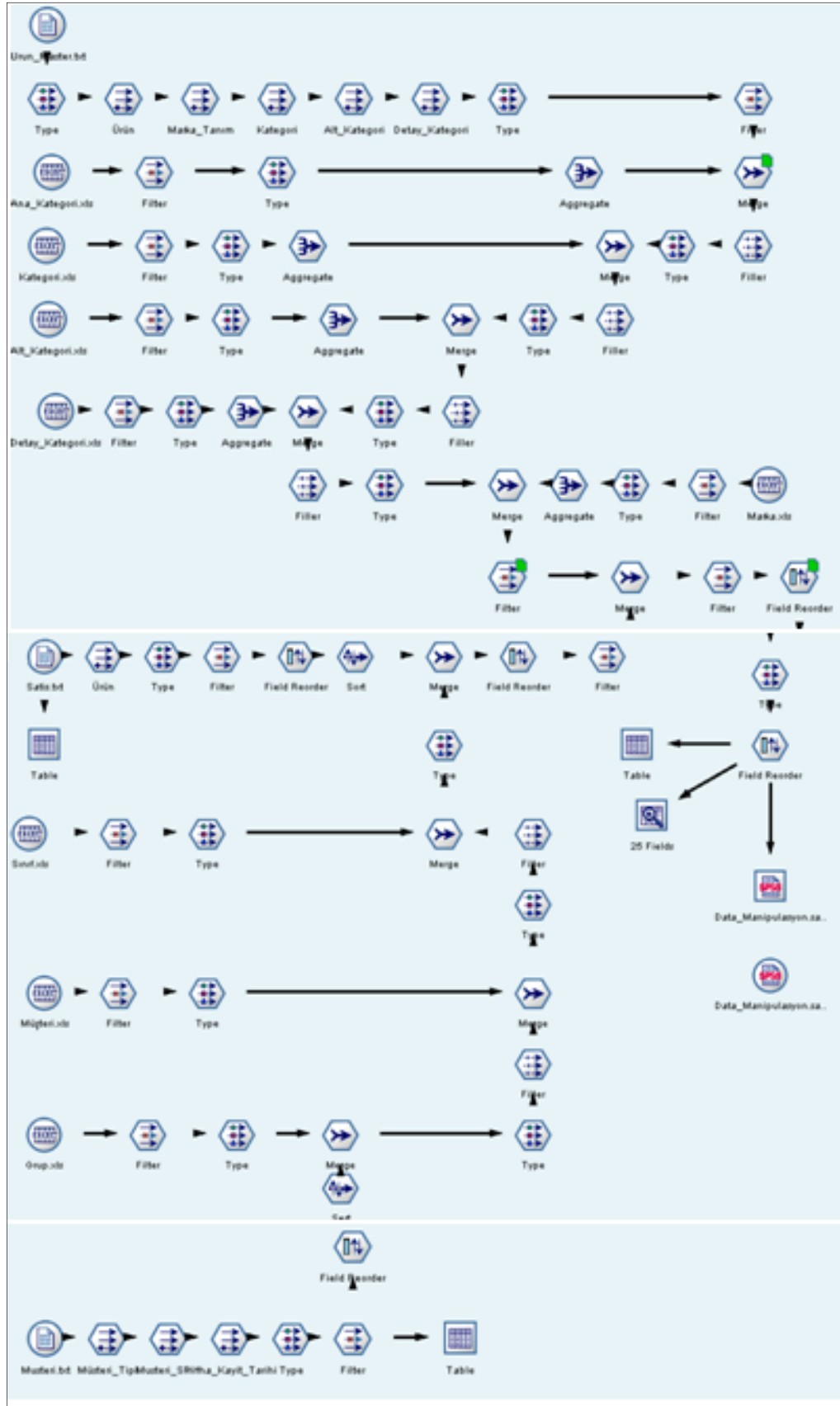
Çalışmada kullanılacak satış tablosunun ham hali aşağıda Şekil 5.11’ de gösterilmektedir.

	Def	Cari	Sube	Srk1	Srk2	BirFia	Net	VerMik	SatBrm	EkbMik	EkBrm	FatSer	FatNo	FatYil	FatAy	FatGun
1	M	218521	3201	20	26	0,630	1,260	2,000 AD		2,000 AD		1	18	2009	1	1
2	M	219107	3408	20	26	0,630	7,560	12,000 AD		12,000 AD		4	12	2009	1	1
3	M	392266	3408	20	26	0,630	1,890	3,000 AD		3,000 AD		2	19	2009	1	1
4	M	186361	5801	20	26	30,2...	30,2...	1,000 KL		48,000 AD		2	1	2009	1	1
5	M	415657	2001	20	26	0,630	2,520	4,000 AD		4,000 AD		3	14	2009	1	1
6	M	420386	2001	20	26	0,630	1,890	3,000 AD		3,000 AD		2	52	2009	1	1
7	M	223339	2001	20	26	0,630	8,820	14,000 AD		14,000 AD		2	56	2009	1	1
8	M	388216	5201	20	26	0,630	2,520	4,000 AD		4,000 AD		2	2	2009	1	1
9	M	444694	2001	20	26	0,630	2,520	4,000 AD		4,000 AD		3	22	2009	1	1
10	M	154263	3503	20	26	0,630	3,150	5,000 AD		5,000 AD		1	6	2009	1	1
11	M	361015	3503	20	26	0,630	3,150	5,000 AD		5,000 AD		1	12	2009	1	1
12	M	305557	2601	20	26	30,2...	30,2...	1,000 KL		48,000 AD		5	16	2009	1	1
13	M	118439	3404	20	26	0,630	1,260	2,000 AD		2,000 AD		1	23	2009	1	1
14	M	425345	4202	20	26	0,630	9,450	15,000 AD		15,000 AD		1	3	2009	1	1
15	M	303651	4202	20	26	0,630	1,890	3,000 AD		3,000 AD		1	9	2009	1	1
16	M	451870	3301	20	26	0,630	5,040	8,000 AD		8,000 AD		2	6	2009	1	1
17	M	52860	3801	20	26	30,2...	60,4...	2,000 KL		96,000 AD		3	4	2009	1	1
18	M	313588	6401	20	26	0,630	0,630	1,000 AD		1,000 AD		2	14	2009	1	1
19	M	252501	902	20	26	0,630	3,150	5,000 AD		5,000 AD		2	31	2009	1	1
20	M	486028	5301	20	26	0,630	15,1...	24,000 AD		24,000 AD		2	2	2009	1	1
21	M	336160	6301	20	26	0,630	3,150	5,000 AD		5,000 AD		1	23	2009	1	1

Şekil 5.11.Satış tablosuna ait veri seti

Bu dosya üzerinde de, Srk1 ve Srk2 alanları birleştirilerek ürün kodu elde edilmiştir. Faturaya ait yıl, ay ve gün alanları tek bir sütunda birleştirilmiş ve bu alan tarih formatına çevrilerek müşterilere ait alışveriş tarihleri elde edilmiştir.

Ürün, müşteri ve satış dosyalarına ait manipülasyonların en son halini gösteren akış aşağıdaki Şekil 5.12’ deki gibidir ve manipülasyon sonucu bundan sonraki analiz çalışmalarında kullanılmak üzere *Data_Manipulasyon.sav* adında tek bir dosya olarak kaydedilmiştir.



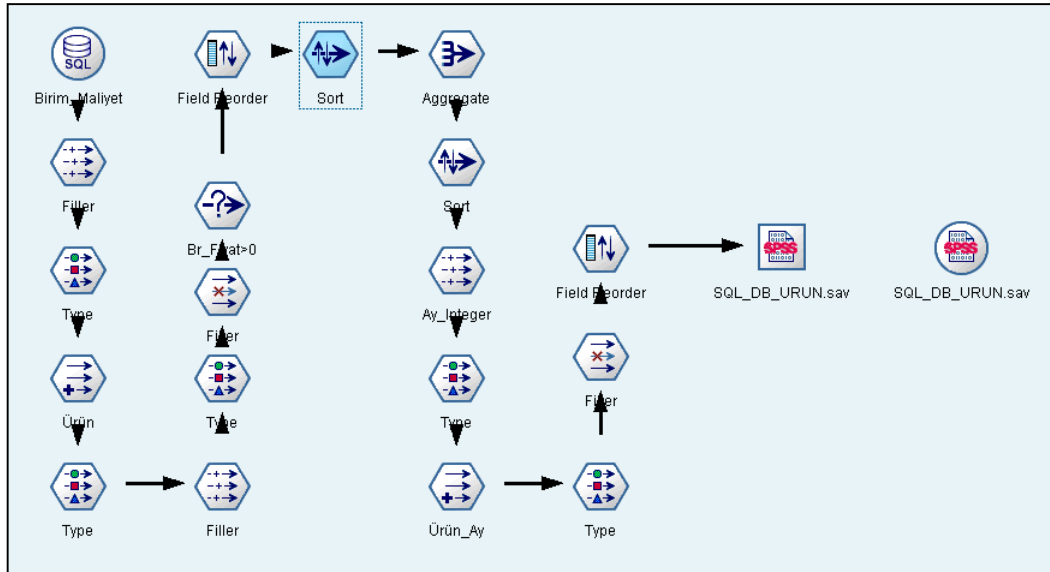
Şekil 5.12.Ürün, müşteri ve satış dosyalarına ait manipölasyonların en son halini gösteren akış

Birim maliyetlerin elde edilebilmesi için *Birim_Maliyet.dbo* dosyası üzerinde yapılan manipölasyon işlemleri adımsal olarak açıklanmadan önce, analiz sürecinde kullanılacak ürün tablosunun ham hali aşağıda Şekil 5.13’ de gösterilmektedir.

	Ürün-1 Kod	Ürün-2 Kod	Ay	Birim Maliyet
1	00077	07	8.000	0.620
2	00177	47	9.000	2.559
3	00169	01	9.000	0.670
4	00136	82	8.000	4.120
5	00076	65	9.000	69.610
6	00147	04	7.000	2.720
7	00133	14	9.000	9.677
8	00137	81	2.000	3.983
9	00151	16	1.000	14.680
10	00160	58	7.000	1.960
11	00027	39	7.000	1.056
12	03257	02	5.000	2.660
13	00075	25	9.000	22.500
14	00108	20	8.000	0.790
15	00013	13	3.000	0.170
16	00157	13	9.000	1.460
17	00005	56	5.000	2.264
18	00081	61	7.000	3.270
19	00148	20	9.000	20.010
20	00020	26	1.000	0.548

Şekil 5.13.Birim maliyet dosyasına ait veri seti

Birim_Maliyet.dbo dosyası üzerinde gerçekleştirilen manipölasyon işlemlerinin akışı Şekil 5.14’ de gösterilmektedir.



Şekil 5.14.Birim maliyet dosyası için gerçekleştirilen manipölasyona ait akış

Öncelikle decimal formatta tutulan ay alanı integer formatına çevrilmiş, daha sonra Ürün-1 ve Ürün-2 kodlarının birleşiminden oluşan yeni ürün kodu ile ay alanı birleştirilmiştir. Oluşan yeni tablo Şekil 5.15’ de gösterilmektedir.

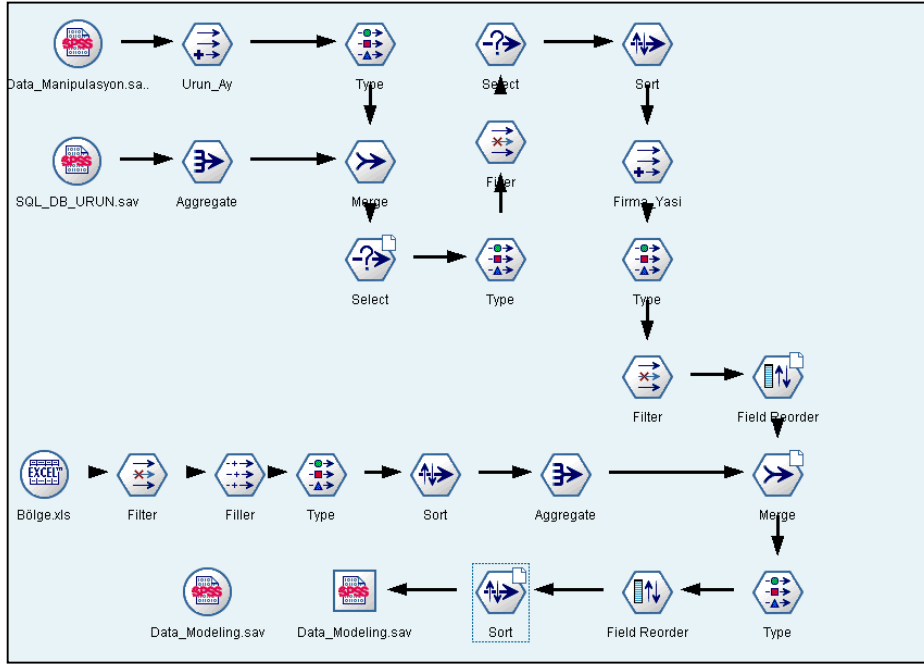
Böylece o ürüne ait ilgili aylar için tekil bir anahtar oluşturulmuştur. Oluşturulan bu yeni alan daha sonraki manipülasyon işlemlerinde ürünlere ait ay bazında birim maliyetlerini getirmek için kullanılmıştır.

En son elde edilen bilgiler bundan sonraki manipülasyon işlemlerinde kullanılmak üzere SPSS dosya formatında *SQL_DB_URUN.sav* olarak kaydedilmiştir.

	Ürün	Ay	Urun_Ay	Birim_Maliyet
1	148	1	1 1481	0.190
2	148	2	2 1482	0.184
3	148	3	3 1483	0.188
4	148	4	4 1484	0.139
5	148	5	5 1485	0.190
6	148	6	6 1486	0.187
7	148	7	7 1487	0.186
8	148	8	8 1488	0.181
9	148	9	9 1489	0.190
10	555	1	1 5551	3.152
11	555	2	2 5552	3.306
12	555	3	3 5553	3.658
13	555	4	4 5554	3.657
14	555	5	5 5555	3.659
15	555	6	6 5556	3.652
16	555	7	7 5557	3.815
17	555	8	8 5558	4.240
18	555	9	9 5559	4.306
19	556	1	1 5561	1.880
20	556	2	2 5562	1.880

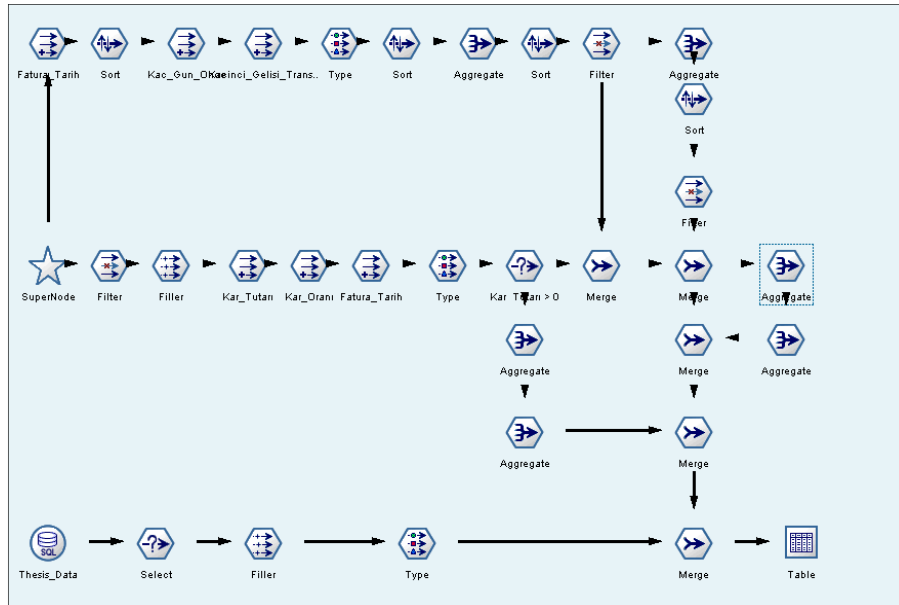
Şekil 5.15.Ürünlere ait birim maliyetlerin aylara göre elde edilmesi

Sonuç olarak, Şekil 5.16’ da gösterildiği gibi manipülasyon çalışmaları sonucunda oluşturulan yeni iki dosyanın birleştirilmesi ile bir müşteriye ait farklı yerlerde tutulan bilgiler toplanarak analiz edilebilir hale getirilmiştir.



Şekil 5.16.Manipülasyon çalışmaları sonucunda oluşturulan yeni iki dosyanın birleştirilmesi

Analiz edilebilir hale getirilen bu veriden, müşterilerin kaç defa mağazalardan alışveriş yaptığı, alışveriş yaptığı ay sayısı, müşterilere ait ortalama ürün kalem sayısı ve ürün çeşit sayısı gibi nicel bilgiler elde edilmiştir. Bu bilgilerin elde edildiği akış aşağıda Şekil 5.17 'de gösterilmiştir.



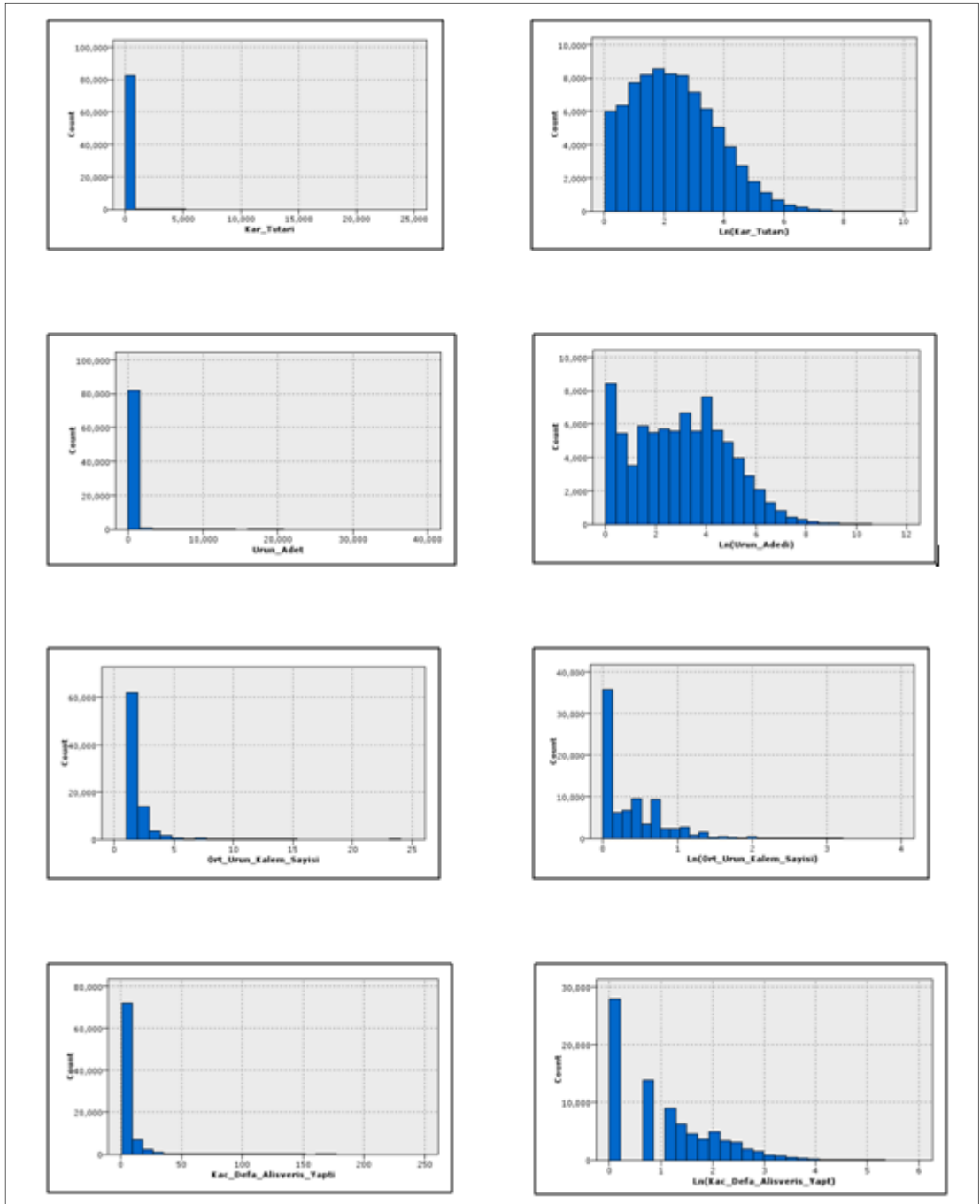
Şekil 5.17.Analiz edilebilir hale getirilen veriden çeşitli nicel bilgilerin elde edilmesi

Veri seti içerisinde, ürünlerin en küçük birim miktarlarına bakıldığında, satışı koli şeklinde yapılmış olan 58.662 adet ürün, satışı adet şeklinde yapılan 706.296 adet ürün olduğu görülmektedir. Toplam ürün içerisindeki paylarına bakıldığında; ürünlerin %7,62 si koli şeklinde, %92,33 de adet şeklinde satışı yapıldığı görülmektedir.

Analiz yorumlamalarında kolaylık sağlamak amacıyla %7,62 lik kısma gelen koli şeklinde satışı yapılan ürünler veri seti içerisinde çıkarılmış, sadece adet cinsinden satışı yapılan ürünler veri seti içerisinde bırakılmıştır. Bundan sonraki yapılacak olan modelleme çalışmalarına bu veri seti üzerinden devam edilmiştir.

5.2. MODELLEME ÖNCESİ VERİ ANALİZİ

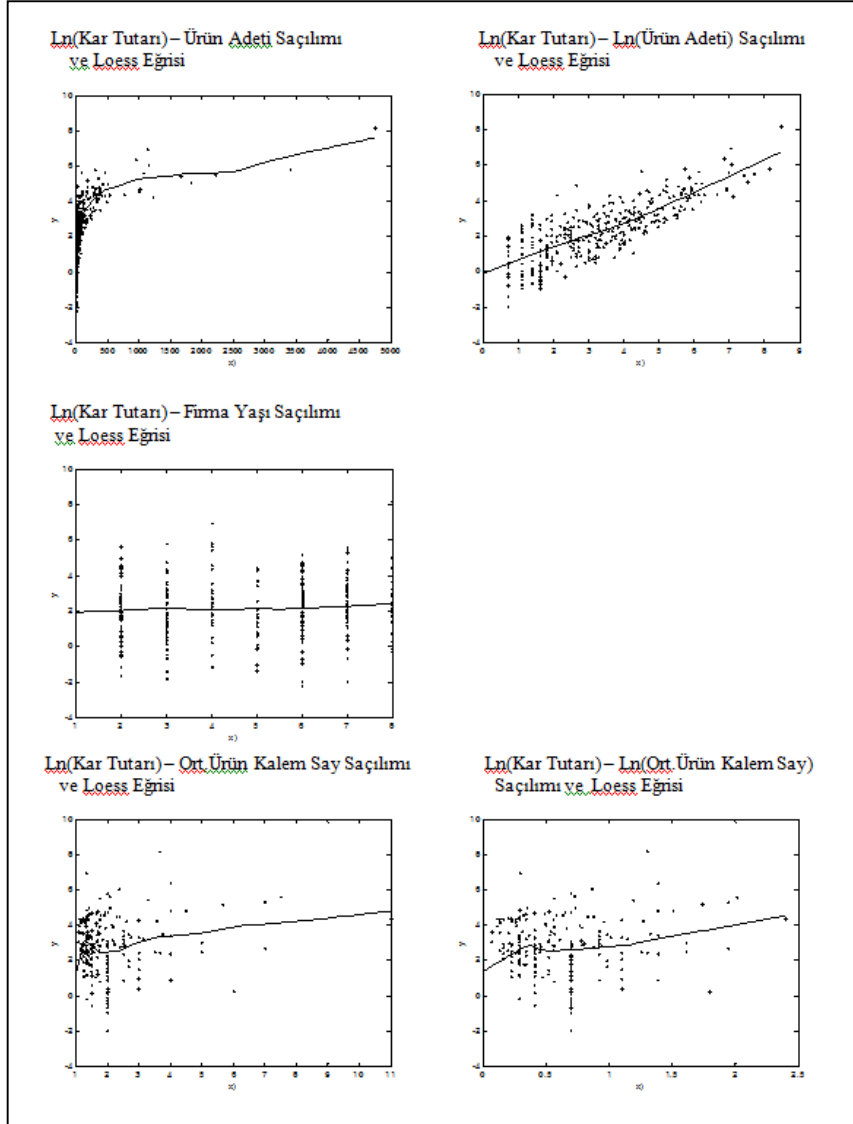
Modellemeye sokulacak nicel değişkenler incelendiğinde, aşırı çarpık dağılımlara sahip oldukları gözlemlenmiştir. Şekil 5.18’ de görüldüğü gibi bu değişkenler üzerinde logaritmik dönüşümler yapılarak, değişkenlere ait dağılımlar daha simetrik bir yapıya getirilmeye çalışılmıştır.



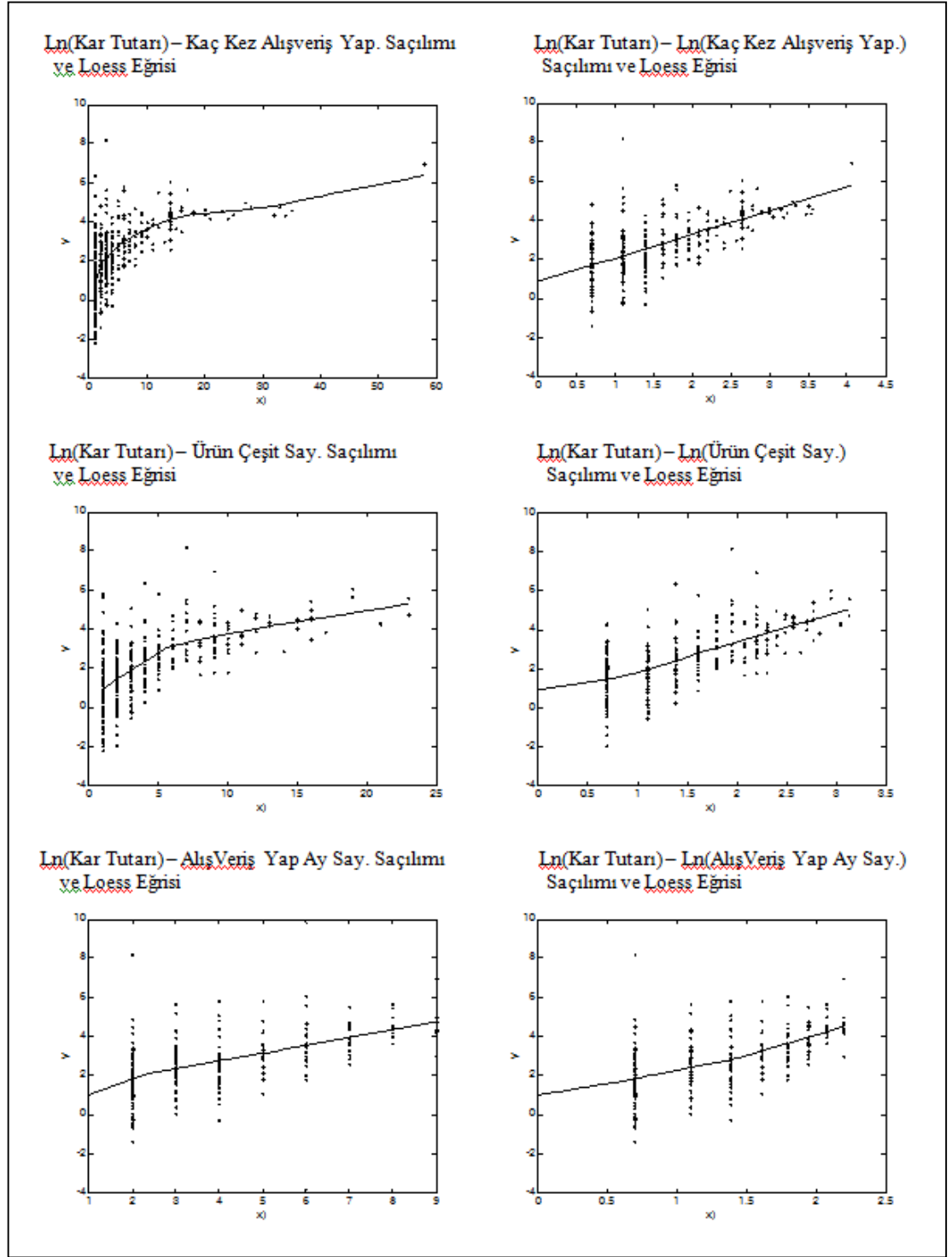
Şekil 5.18.Nicel değişkenlere ait dağılış histogramları

Veri seti üzerinden 500 lük ve 2000 lik rastgele örneklem çekilerek, bağımlı değişken olan kar tutarı ile bağımsız değişkenler arasındaki ilişkiler, logaritmik dönüşüm yapılp ve yapılmaksızın Matlab programı kullanılarak incelenmiştir. Şekil 5.19, Şekil 5.20, Şekil 5.21 ve Şekil 5.22’ de gösterildiği gibi aşağıdaki sonuçlar elde edilmiştir.

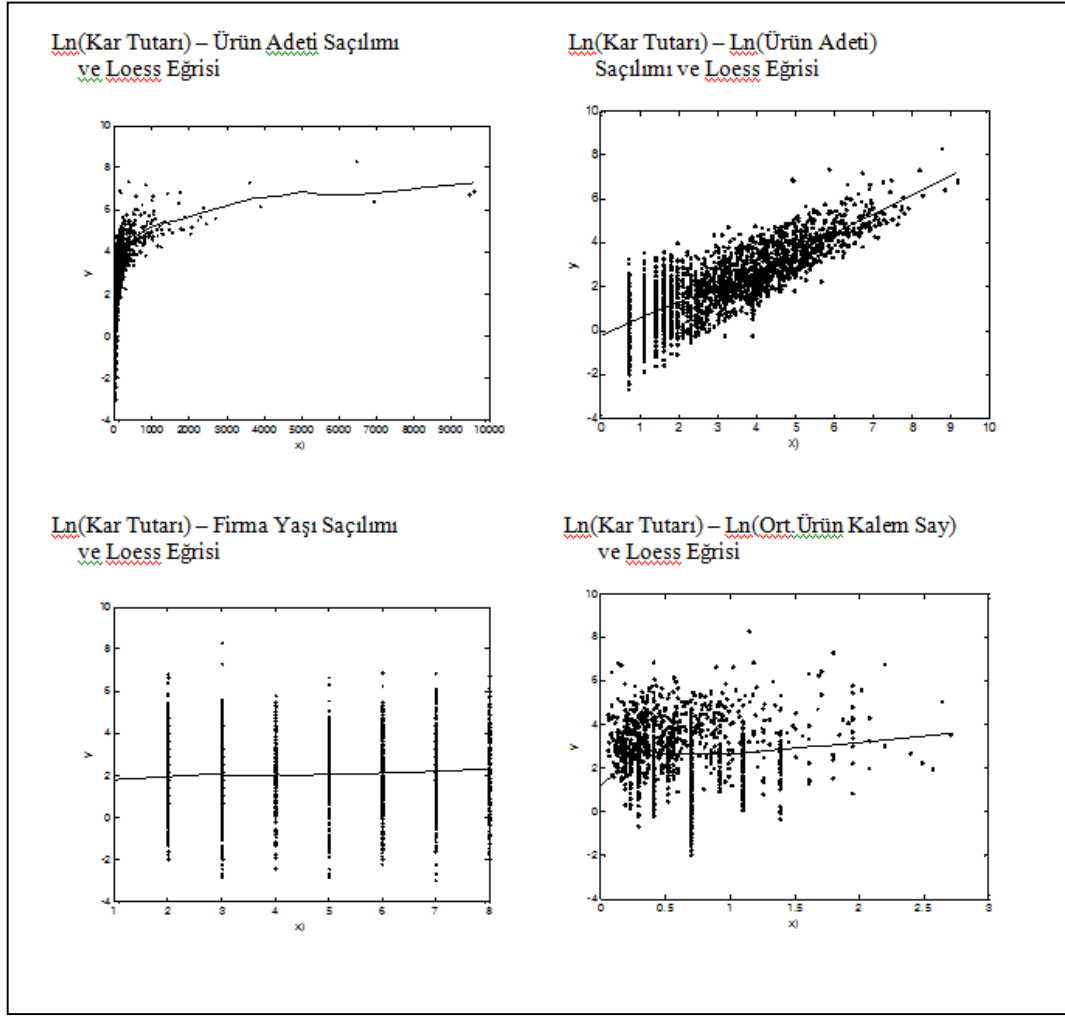
Bu sonuçlara göre, yaş değişkeninin bir etkisi yokken, ürün âdetinin(Ln) kar tutarı(Ln) ile ilişkisi oldukça iyi ve doğrusal olarak gözükmektedir. Diğer çizimlerde de logaritmik dönüşüm yapılmış değişkenler arasındaki yaklaşık doğrusal bağıntılar iyi olarak gözükmektedir. Ancak modellemeye ürün âdeti değişkeni katılmaz ise diğer değişkenler arasında değişen varyanslılık olduğu bu grafiklerden açıkça görülmektedir.



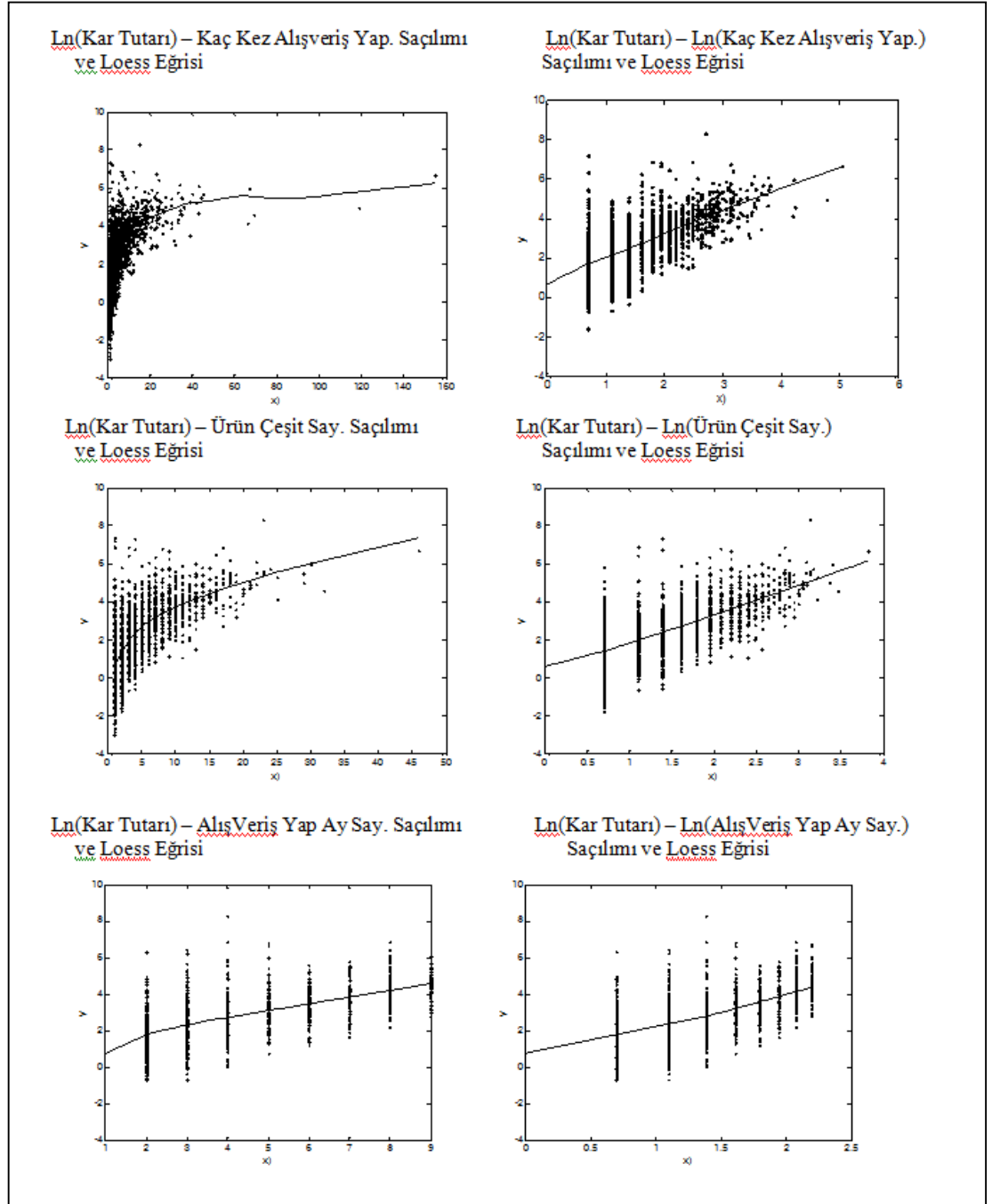
Şekil 5.19.Bağımlı ve bağımsız değişkenler arasındaki ilişkilerin incelenmesi (500 örneklem için)



Şekil 5.20.Bağımlı ve bağımsız değişkenler arasındaki ilişkilerin incelenmesi (500 örneklem için)

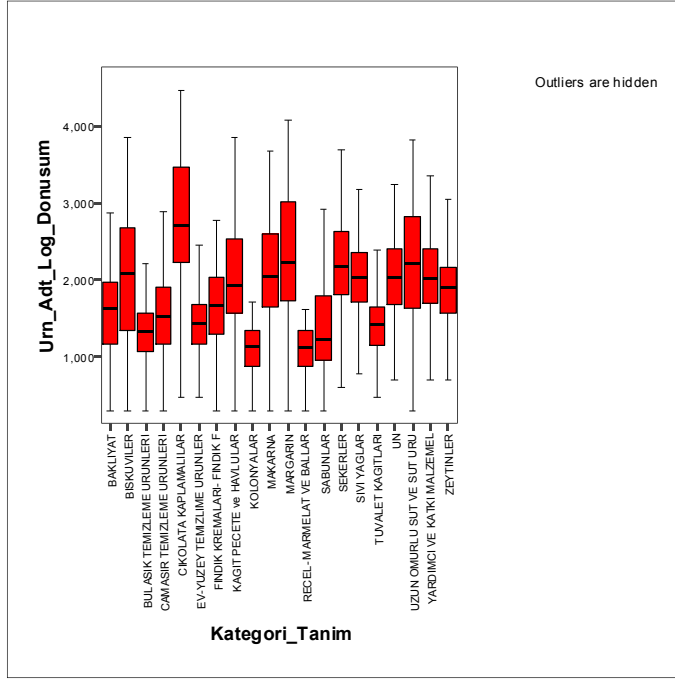


Şekil 5.21.Bağımlı ve bağımsız değişkenler arasındaki ilişkilerin incelenmesi (2000 örneklem için)



Şekil 5.22.Bağımlı ve bağımsız değişkenler arasındaki ilişkilerin incelenmesi (2000 örneklem için)

Kategorik değişken olan ürün kategorisi ve bölge tanımlamaları analize dummy değişken olarak katılması amaçlanmıştır. Bu değişkenler analize katılmadan önce boxplot grafikleri ile incelenmiş ve değişkenler üzerinden düzey / grup birleştirilmesine gidilmiştir. Ürün kategorileri, aşağıda Şekil 5.23' de gösterilen boxplot grafiği ve iş bilgisi kullanılarak birleştirilmiştir.

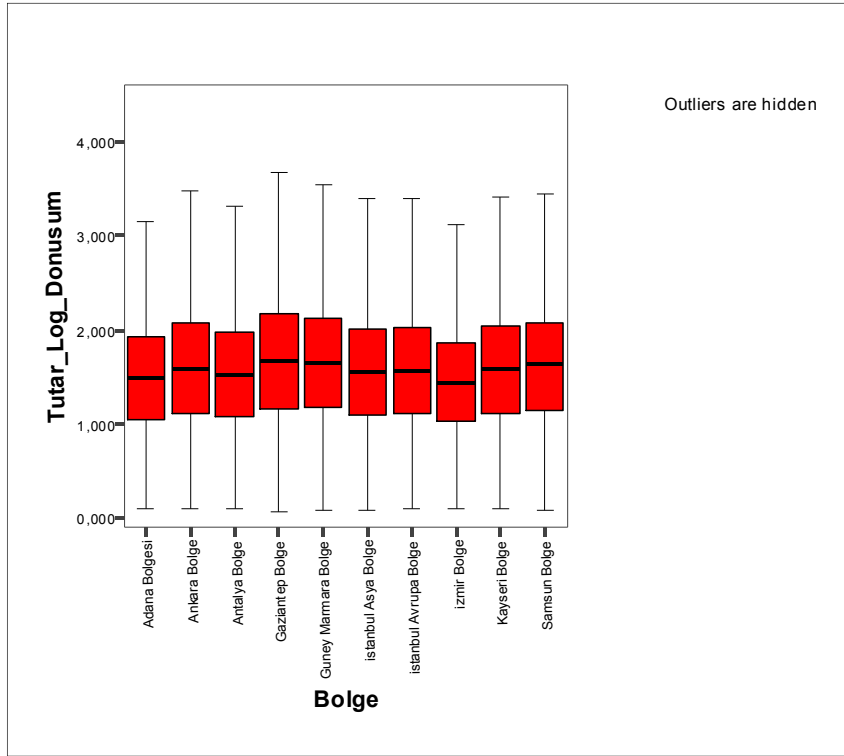


Şekil 5.23.Ürün kategorilerine ait boxplot grafiği

Düzy / grup birleştirilmesi sonucunda; ürün kategorisine ait değişkenler 7 gruba indirgenmiştir. Ürün kategorilerine ait oluşan yeni gruplar ve tanımlamaları aşağıdaki gibidir:

- 1) Bakliyatlar, sıvı yağlar
- 2) Makarna, un, yardımcı ve katkı malzemeleri
- 3) Margarin, şeker, bisküvi, uzun ömürlü süt ve süt ürünleri
- 4) Çikolata kaplamalılar
- 5) Zeytinler, fındık kremaları, reçel ve marmelat
- 6) Çamaşır, bulaşık, ev ve yüzey temizleme ürünleri
- 7) Kağıt peçete ve havlular, tuvalet kağıtları, sabunlar, kolonyalar

Aynı biçimde bölge tanımları da aşağıda Şekil 5.24' de gösterilen boxplot grafiği kullanılarak birleştirilmiştir.



Şekil 5.24.Bölgelere ait boxplot grafiği

Düzey / grup birleştirilmesi sonucunda; bölgeler 4 grupta toplanmıştır. Oluşan yeni gruplar ve tanımlamaları aşağıdaki gibidir:

- 1) Antalya, Adana, İzmir
- 2) Kayseri, Ankara
- 3) Samsun, Güney Marmara, Gaziantep
- 4) İstanbul Avrupa, İstanbul Asya

Kategorik değişkenler üzerinde düzey / grup birleştirilmesi yapıldıktan sonra oluşan yeni gruplar, faktör analizi modeline dummy değişken olarak katılmıştır. Nicel değişkenler de ise aşırı çarpık dağılımlar gözlemlendiğinden dolayı, modele katılmadan önce EDA’ da kullanılan dönüşüm yöntemleri uygulanıp ve uygulanmaksızın, modele katılıp analiz sonuçları gözlemlenmiştir.

5.3. MODELLEME

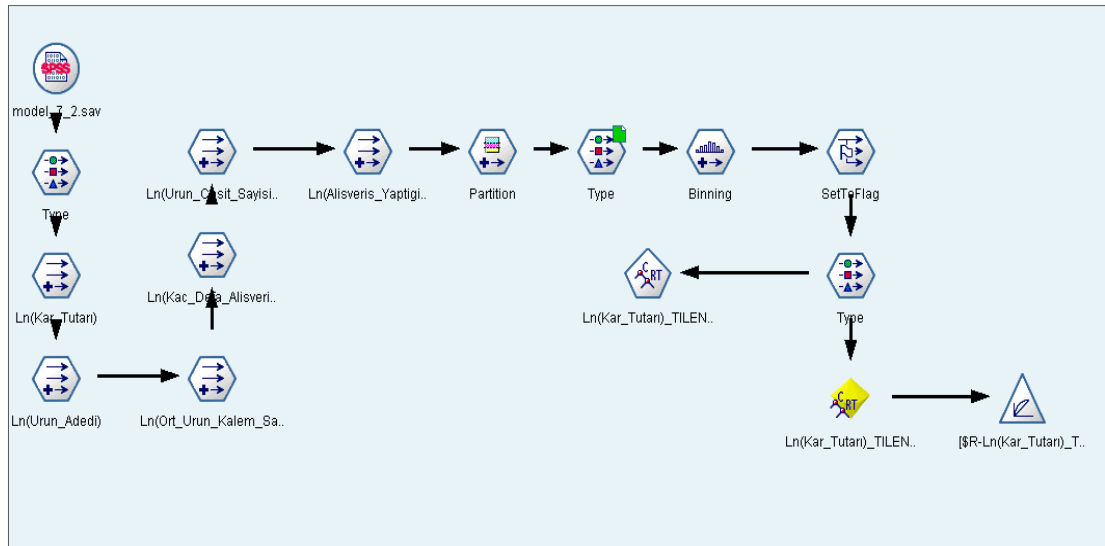
Modelleme aşamasında daha önce analize hazır hale getirilen verilerden bilgi elde edilebilmesi için kullanılacak modelleme yöntemleri ve bunların algoritmaları belirlenmiştir. Veri madenciliği sonuçlarının elde edildiği ve kararların bu aşamada elde edilen bilgiler üzerine verildiği düşünüldüğünde bu aşama oldukça önemlidir.

5.3.1.Karar Ağaçları Yöntemleri İle Sınıflandırma Teknikleri

Bu bölümde, müşteriler kar tutarı durumlarına göre üç farklı grup altında toplanmıştır (düşük kar tutarlı, orta kar tutarlı, yüksek kar tutarlı). Bu çalışmada EDA teknikleri uygulanıp/uygulanmaksızın, karar ağaçları algoritmalarından C&R Tree ve CHAID algoritmaları kullanılarak “kar tutarı yüksek” müşterilerin bu grupta toplanmasında hangi faktörlerin etkili olduğu belirlenmeye çalışılmıştır.

5.3.1.1.EDA Kullanılarak C&R Tree Yöntemi İle Modelleme

EDA sonuçları kullanılarak nicel değişkenler üzerinde uygun logaritmik dönüşümler yapılarak modellemeye girilmiştir. C&R Tree kullanılarak oluşturulan modelleme sürecine ait akış Şekil 5.25’ de gösterilmektedir.



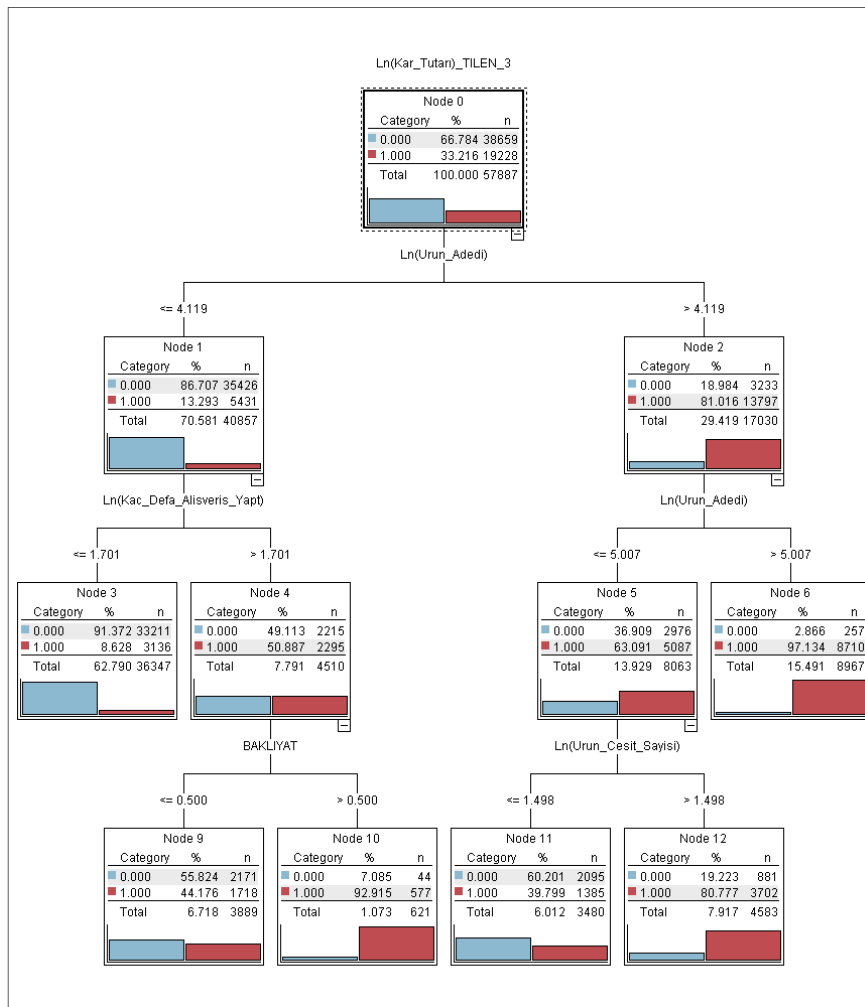
Şekil 5.25. C&R Tree Yöntemi İle Modellemeye ilişkin akış

C&R Tree algoritması sonucunda oluşan karar ağacı Şekil 5.26’ da gösterildiği gibidir. Ağaç yapısı incelendiğinde, kar tutarı yüksek olan müşterilerin belirlenmesinde en önemli faktörün “ürün adedi” olduğu gözükmemektedir.

Buna göre:

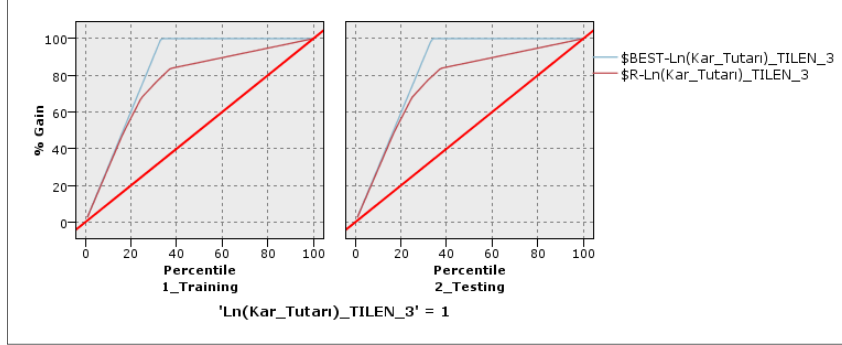
- Ürün adet alımı 61' in üzerindeki müşterilerin kar tutarları daha yüksek olduğu gözlemlenmektedir.
- Bu müşterilerin içerisinde ürün adet alımı 149' un üzerinde olan müşteriler kar tutarı yüksek olan müşteri grubuna girmektedirler.
- Ürün adet alımı 61 ve 61' in altındaki müşterilere bakıldığında ise yüksek kar tutarını belirlemede en önemli faktörün "ürün çeşit sayısı" olduğu gözlemlenmektedir. Bu müşteriler içerisinde ürün çeşit sayısı 4' ün üzerinde olan müşteriler ise yüksek kar tutarlı gruba girmektedir.

Buradaki açıklamada belirtilen değerler, ağaç yapısında verilen sayısal değerlerin anti logaritması alınarak verilmiştir.



Şekil 5.26.EDA kullanılarak oluşturulan C&R Tree algoritmasına ait karar ağacı

Training ve Testing olarak ayrılmış olan veri seti için model performansları incelendiğinde, model performansları hemen hemen birbirine benzerdir ve Şekil 5.27’ de gösterildiği gibi model performansları oldukça iyi olarak gözükmemektedir.



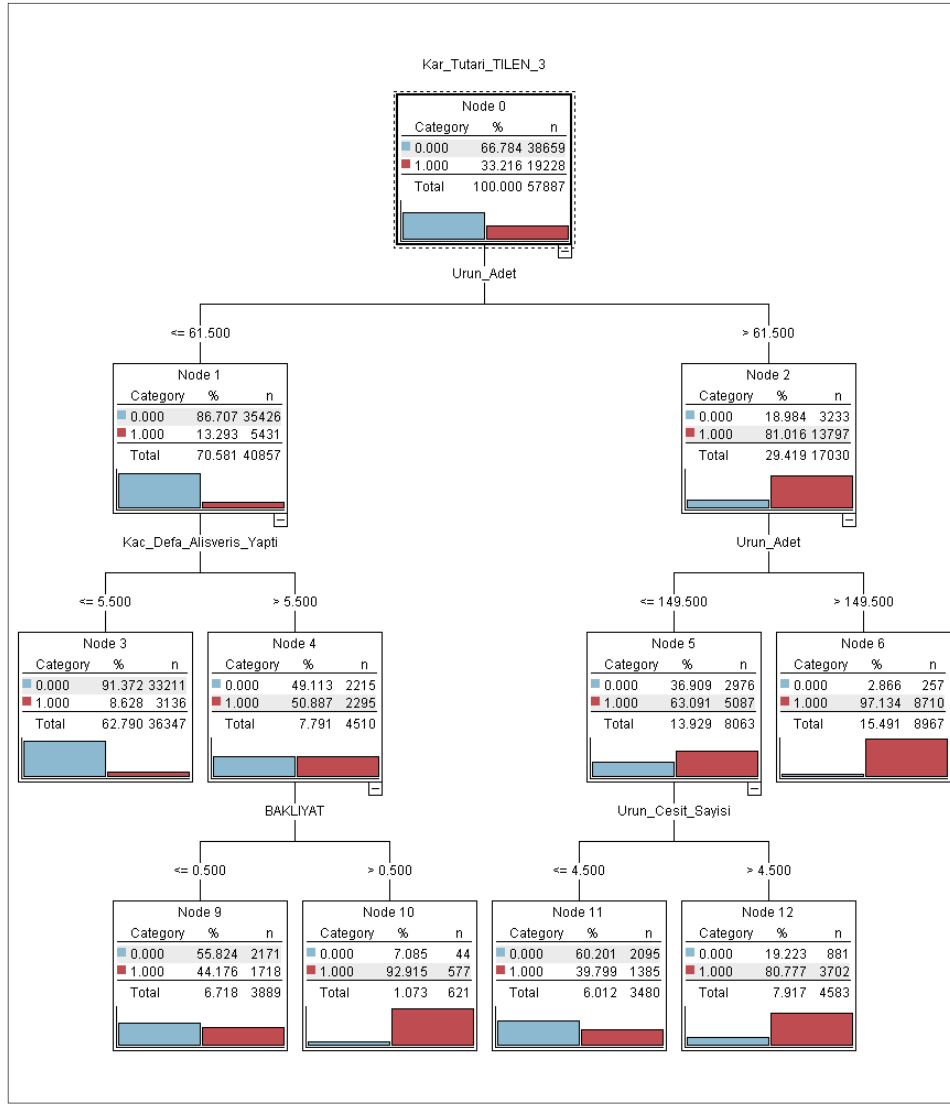
Şekil 5.27 EDA kullanılarak oluşturulan C&R Tree algoritmasına ait Evaluation grafiği

5.3.1.2.EDA Kullanılmadan C&R Tree Yöntemi İle Modelleme

EDA sonuçları kullanılmadan C&R Tree yöntemi ile modellemeye tekrar girilmiştir. Burada modellemede kullanılacak nicel değişkenler üzerinde logaritmik dönüşümler yapılmamış ve model sonucu tekrardan incelenmiştir.

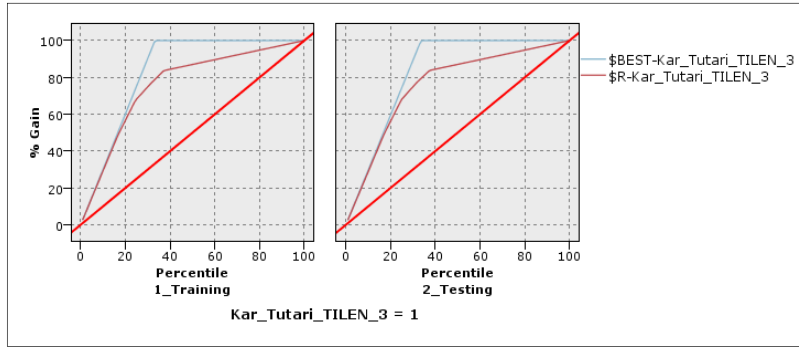
Buna göre Şekil 5.28’ de gösterilen ağaç yapısı tekrar incelendiğinde:

- Ürün adet alımı 61’ in üzerindeki müşterilerin kar tutarları daha yüksek olduğu gözlemlenmektedir.
- Bu müşterilerin içerisinde ürün adet alımı 149’ un üzerinde olan müşteriler kar tutarı yüksek olan müşteri grubuna girmektedirler.
- Ürün adet alımı 61 ve 61’ in altındaki müşterilere bakıldığında ise yüksek kar tutarını belirlemede en önemli faktörün “ürün çeşit sayısı” olduğu gözlemlenmektedir. Bu müşteriler içerisinde ürün çeşit sayısı 4’ ün üzerinde olan müşterilerin ise yüksek kar tutarlı guruba girdikleri gözlemlenmektedir.



Şekil 5.28.EDA kullanılmadan oluşturulan C&R Tree algoritmasına ait karar ağacı

Training ve Testing olarak ayrılmış olan veri seti için model performansları incelendiğinde, model performansları hemen hemen birbirine benzerdir ve Şekil 5.29’ da gösterildiği gibi model performansları oldukça iyi olarak gözükmetedir.



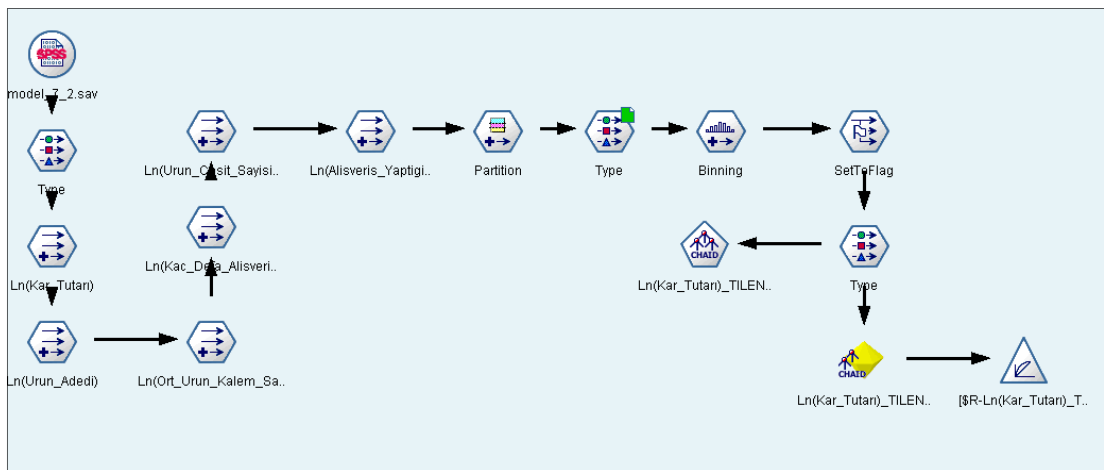
Şekil 5.29.EDA kullanılmadan oluşturulan C&R Tree algoritmasına ait Evaluation grafiği

Bu veriler için, dönüşüm yapmanın, modelleme tekniklerinden olan C&R Tree algoritması üzerinde etkili olmadığı görülmüştür. C&R Tree, verileri oldukça homojen alt guruplara indirgediğinden, verilerin yapısının ve değişkenler arasındaki ilişki yapısının C&R Tree gibi yöntemlerde etkili olmayacağı görülmektedir.

5.3.1.3.EDA Kullanılarak CHAID Yöntemi İle Modelleme

EDA yöntemleri kullanılarak nicel değişkenler üzerinde uygun logaritmik dönüşümler yapılarak modellemeye girilmiştir.

CHAID kullanarak oluşturulan modelleme sürecine ait akış Şekil 5.30’ da gösterilmektedir.

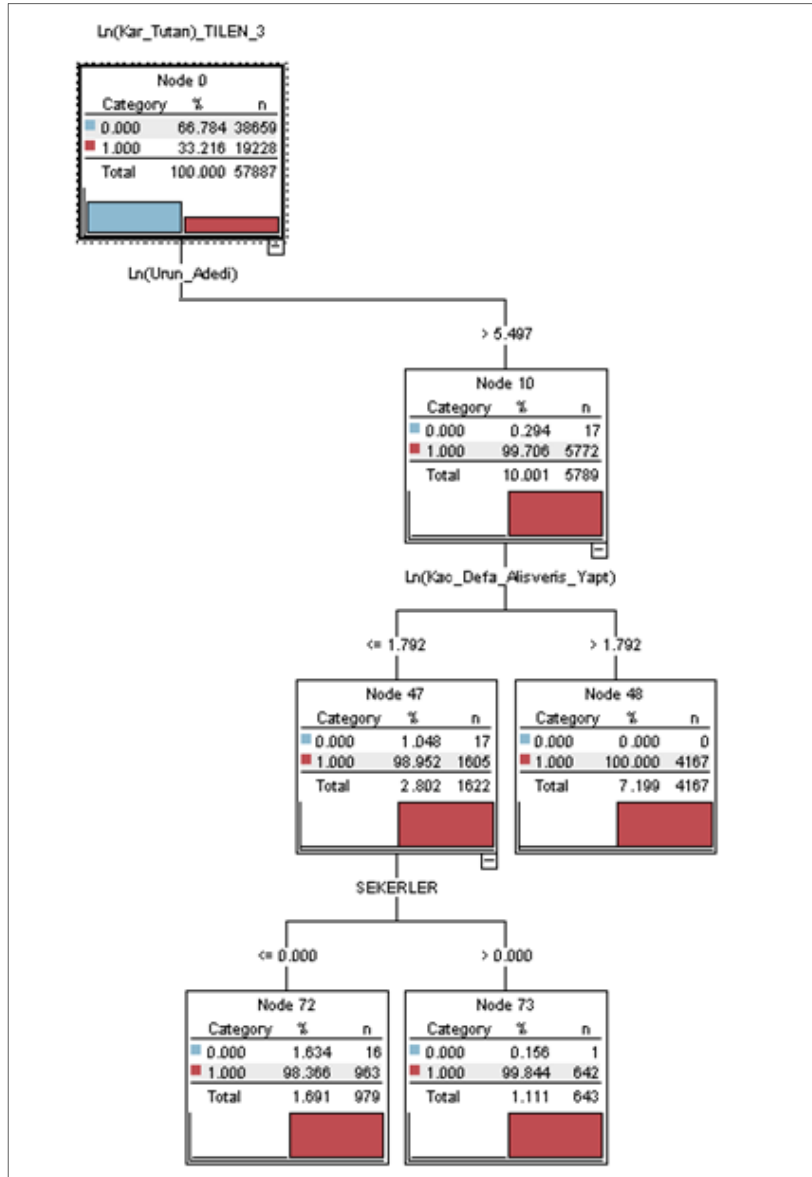


Şekil 5.30 CHAID yöntemi ile modellemeye ilişkin akış

Şekil 5.31’ de gösterilen ağaç yapısı incelendiğinde; kar tutarı yüksek olan müşterilerin belirlenmesinde en önemli faktörün “ürün adedi”, ikinci önemli faktörün ise müşterinin 9 ay boyunca kaç defa alışveriş yaptığı bilgisinin olduğu gözükmemektedir.

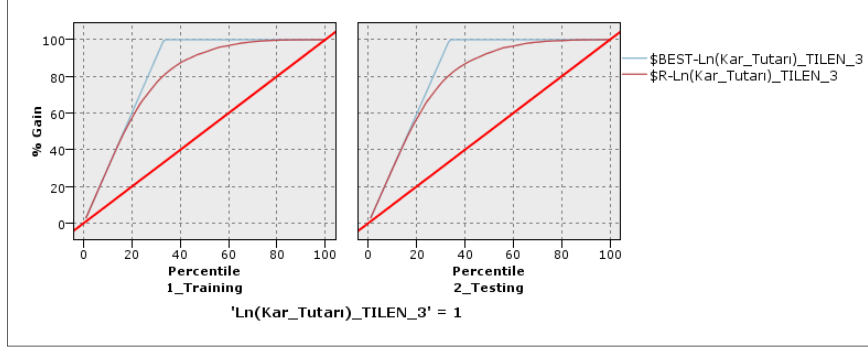
Dokuz ay boyunca toplam alınan ürün adedi 244’ ün üzerindeki müşterilerin kar tutarları daha yüksek olduğu gözlemlenmektedir.

Buradaki açıklamada belirtilen değerler, ağaç yapısında verilen sayısal değerlerin anti logaritması alınarak verilmiştir.



Şekil 5.31.EDA kullanılarak oluşturulan CHAID algoritmasına ait karar ağacı

Training ve Testing olarak ayrılmış olan veri seti için model performansları incelendiğinde, model performansları hemen hemen birbirine benzerdir ve Şekil 5.32’ de gösterildiği gibi model performansları oldukça iyi olarak gözükmemektedir.



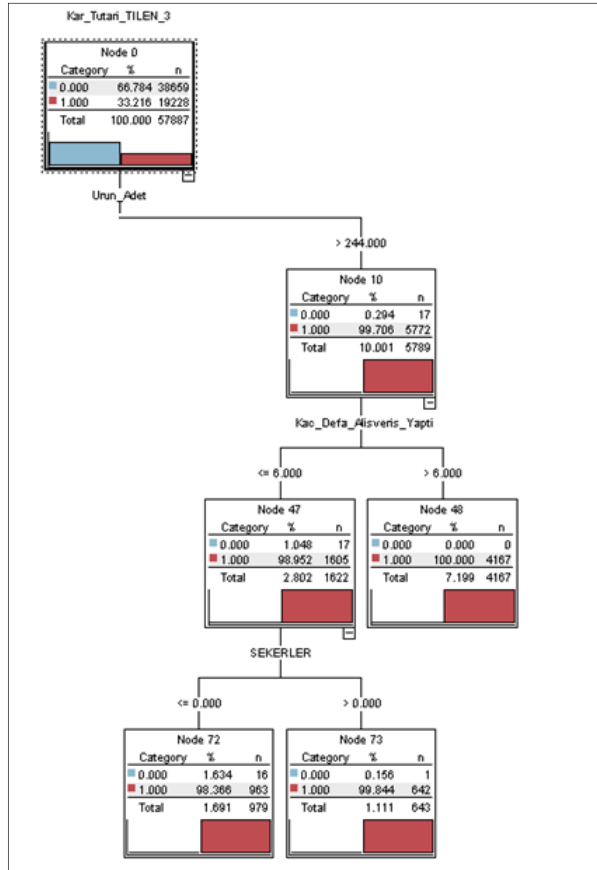
Şekil 5.32.EDA kullanılarak oluşturulan CHAID algoritmasına ait Evaluation grafiği

5.3.1.4.EDA Kullanılmadan CHAID Yöntemi İle Modelleme

EDA kullanılmadan CHAID yöntemi ile modellemeye tekrar girilmiştir. Burada modellemede kullanılacak nicel değişkenler üzerinde logaritmik dönüşümler yapılmamış ve model sonucu tekrardan incelenmiştir.

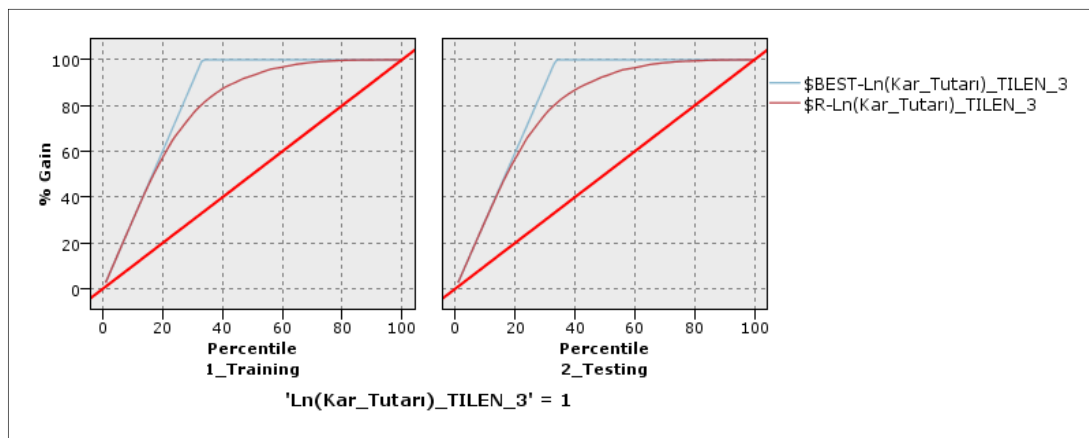
Buna göre Şekil 5.33’ de gösterilen ağaç yapısı tekrar incelendiğinde; kar tutarı yüksek olan müşterilerin belirlenmesinde en önemli faktörün “ürün adedi”, ikinci önemli faktörün ise müşterinin 9 ay boyunca kaç defa alışveriş yaptığı bilgisinin olduğu gözükmemektedir.

Dokuz ay boyunca toplam alınan ürün adedi 244’ ün üzerindeki müşterilerin kar tutarları daha yüksek olduğu gözlemlenmektedir.



Şekil 5.33.EDA kullanılmadan oluşturulan CHAID algoritmasına ait karar ağacı

Training ve Testing olarak ayrılmış olan veri seti için model performansları incelendiğinde, model performansları hemen hemen birbirine benzerdir ve Şekil 5.34’ de gösterildiği gibi model performansları oldukça iyi olmaktadır.



Şekil 5.34.EDA kullanılmadan oluşturulan CHAID algoritmasına ait Evaluation grafiği

Sonuç olarak, kar tutarı yüksek olan müşteri gruplarının tespit edilmesinde, sınıflama tekniklerinden biri olan Karar Ağaçları algoritmalarından, C&RT ve CHAID yöntemleri kullanılmıştır. Her iki yöntem ile de EDA yöntemleri uygulanıp ve uygulanmaksızın modele girilip sonuçlar gözlemlenmiştir. EDA yöntemlerinin karar ağaçları ile uygulanan sınıflama tekniklerinde bir etkisinin olmadığı gözlemlenmiştir.

5.3.2. Müşteri Segmantasyonu

Bu bölümde müşteri segmentleri belirlenmeye çalışılmıştır. Çok fazla girdi değişken olduğundan dolayı faktör analizi kullanılarak, segmantasyon çalışmasında modelleme için kullanılabilecek çok fazla sayıda alan indirgenerek ilgili değişkenlerden ortak faktörler oluşturulmuştur.

Faktör analizi sonucunda, faktör skorları elde edilmiş ve sonrasında yeni oluşturulan bu skor değişkenleri segmantasyon çalışması için girdi değişkeni olarak kullanılmıştır.

Kategorik değişkenler üzerinde düzey / grup birleştirilmesi yapıldıktan sonra oluşan yeni gruplar, faktör analizi modeline dummy değişkeni olarak katılmıştır. Nicel değişkenler de ise aşırı çarpık dağılımlar gözlemlendiğinden dolayı, modele katılmadan önce EDA’ da kullanılan dönüşüm yöntemleri uygulanıp ve uygulanmaksızın, modele katılıp analiz sonuçları gözlemlenmiştir.

Buna göre; EDA kullanılmadan değişkenler faktör modeline katıldığında, aşağıda Şekil 5.35’ de görüldüğü gibi 6 faktör oluşmuştur ve bu faktörlerin tümü toplam varyansın %61,77’ sini açıklamaktadır.

Total Variance Explained									
Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2,964	18,526	18,526	2,964	18,526	18,526	2,298	14,363	14,363
2	1,857	11,603	30,129	1,857	11,603	30,129	2,067	12,921	27,284
3	1,420	8,877	39,006	1,420	8,877	39,006	1,497	9,354	36,638
4	1,351	8,441	47,447	1,351	8,441	47,447	1,405	8,778	45,416
5	1,246	7,786	55,232	1,246	7,786	55,232	1,350	8,438	53,854
6	1,046	6,538	61,771	1,046	6,538	61,771	1,267	7,916	61,771
7	,984	6,152	67,923						
8	,978	6,110	74,033						
9	,912	5,701	79,734						
10	,805	5,033	84,767						
11	,784	4,901	89,668						
12	,744	4,649	94,318						
13	,524	3,273	97,590						
14	,195	1,220	98,810						
15	,177	1,109	99,919						
16	,013	,081	100,000						
Extraction Method: Principal Component Analysis.									

Şekil 5.35.EDA kullanılmadan oluşturulan, Özdeğer İstatistiğine bağlı faktör sayısı ve açıklanan varyans yüzdesi

EDA kullanılarak değişkenler faktör modeline katıldığında ise, aşağıda Şekil 5.36’ da görüldüğü gibi 6 faktör oluşmuştur ve bu faktörlerin tümü toplam varyansın %66,43’ nü açıklamaktadır.

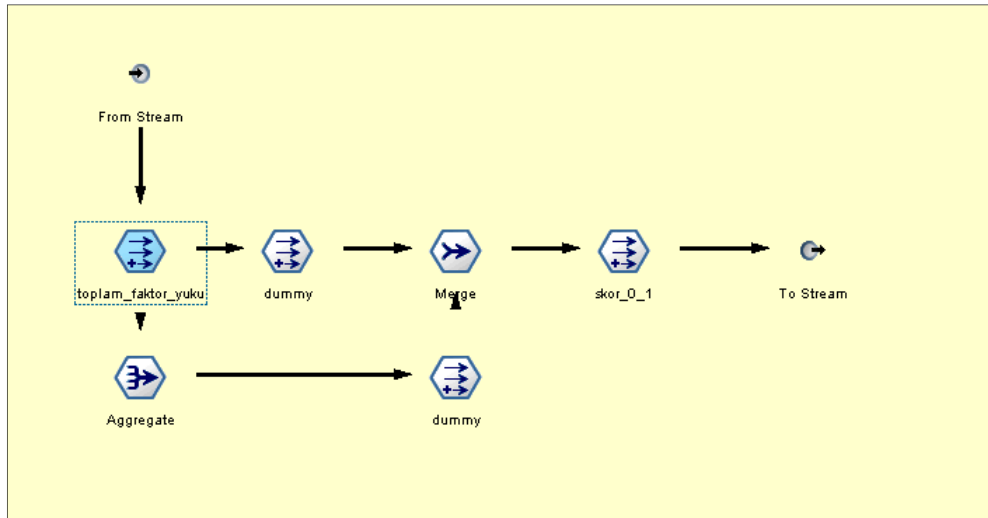
Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4,321	27,006	27,006	4,321	27,006	27,006	2,806	17,539	17,539
2	1,425	8,907	35,914	1,425	8,907	35,914	2,424	15,147	32,686
3	1,363	8,519	44,433	1,363	8,519	44,433	1,405	8,784	41,470
4	1,259	7,867	52,300	1,259	7,867	52,300	1,384	8,648	50,118
5	1,211	7,566	59,866	1,211	7,566	59,866	1,341	8,381	58,499
6	1,049	6,557	66,423	1,049	6,557	66,423	1,268	7,924	66,423
7	,980	6,125	72,548						
8	,959	5,993	78,541						
9	,915	5,720	84,261						
10	,807	5,046	89,307						
11	,786	4,913	94,220						
12	,741	4,633	98,853						
13	,108	,677	99,530						
14	,051	,317	99,846						
15	,013	,080	99,927						
16	,012	,073	100,000						
Extraction Method: Principal Component Analysis.									
Component Matrix(a)									

Şekil 5.36.EDA kullanılarak oluşturulan, Özdeğer İstatistiğine bağlı faktör sayısı ve açıklanan varyans yüzdesi

EDA kullanılarak oluřturan faktörlerin, toplam varyansı daha yüksek oranda açıkladıđı gözlemlendiđinden dolayı sürece nicel deđiřkenler üzerinde gerekli logaritmik dönüřümler uygulanarak devam edilmiřtir. O nedenle, kötü sonuç veren EDA' sız sonuçlar burada verilmemiřtir.

řekil 5.37' de gösterilen akıřta; faktör analizi sonucunda oluřan faktör yükleri toplanarak toplam faktör yükü hesaplanmıřtır. Toplam faktör yükünün, en küçük ve en büyük deđerleri tespit edilerek ařađıdaki belirtilen formül kullanılarak faktör skorları elde edilmiřtir.

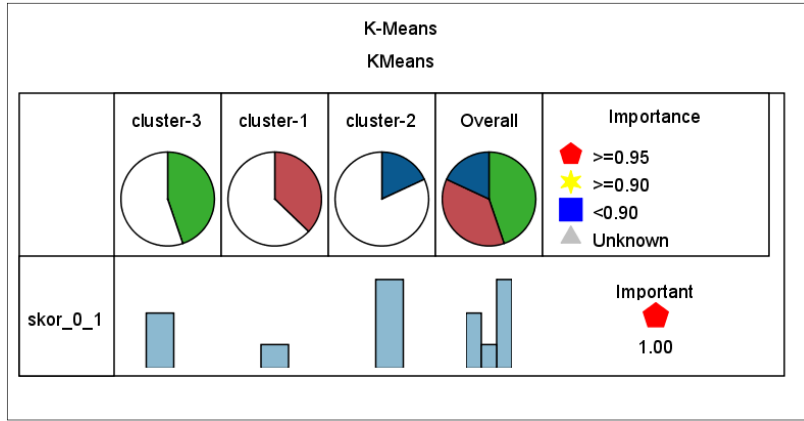
$$\text{Faktör Skor} = (\text{Toplam Faktör Yüğü} - \text{En Küçük Faktör Yüğü}) / (\text{En Büyük Faktör Yüğü} - \text{En Küçük Faktör Yüğü}) \quad (5.1)$$



řekil 5.37.Faktör skorlarının elde edildiđi akıř

Elde edilen faktör skorları, K-Means algoritmasında girdi deđiřken olarak kullanılarak veri seti üzerinden 3 küme elde edilmiřtir (Kırođlu vd, 2008).

Ařađıda řekil 5.38' de oluřturulan kümelerin boyutlarına ait ayrıntılı bilgi gösterilmektedir.



Şekil 5.38.K-Means yöntemi sonucunda oluşan kümeler

Şekil 5.39’ da gösterildiği gibi, elde edilen kümeler skor değerlerine göre yeniden sınıflandırılmıştır.

Original value	New value
cluster-1	3
cluster-2	1
cluster-3	2

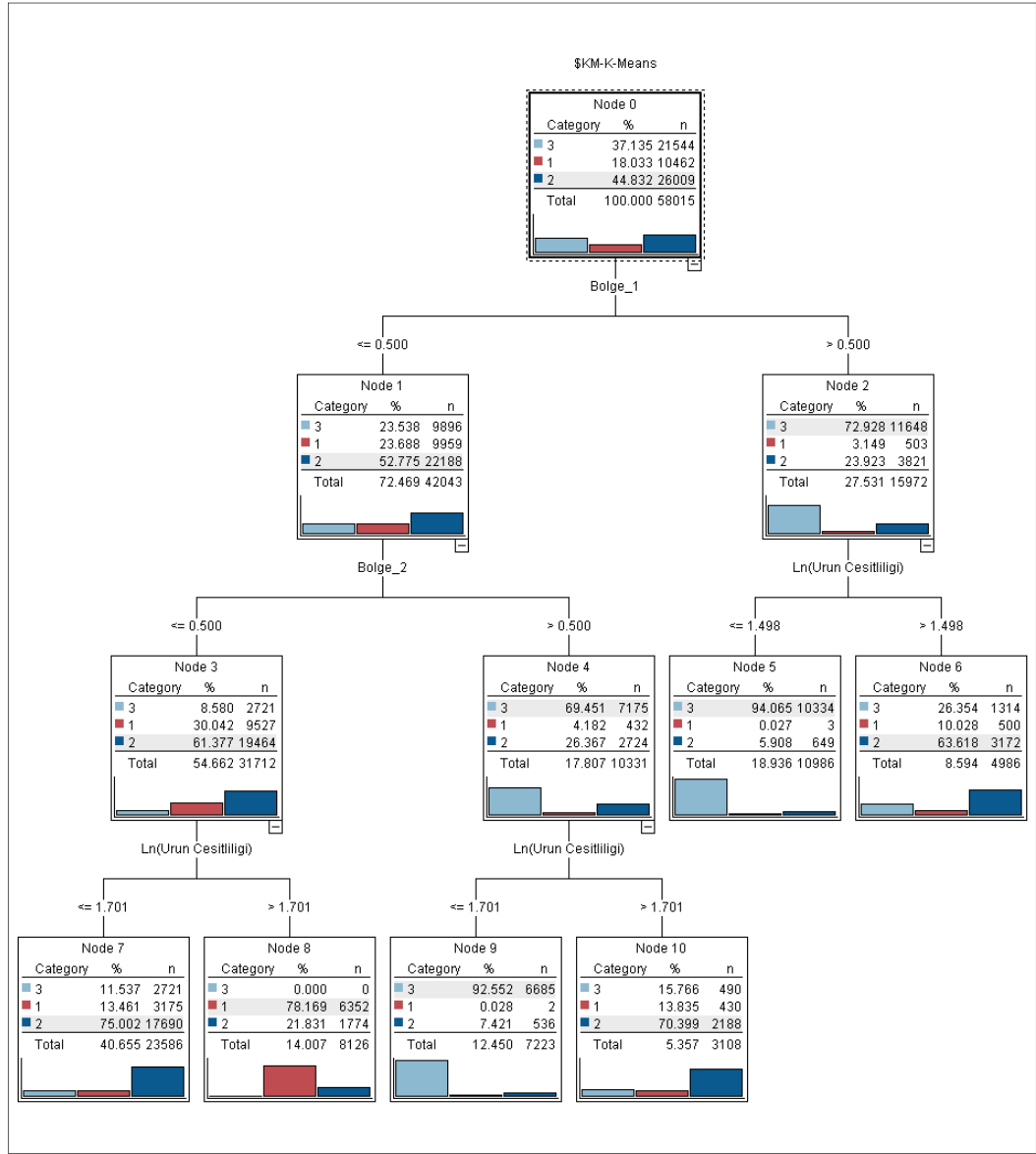
Şekil 5.39.Kümelerin Skor değerlerine göre yeniden sınıflandırılması

Karar ağacı sınıflandırma yöntemlerinden C&R Tree algoritması kullanılarak oluşturulan yeni kümeler için kurallar elde edilmiştir.

Buna göre, Şekil 5.40’ da gösterilen ağaç yapısı incelendiğinde, aşağıda belirtilen kurallar bulunmuştur:

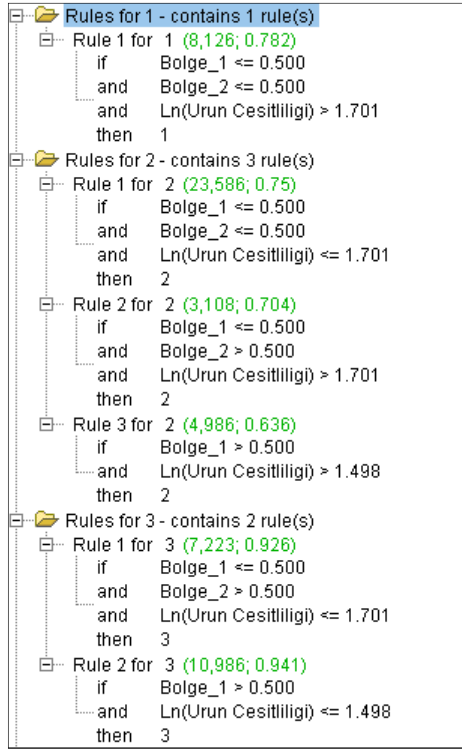
- Hem “Bölge_1” hem de “Bölge_2” den alışveriş yapanların, ürün çeşit sayısı 5 den büyük ise bu müşteriler skoru en yüksek olarak hesaplanmış olan birinciye kümeye girmektedirler.
- Hem “Bölge_1” hem de “Bölge_2” den alışveriş yapanların, ürün çeşit sayısı 5 ve 5 den küçük ise bu müşteriler skoru ikinci yüksek olarak hesaplanmış olan ikinci kümeye girmektedirler.
- “Bölge_1” den alışveriş yapmayan ve ürün çeşit sayısı 4 ve 4 den küçük olan müşteriler ise skorlaması en düşük olarak hesaplanan üçüncü kümeye girmektedirler.

Buradaki açıklamada belirtilen değerler, ağaç yapısında verilen sayısal değerlerin anti logaritması alınarak verilmiştir.



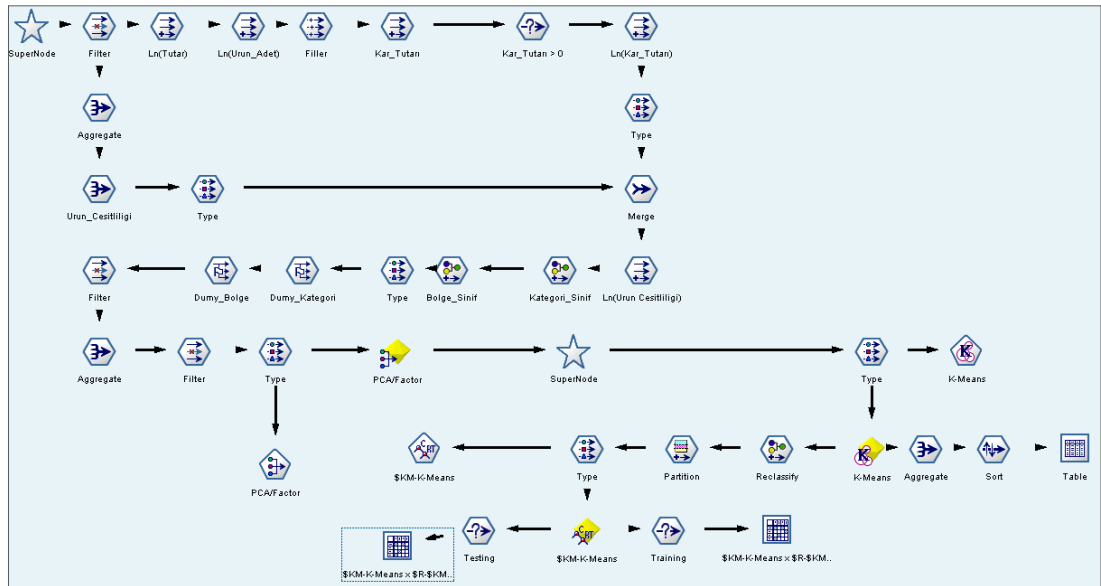
Şekil 5.40. Kümeler için oluşturulan C&R Tree algoritmasına ait ağaç yapısı

Şekil 5.41' de, C&R Tree algoritmasının sonuçları her hedef kategorisi için kural kümeleri olarak gösterilmektedir.



Şekil 5.41.Hedef kategorileri için kural kümeleri

Müşteri segmentasyon çalışmasına ait akış diyagramı Şekil 5.42’ de gösterilmektedir.



Şekil 5.42.Müşteri segmentasyon çalışmasına ait akış

5.3.3. Birliktelik Kuralları ve Sepet Analizi ile Müşteri Satın Alma Davranışı Modellemesi

Birliktelik kuralları veri madenciliğinin en iyi örneklerinden biridir. Birliktelik kuralları verideki potansiyel ilişkileri tanımlar. Birliktelik kuralları analizi ile büyük veri tabanlarından, müşterilerin alışveriş davranışları keşfedilmeye çalışılır. Birliktelik kuralları bir ürün satın alındığında, o ürünle beraber başka hangi ürünlerin satın alındığının belirlenmesi amacıyla kullanılır.

Birliktelik kurallarının kullanıldığı en tipik örnek “Sepet Analizi”dir. Sepet analizi ile müşterilerin yaptıkları alışverişlerdeki ürünle arasındaki birliktelikler bulunarak, müşterilerin satın alma alışkanlıkları belirlenir. Bu tip birlikteliklerin belirlenmesi, müşterilerin hangi ürünleri bir arada aldıkları bilgisini ortaya çıkarır.

Veri setindeki ürünler için satın alındı/alınmadı bilgisinin oluşturulması ile sepet analizinde ürünler arasındaki bağıntı, destek ve güven kriterleri aracılığı ile hesaplanır.

Destek: $P(X \text{ ve } Y) : X \text{ ve } Y \text{ mallarını satın almış müşteri sayısı} / \text{Toplam müşteri sayısı}$ (5.2)

X, Y: Ürünlerini içeren hareket sayısı

Güven: $P(X/Y)=P(X \text{ ve } Y) / P(Y)$ (5.3)

$P(X \text{ ve } Y) : X \text{ ve } Y \text{ mallarını satın almış müşteri sayısı}$

$P(Y)$: Y malını satın almış müşteri sayısı (X ürünün bulunduğu yerde X ve Y ürün setinin bulunma olasılığı)

Destek kriteri veride bu bağıntının ne kadar sık olduğunu, güven kriteri ise Y malını satın almış bir kişinin hangi olasılıkla X malını alacağını belirtmektedir.

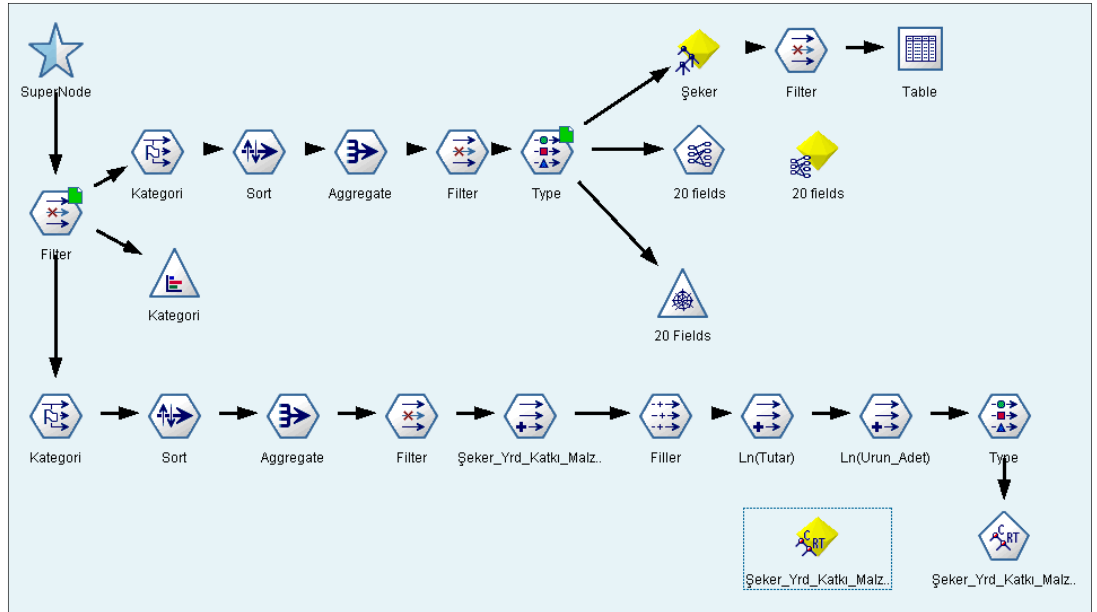
Birliktelik kuralları analizinin yapılabilmesi için verinin kategori bazında 0-1 olarak kodlanması gerekmektedir. Müşterilerin satın aldıkları ürünler için “1”, satın

almadıkları ürünler için “0” kullanılmıştır. Bu uygulamanın gerçekleştirilebilmesi için veri matrisinin devriği alınarak bütün kategoriler ikili alanlara dönüştürülmüştür. Şekil 5.43’ deki, bu ikili alanlar müşterinin ziyaretinde o ürünü satın alıp almadığını göstermektedir.

	ID	SIVI YAĞLAR	MARGARİN	MAKARNA	UN	SABUNLAR	REÇEL- MARMELAT VE BALLAR	TUVALET KAĞITLARI	YARDIMCI VE KATKI MALZEMELERİ
1	37	0	1	0	0	0	0	0	0
2	57	1	0	1	0	0	0	0	1
3	77	0	1	0	0	0	0	0	1
4	83	0	0	0	0	1	0	0	0
5	86	0	0	0	1	0	0	0	1
6	97	0	0	0	0	0	0	0	0
7	100	0	0	0	0	0	0	0	0
8	112	0	0	0	0	0	1	1	0
9	122	0	1	0	0	0	0	0	0
10	123	0	1	0	0	0	0	0	0
11	128	0	0	0	0	0	0	1	0
12	131	0	0	0	0	0	0	0	0
13	147	0	0	0	0	0	0	0	0
14	148	0	0	0	0	0	0	0	1
15	160	0	0	0	1	0	0	0	0
16	163	0	0	1	0	0	0	0	0
17	172	1	0	1	0	0	1	1	0
18	175	0	1	0	0	0	0	0	0
19	199	0	0	0	0	0	0	0	1
20	206	0	0	0	0	0	0	0	0
21	213	1	1	0	0	1	0	0	1
22	216	0	1	1	1	0	0	0	0
23	224	1	0	0	0	0	0	0	1
24	253	1	0	0	0	0	0	0	1
25	256	0	1	1	1	0	1	1	0
26	265	1	0	1	1	0	0	0	0

Şekil 5.43.Birliktelik kuralları analizi için verinin kodlanması

Analize konu değişkenlerin kural çıkarımı için hazırlanması gerekmektedir. Bunun için SPSS Clementine programında gerekli düzenlemeler yapılmış ve düzenlemelere ait akış aşağıda Şekil 5.44’ de gösterilmektedir.



Şekil 5.44.Birliktelik analizine ait akış

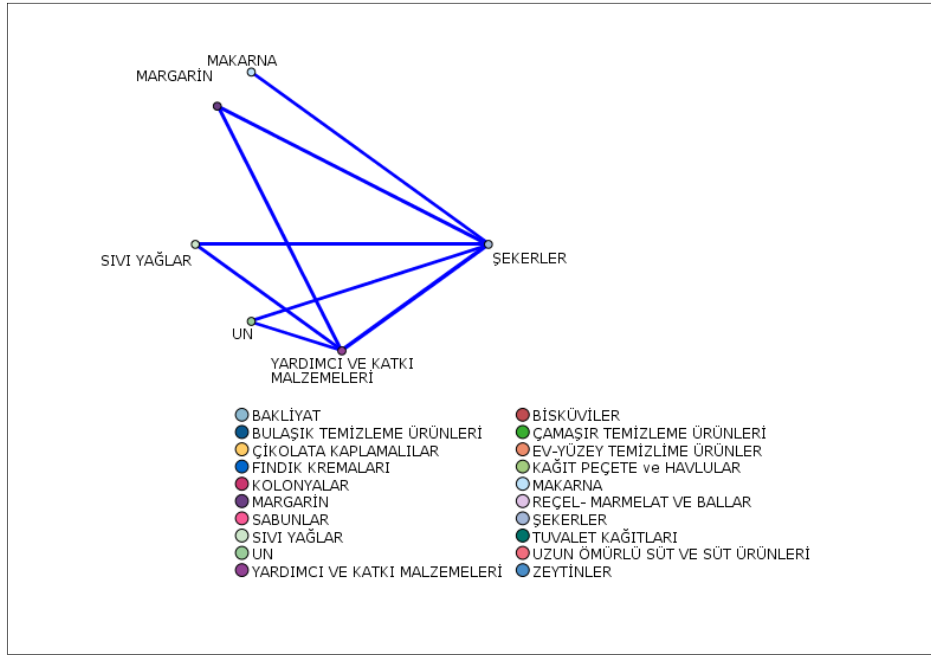
Veri setinde yer alan ürün kategorilerinin satın alınma oranları ve satın alınma frekansları aşağıdaki Şekil 5.45’ de görülmektedir.

Value	Proportion	% ▾	Count
ŞEKERLER		16.87	112553
MARGARİN		14.34	95692
YARDIMCI VE KATKI MALZEMEL		11.2	74731
SIVI YAĞLAR		7.95	53048
MAKARNA		7.91	52778
UN		7.73	51615
UZUN ÖMÜRLÜ SÜT VE SÜT Ü...		6.59	43967
KAĞIT PEÇETE ve HAVLULAR		5.59	37281
EV-YÜZEY TEMİZLİME ÜRÜNLER		4.74	31621
ÇAMAŞIR TEMİZLEME ÜRÜNLE...		3.3	21999
ZEYTİNLER		2.81	18722
BULAŞIK TEMİZLEME ÜRÜNLERİ		2.26	15092
SABUNLAR		1.65	11000
FINDIK KREMALARI- FINDIK F		1.6	10672
BAKLİYAT		1.47	9802
TUVALET KAĞITLARI		1.2	8030
BİSKÜVİLER		1.13	7527
REÇEL- MARMELAT VE BALLAR		1.07	7132
ÇİKOLATA KAPLAMALILAR		0.43	2839
KOLONYALAR		0.19	1251

Şekil 5.45.Ürün kategorilerine ait frekanslar

Şekil 5.45 incelendiğinde en çok %16,87 oranı ile şeker kategorisinin satın alındığı görülmektedir. Daha sonra %14,34 oranı ile margarin kategorisi ve %11,2 oranı ile yardımcı ve katkı malzemeleri kategorisi satın alınmaktadır.

Birliktelik kurallarının belirlenebilmesi için ilk olarak Web Grafiği ile kategoriler arası ilişkiler incelenmiştir. Kategoriler arasında sıklığı yüksek olan bağıntılar aşağıdaki Şekil 5.46’ da verilmiştir. En yüksek sıklık Şekerler ile Yardımcı ve Katkı Malzemeleri arasında olduğu görülmektedir.



Şekil 5.46.Sıklığı yüksek olan bağıntılar

Bu uygulamada, birliktelik tespiti algoritmalarından “Apriori” algoritması kullanılmıştır. Apriori, birliktelik tespitinde kullanılan en yaygın yöntemlerden bir tanesidir fakat sadece kategorik değişkenlerin girdi olarak kullanılması gibi bir kısıtı vardır.

Apriori varsayılan ayarlardan dolayı kuralları üretirken minimum support (destek) değerini % 10 ve minimum confidence (güven) değerini %80 olacak şekilde kullanmaktadır. Bu değerler değiştirilebilmektedir. Pratikte, doğru değerleri bulurken bazı deneme yanılma işlemleri yapmak gerekmektedir (değerler çok yüksek tutulursa hiçbir kural üretilmeyebilir, değerler çok düşük tutulursa çok fazla kural üretilabilir), bu değerler eşik değeri olarak adlandırılmakta ve uygulama alanından uygulama alanına değişiklik göstermektedir.

Bu uygulamada minimum support (destek) değeri %10, minimum confidence (güven) değeri %65 olarak belirlenmiştir ve analiz sonucu aşağıda Şekil 5.47’ de gösterilmektedir.

Consequent	Antecedent	Support %	Confidence %	Rule Support %	Lift
ŞEKERLER	MAKARNA YARDIMCI VE KATKI MALZEMELERİ	12.184	75.525	9.202	1.667
ŞEKERLER	UN SIVI YAĞLAR	11.085	73.892	8.191	1.631
YARDIMCI VE KATKI MALZEMELERİ	UN SIVI YAĞLAR	11.085	73.468	8.144	2.055
ŞEKERLER	SIVI YAĞLAR YARDIMCI VE KATKI MALZEMELERİ	14.836	73.351	10.882	1.619
ŞEKERLER	UN YARDIMCI VE KATKI MALZEMELERİ	13.61	71.978	9.796	1.589
ŞEKERLER	KAĞIT PEÇETE ve HAVLULAR YARDIMCI VE KATKI MALZEMELERİ	11.021	69.58	7.668	1.536
YARDIMCI VE KATKI MALZEMELERİ	SIVI YAĞLAR ŞEKERLER	15.938	68.28	10.882	1.91
YARDIMCI VE KATKI MALZEMELERİ	UN ŞEKERLER	14.44	67.84	9.796	1.897
ŞEKERLER	ZEYTİNLER	13.005	67.79	8.816	1.496
YARDIMCI VE KATKI MALZEMELERİ	MAKARNA ŞEKERLER	13.593	67.7	9.202	1.893
ŞEKERLER	UZUN ÖMÜRLÜ SÜT VE SÜT ÜRÜNLERİ YARDIMCI VE KATKI MALZEMELERİ	10.967	66.557	7.299	1.469

Şekil 5.47. Birliktelik kuralları analizi sonucu

Yukarıdaki Şekil 5.47’ de görülen alanlar şöyle yorumlanmaktadır. Yorumlar ilk üç satırdaki veriler dikkate alınarak yapılmıştır.

- Müşterilerin % 12,18’i hem makarna hem de yardımcı katkı malzemesi almaktadır. Makarna ve yardımcı katkı malzemesi alanların %75,52 si aynı zamanda şekerde almaktadır. Müşterilerin %9.2 si makarna, yardımcı katkı ve şeker almaktadır. Lift değerinin 1,6 olması “makarna ve yardımcı katkı malzemesi alan müşterilerin tüm müşterilere göre yaklaşık 1,5 kat daha fazla şeker almasını beklemekteyiz” ifadesini açıklamaktadır.
- Müşterilerin % 11,08’i hem un hem de sıvı yağ almaktadır. Un ve sıvı yağ alanların %73,89’u aynı zamanda şekerde almaktadır. Müşterilerin %8.19’u un, sıvı yağ ve şeker almaktadır. Lift değerinin 1,6 olması “un ve sıvı yağ alan müşterilerin tüm müşterilere göre yaklaşık 1,5 kat daha fazla şeker almasını beklemekteyiz” ifadesini açıklamaktadır.
- Müşterilerin % 11,08’i hem un hem de sıvı yağ almaktadır. Un ve sıvı yağ alanların %73,46 si aynı zamanda yardımcı ve katkı malzemesi almaktadır. Müşterilerin %8.14’ü un, sıvı yağ ve yardımcı ve katkı malzemesi almaktadır. Lift değerinin 2 olması “un ve sıvı yağ alan müşterilerin tüm müşterilere göre yaklaşık 2 kat daha fazla yardımcı ve katkı malzemesi almasını beklemekteyiz” ifadesini açıklamaktadır.

Kategoriler arası ilişkiler incelendiğinde en yüksek sıklığın şeker kategorisi ile yardımcı ve katkı malzemeleri arasında olduğu belirlenmişti. Bu kategorilerden her

ikisini de birlikte alan müşteriler için “Şeker_Yrd_Katkı_Malzemeleri” adında yeni bir alan oluşturulmuştur. Bu alanda her iki kategoriden de alan müşteriler 1, diğer müşteriler ise 0 ile temsil edilmektedir.

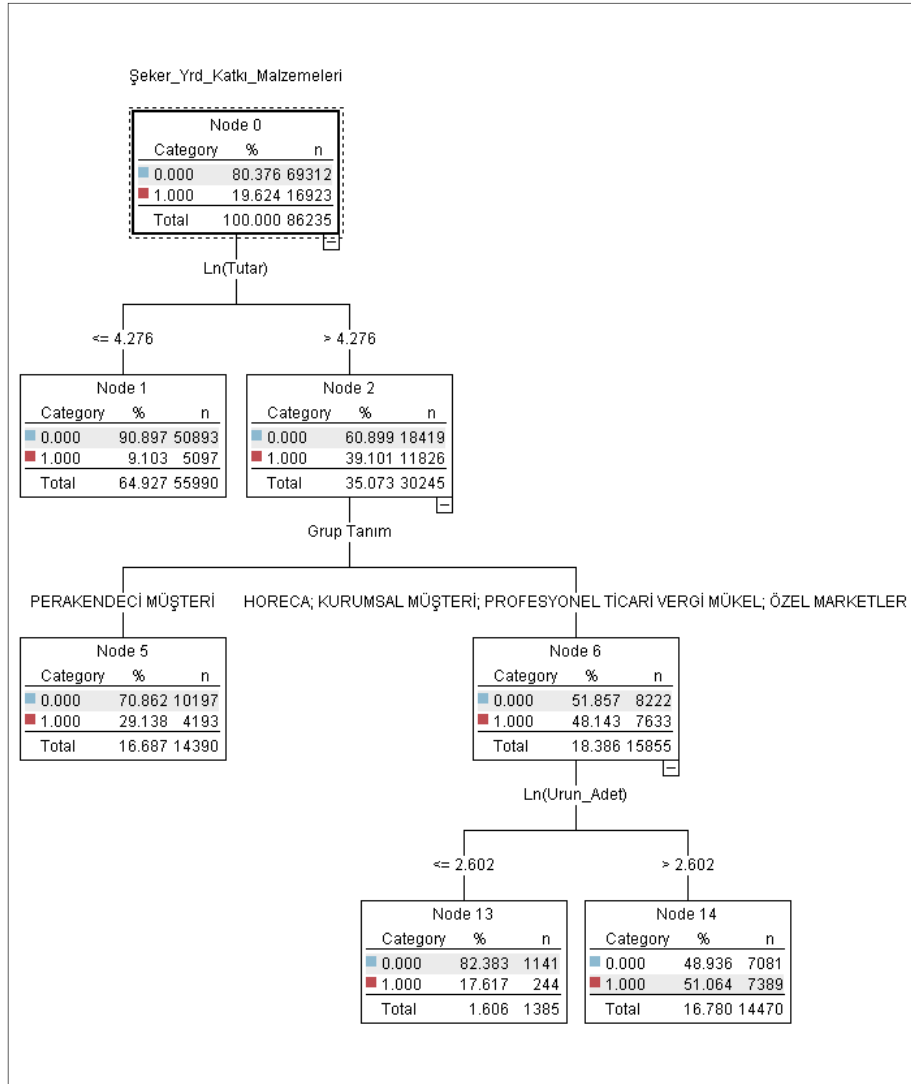
Şeker ile yardımcı ve katkı malzemesi kategorilerinden her iki kategoriden de alan müşterilerin profillerini belirlemek için karar ağacı algoritmalarından C&R Tree kullanılmıştır. Hedef değişken olarak “Şeker_Yrd_Katkı_Malzemeleri” alanı kullanılmıştır, nicel alanlar logaritmik dönüşüm uygulandıktan sonra modele katılmışlardır.

C&R Tree algoritması sonucunda, Şekil 5.48’ de gösterilen ağaç yapısı incelendiğinde; şeker ile yardımcı ve katkı malzemesi kategorilerinin her ikisinden de alan müşterilerin belirlenmesinde en önemli faktörün “Tutar” olduğu görülmektedir.

Alışveriş tutarları 72 TL den büyük olan müşterilerin bu iki kategoriden alma oranlarının daha yüksek olduğu gözlemlenmektedir. Bu iki kategorinin beraber alımında en önemli ikinci faktör ise müşteri grup tanımlamalarıdır.

Müşteri gruplarından, “Horeca”, “Kurumsal Müşteri”, “Profesyonel Ticari Vergi Mükellefleri” ve “Özel Marketler” grubunda bulunan müşterilerin, şeker ve yardımcı katkı malzemesi kategorilerinin her ikisinden birlikte alma oranları daha yüksek olduğu bilgisi bulunmuştur. Bu müşterilerin içerisinde, ürün adedi 13 den büyük olan müşterilerin daha yüksek oranda bu iki kategoriden aldıkları bilgisi elde edilmiştir.

Buradaki açıklamada belirtilen değerler, ağaç yapısında verilen sayısal değerlerin anti logaritması alınarak verilmiştir.



Şekil 5.48.Sıklığı en yüksek olan kategorilere ilişkin C&R Tree algoritmasına ait ağaç yapısı

Çalışma sonucunda elde edilen bilgiler ile müşteri beklentileri analiz edilerek, satın alma davranışlarına göre farklı ürünleri satın almaları sağlanarak, firmanın kar marjı yükseltilebilecektir. Mevcut müşterilerin yanı sıra hangi müşteri gruplarının hedeflenmesi gerektiği de yine analizler sonucunda elde edilecek faydalı diğer bilgilerdendir.

6.SONUÇ VE ÖNERİLER

Bilginin firmalar için hayati bir dönem taşıdığı günümüzde, Veri Madenciliği yöntemleri ile karlılığı yüksek olan müşterilerin belirlenmesi sağlanmaktadır. Bu çalışmalar ile elde edilen veriler firmalar için stratejik bir öneme sahiptirler. Bu şekilde elde edilen sonuçlar ile rakiplere göre önemli derecede rekabet avantajı sağlamaktadır.

Bu tez çalışmasında bir perakende işletmenin belirli bir dönem boyunca alışveriş yapmış müşteri verileri üzerinde, öncelikle Veri Manipülasyonu yöntemleri uygulanmış ve sonrasında EDA yöntemlerinden faydalanılarak, veri madenciliği modellemelerinden Sınıflama ve Regresyon Modelleri, Kümeleme ve Birliktelik Kuralları incelenmiştir.

Öncelikle, veri madenciliğinin en uzun süreci olan verilerin elde edilmesi, düzenlenmesi, temizlenmesi, eksik ve aykırı gözlemlerin belirlenerek uygun tekniklerle gerekli işlemlerin yapılarak veri setinin analize hazır hale getirilmesini kapsayan manipülasyon süreci uygulanmıştır.

Veri manipülasyonu süreci sonunda, veri setinde bulunan müşterilere ait, farklı sistemlerde tutulan bilgiler toplanarak analiz edilebilir hale getirilmiştir. Analiz edilebilir hale getirilen bu veriden, müşterilerin kaç defa mağazalardan alışveriş yaptığı, alışveriş yaptığı ay sayısı, müşterilere ait ortalama ürün kalem sayısı ve ürün çeşit sayısı gibi nicel bilgiler veri manipülasyonu süreci sonucunda elde edilmiştir.

Modelleme öncesinde, modellemelere katılacak olan nicel değişkenler EDA yöntemleri kullanılarak incelenmiş ve aşırı çarpık dağılımlara sahip olduğu gözlemlenen bu değişkenler üzerinden uygun logaritmik dönüşümler yapılarak, değişken dağılımları daha simetrik bir yapıya getirilmiştir.

Veri seti üzerinden rastgele örneklemeler çekilerek, bu örneklemeler üzerinden bağımlı değişken olan kar tutarı ile bağımsız değişkenler arasındaki ilişkiler, logaritmik dönüşüm yapıp ve yapılmaksızın Matlab programı kullanılarak incelenmiştir. Yapılan bu incelemenin sonunda, bağımlı değişken olan kar tutarı üzerinde firma yaşı değişkeninin bir etkisi yok iken ürün âdetinin(Ln) kar tutarı(Ln) ile ilişkisinin oldukça iyi ve doğrusal olduğu belirlenmiştir.

Analize hazır hale getirilen verilerden bilgi elde edilebilmesi için, kullanılacak modelleme yöntemleri ve bunların algoritmaları belirlenmiştir. Veri madenciliği sonuçlarının elde edildiği ve kararların bu aşamada elde edilen bilgiler üzerine verildiği düşünüldüğünde bu aşama oldukça önemlidir.

Müşterilere ait kar tutarı durumları üç farklı grup altında toplanarak (düşük kar tutarlı, orta kar tutarlı, yüksek kar tutarlı) incelenmiştir.

EDA teknikleri uygulanıp/uygulanmaksızın, karar ağaçları algoritmalarından C&R Tree ve CHAID algoritmaları kullanılarak “kar tutarı yüksek” müşterilerin bu grupta toplanmasında hangi faktörlerin etkili olduğu belirlenmeye çalışılmış ve EDA tekniklerinin uygulanıp/uygulanmamasına göre bu sonuçlar incelenmiştir. Elde edilen bu sonuçlara göre, EDA yöntemlerinin karar ağaçları ile uygulanan sınıflama tekniklerinde bir etkisinin olmadığı gözlemlenmiştir.

Faktör analizi kullanılarak, müşterilere ait faktör skorları elde edilmiş ve bu skorlar modellemede girdi değişkeni olarak kullanılmıştır. Elde edilen bu skorlar kullanılarak K-Means algoritması ile müşteriler segmentlere ayrılmıştır. Karar ağacı sınıflandırma yöntemlerinden C&R Tree algoritması kullanılarak oluşturulan yeni segmentler için kurallar elde edilmiştir.

Faktör analizinde kullanılacak nicel değişkenler, EDA yöntemleri uygulanıp/uygulanmaksızın analize sokulmuş ve analiz sonuçları incelenmiştir. Analiz sonuçlarına göre, EDA kullanılarak oluşturulan faktörlerin, toplam varyansı daha yüksek oranda açıkladığı belirlenmiştir ve dolayısıyla, müşteri segmentasyonu sürecine nicel değişkenler üzerinde EDA yöntemleri kullanılarak devam edilmiştir.

Karar ağacı sınıflandırma yöntemlerinden C&R Tree algoritması kullanılarak oluşturulan yeni segmentler için; Adana, Ankara, Antalya, İzmir ve Kayseri bölgelerinden alışveriş yapan müşterilerin, ürün çeşit sayıları 5 ten büyük ise bu müşterilerin skoru en yüksek olarak hesaplanmış birinci kümeye dâhil oldukları belirlenmiştir.

Birliktelik Kuralları ve Sepet Analizi kullanılarak, müşterilerin satın alma davranışları incelenmiştir. Müşterilerin satın aldıkları ürün kategorilerinin frekanslarına bakılmış; en çok %16,87 oranı ile şeker kategorisinin satın alındığı, ikinci olarak %14,34 oranı ile margarin kategorisi ve ardından %11,2 oranı ile yardımcı ve katkı malzemeleri kategorisi satın alındığı gözlemlenmiştir.

Web Grafiği kullanılarak kategoriler arası ilişkilere bakılmış ve kategoriler arasındaki en yüksek sıklığın, Şekerler ile Yardımcı ve Katkı Malzemeleri arasında olduğu belirlenmiştir.

Birliktelik tespitinde, en yaygın yöntemlerden bir tanesi olan “Apriori” algoritması kullanılmıştır. Birliktelik kuralları ve sepet analizi ile müşterilerin hangi ürünleri hangi ürünlerle birlikte satın aldıkları belirlenmiştir. Daha sonra müşterilerin satın alma davranışını en çok etkileyen değişkenler karar ağaçlarıyla tespit edilmiştir.

Firma sepet analizi ile elde edilen bu sonuçları; reklam stratejileri belirlemede, satış işlemlerini planlamada kullanabilir. Birlikte sık olarak satın alınan ürünlerini yakın raflara koyarak satışlarını arttırabilir.

KAYNAKLAR

- Akbulut, S., 2004.** Veri Madenciliği Teknikleri ile Bir Kozmetik Markanın Ayrılan Müşteri Analizi ve Müşteri Segmentasyonu, *Yüksek Lisans Tezi*.
- Akpınar, 2000.** Veritabanlarında bilgi keşfi ve veri madenciliği, *İstanbul Üniversitesi İşletme Fakültesi Dergisi*, 29: 1-22
- Albayrak, S. ve Yılmaz, Ş.K., 2009.** Veri madenciliği karar ağacı algoritmaları ve İMKB verileri üzerine bir uygulama, *Süleyman Demirel Üniversitesi, İktisadi ve İdari Bilimler Dergisi*, C.14, S.1 s.31-52.
- Aydoğan, F., 2003.** E-ticarete veri madenciliği yaklaşımlarıyla müşteriye hizmet sunan akıllı modüllerin tasarımı ve gerçekleştirimi", *Yüksek Lisans Tezi*, Hacettepe Üniversitesi Fen Bilimleri Enstitüsü, Ankara, 12-16.
- Çelik, M., 2009.** Veri Madenciliğinde Kullanılan Sınıflandırma Yöntemleri ve Bir Uygulama, *Yüksek Lisans Tezi*, İstanbul Üniversitesi Sosyal Bilimler Enstitüsü,
- Erar, A. Ve Kiroğlu, G. 2008.** Veri Çözümleme Ders Notları, *Mimar Sinan G.S. Üniversitesi İstatistik Bölümü*
- Giudici, P., 2003.** Applied Data Mining: Statistical Methods for Business and Industry, *John Wiley & Sons Inc.*, Chichester, 85-100.
- Han, J. ve Kamber, M., 2006.** Data Mining Concepts & Techniques, *San Francisco, Morgan Kauffmann Publishers*, s. 10.

Kıroğlu,G., Erar,A., Göçer,Z., Göncü,E., Zekier,F., Çilengiroğlu,E., Deliloğlu,S., Şatır,U., 2008, Implementation of Data Mining in Automotive Sector., *International Conference On Multivariate Statistical Modelling and High Dimensional Data Mining, 19-23 Haziran,* Kayseri.

Kök, B.V. ve Kuloğlu, N., 2005. Sollaama Esnasında Taşıt ve Yol ile İlgili Faktörlerin Karar Ağacı Yöntemi ile İrdelenmesi, *Erciyes Üniversitesi Fen Bilimleri Enstitüsü Dergisi.*

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

Mattison R.,1997. Data Warehousing and Data Mining for Telecommunications, *Norwood, Artech House,* s. 254.

Oğuzlar A.,2003. Veri ön işleme, *Erciyes Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi,* Sayı: 21, ss. 67-76.

Tüzüntürk, S., ve Oğuzlar A. 2009. Borsada İşlem Gören Şirketlerin Finansal Göstergelerinin Analizi, *10. Türkiye Ekonometri ve İstatistik Sempozyumu, Atatürk Üniversitesi, Palandöken / Erzurum.*

Özkan, Y.,2008. Veri Madenciliği Yöntemleri, *İstanbul Papatya Yayıncılık Eğitim*

Şentürk A., 2006. Veri Madenciliği Kavram ve Teknikler, syf 6.

Tezcanlar, P. ve Timor M.,2007.Müşteri ilişkileri yönetimi, veri madenciliği ve bir uygulama, *İstanbul Üniversitesi Sosyal Bilimler Enstitüsü,Yüksek lisans tezi*

Yarahoğlu, K.,2004. *Uygulamada Karar Destek Yöntemleri,* İzmir, İlkem Ofset.

ÖZGEÇMİŞ

Adı Soyadı : Fahrettin CEVAHİR

Doğum Yeri : İzmir

Doğum Tarihi : 23.04.1980

Medeni Hali :Bekâr

Eğitim Durumu

Lise : Isparta Anadolu Öğretmen Lisesi (1998)

Ön Lisans : Dumlupınar Üniversitesi, Tavşanlı Meslek Yüksek Okulu, Bilgisayar ve Enformasyon Sistemleri (2002)

Lisans : Ondokuz Mayıs Üniversitesi, Fen Edebiyat Fakültesi, İstatistik Bölümü (2007)

Yabancı Dil : İngilizce

İş Tecrübesi

2007 – 2009 : Komtaş Bilgi Yönetimi Danışmanlık ve Tic. A.Ş, Yazılım Uzm. Yrd.

2009 – 2010 : Bizim Toplu Tüketim Paz. San. Ve Tic. A.Ş, CRM Uzmanı

2011 - : Türk Telekomünikasyon A.Ş, Gelir Güvence Uzmanı