

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

**KREDİ KARTI BAŞVURU AŞAMASINDA
SAHTECİLİK TESPİTİ İÇİN BİR
VERİ MADENCİLİĞİ MODELİ**

YÜKSEK LİSANS TEZİ

Mak. Müh. Mustafa Aykut GÖRAL

Anabilim Dalı: Endüstri Mühendisliği

Programı: Endüstri Mühendisliği

OCAK 2007

**KREDİ KARTI BAŞVURU AŞAMASINDA
SAHTECİLİK TESPİTİ İÇİN BİR
VERİ MADENCİLİĞİ MODELİ**

**YÜKSEK LİSANS TEZİ
Mak. Müh. Mustafa Aykut GÖRAL
507021131**

**Tezin Enstitüye Verildiği Tarih : 25 Aralık 2006
Tezin Savunulduğu Tarih : 29 Ocak 2007**

**Tez Danışmanı : Doç.Dr. Cengiz GÜNGÖR (İ.T.Ü.)
Diğer Jüri Üyeleri Öğr.Gör.Dr. Gülgün KAYAKUTLU (İ.T.Ü.)
Doç.Dr. Özgür KAYALICA (İ.T.Ü.)**

OCAK 2007

ÖNSÖZ

Veri Madenciliği konusunda araştırma olanağı sağlayan, bu çalışma sırasında ilgisini ve desteğini esirgemeyen tez danışmanım Sayın Doç. Dr. Cengiz GÜNGÖR'e, bu çalışmada XYZ Bank'ta yapılan projeyi kullanmama izin veren ve gerekli bilgileri benimle paylaşan XYZ Bank yöneticileri Sayın Pınar İŞMEN ve Sayın Başak KAYALIK'a, XYZ Bank eski yöneticileri Sayın Kunter KUTLUAY ve Ayla GÜLTÜRK'e, çalışmamda destek olan Kratis Bilgi Teknolojileri'ndeki eski yöneticim Sayın Değerhan USLUEL'e ve eski çalışma arkadaşım Sayın Tuba YAVAŞ'a ve son olarak da bana ömrümün ilk gününden beri her konuda destek olan aileme sonsuz teşekkürlerimi sunarım.

Aralık 2006

Mustafa Aykut GÖRAL

İÇİNDEKİLER

KISALTMALAR	vi
TABLO LİSTESİ	vii
ŞEKİL LİSTESİ	vii
ÖZET	ix
SUMMARY	xi
1. GİRİŞ	1
1.1. Giriş ve Çalışmanın Amacı	1
2. VERİNİN YARARLI BİLGİYE DÖNÜŞÜM SÜRECİ	4
2.1. Veri	4
2.2. Enformasyon	5
2.3. Analitik	6
2.4. Bilgi	6
2.5. Yararlı Bilgi	7
2.6. Verinin Anlam Düzeylerine Göre Kullanılan Teknolojiler	7
3. VERİ MADENCİLİĞİ	9
3.1. İş Zekası	9
3.2. Veri Tabanlarında Bilgi Keşfi	10
3.3. Veri Madenciliğinin Tanımı	12
3.4. Veri Madenciliğine Katkıda Bulunan Disiplinler	13
3.5. Veri Madenciliğinin Gelişimi	14
4. VERİ MADENCİLİĞİ PROJESİ SAFHALARI	17
4.1. İş Sorusunu Anlama	18
4.2. Veriyi Anlama	19
4.3. Veri Hazırlığı	19
4.3.1. Örnekleme	19
4.3.2. Veri Temizliği	20
4.3.3. Aykırı Değer Analizi	21
4.3.4. Eksik Değerler	22
4.3.5. Veri Dönüştürme	23
4.3.6. Değişken Seçimi	23
4.4. Modelleme	24
4.5. Değerlendirme	25
4.6. Uygulama	27
4.7. İzleme	28

5. VERİ MADENCİLİĞİ MODELLERİ	29
5.1. Sınıflandırma	30
5.2. Tahmin	33
5.3. Öngörme	34
5.4. Zaman Serisi Analizi	34
5.5. Kümeleme	35
5.6. Birliktelik Kuralları	36
5.7. Ardışlık Keşfi	37
6. VERİ MADENCİLİĞİ TEKNİKLERİ	38
6.1. Pazar Sepeti Analizi	38
6.2. Karar Ağaçları	45
6.3. Yapay Sinir Ağları	48
6.3.1. Yapay Sinir Ağlarının Temel Özellikleri	50
6.3.2. Öğrenme Şekillerine Göre Yapay Sinir Ağları	51
6.3.3. Ağ Yapılarına Göre Yapay Sinir Ağları	53
6.3.4. Yapay Sinir Ağlarının Kuvvetli ve Zayıf Yönleri	54
6.4. Genetik Algoritmalar	54
6.5. Yapısal Risk Minimizasyonu	56
6.6. Destek Vektör Makinaları	61
7. FARKLI ENDÜSTRİLERDEKİ VERİ MADENCİLİĞİ UYGULAMALARI	65
7.1. Bankacılık ve Finans Endüstrilerinde Veri Madenciliği Uygulamaları	65
7.2. Perakende Sektöründe Veri Madenciliği Uygulamaları	66
7.3. Telekomünikasyon Sektöründe Veri Madenciliği Uygulamaları	67
7.4. Üretim Sektöründe Veri Madenciliği Uygulamaları	68
7.5. Diğer Sektörlerdeki Veri Madenciliği Uygulamaları	69
8. LİTERATÜRDE BANKACILIK VE FİNANS SEKTÖRÜNE AİT UYGULAMA ÖRNEKLERİ	71
8.1. CHAID Algoritmasının Kullanıldığı Bir Kredi Değerlendirme Problemi	71
8.2. C&RT Algoritmasının Kullanıldığı Bir Kredi Kartı Başvurusu Değerlendirme Problemi	72
8.3. Destek Vektör Makinaları ve Sinir Ağları Kullanımı ile Kredi Değerlendirme Analizi	75
8.3.1. Kurulan Modeller ve Öngörü Doğruluk Dereceleri	78
9. TÜRKİYE’DE KREDİ KARTI KULLANIMI VE KREDİ KARTI DOLANDIRICILIK TİPLERİ	80
9.1. Türkiye’de Kredi Kartı Kullanımı	80
9.2. Kredi Kartı Dolandırıcılık Tipleri	83
9.2.1. Kayıp/Çalıntı Kart Kullanımı	83
9.2.2. Ele Geçmeyen Kartlar	84
9.2.3. Sahte Başvuru	84
9.2.4. Posta, İnternet ve Telefon ile Yapılan Mal Sipariş Dolandırıcılığı	85
9.2.5. ATM Dolandırıcılıkları	85

9.2.6. Sahte Kart	86
9.2.7. Nakit Temin Etmeye Yönelik Kullanım	86
10. UYGULAMA: XYZ BANK'TA KREDİ KARTI BAŞVURU AŞAMASINDA VERİ MADENCİLİĞİ İLE SAHTEKARLIK TESPİTİ	87
10.1. Giriş	87
10.2. XYZ Bank	89
10.3. XYZ Bank'ın Problemi	89
10.4. Sahtecilik Tespiti	89
10.4.1. Bilinen Sahtecilikler İçin Süreç	90
10.4.2. Bilinmeyen Sahtecilikler İçin Süreç	91
10.5. Proje Konusu	92
10.5.1. Projenin Amacı	92
10.6. XYZ Bank'daki Mevcut Durum	92
10.7. Verilerin Hazırlanması	96
10.7.1. Eksik Değerler	101
10.7.2. Verileri Dönüştürme	102
10.7.3. Değişken Seçimi	104
10.8. Modelin Kurulması	105
10.9. Kurulan Modelin Değerlendirilmesi	106
10.10. Modelin Kullanılması	113
10.11. Modelin İzlenmesi	113
10.12. Sonuçların Değerlendirmesi	114
11. SONUÇ	119
KAYNAKLAR	121
EK A :KXEN ANALYTIC FRAMEWORK PAZAR SEPETİ ANALİZİ SONUÇLARI	124
ÖZGEÇMİŞ	146

KISALTMALAR

OLAP	: Online Analytical Processing
OLTP	: Online Transaction Processing
VTBK	: Veri Tabanlarında Bilgi Keşfi
COBOL	: Common Business Oriented Language
PCA	: Principal Component Analysis
CRISP-DM	: Cross Industry Standard Process for Data Mining
AID	: Automatic Interaction Detector
CHAID	: Chi-Squared Automatic Interaction Detector
C&RT	: Classification and Regression Trees
MARS	: Multivariate Adaptive Regression Splines
QUEST	: Quick, Unbiased, Efficient Statistical Tree
SLIQ	: Supervised Learning in Quest
SPRINT	: Scalable Parallelizable Induction of Decision Trees
YSA	: Yapay Sinir Ağları
YRM	: Yapısal Risk Minimizasyonu
VC	: Vapnik-Chervonenkis
DVM	: Destek Vektör Makinaları
SFI	: Securities and Futures Institute
TRC	: Taiwan Ratings Corporation
ANOVA	: Analysis of Variance
SA	: Sinir Ağları
LojR	: Lojistik Regresyon
BKM	: Bankalararası Kart Merkezi A.Ş.
CRM	: Customer Relationship Management
SB	: Sahte Başvuru
TCMB	: Türkiye Cumhuriyeti Merkez Bankası
SABAS	: Sahte Bilgi/Belge/Beyan/Başvuru Alarm Sistemi
KKB	: Kredi Kayıt Bürosu
PSB	: Potansiyel Sahte Başvuru

TABLO LİSTESİ

	<u>Sayfa No</u>
Tablo 2.1 Veri İşlem Örneği.....	5
Tablo 2.2 Verilerin Enformasyona Dönüşecek Şekilde Toplanması	5
Tablo 2.3 Enformasyonun Analiz İçin Ayırıştırılması.....	6
Tablo 2.4 Bir Satın Alma Örüntüsünün Teşhis Edilmesi.....	7
Tablo 2.5 Verilerin Anlam Düzeylerine Karşılık Gelen Teknolojilerin Sınıflandırılması.....	8
Tablo 3.1 Veri Madenciliğinin Gelişimi.....	15
Tablo 4.1 Yanlış Veri Örneği.....	20
Tablo 4.2 Risk Matrisi.....	27
Tablo 5.1 Iris Çiçeği Veri Kümesi.....	31
Tablo 6.1 Örnek Sepet Analizi Sonuçları.....	40
Tablo 6.2 Fatura Bilgileri.....	41
Tablo 6.3 Gözlem Yüzdesiyle Gösterilen Kurallar.....	42
Tablo 6.4 Gözlem Sayısıyla Gösterilen Kurallar.....	42
Tablo 6.5 Bilinen YSA Mimarilerinin Tarihsel Gelişimi.....	49
Tablo 8.1 Veri Kümesinde Yer Alan Bağımsız Değişkenler.....	71
Tablo 8.2 Kredi Kartı Başvurusuna İlişki Veriler.....	73
Tablo 8.3 Risk Matrisi.....	74
Tablo 8.4 Kazançlar Matrisi.....	74
Tablo 8.5 Kredi Değerlendirme Kategorilerine Ait Dağılımlar.....	76
Tablo 8.6 Veri Kümesinde Kullanılan Finansal Oranlar.....	77
Tablo 8.7 Doğruluk Dereceleri.....	78
Tablo 9.1 Yerli ve Yabancı Kredi Kartlarının Yurtiçi Kullanımı.....	83
Tablo 10.1 Proje Öncesi Ön İzleme Servisi Performansı.....	93
Tablo 10.2 Analiz Öncesi Elde Edilen Değişkenler.....	99
Tablo 10.3 Modellemeye Sokulmayan Değişkenler.....	104
Tablo 10.4 Üç Alt Kümenin Modellemedeki Roller.....	105
Tablo 10.5 Proje Öncesi ve Sonrası Ön İzleme Servisi Performansı.....	114
Tablo 10.6 Proje Sonrası Performansta Gözükten Yüzdesel Artış.....	114
Tablo 10.7 Sahte Başvuru Adet Bazında Analiz.....	115
Tablo A.1 KXEN Analytic Framework Pazar Sepeti Analizi Sonuçları.....	123

ŞEKİL LİSTESİ

	Sayfa No
Şekil 2.1 : Veri Dönüşüm Süreci Aşamaları.....	4
Şekil 3.1 : İş Zekası Elemanları ve Süreci.....	10
Şekil 3.2 : VTBK Süreci.....	11
Şekil 3.3 : Veri Madenciliğine Katkıda Bulunan Disiplinler.....	13
Şekil 4.1 : CRISP-DM'ye Göre Veri Madenciliği Süreci.....	17
Şekil 4.2 : Denetimli Öğrenme.....	25
Şekil 5.1 : Veri Madenciliği Modelleri.....	29
Şekil 5.2 : Taç Yaprağının En ve Boyuna Göre XY-Grafiği.....	32
Şekil 5.3 : Bir Karar Ağacı Örneği.....	32
Şekil 5.4 : Kümeleme.....	35
Şekil 6.1 : Verilerden Karar Ağacına.....	47
Şekil 6.2 : İleri Beslemeli Sinir Ağı Örneği.....	49
Şekil 6.3 : Denetimli Öğrenme Yapısı.....	52
Şekil 6.4 : Denetimsiz Öğrenme Yapısı.....	53
Şekil 6.5 : Takviyeli Öğrenme Yapısı Takviyeli Öğrenme Yapısı.....	53
Şekil 6.6 : İleri Beslemeli ve Geri Beslemeli Ağ Yapıları.....	54
Şekil 6.7 : VC Boyutu.....	58
Şekil 6.8 : Bir Noktalar Kümesinin Bir Doğru ile Bölünmesi Örneği.....	60
Şekil 6.9 : Noktalar Kümesinin Bir Polinom İle Bölünmesi Örneği.....	60
Şekil 8.1 : CHAID Algoritması Uygulanan Bir Karar Ağacı.....	72
Şekil 9.1 : Son Yıllara Ait Kredi Kartı ve Banka Kartı Sayıları.....	82
Şekil 10.1 : Bankalarda Kara Geçmek İçin Gerekli İşlemler.....	88
Şekil 10.2 : Bilinen Sahtecilikler İçin Süreç.....	91
Şekil 10.3 : Bilinmeyen Sahtecilikler İçin Süreç.....	91
Şekil 10.4 : XYZ Bank Proje Öncesi Başvuru Değerlendirme Süreci.....	95
Şekil 10.5 : Üç Aylık Döneme Ait Sahtecilik Vakaları.....	96
Şekil 10.6 : Model Oluşturma.....	106
Şekil 10.7 : Performans Eğrisi.....	107
Şekil 10.8 : Değişkenlerin Modele Etkileri.....	109
Şekil 10.9 : Ev Adresi İl Kodu Değişkeninin Hedefe Olan Etkisi.....	110
Şekil 10.10 : Öğrenim Durumu Değişkeninin Hedefe Olan Etkisi.....	111
Şekil 10.11 : XYZ Bank Proje Sonrası Başvuru Değerlendirme Süreci.....	113
Şekil 10.12 : Sahte Başvuru Tespit Adetleri.....	117
Şekil 10.13 : Tespit Edilen Toplam Sahte Başvuruların Başvurulara Oranı.....	117
Şekil 10.14 : Tespit Edilen Toplam Sahte Başvuruların Onaylanan Kartlara Oranı.....	118

KREDİ KARTI BAŞVURU AŞAMASINDA SAHTECİLİK TESPİTİ İÇİN BİR VERİ MADENCİLİĞİ MODELİ

ÖZET

Veri madenciliği, muazzam boyuttaki veriden şirketlerin daha iyi kararlar almalarına yardımcı olup, pazarda rekabetçi olarak kalmalarını sağlayabilecek ilginç bilgileri keşfetme sürecidir.

Hızla gelişen iş dünyası içerisinde her gün değeri artan, yöneticilerin ileriye dönük olarak doğru kararlar almasında ışık tutan karar destek sistemlerinden birisi olan veri madenciliği kavramının, teknolojilerinin ve kullanım alanlarının ayrıntılı olarak ele alındığı bu tez çalışmasında ayrıca Türkiye’de XYZ Bank’ta yapılan, kredi kartı başvurusu aşamasında sahte başvuru tespitine yönelik bir veri madenciliği projesi anlatılmıştır.

Bankacılık sektöründe, özellikle de kredi kartlarında yaşanan büyük rekabet bankaları, müşteri tabanlarını büyütmeye ve daha riskli segmentler hedeflemeye yönelmektedir. Bankalar daha fazla müşteri kazanma çabalarını sürdürdükçe de, kredi kartı başvuru sayısı giderek artmaktadır. Bütün bunların sonucunda ise hem başvuru aşamasında, hem de işlem aşamasında sahtekarlıkların sayısı hızlı bir artış göstermektedir.

Projede, Yapısal Risk Minimizasyonu algoritmasını kullanan KXEN Analytic Framework yazılımı yardımıyla sahtecileri tespit edebilmek için bir öngörüselsel model kurulmuştur. Bu model tüm başvuruları skorlamaktadır. Modelin sonucunda ortaya çıkan rapor, tüm başvurular için bir sahtekarlık skoru içermektedir. Bu raporu kullanan Güvenlik Birimi çalışanları, tecrübelerini ve zamanlarını sahte olması en muhtemel başvuruları incelemeye harcamaktadır.

Bu proje ile kredi kartı başvurusu sırasında yapılan sahtekarlıkların tespit sayısı günde 7’den 23’e çıkmış ve %228.6’lık bir artış elde edilmiştir. Bu sayı, günde 65,380.1 YTL’lik bir tasarruf anlamına gelmekte ve projeye yapılan tüm yatırım 4 gün içinde karşılanmaktadır.

Yeni sistemi, Kredi Kartı Güvenlik biriminde 13 kiři kullanmaktadır. Birim elemanlarının morallerini de hızla yükselten bu sistemin uygulanması sayesinde, çalışmalar daha verimli gerçekleşmeye başlamıştır. Güvenlik bölümü çalışanları kayıtları elle kontrol ederek sahtekarlıkları yakalamaya çalışmak için gereğinden fazla bir süre harcamaktansa, zamanlarını gerçek sahtekarlıkları yakalayarak harcar hale geldiler. Yakalanan sahtekarlık sayısını bu sayede 3.3 katına çıkarmayı başardılar.

A DATA MINING MODEL FOR FRAUD DETECTION AT CREDIT CARD APPLICATION STAGE

SUMMARY

Data mining is the process of discovering interesting knowledge from large amounts of data that can be used to help companies make better decisions and remain competitive at marketplace.

One of the most popular decision support system that helps managers to take future decisions more clearly is data mining. Data mining, its technologies and application fields are widely analysed in this thesis, moreover a data mining project which is done at XYZ Bank in Turkey in order to detect fraudulent application for credit card is described.

Fierce competition in banking industry, especially in the credit card business, forces banks to grow their customer bases and target lower value segments. Credit card application volumes increase as banks reach out to acquire more customers. As a consequence, the number of fraudulent applications and transactions is rapidly increasing.

For the project at XYZ Bank, a predictive model is generated by using KXEN Analytic Framework which uses Structured Risk Minimization algorithm. This model scores all individual applications. The resulting reports include a fraud score for all individual applications and are forwarded to fraud agents who now apply their experience only on reviewing those applicants most likely to commit fraud.

By this project XYZ Bank increased the number of identified fraudulent applications by %228.6 from 7 to 23 per day. This represents saving of 65,380.1 YTL per day and returned more than the entire project investment with in four days.

The new system is used by 13 fraud agents within the Credit Card Security Department. It had an immediate positive impact on the morale of the team members, who became much more effective at their job. Rather than manually evaluating large

amounts of non-fraudulent data to identify potential candidates, they spend their time working on actual fraud cases. In this manner they increased the number of fraud cases caught by a factor 3.3.

1. GİRİŞ

1.1. Giriş ve Çalışmanın Amacı

Organizasyonlar canlı organizmalara benzer. Bilgiyle beslenir, iletişim ağları üzerinden ilişki kurar, topladığı verileri değerlendirerek bir sonraki adımını planlarlar. Özetle, ham veri işlenerek stratejiye dönüştürülür. Ancak iş verilerinin hızlı artışı karşısında, verinin bilgiye dönüşüm hızı çok önem kazanmıştır. Manuel yöntemlerin etkisini yitirdiği bu zamanlarda, veri madenciliği teknolojisi bu sorunu gidermekte ve çok büyük veri yığınları içinde saklı kalmış bilgileri hızlı bir şekilde gün yüzüne çıkarmaya yardımcı olmaktadır.

Verinin olduğu her ortamda çalışma yapabileceğiniz veri madenciliği teknolojisi ile organizasyonlara fayda sağlayacak bir çok çözüme imza atmak mümkündür. Bankacılık sektöründen sağlık sektörüne, perakende sektöründen üretim sektörüne kadar bir çok sektörde kullanılan veri madenciliği, firmaların müşterilerini analiz etmesinde, onlara doğru hizmetleri sunabilmesinde ve geleceğe yönelik öngörülerde bulunabilmesinde yardımcı olmaktadır.

Firmaların temel amacı müşteri memnuniyetini üst seviyelerde tutarken, aynı zamanda firmayı kara geçirecek faaliyetleri de yürütebilmektir. Kara geçmek için en etkin yollardan biri şüphesiz ki masrafları azaltmaktır. Yoğun olarak masraf yapılan pazarlama ve satış maliyetlerinde yapılacak tasarruf, bu amaç doğrultusunda faydalı olacaktır. Bankacılık sektörü düşünüldüğünde ise, bir bankayı maddi olarak büyük zarara uğratan riskli müşterilerden ve sahtekarlardan sakınmanın, kara geçmek için çözüm aranan ilk konular arasında yer aldığı görülmektedir.

Bu tez çalışmasının amacı, veri madenciliği kavramını, teknolojilerini ve uygulama alanlarını incelemek, Türkiye’de bir Banka’da kredi kartı başvuru aşamasında sahtecilik tespiti için kurulan bir veri madenciliği modelini ve bu modelin sağladığı faydaları paylaşmaktır.

Yapılan çalışma Giriş, Verinin Yararlı Bilgiye Dönüşüm Süreci, Veri Madenciliği, Veri Madenciliği Proje Safhaları, Veri Madenciliği Modelleri, Veri Madenciliği

Teknikleri, Farklı Endüstrilerdeki Veri Madenciliği Uygulamaları, Literatürde Bankacılık ve Finans Sektörüne Ait Uygulama Örnekleri, Türkiye’de Kredi Kartı Kullanımı ve Kredi Kartı Dolandırıcılık Tipleri, Türkiye’de Yapılmış Bir Veri Madenciliği Projesi: Kredi Kartı Başvurusu Aşamasında Sahtekarlık Tespiti ve Sonuç olmak üzere on bir bölümden oluşmaktadır.

İkinci bölümde, verinin dönüşüm süreci ve bu sürecin tüm aşamaları örneklerle açıklanmakta, verinin her anlam düzeyinde kullanılan teknolojiler belirtilmektedir.

Üçüncü bölümde, iş zekası, veri tabanlarında bilgi keşfi ve veri madenciliği tanımları yapılmakta, veri madenciliğine katkıda bulunan disiplinler ve veri madenciliğinin gelişimi anlatılmaktadır.

Dördüncü bölümde, veri madenciliği projesinin yedi safhası, bu safhalarda yapılması ve dikkat edilmesi gereken işlemler anlatılmaktadır.

Veri madenciliğinde kullanılan öngörüsel ve tanımsal modellerin tümüne, bu modellerin kullanım şekillerine ve konularına beşinci bölümde değinilmektedir.

Altıncı bölümde, bazı veri madenciliği tekniklerinden detaylı bir şekilde bahsedilmekte, bu tekniklerin kullanım şekilleri, kuvvetli ve zayıf yönleri anlatılmaktadır. Pazar sepeti analizi için gerçek hayatta bir perakendeciden alınmış fiş bilgileriyle yaptığım bir pazar sepeti analizi sonuçları paylaşılmaktadır.

Yedinci bölümde, veri madenciliğinin farklı endüstrilerdeki kullanım alanları anlatılmakta ve bu endüstrilerdeki firmalardan gerçek örnekler verilmektedir.

Sekizinci bölümde, literatürde bulunan bankacılık ve finans sektörüne ait uygulama örnekleri verilmiş, CHAID algoritmasının kullanıldığı bir kredi değerlendirme problemi, C&RT algoritmasının kullanıldığı bir kredi kartı başvurusu değerlendirme problemi ve destek vektör makinaları ile sinir ağlarının kullanıldığı bir kredi değerlendirme analizi anlatılmıştır.

Dokuzuncu bölümde, Türkiye’deki kredi kartı kullanımı hakkında bilgi verilmiş ve kredi kartı dolandırıcılık tipleri anlatılmıştır.

Onuncu bölümde, kredi kartı başvuru aşamasındaki sahtekarlık tespiti için Türkiye’de özel bir Banka’da yapmış olduğum veri madenciliği projesi anlatılmıştır. Sahtecilik tipleri ve bunlarla ilgili tespit süreçleri, Banka’daki proje öncesi mevcut durum analizleri, verinin hazırlanma aşamaları, modelin kurulması ve

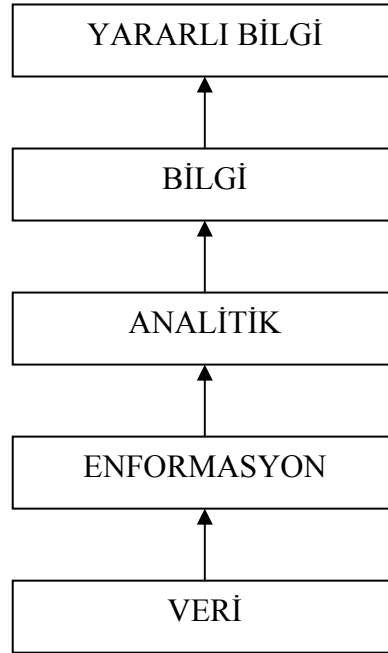
değerlendirilmesi, proje sonrası modelin kullanıma alınmasıyla değişen süreç ve modelin izlenmesi gibi proje içindeki tüm adımlar detaylı olarak belirtilmiştir. Projenin Banka'ya sağladığı faydalar anlatılmış ve rakamsal olarak gösterilmiştir.

Son bölümde, projenin sonuçları ve literatüre olan katkıları anlatılmış, projenin bir sonraki aşaması için öneride bulunulmuştur.

2. VERİNİN YARARLI BİLGİYE DÖNÜŞÜM SÜRECİ

Bir organizasyonun enformasyon sistemlerinde ve veritabanlarında birikmiş verilerde muazzam fırsatlar gizlidir. Bu verileri bilgiye dönüştürebilen organizasyonlar; trendleri belirleyebilir, daha önce fark edilmemiş örüntü (pattern) ve ilişkileri keşfedebilirler. Böylelikle de yeni fırsatlar oluşturarak en önemli rekabet avantajını yaratmış olurlar.

Verinin yararlı bilgiye dönüşüm süreci Şekil 2.1’de gösterildiği gibi birkaç aşamadan oluşur.



Şekil 2.1 : Veri Dönüşüm Süreci Aşamaları [1]

2.1. Veri (Data)

Veritabanı yönetim sistemlerinin keşfedilmesi ve veri saklama teknolojilerindeki ilerleme ile birlikte organizasyonlar, kolay analiz edilebilecek kişi, yer, işlem, kavram ve olaylar ile ilgili muazzam miktarlarda veri toplamaya, işlemeye ve saklamaya başlamışlardır. Verinin büyük bir kısmı organizasyonun fonksiyonel

prosesleri ile ilişkilidir [1]. Örneğin, bir market, müşterisinin almış olduğu ürünleri kasadan geçirerek, ilgili verileri kaydetmektedir. Kasada ürünler sisteme tanıtıldığında, sistem tarafından her bir ürünün fiyatı belirlenir ve ödenecek toplam tutar hesaplanır. Bu işlemle, ürün, miktar, fiyat, tarih, kayıt numarası, satışı gerçekleştiren personel ve müşterinin mağaza kartı bulunuyorsa, kart bilgisi verileri sistem tarafından kaydedilir. Tablo 2.1’de bir işleme ait örnek veriler gösterilmektedir.

Tablo 2.1: Veri İşlem Örneği

Ürün	Miktar	Fiyat (YTL)	Tarih	Kayıt No	Kullanıcı ID	Club Card ID
Çocuk bezi	1	18.00	25/1/2005	001	213	1209

2.2. Enformasyon (Information)

İşlem sayıları arttıkça toplanan veri miktarıda hızla artmaktadır. Her bir veri elemanı, işlemin bir bileşeni olduğundan, tek başına bir anlam ifade etmez. Verinin anlamlı bir şekilde biriktirilmesi enformasyonu sağlar. İş zekası (Business Intelligence) uygulamalarının sorgu ve raporlama yetenekleri sayesinde, veritabanındaki verinin çekilerek enformasyona dönüşümü sağlanır [1]. Örneğin, ürün, miktar ve fiyat toplamaları, satılan ürünlerle bunların miktar ve hacimleri enformasyonu sağlar. Tablo 2.2’de veriler, enformasyona dönüştürülecek şekilde özetlenmiştir.

Tablo 2.2: Verilerin Enformasyona Dönüşecek Şekilde Toplanması

Ürün	Miktar	Fiyat (YTL)	Satış Tutarı (YTL)
Bira	265	1.9	503.50
Makarna	430	0.9	387.00
Ekmek	850	0.35	297.50
Süt	1100	1.4	1,540.00
Çocuk bezi	200	18.00	3,600.00

2.3. Analitik (Analytic)

Enformasyonu yaratma aşamasında veri ile anlamını birleştirmek çok faydalı olurken, enformasyonu ayrıştırmak veya yeniden gruplamak onun değerini artırır. OLAP (online analytical processing) teknolojisi içeren uygulamalar sayesinde, ilişki, örüntü, eğilim ve istisnaların belirlenebilmesi için enformasyonun analiz edilmesi mümkündür [1]. Market örneğinde, bir önceki adımda elde edilen enformasyon Tablo 2.3’de gözüken şekilde dönemlere ayrıştırılarak analiz edilebilir.

Tablo 2.3: Enformasyonun Analiz İçin Ayrıştırılması

Ürün	Dönem 1	Dönem 2	Dönem 3	Dönem 4	Toplam Miktar	Fiyat (YTL)	Satış Tutarı
Bira	35	75	100	55	265	1.9	503.50
Makarna	110	110	100	110	430	0.9	387.00
Ekmek	200	215	235	200	850	0.35	297.50
Süt	200	300	300	300	1100	1.4	1,540.00
Çocuk bezi	10	20	50	120	200	18.00	3,600.00

Dönem bazında satılan ürün miktarlarını gösteren tablodan; bira ve çocuk bezi satışlarının dönemden etkilendiklerini, ancak makarna, ekmek ve süt satışlarının döneme bağlı olarak değişkenlik arz etmedikleri sonucuna varılabilir.

2.4. Bilgi (Knowledge)

Önceki aşamalardan elde edilebilirliği veya mevcut bilginin mantıksal çıkarımları sonucu oluşabilirliği nedeniyle bilgi, veri ve enformasyondan farklıdır. Veri madenciliği (data mining) teknolojisi içeren uygulamalar sayesinde, veri içerisindeki gizli eğilim ve örüntüler belirlenebilir [1]. Market örneğinde veri madenciliği teknikleri kullanılarak, çocuk bezi alan müşterilerin %50’sinin bira satın aldıkları bilgisi elde edilmiştir. Çocuk bezi alan müşterilerin mama da satın alacağını veya bira satın alanların kuruyemiş de satın alacağı tahmin edilebilir, ancak sadece otomatik bir analiz bütün olasılıkları göz önüne alır ve kolay düşünölmeyecek, örneğin çocuk bezi ve bira arasındaki bağlantıları da bulur.

2.5. Yararlı Bilgi (Wisdom)

Yaralı bilgi, toplanan bilginin yarar sağlayacak şekilde kullanılmasıdır [1]. Market örneğinde, alışılmışın dışındaki bira ve çocuk bezi ile ilgili örüntünün keşfedilmesi neticesinde alınabilecek aksiyonların tespiti için, bu sonuç analitik veri kümesinde analiz edilmelidir.

Tablo 2.4: Bir Satın Alma Örüntüsünün Teşhis Edilmesi

Ürün	Dönem 1	Dönem 2	Dönem 3	Dönem 4	Toplam Miktar
Bira	35	75	100	55	265
Çocuk bezi	10	20	50	120	200
İlişkilendirilmiş bira satışları	5	8	27	60	100

3. dönemde, *çocuk bezi alan müşterilerin %50'si bira satın alır* kuralına karşılık gelen miktarın üzerinde, 2 dönemde ise bu kurala karşılık gelen miktarın altında bira satışı gerçekleştirmiştir. 1.ve 4. dönemlerde ise bu satış kuralın öngördüğü şekilde gerçekleşmiştir. Bu sonuçlar neticesinde 4. ve 3. dönemlerdeki bira ile ilgili pazarlama kampanyalarının stratejileri kıyaslanarak, 3. dönemdeki etki ile 4. dönemdeki değişim analiz edilmelidir. Ayrıca diğer dönemlerdeki satışlar da büyüteç altına alınarak, öngörülen miktarın üzerinde gerçekleşen bira satışlarına katkıda bulunan diğer olaylar araştırılmalıdır. Bilginin bu şekilde kullanımı, kendisini oluşturan verinin daha iyi anlaşılmasını sağlar.

Operasyonel sistemdeki verileri biriktiren organizasyonlar, kendilerine değer katacak verideki potansiyeli anlama fırsatına sahip olurlar.

2.6. Verinin Anlam Düzeylerine Göre Kullanılan Teknolojiler

Verilerin anlam düzeyleri ve bunlara karşılık gelen teknolojiler Tablo 2.5’de sınıflandırılmıştır.

Yapay zeka, insanın düşünce işleyişini taklit etmeyi denediği halde, hiçbir teknoloji insan beyninin yerini alamaz. Organizasyonların çoğu veri proses sürecinin analitik aşamasındadırlar. Sadece, verinin ve teknolojinin değerini anlayan organizasyonlar, rekabet avantajı sağlayacak diğer aşamalara geçmiştir.

Tablo 2.5: Verilerin Anlam Düzeylerine Karşılık Gelen Teknolojilerin Sınıflandırılması [1]

Anlam düzeyi	Teknoloji
Veri	OLTP (Online transaction processing) sistemleri
Enformasyon	Sorgu ve raporlama uygulamaları
Analitik	OLAP uygulamaları
Bilgi	Veri madenciliği uygulamaları
Yaralı bilgi	İnsan zekası

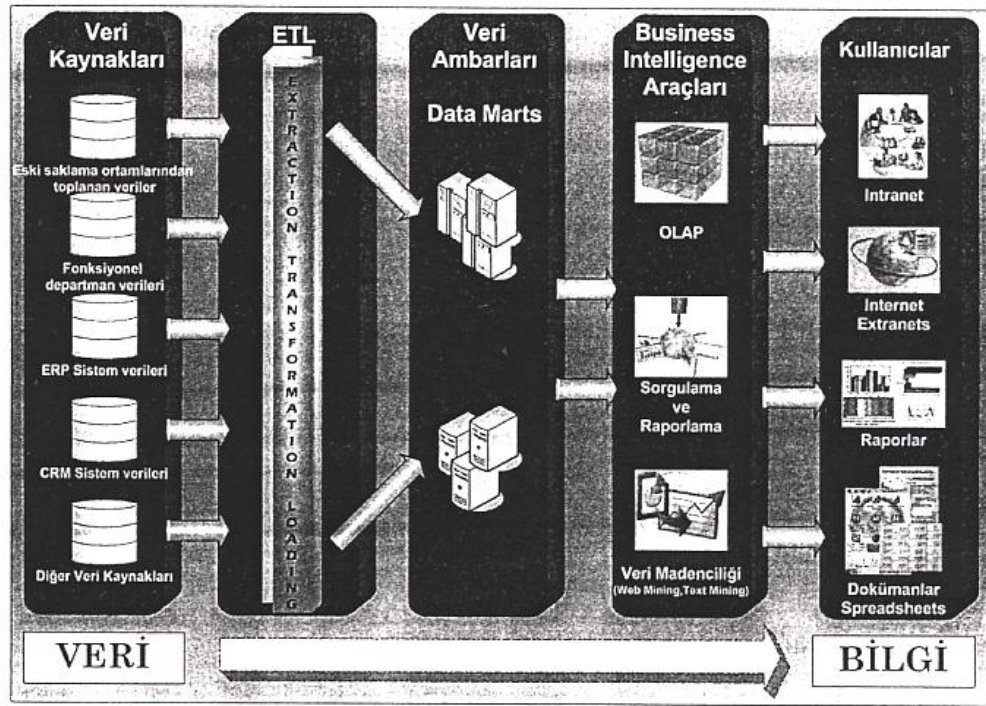
3. VERİ MADENCİLİĞİ

3.1. İş Zekası

Enformasyon sağanağının inanılmaz boyutlara ulaştığı günümüzde, işletme yöneticileri hızla değişen Pazar koşullarında, verilerden yararlanarak etkin stratejiler oluşturmak, belirli analizler sonucu performanslarını ölçmek, bu analizler sonucunda en hızlı ve optimal sonucu almak zorundadırlar. 90'lı yılların başında ilk defa Gartner Group tarafından dile getirilen İş Zekası kavramı bu amaçları hedeflemektedir. İşletme dünyasında olduğu kadar akademik dünyada da yeni bir kavram olarak karşımıza çıkan İş Zekası ile ilgili farklı tanımlara rastlanmaktadır. Gartner Group'a göre İş Zekası, enformasyona bilgi teknolojileri destekli her türlü erişimi ve karar destek amacıyla enformasyonun analizini gerçekleştiren tüm süreçleri ifade etmektedir.

Literatürdeki tüm kavram karmaşasına rağmen İş Zekası, işletmelerin karar verme süreçlerini etkileyen ve optimize eden tüm araçların kullanımını, verilerin toplanmasını, saklanmasını, düzenlenmesini, analiz edilmesini ve görselleştirilmesini, verilerin en etkin ve kolay biçimde yönetilmesini sağlayan tüm süreçleri kapsayan bir anlayıştır. Verilerden bilgiye geçiş sürecinde İş Zekası bir takım yöntem ve teknolojilerle bütünleşik bir yapı sunmaktadır [2].

Verinin saklanması, sorgulanmasından, analiz edilip yönetilmesine kadar bir çok aşamayı kapsayan bilgi kazanımını ve bilginin değerlendirilmesini sağlayan İş Zekası elemanları ve akışı Şekil 3.1'de görülmektedir.

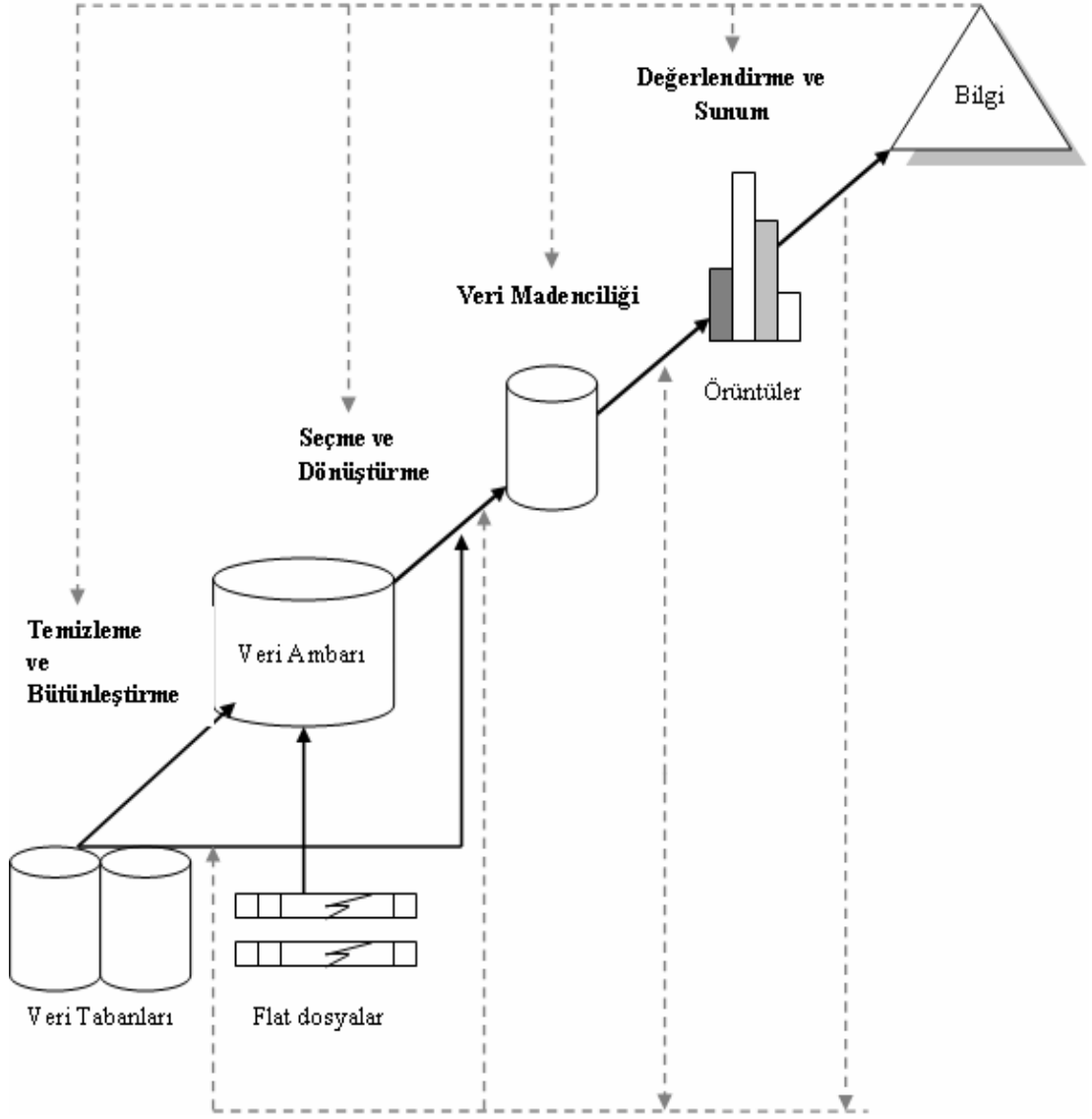


Şekil 3.1 : İş Zekası Elemanları ve Süreci [2]

3.2. Veri Tabanlarında Bilgi Keşfi (VTBK)

Veri tabanları sistemlerinin artan kullanımı ve hacimlerindeki olağanüstü artış, organizasyonları elde toplanan bu verilerden nasıl faydalanabileceği 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. Şekil 3.2’de görülen VTBK süreci içerisinde modelin kurulması ve değerlendirilmesi aşamalarından meydana gelen Veri Madenciliği (Data Mining) en önemli kesimi oluşturmaktadır. Bu önem, bir çok araştırmacı tarafından VTBK ile veri madenciliği terimlerinin eş anlamlı olarak da kullanılmasına neden olmaktadır [2].

VTBK, veri içerisindeki faydalı bilgi ve örüntüleri bulmaya yarayan bir süreç olarak tanımlanır. Veri madenciliği ise, VTBK süreci ile elde edilen bilgi ve örüntüleri seçip çıkarmak için algoritmaların kullanılmasıdır [3].



Şekil 3.2 : VTBK Süreci [4]

VTBK sürecinin adımları kısaca şöyledir.

- *Veri Temizleme (Data Cleaning)*: Kirlı, gürültülü ve tutarsız verinin ayıklanıp veri setinden uzaklaştırıldığı aşamadır.
- *Veri Bütünleştirme (Data Integration)*: Bu aşamada birden fazla veri kaynağı birleştirilebilir.
- *Veri Seçme (Data Selection)*: Analiz edilecek olan konuyla ilgili verinin veri tabanlarından çekilmesidir.
- *Veri Dönüştürme (Data Transformation)*: Verilerin madencilik için uygun biçimlere dönüştürüldüğü aşamadır.

- *Veri Madenciliği (Data Mining)*: Verilen bir veri madenciliği sorgusunun (sınıflama, öngörme vb.) işletilmesidir.
- *Örüntü Değerlendirme (Pattern Evaluation)*: Keşfedilen bilginin geçerlilik, yenilik, yararlılık ve basitlik kıstaslarına göre değerlendirilmesi aşamasıdır.
- *Bilgi Sunumu (Knowledge Presentation)*: Veri madenciliği ile elde edilen bilginin kullanıcıya aktarımı için görsellik ve sunum tekniklerinin kullanılmasıdır [4].

3.3. Veri Madenciliğinin Tanımı

Kendi iç performanslarını arttırdığı ve rekabetçi avantajlar sağladığı için veri madenciliği, organizasyonlar tarafından geniş çapta kabul gören bir süreç haline gelmiştir. Veri madenciliği nispeten yeni bir kavram olduğundan, yakın geçmişte bir çok yazarlar tarafından çeşitli şekillerde tanımlanmıştır.

“Veri madenciliği, VTBK sürecinde bir adımdır ve verideki örüntüleri ortaya çıkarmak için kullanılan algoritmaları kapsar. Ortaya çıkarılan bilgi daha sonra bir öngörü (prediction) veya sınıflandırma (classification) modeli kurmak, eğilimleri ve birliktelikleri belirlemek, mevcut bir modeli yenilemek veya üzerinde madencilik çalışması yapılmış bir veri tabanının özetini çıkarmak için kullanılabilir”[1].

“Veri madenciliği, önceleri bilinmeyen, geçerli ve etkin bilginin büyük veri tabanlarından çekilmesi ve daha sonra bu bilginin son iş kararlarını almak için kullanılmasını kapsayan bir süreçtir” [5].

“Veri madenciliği, aksi halde keşfedilemeyebilecek olan eğilimleri ve örüntüleri bulmak için, çok miktardaki verinin otomatikleştirilmiş analizidir” [6].

“Veri madenciliğini amacı, mevcut veri içindeki geçerli, alışılmamış, kullanışlı ve anlaşılır korelasyonları ve örüntüleri saptamaktır” [7].

“Veri madenciliği, muazzam boyuttaki veriden şirketlerin daha iyi kararlar almalarına yardımcı olup, pazarda rekabetçi olarak kalmalarını sağlayabilecek ilginç bilgileri keşfetme sürecidir”[8].

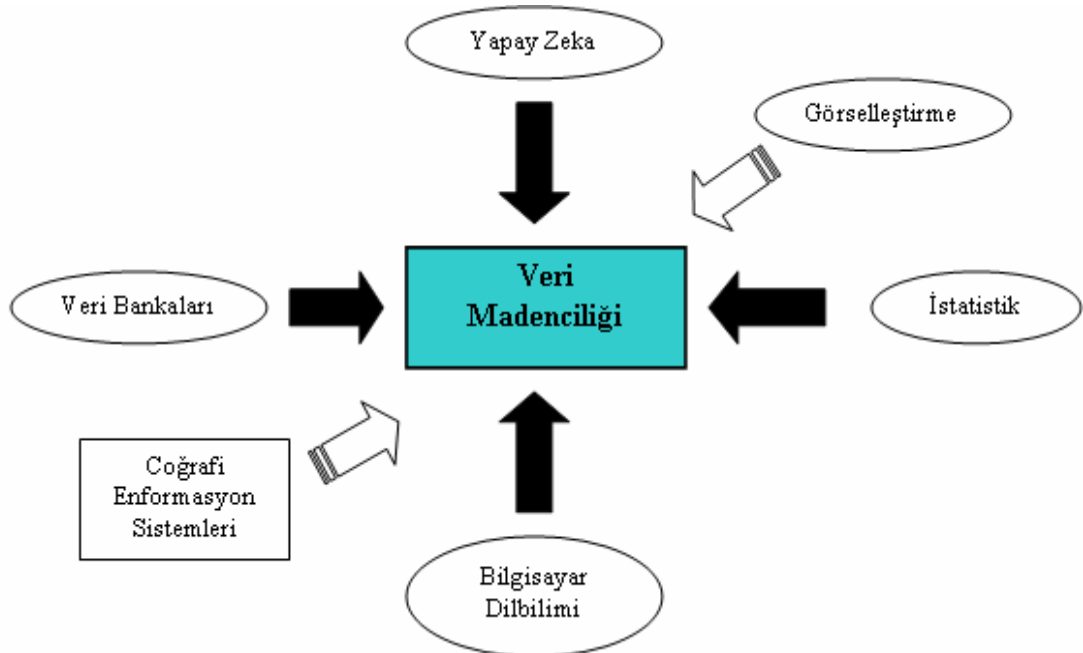
“Veri madenciliği, anlamlı örüntüler ve kurallar keşfetmek için büyük miktardaki veriyi, otomatik veya yarı otomatik yöntemlerle araştırma ve analiz etme sürecidir” [9].

Bu yazarlar tarafından verilen veri madenciliği tanımları biraz farklılık gösterebilir, hepsi aynı fikre sahiptir. Veri madenciliği, mevcut veriden önemli bilgileri açığa çıkartır ve bir organizasyonda daha iyi kararlar alınmasına olanak sağlar. Veri madenciliği sadece karar verme yetisini geliştirmez, aynı zamanda aşırı bilgi yüklemesinde önüne geçebilir.

Veri madenciliği, organizasyonların veri tabanlarında bulunan en önemli bilgilere odaklanabilmesine olanak sağlar. Bu sayede yöneticiler gelecekteki eğilimleri ve davranışları öngörerek daha bilgili kararlar alabilirler [10]. Veri madenciliği kullanılmadan önce, yöneticiler bu kadar bilgili kararlar alamıyorlardı. Bunun sebepleri ise, büyük miktarlarda verinin incelenmesinin çok pahalı ve zaman kaybettirici bir iş oluşuydu.

3.4. Veri Madenciliğine Katkıda Bulunan Disiplinler

Veri madenciliği bir çok bilim dalının katkılarıyla gelişen ve gelişmeye devam eden çok disiplinli bir daldır. Bu dalların içerisinde ana rolü Şekil 3.3’de görüldüğü gibi istatistik, dilbilim, veri tabanları ve yapay zeka üstlenmiştir. Ayrıca görselleştirme (Visualization) ve coğrafi enformasyon sistemleri diğer ana oyuncularındır.



Şekil 3.3 : Veri Madenciliğine Katkıda Bulunan Disiplinler

Veri madenciliği bir taraftan klasik ve modern istatistik yöntemlerini kullanırken, diğer taraftan özellikle yapay zeka başlığı altında gelişen makina öğreniminden çok

değerli kazanımlar elde etmektedir. Ancak ilgili analiz çalışmalarının çok güçlü veri tabanı yönetim sistemleri ile gerçekleştirilebileceği hiç bir zaman akıldan çıkartılmamalıdır. Çok boyutlu verilerin gösterilebilmesi ise etkin görselleştirme teknikleri ile mümkün olacaktır. Diğer taraftan veri madenciliği analizlerinin coğrafi veriler de dikkate alınarak yapılması, veri madenciliği alanında coğrafi veri tabanlarının (Spatial Database) ve coğrafi enformasyon sistemlerinin önemini arttırmıştır [2].

İstatistik alanında regresyon, faktör, kümeleme, ayırma (Discriminant) ve zaman serileri analizleri; yapay zeka alanında makina öğrenimi, yapay sinir ağları, genetik algoritmalar, zeki ajan sistemleri (Intelligent Agent Systems), bayes ağları, örüntü tanıma (Pattern Recognition) modelleri veri madenciliğine önemli katkılarda bulunmaktadır. Bilgisayar dilbilimi (Computer Linguistics) alanında ise web madenciliği (Web Usage Mining), metin madenciliği (Text Mining) ve vaka temelli çıkarım (Case Based Reasoning) veri madenciliğinde önemli rol oynayan alanlardır [2].

3.5. Veri Madenciliğinin Gelişimi

Veri madenciliğinin gelişimi, etkisinde olduğu disiplinlerdeki gelişmelerle birlikte boy göstermiştir. Yapay zeka, veri tabanları ve istatistik gibi veri madenciliğinin şu anki görünümüne öncülük eden alanlardaki bazı gelişmeler Tablo 3.1’de görülmektedir.

Gerek bilişim teknolojilerinin gerekse bu teknolojilerin modern hayatta kullanım alanlarının artmasıyla, kurumlarda saklanan veri miktarları da büyük bir hızla artmaya başladı. Büyük boyutlardaki bu veriyi saklamak için veri tabanları, ardından da veri ambarları yaygın olarak kullanılmaya başlandı. Verinin boyutu büyüdükçe veriden anlamlı bilgiye ulaşmak da giderek zorlaştı. İnsan gözüyle veya elle analizin imkansız olması, bilgiye erişmek için bilişim teknolojilerinin yoğun bir şekilde kullanılmasına yol açtı.

Tablo 3.1: Veri Madenciliğinin Gelişimi [3]

Zaman	Alan	Katkı
1700'lerin sonu	İstatistik	Olasılığa ait Bayes teoremi
1900'lerin başı	İstatistik	Regresyon Analizi
1920'lerin başı	İstatistik	Maksimum olasılık tahmini
1940'ların başı	Yapay zeka	Sinir ağları
1950'lerin başı		En yakın komşu
1950'lerin başı		Tek bağlantı
1960'ların başı	Veri tabanı	Toplu raporlar
1960'ların ortaları		Karar ağaçları
1960'ların ortaları	İstatistik	Sınıflama için lineer modeller Kümeleme
1960'ların sonları	Veri tabanı	İlişkisel veri modeli
1970'lerin ortaları	Yapay zeka	Genetik algoritmalar
1970'lerin sonları	İstatistik	Eksik veri ile tahmin
1970'lerin sonları	İstatistik	K-Ortalama kümelemesi (K-Means)
1980'lerin başları	Yapay zeka	Kohonen kendini düzenleyen haritalar
1980'lerin ortaları	Yapay zeka	Karar ağacı algoritmaları
1990'ların başı	Veri tabanı	Birliktelik kuralı algoritmaları Web ve arama motorları
1990'lar	Veri tabanı	Veri depolama (Data warehousing)
1990'lar	Veri tabanı	OLAP

Bera 2001, bu gelişim sürecini şöyle açıklamaktadır. “Bilgisayar destekli veri analizi ilk olarak 1960’larda anabilgisayarların büyük miktarlarda veri oluşturmasıyla gündeme gelmiştir. Anabilgisayarlar ve COBOL (Common Business Oriented Language) ile üretilen listeler bilgisayar destekli analizin ilk örnekleri olarak gösterilebilir.

İlerleyen yıllarda OLAP ile son kullanıcıya bir takım çözümleyici (analytical) araçlar verilmiştir. OLAP teknolojisinde veri, her boyutuna verinin bir alanı karşılık gelen çok boyutlu bir küpe benzetilir ve kullanıcı veride istediği seviyeye kadar inerek inceleme yapabilir. COBOL listelerinden daha fazla bilgi vermesine rağmen, OLAP teknolojisi ile sadece geçmiş veri incelenebildiği ve elde edilen sonuçlar kullanıcının aklına gelen sorgularla sınırlı olduğu için günümüz koşullarında yeterli olmamaktadır.

Veri analizinde bir sonraki adım 1980’lerde geleneksel algoritmalara dayalı istatistik araçları ile ortaya çıktı. Bu araçlar mükemmel sonuçlar üretmelerine rağmen, sadece

deneyimli istatistik uzmanları tarafından kullanılabilmekteydiler. Bu yüzden bazı firmalar kolay kullanımlı veri madenciliği araçları üretmeye çalıştılar. Ancak geleneksel algoritmalara dayalı bu araçların kullanımını kolaylaştırmak için bazı özelliklerinden vazgeçmek veya renkli kullanıcı ara yüzleri koymaktan daha ileriye gidemediler. Geleneksel yöntemlerin bir diğer sorunu da veri boyutu büyüdükçe modellerin güvenilirliğinin azalması idi.

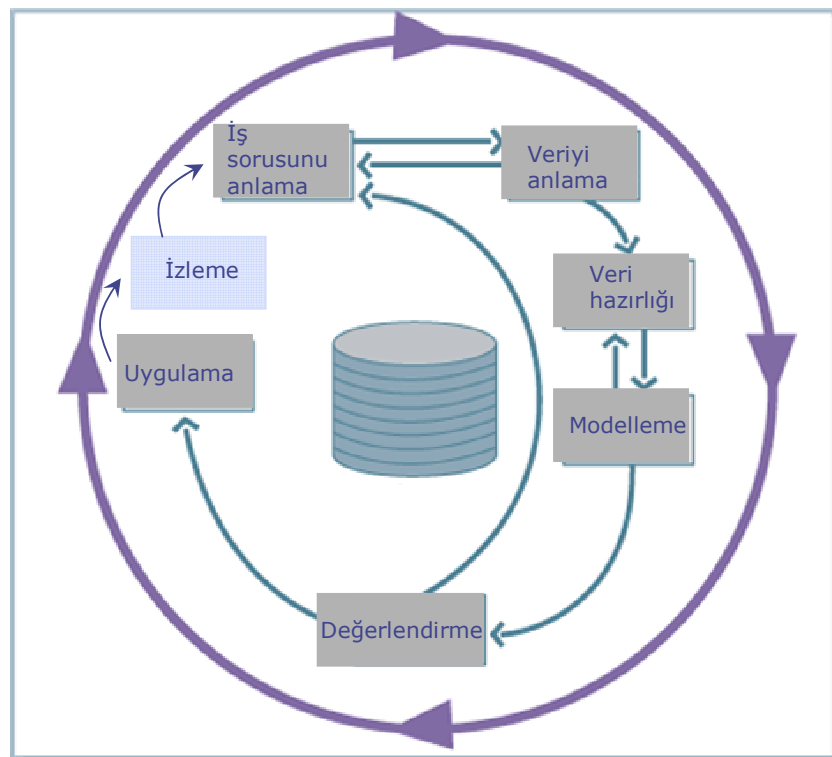
Daha sonra ortaya çıkan PCA (Principal Component Analysis) veya yapay sinir ağları (artificial neural networks) gibi belli bir kurama dayalı olmayan yöntemler yüksek boyutlu veride çok iyi sonuçlar veriyorlardı. Ancak bu yöntemlerin kapalı kutu olmaları ve sonuçları kesin olarak açıklayamamaları, istatistikçilerin bu yöntemlere şüpheyle yaklaşmalarına neden olmuştur.

Bilgisayar destekli veri analizinde bu gelişmeler yaşanırken, 1970'lerde Vladimir Vapnik ve bir grup Rus bilim adamı, İstatistik Öğrenme Teorisi'nin temelinde olan Vapnik-Chervonenkis (VC) boyutunu icat ettiler. Modellenen verinin dağılımına bağlı olmaksızın, VC boyutu ile modellerin güvenilirliği arasında bir bağlantı kurdular. Vapnik önce sınıflandırma (classification), daha sonra da regresyon problemlerini ele almış ve VC boyutunu kontrol ederek modelin kullanılan veriye uygunluğu ve yeni veride doğru sonuç vermesi arasında en iyi dengeyi sağlayan bir teknik geliştirmiştir [11].” Yapısal Risk Minimizasyonu (Structured Risk Minimization) adı verilen bu teknik Bölüm 6.5’de detaylı olarak anlatılmıştır.

4. VERİ MADENCİLİĞİ PROJESİ SAFHALARI

Veri madenciliğinin bir çok disiplini barındıran yapısı ve farklı uygulama alanlarındaki görevlerle prosedürlerin çeşitliliği, standart bir endüstri metodolojisi oluşturma yolunda karşımıza çıkan başlıca problemlerdir. Standart bir uygulama metodolojisi, teknoloji uygulamasını daha ucuz, daha güvenilir, daha kullanışlı ve daha hızlı bir hale getirebilir. Bunların dışında bir metodoloji, veri madenciliği teknolojisini daha kolay uyum sağlayabilir ve anlaşılabilir kılacaktır [12].

CRISP-DM (Cross Industry Standard Process for Data Mining) projesi, bir süreç modeli tanımlayarak bu problemlere hitap eder. CRISP-DM süreç modeli, Daimler Chrysler AG, SPSS, NCR ve OHRA gibi lider veri madenciliği kullanıcıları ve tedarikçilerinden oluşan bir konsorsiyum tarafından geliştirilmiştir.



Şekil 4.1 : CRISP-DM’ye Göre Veri Madenciliği Süreci [13]

CRISP-DM modeli, veri madenciliği projesini yaşam döngüsüne bir genel bakış sağlar. Bu model, projenin safhalarını, görevlerini ve sonuçlarını içerir. Veri madenciliği projesinin yaşam döngüsü, Şekil 4.1’de gözüktüğü gibi yedi safhaya ayrılmıştır.

4.1. İş Sorusunu Anlama

İlk safha proje amaçlarını ve iş gereksinimlerini anlamak, sonrasında da bu bilgiyi veri madenciliği problemi tanımına ve amaçlara ulaşmak için oluşturulan ilk proje planına dönüştürme üzerine odaklanır. Veri madenciliği projesindeki en zor kısmı olan bu safhada, ne tür bir analiz yapılması gerektiğinin kesinlikle iyi bir şekilde anlaşılması gerekir. Aksi takdirde tüm proje yanlışlıklar üzerine kurulmuş olacak ve bulunan sonuçlar amacı temsil etmeyecektir. Bu yüzden ilk ve en önemli adım, amacı açıkça belirlemek ve amaca giden süreci geliştirmektir. Amacı tanımlarken, neyi ölçmeye veya öngörmeye çalıştığımıza karar vermek gerekir.

Modelleri öngörüsöl ve tanımsal olmak üzere ikiye ayırmak mümkündür. Öngörüsöl modeller gelecekteki faaliyetleri gösteren bazı değerleri hesaplarlar. Bu değerler, satın alma tutarı veya bir teklife olumlu cevap verme olasılığı gibi sürekli değerlerdir. Tanımsal modellemeler ise adından da anlaşılacağı gibi belli tanımlamaları yapabilmek için kullanılır. Bu modeller, nesneleri farklı kategorilere de gruplayabilmek için kurallar oluştururlar. Günümüzde finans, perakende ve telekomünikasyon sektörlerinde kullanılan genel analitik amaçlardan bazıları aşağıda belirtilmiştir.

- Profil Analizi
- Segmentasyon
- Kampanya Optimizasyonu
- Risk Yönetimi
- Çapraz Satış ve Dikey Satış (Cross Sell & Up Sell)
- Sadakat (Churn) Analizi
- Sahtekarlık Tespiti
- Sepet Analizi

- Satış Tahminleri
- Müşteri Ömür Değeri

4.2. Veriyi Anlama

Veriyi anlama safhası öncelikle veriyi toplamakla başlar ve veri kümesinin içinde hangi değişkenlerin olduğunun saptanması, bu değişkenlerin ve değerlerinin neleri ifade ettiklerini anlamakla devam eder. Eğer analizci veriye hakim değil ise, projenin diğer aşamalarına geçmeden önce verilere hakim bir kişiden muhakkak yardım alıp veriyi anlamalıdır. Aksi takdirde, yanlış bir model oluşturmak söz konusu olabilir.

Modelleme için gerekli olan veri, iç ve dış olmak üzere iki kategoriye ayrılan veri kaynaklarından temin edilebilir. İç kaynaklar için müşteri veri tabanları, işlem veri tabanı, geçmiş teklifleri, siparişleri ve satın almaları içeren veri tabanları kullanılır. Dış veri kaynakları olarak ise nüfus sayımı, hava durumu, merkez bankası kara listesi gibi veri tabanlarından veya veri pazarlayan kuruluşların veri tabanlarından yararlanılabilir.

4.3. Veri Hazırlığı

Veri hazırlığı aşaması, ham veriden veri madenciliği aracında kullanılacak en son veri kümesini oluşturmak için yapılan tüm işlemleri kapsamaktadır. Veri madenciliğinin en önemli aşamalarından bir tanesi olan verinin hazırlanması aşaması analistin toplam zaman ve enerjisinin %70 - %80'ini harcamasına neden olmaktadır. Bu aşamadaki görevlerden bazıları; tablo oluşturma, kayıt ve değişken seçimi, veri temizliği, yeni değişkenler oluşturma ve modelleme araçları için verileri dönüştürme (transformation) işlemleridir.

4.3.1. Örnekleme (Sampling)

Bilgisayar teknolojisindeki gelişmeler örnekleme önemi azaltmıştır. Örneklem yapmadan da birçok analiz yapılabilmekte, ancak bu daha gelişmiş yazılımlara ve bilgisayar donanımlarına ihtiyaç doğurmaktadır. Örneklem oluşturmamak sürecin işleme ve zaman maliyetlerini de arttıracaktır. Örnekleme, süreci hızlandırdığı ve genellikle aynı sonuçları verdiği için yapmaktan kaçınılmaması gerek bir işlemdir.

Analiste kısa süre sonunda, projenin bitiminde ne tür sonuçlarla karşılaşacağı, veride hata olup olmadığı ve hangi değişkenlerin önemli olabileceği hakkında ön bilgi verir.

4.3.2. Veri Temizliği (Data Cleaning)

Veri nadiren temizdir. Her sektörde, üzerlerinde çalışarak belli kararlar almayı sağlayan verilerin kalitesinden şüphe edilebilir. Veri madenciliği sonuçlarının güvenilir oluşu, kullanılan verinin kalitesine bağlıdır. Veriyi aşağıda listelenen problemlerden arındırma işlemine veri temizleme işlemi denir.

- *Mükerrer veri kayıtları:* Sıkça karşılaşılan önemli bir sorundur. Örnek olarak 100,000 üyesi olan haftalık bir dergi üzerinde duralım. Derginin üyeleri arasında, Mustafa Aykut Göral, Mustafa A. Göral, Aykut Goral ve Mustafa Göral adlarında farklı kayıtların olduğunu ama aslında bu kayıtların tek bir kişiye ait olduğunu düşünelim. Mustafa Aykut Göral için yaratılan bu mükerrer kayıtlar diğer müşteriler içinde yaratılmışsa, dergiye her hafta gereksiz bir postalama masrafı yüklenecektir. Sonuç olarak bu kayıtlar tespit edilip, düzeltilmezse bu kayıpların önüne geçmekte imkansız olacaktır.
- *Yanlış veya Tutarsız Veri:* Veri setindeki yanlış ve tutarsız bilgilerdir. Örneğin, veri setinde müşteri ismi yerine firma adı, firma adı yerine müşteri adına yazılmışsa, böyle bir hatadan söz edebiliriz

Tablo 4.1: Yanlış Veri Örneği

<i>Müşteri adı</i>	<i>Yaş</i>	<i>Firma Adı</i>	<i>...</i>
<i>General Electric</i>	27	<i>Aykut</i>	...
Ahmet	65	Vestel	...
Mesut	35	Arçelik	...

- *Yazım Hataları:* Bilgisayar veriyi yazıldığı gibi algılar. İnsanoğlu yazım hatalarını okurken düzeltebilir ancak bir bilgisayarın bunu yapabilmesi için karmaşık algoritmalara ve bilgi tabanına ihtiyacı vardır. Bir çok veri tabanı hassastır ve kullanılan büyük harfler bile sorunlar çıkartabilir. Örneğin parantez içindeki veriler farklı yazım hatalarını göstermektedir. (Annkara, Ankara, ANKARa, anlara)

- *Eskimiş Veri:* Aktif olarak değişen verilerdir. Adres ve yaş değişkenleri eskimiş verinin tipik örnekleridir. Dünyadaki koşulların değişmesiyle birlikte de, elimizde bu tarz veriler oluşmaya başlar. Örneğin müşteri davranışları ve eğilimleri de zaman içinde değişiklik göstermektedir.
- *Terimleri Tanımlamadaki Farklılık:* Eğer veri farklı kaynakların birleşiminden oluşuyorsa, veri alanlarının tanımlarında farklılıklar gözükabilir. Örneğin verinin aynı ürünleri üreten iki farklı fabrikadan toplandığını kabul edelim. Böyle bir durumda, “çevrim süresi” alanı farklı prosedürler ve tekniklerle hesaplanmış olabilir. Bu sebeple bu süreler karşılaştırılmazlar.

4.3.3. Aykırı Değer Analizi (Outlier Analysis)

Bir veri tabanı verinin modeliyle veya genel yapısıyla uyuşmayan örnekler içerebilir. Veri kümesinden tamamen farklı veya onunla tutarsızlık gösteren bu tür örneklerle aykırı değerler (outlier) denir.

Aykırı değerler, ölçme veya uygulama hatalarından kaynaklanabilirler. Örneğin veri kümesinde bir kişinin yaşının “999” olarak gösterilmesinin sebebi, boş bırakılmış yaş bilgisinin programın başlangıç ayarlarından dolayı bu sayıyla doldurulması olabilir. Alternatif olarak aykırı değerler, verinin doğal değişkenliğini bir sonucuda olabilirler. Şöyle ki, bir firmanın genel müdürünün maaşı firmanın diğer çalışanlarının maaşlarıyla birlikte incelendiğinde aykırı bir değer gibi durabilir.

Bir çok veri madenciliği algoritması aykırı değerlerin etkisini minimize etmeye (enküçüklemeye) veya hepsini elemeye çalışmıştır. Ancak bu saklı halde bulunan önemli bilgilerin kaybıyla sonuçlanabilir çünkü aykırı değerlerin kendileri olağan dışı bir şekilde faydalı olabilirler. Şöyle ki, sahtekarlık tespiti durumlarında aykırı değerler sahte hareketleri işaret edebilirler. Bu sebeple aykırı değer tespiti ve analizi önemli bir veri madenciliği görevidir.

Aykırı değer analizinin geniş bir kullanım alanı vardır. Kredi kartlarının veya telekomünikasyon servislerinin alışılmadık kullanımlarını tespit ederek, sahtecilik tespitinde kullanılabilirler. Buna ek olarak, çok düşük gelirli veya çok yüksek gelirli müşterilerin harcama alışkanlıklarını belirlemek için pazarlama faaliyetlerinde çok kullanışlıdır. Ayrıca çeşitli tıbbi tedavilere verilen olağan dışı cevapları bulmak için tıbbi analizlerde de kullanılır [4].

4.3.4. Eksik Değerler (Missing Values)

Hemen hemen her veri kümesinde eksik veriler mevcuttur. Dikkat edilmesi gereken husus, değişken değerlerinin eksik olmasına rağmen anlam ifade edebileceğidir. Örneğin aylık maaş değişkeninde bir kişinin maaş bilgisinin eksik olması, bu bilginin müşteriden alınmamış olduğunu gösterebileceği gibi bu müşterinin şu an için çalışmadığını da gösterebilir. Bunun dışında sahte kimliklerle yapılan kredi kartı başvurularının tespiti sırasında kullanılan veri kümesinde bir kişiye ait eksik değerlerin oluşu, bu kişiyi daha risksiz kılar. Bunun sebebi, sahtekarların genelde herhangi bir sorun çıkmaması için tüm bilgilerini eksiksiz olarak bankaya sunmasıdır.

Bazı veri madenciliği teknikleri eksik değerlere sahip veri kümeleri üzerinde çalışabilirken, bir çok yazılım paketi eksik değerlere sahip kayıtları analize sokmamaktadır. Veriyi analize sokmadan önce yapılabilecek, eksik veriyle uğraşma seçenekleri aşağıda sunulmuştur [14].

- *Eksik değerli kayıtları atmak:* Bu metot eğer tüm örneklerin küçük bir yüzdesi eksik veri içeriyorsa uygundur.
- *Eksik değerleri sınıf ortalamasıyla değiştirmek:* Bir çok durumda sayısal değişkenler için makul bir yaklaşımdır. Eksik sayısal veriyi, sıfır ile veya keyfi olarak seçilmiş çok büyük yahut da çok küçük bir sayı ile değiştirme seçenekleri genellikle zayıf bir tercihtir.
- *Eksik değişken değerlerini kendisine çok benzeyen diğer kayıtlardaki değerlerle değiştirmek*

Bazı veri madenciliği teknikleri eksik değer içeren kayıtlarla çalışabilmektedir. Aşağıda bu tekniklerin öğrenme esnasında eksik veri ile nasıl uğraştıkları belirtilmiştir [14].

- *Eksik değerleri önemsememek:* Yapay sinir ağlarında içinde olduğu çeşitli bir çok veri madenciliği algoritması bu yaklaşımı kullanmaktadır.
- *Eksik değerleri birbirine eşit olarak ele almak:* Bu teknik, gürültülü (noisy) veri söz konusu olduğunda tehlikelidir. Birbirine benzemeyen kayıtların, birbirine oldukça benzer olarak görülme riski vardır.

- *Eksik değerleri birbirinden tamamen farklı olarak ele almak:* Bu kötümser bir yaklaşım olmasına rağmen bazı durumlarda uygun olabilmektedir. Birçok eksik değer içeren birbirine benzer kayıtlar, benzer değillermiş gibi gözükecektir.

4.3.5. Veri Dönüştürme (Data Transformation)

Kullanılacak model ve algoritma çerçevesinde verilerin tanımlama ve gösterim şeklinin de değiştirilmesi gerekebilir. Dönüştürme tekniklerinden biri normalizasyondur. Veri normalizasyonu, sayısal verileri belirlenmiş aralıkta tanımlayabilmek için yapılan değiştirme işlemidir. Yapay sinir ağları gibi sınıflandırma algoritmaları, 0 ile 1 arasında ölçeklenmiş sayısal verilerle daha iyi çalışmaktadırlar. Örnek olarak kişilerin gelir seviyesini verebiliriz [14].

Bunun dışında sinir ağlarını ve bazı istatistik algoritmalarını içeren bir çok veri madenciliği aracı, kategorik veriyi işleyememektedir. Bu yüzden kategorik verileri sayısal eşleniklerine çevirmek genel bir veri dönüştürme işlemidir.

4.3.6. Değişken Seçimi

Klasik veri madenciliği algoritmalarının birlikte çalışabildikleri değişken sayısı sınırlıdır. Bununla birlikte veri kümesine eklenen her bir değişken, analiz sürecini karmaşıktırıp, modelleme süresinin uzamasına sebep olacaktır. Dolayısıyla klasik algoritmalarla çalışan veri analistleri, bu tür problemlerle karşılaşmamak için analiz öncesinde değişkenler üzerinde kapsamlı çalışmalar yaparlar. Veri madenciliği algoritmalarının hedef üzerinde açıklayıcı etkisi az olan değişkenlerle çalışırken genellikle iyi performans göstermedikleri bilinmektedir. Bu yüzden ilk olarak bu değişkenler belirlenip, analize sokulmazlar.

Bu değişkenlerin dışında veri kümesinde birbirleriyle yüksek korelasyona sahip değişkenlerde bulunabilmektedir. Bir çok veri madenciliği aracı, yüksek korelasyona sahip değişkenler kümesinden sadece bir değişkeni girdi olarak alıp çalıştığında daha iyi modeller oluşturmaktadır. Dolayısıyla veri analistlerinin analiz öncesinde uğraştıkları bir diğer iş de, aralarında yüksek korelasyon bulunan değişkenleri bulmak ve birini seçerek diğerlerini analiz dışına atmaktır.

İyi bir model oluşturabilmek için veri kümesinden bazı değişkenleri atmanın yanında veri kümesine farklı değişkenleri eklemekte gerekebilir. Özellikle davranış öngörüsü

için kurulacak modellerde bu yeni verilere ihtiyaç duyulacaktır. Örneğin müşteri işlem (transaction) verisinden elde edilecek olan, son harcamadan bugüne kadar geçen süre, son harcama tutarı, son bir aydaki harcama adedi, son iki ayda harcamalarda gözüken yüzdesel artış/azalış gibi değişkenlerin türetilmesi, oluşturacak modele güç katacaktır. Eklenen bu değişkenler, veri madenciliği yazılımının faydalı ilişkiler keşfedebileceği yeni parametreler sağlamaktadır.

4.4. Modelleme

Bu safhada çeşitli modelleme teknikleri seçilip, uygulanır ve parametreler en uygun değerlere ayarlanır. Aynı tip veri madenciliği problemleri için çeşitli teknikler mevcuttur. Bazı teknikler belli veri formatlarına ihtiyaç duymaktadır. Bu yüzden genellikle veri hazırlama safhasına geri dönüş gerekir.

Klasik teknikler kullanıldığında 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.

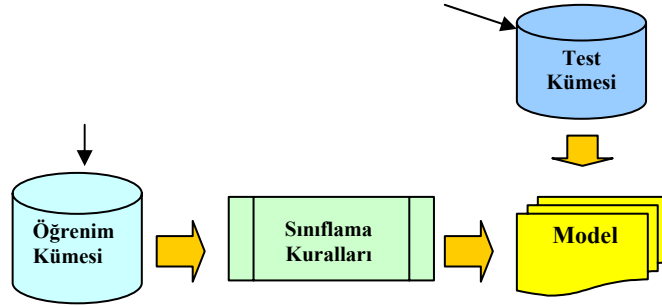
Artık günümüzde son teknolojileri bünyesinde barındıran bazı yazılımlar, karşılaştırılmaya gerek duyulmayan ve en iyi olduğu düşünülen tek bir modeli kullanıcıya sunabilmektedirler.

Model kuruluş süreci denetimli (*Supervised*) ve denetimsiz (*Unsupervised*) öğrenimin kullanıldığı modellere göre farklılık göstermektedir [15].

Örnekten öğrenme olarak da isimlendirilen denetimli öğrenimde, bir denetçi tarafından ilgili sınıflar önceden belirlenen bir kritere göre ayrılarak, her sınıf için çeşitli örnekler verilmektedir. Sistemin amacı verilen örneklerden hareket ederek her bir sınıfa ilişkin özelliklerin bulunması ve bu özelliklerin kural cümleleri ile ifade edilmesidir [16].

Öğrenme süreci tamamlandığında, tanımlanan kural cümleleri verilen yeni örneklerle uygulanır ve yeni örneklerin hangi sınıfa ait olduğu kurulan model tarafından belirlenir. Denetimli öğrenme süreci Şekil 4.2’de gösterilmiştir.

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



Şekil 4.2 : Denetimli Öğrenme

4.5. Değerlendirme

Projenin bu aşamasında analistin elinde kalitesi yüksek bir model mevcuttur. Modelin uygulama aşamasına geçmeden önce modelin eksiksiz olarak değerlendirilmesi ve iş amaçlarına ulaşıp ulaşılmadığından emin olmak için model oluşturulana kadar yürütülen adımların tekrar gözden geçirilmesi büyük önem taşımaktadır. Buradaki temel amaç, analiz süresince gözden kaçan önemli bir noktanın var olup olmadığını belirlemektir.

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) belirlenmektedir.

Bir modelin doğruluğunun test edilmesi için kullanılan en basit yöntem, basit geçerlilik (Simple Validation) testidir. Bu yöntemde verilerin %5'i ile %33'ü arasındaki bir kısmı test verileri olarak ayrılır ve kalan kısım üzerinde modelin öğrenimi gerçekleştirildikten sonra, bu veriler 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 ise doğruluk oranı hesaplanmaktadır [16].

$$\text{Doğruluk Oranı} = 1 - \text{Hata Oranı} \quad (4.1)$$

Sınırlı miktarda veriye sahip olunulması durumunda, kullanılabilecek diğ er bir y ntem  apraz ge erlilik (Cross Validation) testidir. Bu y ntemde veri k mesi tesad fi olarak iki e it par aya ayrılır. İlk a amada A par ası  zerinde model e itimi ve B par ası  zerinde test i lemi; ikinci a amada ise B par ası  zerinde model e itimi ve A par ası  zerinde test i lemi yapılarak, elde edilen hata oranlarının ortalaması kullanılmaktadır [16].

Bir ka  bin veya daha az satırdan meydana gelen k   k veri tabanlarında, verilerin n gruba ayrıldığı n katlı  apraz ge erlilik (N-Fold Cross Validation) testi tercih edilebilir. Verilerin  rne in 10 gruba ayrıldığı bu y ntemde, ilk a amada birinci grup test, di er gruplar ise   renim i in kullanılır. Bu s re  her defasında bir grubun test, di er grupların   renim ama lı kullanılması ile s rd r l r. Sonu ta elde edilen on hata oranının ortalaması, kurulan modelin tahmini hata oranı olacaktır [16].

Bootstrapping, k   k veri k meleri i in modelin hata d zeyinin tahmininde kullanılan bir ba ka tekniktir.  apraz ge erlilikte oldu u gibi model b t n veri k mesi  zerine kurulmaktadır. Daha sonra en az 200, bazen 1000'in  zerinde olmak  zere  ok fazla sayıda   renim k mesi tekrarlı  rneklemelemlerle veri k mesinden olu turularak hata oranı hesaplanmaktadır [16].

Model kurulu u  alı malarının sonucuna ba lı olarak, aynı teknikle farklı parametrelerin kullanıldığı veya ba ka algoritma ve ara ların denendi i de i ik modeller kurulabilir. Model kurulu   alı malarına ba lamadan  nce, imkansız olmasa da hangi tekni in en uygun oldu una karar verebilmek g  t r. Bu nedenle farklı modeller kurarak, do ruluk derecelerine g re en uygun modeli bulmak  zere sayısız deneme yapılmasında yarar bulunmaktadır.

 zellikle sınıflama problemleri i in kurulan modellerin do ruluk derecelerinin de erlendirilmesinde basit ancak faydalı bir ara  olan risk matrisi kullanılmaktadır. Tablo 4.2'de bir  rne i g r len bu matriste s tunlarda fiili, satırlarda ise tahmini sınıflama de erleri yer almaktadır.  rne in, fiilen B sınıfına ait olması gereken 46 elemanın, kurulan model tarafından 2'sinin A, 38'inin B, 6'sının ise C olarak sınıflandırıldığı matrisde kolayca g r lebilmektedir.

Tablo 4.2: Risk Matrisi

	Fiili		
Tahmini	A Sınıfı	B Sınıfı	C Sınıfı
A Sınıfı	45	2	3
B Sınıfı	10	38	2
C Sınıfı	4	6	40

Önemli diğer bir değerlendirme kriteri modelin anlaşılabilirliğidir. Bazı uygulamalarda doğruluk oranlarındaki küçük artışlar çok önemli olsa da, bir çok işletme uygulamasında ilgili kararın niçin verildiğinin yorumlanabilmesi çok daha büyük önem taşıyabilir. Çok ender olarak yorumlanamayacak kadar karmaşıklarsalar 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 [15]. Örneğin kredi kartını muhtemelen iade edecek müşterilerin belirlenmesi amacını taşıyan bir uygulamada, kullanılan modelin belirlediği 100 kişinin 40'ı gerçekten bir süre sonra kredi kartını iade ediyorsa ve tesadüfi olarak seçilen 100 müşterinin aynı zaman diliminde sadece 5'i kredi kartını iade ediyorsa kaldıraç oranı 8 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.

4.6. Uygulama

Modeli yaratmak genellikle veri madenciliği projesinin sonu değildir. Elde edilen bilginin düzenlenmesi ve müşterinin kullanacağı bir şekilde ifade edilmesi

gerekmektedir. İhtiyaçlara göre uygulama safhası, bir rapor üretimi kadar basit veya oluşturulan modelin başka sistemlerin içine entegre edilmesi kadar karmaşık olabilir. Birçok durumda, uygulama adımlarını gerçekleştirecek kişi veri analisti değil, kullanıcı olmaktadır.

Kurulan modeller risk analizi, kredi değerlendirme, dolandırıcılık tespiti gibi işletme uygulamalarında doğrudan kullanılabileceği 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.

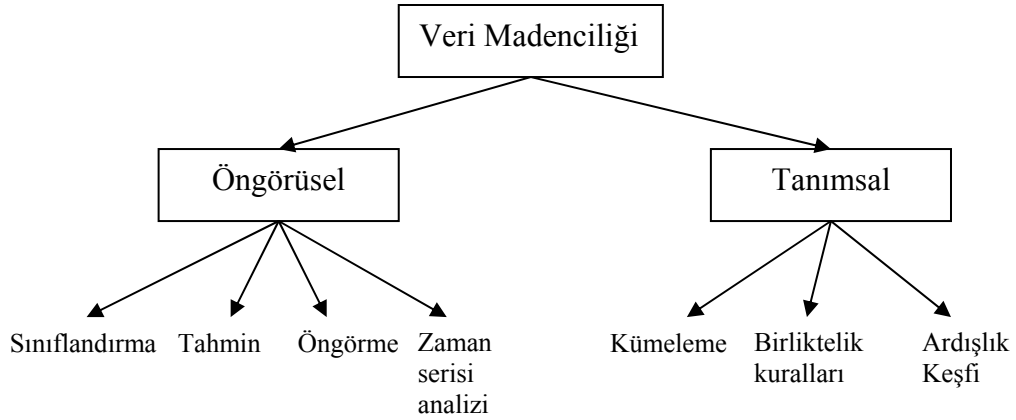
4.7. İzleme

Zaman içerisinde bütün sistemlerin özelliklerinde ve dolayısıyla ürettikleri verilerde değişiklikler ortaya çıkacaktır. Böyle bir durum karşısında modelin güncellenmesi gerekecektir. Günümüzde model güncelleme işleminin uzunca bir zaman alması ve efor gerektirmesi nedeniyle, işletmelerin çoğu bu tarz bir çalışma yapmamakta ve oluşturulmuş modelleri uzun yıllar boyunca kullanmaktadır. Sahtecilik tespit projesi için oluşturulan bir model düşünüldüğünde, bu modelin 1 yıl boyunca hatta 2 ay boyunca bile sürekli kullanılmaması gerektiği bir gerçektir. Bunun sebebi, sahtekarların yakalandıkça taktiklerini değiştirmeye başlayacak olmalarıdır. Dolayısıyla yeni taktiklerle gelen yeni sahtekarların profilleri, oluşturulan modelde bulunan sahtekar profiline uymayacak ve model bu sahtekarları yakalayamaz duruma gelecektir. Böyle bir durum ile karşılaşmamak için sahtecilik modellerinin belirli aralıklarla güncellenmesi gerekmektedir.

Sonuç olarak güncelliğini yitirmiş modellerin tespiti için bu modellerin sürekli olarak izlenmesi gerekmektedir. Tahmin edilen ve gözlenen değişkenler arasındaki farklılığı gösteren grafikler model sonuçlarının izlenmesinde kullanılan yararlı bir yöntemdir.

5. VERİ MADENCİLİĞİ MODELLERİ

Veri madenciliğinde kullanılan modeller, öngörüsöl (predictive) ve tanımsal (descriptive) olmak üzere iki ana başlık altında incelenmektedir.



Şekil 5.1 : Veri Madenciliği Modelleri [3]

Öngörüsöl modellerde, ilk olarak sonuçları bilinen verilerden bir model geliştirilmesi ve sonrasında kurulan bu modelden yararlanılarak, sonuçları bilinmeyen veri kümeleri için sonuç değerlerinin tahmin edilmesi amaçlanmaktadır. Örneğin bir banka önceki dönemlerde vermiş olduğu kredilere ilişkin gerekli tüm verilere sahip olabilir. Bu verilerde bağımsız değişkenler (girdiler, açıklayıcı değişkenler) kredi alan müşterinin özellikleri, bağımlı değişken (çıktı, hedef 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.

Tanımsal 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 birbirine benzerlik gösterdiğinin belirlenmesi tanımlayıcı modellere bir örnektir.

Veri madenciliği modelleri, gördükleri işlevlere göre aşağıdaki şekilde sıralanabilir:

- Sınıflandırma (Classification)
- Tahmin (Estimation)
- Öngörme (Prediction)
- Zaman Serisi Analizleri (Time Series Analysis)
- Kümeleme (Clustering)
- Birliktelik Kuralları (Association Rules)
- Ardışlık Keşfi (Sequence Discovery)

Veri madenciliği stratejileri, denetimli (supervised) ve denetimsiz (unsupervised) olarak da ikiye ayrılmaktadır. Denetimli öğrenme, hedef değişken değerlerini tahmin etmek için açıklayıcı değişkenleri kullanarak modelleri oluşturur. Sınıflandırma, tahmin ve öngörme modellerinde denetimli öğrenme söz konusudur. Denetimsiz öğrenmede ise, hedef değişken bulunmamaktadır. Dolayısıyla modeli kurmak için kullanılan tüm değişkenler, açıklayıcı değişkenlerdir. Kümeleme analizi denetimsiz öğrenme ile yapılmaktadır.

5.1. Sınıflandırma (Classification)

Sınıflandırma, veri madenciliği işlevleri arasında en yaygın olanlarından. İnsanoğlu dünya üzerindeki maddeleri daha iyi anlamak ve başkalarına anlatmak için hemen hemen herşeyi sürekli sınıflandırmakta, kategorilere ayırmakta ve derecelendirmektedir. Örneğin tüm canlıları türlerine ve cinslerine göre, maddeleri elementlerine göre kategorilere ayırmaktadır.

Sınıflandırma işleminde öğrenme denetimlidir ve hedef değişken değerleri kategorik veya ikili (0/1) değerlerden oluşur. Sınıflandırma işleminde amaç, yeni karşılaşılan bir girdinin özelliklerinin incelenip, bu girdinin daha önce tanımlanmış olan sınıflardan hangisine atanacağına karar vermektir.

Sınıflandırma işlemine örnek olarak verilebilecek bazı işlemler aşağıda sıralanmıştır:

- Kredi başvurularını risk derecelerine göre sınıflandırma
- Kredi kartı başvurusunun sahte olduğunu belirleme

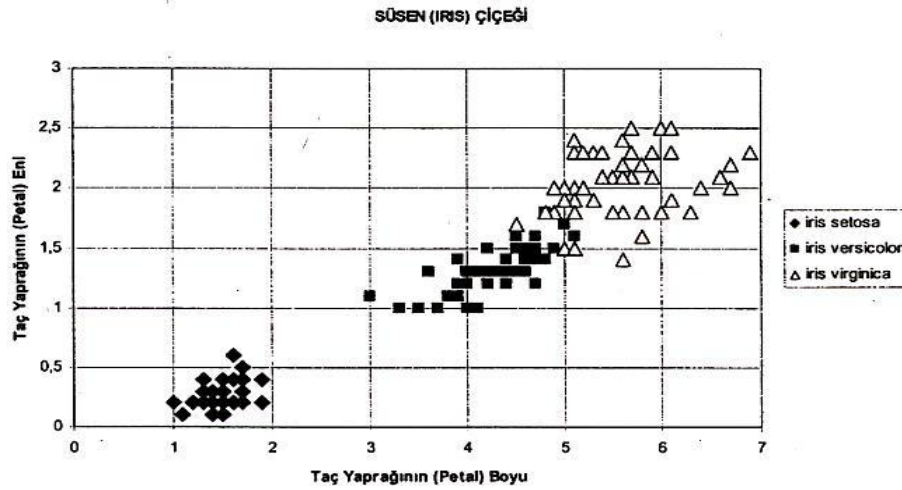
- Kampanyaya olumlu geri dönüş yapacak müşterileri seçme
- “Zengin” bir insanın profilini belirleme
- Kalp krizi geçiren bireylerin kalp krizi geçirmeyenlere göre farklılık gösteren özelliklerini belirleme

Fisher 1930’lu yıllarda yaptığı çalışmalarda, diskriminant analizinin prensiplerini gösterebilmek amacıyla 50 adet Iris Setosa, 50 adet Iris Versicolor ve 50 adet Iris Virginica isimli çiçeklerin taç ve çanak yapraklarının en ve boy uzunluklarına ilişkin veriyi kullanmıştır. Bu verinin bir kısmı Tablo 5.1’de gösterilmiştir.

Tablo 5.1: Iris Çiçeği Veri Kümesi [2]

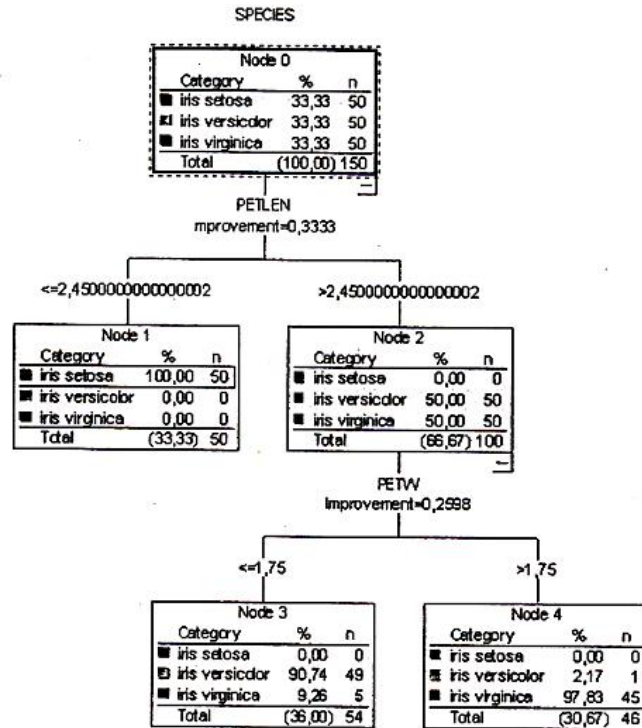
Tür	Taç Boy	Taç En	Çanak Boy	Çanak En
iris setosa	1,4	0,2	5,1	3,5
iris setosa	1,4	0,2	4,9	3
iris setosa	1,3	0,2	4,7	3,2
iris versicolor	4,7	1,4	7	3,2
iris versicolor	4,5	1,5	6,4	3,2
iris versicolor	4,9	1,5	6,9	3,1
iris versicolor	4	1,3	5,5	2,3
iris virginica	6	2,5	6,3	3,3
iris virginica	5,1	1,9	5,8	2,7
iris virginica	5,9	2,1	7,1	3
iris virginica	5,6	1,8	6,3	2,9

Bu veri kümesinde sadece taçyaprağının boyu ve taç yaprağının eni değişkenleri kullanıldığında elde edilecek olan XY-Grafiği Şekil 5.2’de sunulmuştur. Bu grafikte görüldüğü gibi Iris Setosa’yı diğer gruplardan kolayca ayırabilmek mümkündür. Bununla birlikte aynı ayırımı Iris Versicolor ve Iris Virginica arasında aynı kolaylıkta yapabilmek mümkün olmayacaktır.



Şekil 5.2 : Taç Yaprığının En ve Boyuna Göre XY-Grafığı [2]

Şekil 5.3'te bir karar ağacı yazılımı olan SPSS Answertree ile elde edilen sınıflandırma sonucu görülmektedir. Grafikten kolayca takip edilebileceği gibi ilk ayırım taç yaprağı boyuna (Petlen) göre yapılmış ve taç yaprağı boyu $\leq 2,45$ olan çiçekler sıfır hata Iris Setosa olarak sınıflandırılmıştır. Bir sonraki aşamada ise taç yaprağının enine (Petw) göre sınıflandırmaya devam edilmiş ve taç yaprağı eni $\leq 1,75$ olan çiçekler genelde Iris Versicolor olarak sınıflandırılmıştır. Ancak görüldüğü gibi bu seferki sınıflandırma bir önceki gibi hatasız olmayıp 5 adet Iris Verginica hatalı olarak bu gruba ayrılmıştır.



Şekil 5.3 : Bir Karar Ağacı Örneği [2]

5.2. Tahmin (Estimation)

Tahmin modelinin amacı, sınıflandırma işlevi gibi, bilinmeyen bir hedef değişken için bir değer belirlemektir. Aradaki fark, sınıflandırma işlevinde hedef değişkenler kategorik bir değere sahip iken, tahmin işlevinde söz konusu değerlerin süreklilik göstermesidir. Girdi olarak kullanılan veriden tahmin işlemi sonucunda gelir, boy veya kredi kartı bakiyesi gibi bilinmeyen ancak süreklilik arz eden değişkenlik arz eden değişkenler için değer üretilir.

Tahmin işlevinin pratikte kullanımı ile ilgili bazı örnekler:

- Bir ailedeki toplam gelirin tahmini
- Bir ailedeki toplam çocuk sayısının tahmini
- Bir kredi kartının çalınmış olma olasılığının tahmini
- Bir müşteri ile devam eden ilişkinin ömür değeri tahmini
- Müşterilerin yeni banka hesap tipleri seçme olasılığının tahmini
- Bir yıldırım fırtınasının belirlenen bir yere kaç dakika sonra ulaşacağını tahmini

Denetimli veri madenciliği tekniklerinin bir çoğu, sınıflandırma veya tahmin problemlerinden birini çözmekte, fakat ikisine birden aynı anda çözüm bulamamaktadır. Günümüzde son teknolojilerini barındıran veri madenciliği araçlarıyla bu sorun biraz olsun aşılmış olsa da, eğer veri madenciliği aracı bu stratejilerden sadece birini destekliyor ise genellikle problem bu iki stratejiden birine göre düzenlenir [14]. Örneğin yukarıda örnek olarak verilen çalıntı kredi kartı problemiyle ilgili veri kümesindeki hedef değişken değerlerinin sayısal olduğunu düşünelim. Ayrıca hedef değişken değerlerinin 0 ile 1 arasında değiştiğini ve 1 değerine sahip olmanın kesinlikle çalıntı kartı ifade ettiğini kabul edelim. Bu durumda hedef değişken değerleri 0.0 ve 0.3 arasında olanlar “risksiz”, 0.3 ve 0.7 arasında olanlar “riskli” ve 0.7 ‘den büyük olanlar “en riskli” olarak değiştirilerek farklı kategoriler yaratılabilir. Böylelikle kullanılan teknik sadece sınıflandırma problemlerini çözebiliyorsa, bu probleme uygun bir veri hazırlanmış olur.

5.3. Öngörme (Prediction)

Öngörme işlevini sınıflandırma ve tahmin işlevlerinden ayırmak çok kolay değildir. Aralarındaki en önemli fark ise öngörü modellerinde temel amacın, mevcut davranışları değil gelecekteki davranış ve değerleri belirlemesi oluşudur. Bu modellerde hedef değişken değerleri kategorik veya sayısal olabilir.

Öngörme işlevinin pratikte kullanımı ile ilgili bazı örnekler:

- İlk altı ayda sizi terk etme olasılığı yüksek olan müşterilerin öngörülmesi
- Telefon bankacılığını kullanacak müşterilerin öngörülmesi
- EFT yapacak müşterilerin öngörülmesi
- Yeni ürün talep edecek müşterilerin öngörülmesi
- Kredi limiti artırımı isteyebilecek müşterilerin öngörülmesi

5.4. Zaman Serisi Analizi (Time Series Analysis)

Zaman serisi, zaman içinde gözlemlenen geçmiş ölçümlerin bir dizisidir [3].

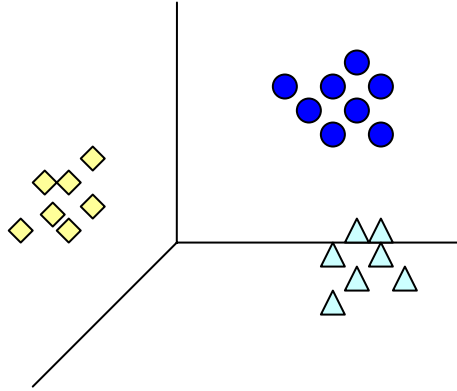
Zaman serisi analizi, bir değişkenin zamana bağlı olarak değişen değerlerini inceleyerek gelecekte alacağı değerleri tahmin etmektedir. Modelin kurulduğu öğrenme veri kümesindeki hedef değişkenin değerleri (sinyal), belli periyotlara (yıl, ay, gün, saat, dakika vb.) göre verilir. Analiz sonrasında öngörülen sonuçlar da aynı periyotla elde edilir. Örneğin haftalık verilere sahipseniz, oluşturduğunuz model size haftalık tahminleri verecektir. Zaman serisi analizi ayrıca, eğer veri setinde mevcut ise eğilimleri, mevsimsellikleri ve devresellikleri tespit eder.

Zaman serisi analizinin pratikte kullanımı ile ilgili bazı örnekler:

- ATM'lerdeki nakit yönetimi
- Ürünlerin, bayilerin, tedarikçilerin, rakiplerin satışlarının tahmini
- Pazar payı, maliyet, kar tahmini
- Makine bozulma zamanları tahmini
- Önleyici bakım zamanlarının tahmini

5.5. Kümeleme (Clustering)

Kümeleme modellerinde amaç, tanımlanan veriler ışığında nesnelerin birbirlerine olan benzerlik ve farklılıklarına göre kümelere ayrılmasıdır. Şekil 5.4’de görüldüğü gibi analizde aynı küme içerisinde yer alan küme elemanlarının olabildiğince birbirine benzer (homojen), farklı küme elemanlarının ise olabildiğince birbirine benzemez (heterojen) olarak gruplandırılması sağlanır.



Şekil 5.4 : Kümeleme

Kümeleme işlemi sınıflandırma işlemine benzerlik gösterir. Aralarındaki fark, kümeleme işleminde önceden tanımlanmış sınıfların olmayışındır. Kümeleme alternatif olarak denetimsiz öğrenme veya segmentasyon (segmentation) olarak adlandırılır [3].

Denetimli öğrenmenin tipik temsilcisi olan sınıflandırma modellerinde sınıfların sayısı bellidir. Bölüm 5.1’de verilen Iris çiçeği örneğinde, herbir kayıta ilgili ölçülerin hangi çiçeğe ait olduğu belirtilmiştir. Böylece sınıflandırma modelinin amacı tanımlanan grup için profil ölçülerinin belirlenmesi olacaktır. Buna karşılık kümeleme analizinde önceden belirlenen sınıflar yoktur. Kimi kümeleme modellerinde sınıf sayısının baştan kullanıcı tarafından belirlenmesi de istenebilir. Iris örneği kümeleme analizinde kullanılacak olursa bu defa sadece çiçeklerin taç ve çanak yapraklarının en ve boy uzunlukları verilecek, buna karşılık sınıflandırmada olduğu gibi bu ölçülerin hangi çiçeğe ait olduğu tanımlanmayacaktır.

Kümeleme işlemi, çoğunlukla veri madenciliğinin diğer yöntemleri veya modelleme çeşitleri için bir başlangıç niteliğinde kullanılır. Örneğin pazar segmentasyonu çalışmasında kümeleme ilk adım olabilir. “Müşteriler en çok hangi promosyona ilgi gösterebilirler?” sorusu yerine, müşterileri benzer alışveriş alışkanlıklarına göre

kümelere ayırdıktan sonra, “Her bir müşteri kümesi için en uygun promosyon hangisidir?” sorusu sorulabilir.

5.6. Birliktelik Kuralları (Association Rules)

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, pazarlama amaçlı olarak pazar sepeti analizi (Market Basket Analysis) adı altında veri madenciliğinde yaygın olarak kullanılmaktadır.

Birliktelik kuralları aşağıda sunulan örneklerde görüldüğü gibi eş zamanlı olarak gerçekleşen ilişkilerin tanımlanmasında kullanılır.

- Müşteriler gömlek satın aldığında, %75 ihtimalle kravat da satın alırlar,
- Az yağlı peynir ve yağsız yoğurt alan müşteriler, %85 ihtimalle diet süt de satın alırlar.

Pazar sepeti analizi perakende sektöründe pazarlama, çapraz satış, stok kontrolü ve raf düzenleme gibi işlemler için sıkça kullanılmaktadır. Bu analiz perakende dışında bir çok sektörde de uygulanabilir.

- Aynı kredi kartı üzerinden yapılan alışverişlerin incelenmesi ve müşterilerin yapacakları bir sonraki potansiyel harcama kalemlerini bulmak
- Bireysel müşterilerin kullandıkları ev kredisi, döviz hesabı vb. bankacılık hizmetlerini inceleyerek kurallar yaratıp, bu kurallarla müşterilerin şu an için kullanmadıkları hangi hizmetlere ilgi gösterebileceklerini bulmak
- Cep telefonundaki opsiyonel hizmetlerin (GPRS, WAP, telesekreter, burç yorumu vb.) müşteriler tarafından tercih edilmelerine göre, karı arttırmak için hangi ürünlerin birlikte kampanyaya girmesi gerektiğini belirlemek
- Hastaların sağlık kayıtlarından, bir arada yapılan tedavilerin kaçınılmaz yan etkilerini belirlemek.

5.7. Ardışık Keşfi (Sequence Discovery)

Ardışık keşfi, verideki ardışık zamanlı örüntüleri (Sequential Patterns) belirlemek için kullanılır. Bu örüntüler birlikteliklere benzerlik gösterirler ancak aralarındaki ilişki zamana dayanmaktadır. Ürünlerin aynı zamanda alınmasına gereksinim duyan pazar sepet analizinin aksine, ardışık keşfinde ürünler zaman içerisinde herhangi bir sırada satın alınır.

Ardışık keşfinin pratikte kullanımı ile ilgili bazı örnekler:

- Web sitesi davranışlarını öngörmek:

Müşteri şimdi ne yapacak?

Müşterinin bir sonraki en muhtemel alışverişi ne olacak?

- Segmentler arası geçişleri inceleyerek müşterinin hayat boyu değerini hesaplamak

Firmadan ayrılmak üzere olan bir müşterinin bulunduğu duruma gelmesine nasıl bir olaylar sırası neden oldu?

- Sahtekarlık tespiti

Sahtekarlık tespitinde en anlamlı işlem sırası nedir?

Keşfedilebilecek örüntüleri neler olabileceğini ve bunlardan nasıl faydalanabileceğimizi gözümüzde canlandırabilmek için bir örnek olay düşünelim. Bu örnekte ABC şirketinin bilgi işlemcisi, müşterilerin ABC şirketinin web sayfalarından kendilerine nasıl ulaştıklarını anlamak için periyodik olarak web kullanım verisini inceliyor olsun. Kişinin asıl hedefi sıkça ziyaret edilen sayfaların ardışıklık ilişkisini bulmaktır. Kişi yaptığı analiz sonunda görmüştür ki, A sayfasını ziyaret eden kullanıcıların %70'i (A, B, C) veya (A, D, B, C) veya (A, E, B, C) sırasındaki sayfaları kullanmaktadır. Bulunan bu örüntülere göre kişinin bundan sonraki adımı, A sayfasına bu sayfadan C sayfasına direk geçiş sağlayacak bir link koymak olacaktır.

6. VERİ MADENCİLİĞİ TEKNİKLERİ

6.1. Pazar Sepeti Analizi

Pazar sepeti analizi tekniği, çok basit olarak hangi ürünlerin hangi ürünlerle satıldığını, hangi ürünlerin promosyona girmesi gerektiğini ve benzeri bilgileri ortaya çıkarır. Pazar sepet analizi tekniğinin kullanım alanlarıyla ilgili bilgiler Bölüm 5.6’da belirtilmiştir.

Sepet analizi çoğunlukla ticari anlam taşıyan verilerin var olduğu ancak bu veri üzerinde hangi örüntülerin aranılacağı bilinmediği durumlarda bir başlangıç noktası olarak kullanılır. Bu veri içerisindeki bazı kalıplar sayesinde kazancı artırmak üzere bazı aksiyonlara gidilebilir.

Örneğin yurtdışında yapılan sepet analizi tekniklerine göre Perşembe günleri bira ve çocuk bezi satışlarının çok fazla sayıda olduğu görülmüştür [17]. Bunun temel nedeni olarak ise, evli çiftlerin hafta sonunu evde geçirmek istemeleri ve bu süre içerisinde gerekli olması muhtemel bira ve çocuk bezini hafta sonu gelmeden almak istemeleri olarak gösterilmiştir.

Ticari anlam taşıyan veriler üzerinde belirli bir ürün kombinasyonunun kaç defa geçtiğinin bulunması işlemi tek başına yeterli değildir. Bu kombinasyonu işletme açısından anlamlı hale getirecek olan kilit nokta, bu kombinasyonu oluşturan kuralı bulmaktır.

Kural tanımı, koşul kısmı ve sonuç kısmı olmak üzere iki kısımdan oluşmaktadır.

Eğer KOŞUL doğru ise, SONUÇ da doğrudur.

Örneğin “gömlek satın alan bir müşteri, kravat da satın almıştır” kuralı kısaca şu şekilde gösterilir:

Gömlek $\Rightarrow$ Kravat

Pratikte işletme açısından eyleme dönüştürülebilecek kuralların sonuç kısımlarında sadece bir adet parça / ürün bulunur. Yani

Çocuk bezi & Perşembe günü $\Rightarrow$ *Bira* kuralı,

Perşembe günü $\Rightarrow$ *Bira & Çocuk bezi* kuralından daha çok faydalıdır. Çünkü sadece günün Perşembe olmasından dolayı müşteriye çocuk bezi veya bira satmaya çalışmak anlamsız olacaktır. Aksine eğer günlerden Perşembe ve müşteri çocuk bezi almış ise bu müşterinin bira alma olasılığı çok yüksek demektir. Bu sebeple işletme bira satışlarını arttırmak için Perşembe günleri çocuk bezi ürünleri ile biraları beraber satmak üzere promosyona girebilir. Dolayısıyla ikinci kural işletme açısından çok daha anlamlı ve eyleme dönüştürülecek yapıya sahiptir.

Oluşan kuralların kullanılabilirliği ve kalitesi ise destek (support) ve güven (confidence) ölçütleri ile tanımlanır. $X \Rightarrow Y$ kuralını düşündüğümüzde

Destek, tüm işlemler içinde X ve Y ürünlerinin birlikte gözlemlendiği işlemlerin yüzdesini ifade etmektedir [3].

$$\text{Destek } (X \Rightarrow Y) = P(X \cup Y) \quad (6.1)$$

Güven ise, X ürününü içeren bir işlemin aynı zamanda Y ürününü de içermesi olasılığını göstermektedir [3].

$$\text{Güven } (X \Rightarrow Y) = P(Y|X) = \frac{P(Y \cap X)}{P(X)} \quad (6.2)$$

Kuralın kullanılabilir olması için, kullanıcı veya uzman tarafından belirlenmiş minimum destek ve minimum güven eşik değerlerinin elde edilmiş olması gerekir [4].

Pazar sepet analizi tekniği kullanılıp kurallar oluşturulduktan sonra, bu kuralların kullanımı genelde işletmeciye bırakılmaktadır. Şöyle ki, bir giyim mağazası için analiz yapıldığını düşünelim ve burada elde edilen güvenilir kurallardan biri bize “gömlek alan müşterilerin kravat da aldığını” ($Gömlek \Rightarrow Kravat$) anlatsın. Mağaza işletmecisi bu kurala göre değişik stratejiler belirleyebilir. İlk strateji olarak, bu ürünlerin birlikte alımını sağlamak için, gömlek ve kravat reyonları bir birine çok yakın bir şekilde oluşturulabilir. İkinci bir strateji olarak ise, bu ürünler mağazaların farklı köşelerine koyulup, iki reyon arasına müşterilerin alması istenilen diğer ürünler

yerleştirilebilir. Örneğin bu mağazada gömlek ve kravat reyonları arasına, kemer, kol düğmesi ve kazak reyonları koyulabilir. Böylelikle, gömlek aldıktan sonra kravat almayı isteyen müşteriler, kravat reyonuna giderken diğer ürünleri de görüp almak isteyebilirler.

Türkiye’de bulunan bir süpermarket zinciri için yaptığım pazar sepeti analizinde, bir veri madenciliği yazılımı olan KXEN Analytic Framework’ü kullandım. Yapılan bu analizde 4,700 adet alış veriş fişindeki bilgiler kullanılmış ve her fişe ait ürünler Tablo 6.2’de de gözüktüğü gibi gruplandırılmıştır.

Analizi yapmadan önce belli parametreleri aşağıdaki gibi belirledim:

- *Minimum Destek (Minimum Support) = 100 adet*

“Reyon 1 ve Reyon 2 $\Rightarrow$ Reyon 3” diye bir kuralın, Reyon 1, Reyon 2 ve Reyon 3 bilgilerinin en az 100 adet farklı faturada birlikte gözükmesi halinde oluşacağını ifade eder.

- *Minimum Güven (Minimum Confidence) = %50*

“Reyon 1 ve Reyon 2 $\Rightarrow$ Reyon 3” diye bir kuralın sadece, Reyon 1 ve Reyon 2’den alış veriş yapmış müşterilerin en az %50’sinin Reyon 3’den de alış veriş yapması halinde oluşacağını ifade eder.

- *Maksimum Uzunluk (Maximum Length) = 6*

Oluşturulan kurallar en fazla 6 adet reyondan oluşur. Bir başka deyişle, bunlardan sadece 1 tanesi sonuç kısmında olacağından, koşul kısmında en fazla 5 adet reyon yer alabilir.

Analiz sonuçlandığında ise Tablo 6.1’deki bilgilere ulaşılmıştır.

Tablo 6.1: Örnek Sepet Analizi Sonuçları

Association Rules: Modeling Results	
Rules Found:	375
Items Found:	32
Frequent ItemSets:	248
Sessions Treated:	4700
Transactions Found:	28763

Bu sonuçlara göre analiz sonucunda, parametrelere uyan 375 adet kural bulunmuştur. Bunun dışında 248 tane kural, güven eşik değerini geçemediği için raporlanamamıştır. Ayrıca analiz edilen veri setinde 32 tane farklı Reyon bulunduğu,

analiz için 4,700 adet fatura ve bu faturalara ait 28,763 adet işlemin incelendiği bilgilerine ulaşılmıştır.

Tablo 6.2: Fatura Bilgileri

Fiş No	Reyon Adı
1	BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER
1	KIRMIZI ET
1	KURUYEMİŞLER
1	MANDIRA
1	SÜTLÜK
1	ZÜCCACİYE/HEDİYELİK EŞYA/MUHTELİF EV EŞYALARI
2	BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER
2	MANAV
2	MANDIRA
2	SICAK İÇECEKLER
2	SİGARA
2	SÜTLÜK
2	ŞEKER VE ŞEKERLİ MAMÜLLER
2	UNLU MAMULLER
3	MANAV
3	MANDIRA
3	SICAK İÇECEKLER
3	ŞEKER VE ŞEKERLİ MAMÜLLER
3	UNLU MAMULLER
4	BAHARATLAR
4	BEBEK_GIDA
4	BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER
4	KATKI MADDELERİ
4	KOZMETİK VE KİŞİSEL BAKIM
4	SICAK İÇECEKLER
4	SÜTLÜK
4	TEMİZLİK ÜRÜNLERİ
4	ZÜCCACİYE/HEDİYELİK EŞYA/MUHTELİF EV EŞYALARI
5	KÜMES HAYVANLARI
5	MANAV
5	MANDIRA
5	SÜTLÜK
6	SICAK İÇECEKLER
6	SÜTLÜK
6	ŞEKER VE ŞEKERLİ MAMÜLLER
7	BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER
7	KATKI MADDELERİ
7	KOZMETİK VE KİŞİSEL BAKIM
7	KURU GIDALAR
7	MANDIRA
7	ŞEKER VE ŞEKERLİ MAMÜLLER
7	TEMİZLİK ÜRÜNLERİ
8	KONSERVE GIDALAR
8	KURUYEMİŞLER

8	KÜMES HAYVANLARI
8	MANAV

4693	BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER
4693	KONSERVE GIDALAR
4693	MANAV
4693	SÜTLÜK
4693	UNLU MAMULLER
4694	BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER
4694	KIRMIZI ET
4694	MANAV
4694	MANDIRA
4694	SİGARA
4694	SÜTLÜK
4694	ŞEKER VE ŞEKERLİ MAMÜLLER
4694	UNLU MAMULLER
4695	KONSERVE GIDALAR
4695	SÜTLÜK
4695	TEMİZLİK ÜRÜNLERİ
4695	ZÜCCACİYE/HEDİYELİK EŞYA/MUHTELİF EV EŞYALARI
4696	KÜMES HAYVANLARI
4696	MANAV
4696	MANDIRA
4696	SÜTLÜK
4696	UNLU MAMULLER
4697	BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER
4697	MANDIRA
4697	SÜTLÜK
4697	TEMİZLİK ÜRÜNLERİ
4698	BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER
4698	MANAV
4698	MANDIRA
4698	SÜTLÜK
4698	UNLU MAMULLER
4698	ZÜCCACİYE/HEDİYELİK EŞYA/MUHTELİF EV EŞYALARI
4699	SOĞUK İÇECEKLER
4699	SÜTLÜK
4699	TEMİZLİK ÜRÜNLERİ
4699	UNLU MAMULLER
4699	ZÜCCACİYE/HEDİYELİK EŞYA/MUHTELİF EV EŞYALARI
4700	BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER
4700	SOĞUK İÇECEKLER
4700	ŞEKER VE ŞEKERLİ MAMÜLLER

Analiz sonucu elde edilen 375 adet kuraldan bazıları gözlemlenme yüzdelere göre Tablo 6.3’de, gözlemlenme sayılarına göre Tablo 6.4’de gösterilmiştir. Oluşturulan kuralların tümü Ek A’da sunulmuştur.

Tablo 6.3: Gözlem Yüzdesiyle Gösterilen Kurallar

Rules		Confidence	KI	Rule Support	Consequent Support	Antecedent Support
Rule 4	MANDIRA -> SÜTLÜK	79,40%	0,2908	38,04%	65,72%	47,91%
Rule 26	MANDIRA -> MANAV	71,85%	0,1818	34,43%	63%	47,91%
Rule 23	SÜTLÜK -> MANAV	69,25%	0,1761	45,51%	63%	65,72%
Rule 92	UNLU MAMULLER -> MANDIRA	56,65%	0,1719	27,83%	47,91%	49,13%
Rule 182	SÜTLÜK & MANAV & BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER -> SOĞUK İÇECEKLER	50,07%	0,1626	14,66%	37,11%	29,28%
Rule 3	UNLU MAMULLER -> SÜTLÜK	72,50%	0,1477	35,62%	65,72%	49,13%
Rule 112	SÜTLÜK & TEMİZLİK ÜRÜNLERİ -> SOĞUK İÇECEKLER	50,51%	0,1434	12,62%	37,11%	24,98%
Rule 10	KURU GIDALAR -> SÜTLÜK	80,14%	0,1330	16,66%	65,72%	20,79%
Rule 318	MANDIRA & SOĞUK İÇECEKLER & TEMİZLİK ÜRÜNLERİ -> ŞEKER VE ŞEKERLİ MAMÜLLER	50,87%	0,1317	4,98%	25,38%	9,79%
Rule 58	SİGARA -> BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER	58,38%	-0,0051	10,30%	59,09%	17,64%

Tablo 6.4: Gözlem Sayısıyla Gösterilen Kurallar

Rules		Confidence	KI	Rule Support	Consequent Support	Antecedent Support
Rule 4	MANDIRA -> SÜTLÜK	79,40%	0,2908	1788	3089	2252
Rule 26	MANDIRA -> MANAV	71,85%	0,1818	1618	2961	2252
Rule 23	SÜTLÜK -> MANAV	69,25%	0,1761	2139	2961	3089
Rule 92	UNLU MAMULLER -> MANDIRA	56,65%	0,1719	1308	2252	2309
Rule 182	SÜTLÜK & MANAV & BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER -> SOĞUK İÇECEKLER	50,07%	0,1626	689	1744	1376
Rule 3	UNLU MAMULLER -> SÜTLÜK	72,50%	0,1477	1674	3089	2309
Rule 112	SÜTLÜK & TEMİZLİK ÜRÜNLERİ -> SOĞUK İÇECEKLER	50,51%	0,1434	593	1744	1174
Rule 10	KURU GIDALAR -> SÜTLÜK	80,14%	0,1330	783	3089	977
Rule 318	MANDIRA & SOĞUK İÇECEKLER & TEMİZLİK ÜRÜNLERİ -> ŞEKER VE ŞEKERLİ MAMÜLLER	50,87%	0,1317	234	1193	460
Rule 58	SİGARA -> BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER	58,38%	-0,0051	484	2777	829

Tablo 6.3 ve Tablo 6.4’de ifade edilen Kural Destek (Rule Support) ifadesi kuralın gözlem adedini, Sonuç Destek (Consequent Support) ifadesi kuralın sonuç kısmının gözlem adedini ve Önceki Destek (Antecedent Support) ifadesi de kuralın koşul kısmının gözlem adedini sırasıyla yüzdesel ve adet olarak göstermektedir.

Buna göre 318 numaralı “Mandıra & Soğuk İçecekler & Temizlik Ürünleri $\Rightarrow$ Şeker ve Şekerli Mamüller” kuralı incelendiğinde;

- Mandıra, Soğuk İçecekler ve Temizlik Ürünleri reyonlarının üçünüde ziyaret eden müşterinin, Şeker ve Şekerli Mamüller reyonunu da ziyaret edebileceği,
- 1193 müşterinin (Sonuç Destek) veya müşterilerin %25.38'inin Şeker ve Şekerli Mamüller reyonunu ziyaret ettiği,
- 460 müşterinin (Önceki Destek) veya müşterilerin %9.79'unun Mandıra, Soğuk İçecekler ve Temizlik Ürünleri reyonlarının üçünüde ziyaret ettiği,
- 234 müşterinin (Kural Destek) veya müşterilerin %4.98'inin, bu dört reyonun hepsine uğradığı,
- Mandıra, Soğuk İçecekler ve Temizlik Ürünleri reyonlarının üçünede uğrayan 460 müşterinin 234 tanesinin Şeker ve Şekerli Mamüller reyonuna da uğradığı ve dolayısıyla kuralın güven değerinin %50.87 olduğu anlaşılır.

Bu tablolarda gözüken bir diğer ifade de KI göstergesidir. KI, KXEN tarafından üretilen ekstra bir göstergedir ve Kaldıraç (Lift) değerinin bir fonsiyonudur. Kurallarda, koşul kısmındaki ürünlerle sonuç kısmındaki ürün arasındaki ilişkiyi gösterir. Bir örnek üzerinden gidelim ve $X \Rightarrow Y$ kuralına sahip olduğumuz düşünelim.

X ve Y ürünleri arasındaki ilişkiyi ölçmek için kuralın güven değerini, $P(Y|X)$, sonuç kısmındaki ürün, $P(Y)$, ile kıyaslayabiliriz. Bu ölçüm Kaldıraç değerine veya onun bir fonksiyonu olan KI değerine bakılarak yapılabilir.

$$\text{Lift} = \frac{P(Y|X)}{P(Y)} \quad (6.3)$$

Eğer Kaldıraç değeri;

- < 1 ise X ürünü, Y ürününün satışına negatif etki yapar
- $= 1$ ise X ürünü, Y ürününün satışına etki yapmaz
- > 1 ise X ürünü, Y ürününün satışına pozitif etki yapar

Sayısal bir örnek vermek gerekirse, Gömlek $\Rightarrow$ Kravat kuralını ele alalım ve $P(\text{Kravat}|\text{Gömlek}) = \%80$ olsun. Bir başka deyişle, gömlek alan müşterilerin %80'i kravat alsın.

Böyle bir durumda Kravat'ın tek başına satın alınma oranı, $P(\text{Kratat})$, eğer

- %90 olursa Gömlek almak Kravat alma eğilimi azaltacağından,
- %80 olursa Gömlek almanın Kravat almaya bir etkisi olmayacağından,
- %60 olursa Gömlek almak Kravat almak için pozitif bir etki yapacağından bahsedebiliriz.

Bu kural için çıkabilecek Kaldıraç değerinin ≥ 2 oluşu, Gömlek ve Kravatın birlikte olma olasılığının, Kravatın yalnız olma olasılığından en az 2 kat daha fazla olacağını ifade edecektir.

Kaldıraç değeri ile KI değeri arasındaki ilişkiyi ise aşağıdaki gibi ifade etmek mümkündür.

- Kaldıraç değeri < 1 ise $KI < 0$
- Kaldıraç değeri $= 1$ ise $KI = 0$
- Kaldıraç değeri > 1 ise $KI > 0$

Bu sebeplerden dolayı Tablo 6.3 ve Tablo 6.4'de ifade edilen KI değerleri büyük önem teşkil etmekte ve kuralları bu değere göre de incelemek gerekmektedir. 4 numaralı kurala ait 0.29'luk KI değerinin en yüksek KI değeri olduğu gözükmemektedir. Bu değer, Mandıra reyonundan yapılan alışverişin Sütçük reyonunda yapılan alışverişe pozitif bir etkisi olacağını göstermektedir. 58 numaralı kural incelendiğinde ise, Sigara almanın Bisküvü/Kek/Çikolata/Şeker ürün gruplarından alışveriş yapmaya negatif bir etkisi olacağını görebiliriz.

Sepet analizinin başarılı olduğu noktalar:

- Açık ve anlaşılabilir sonuçlar üretir.
- Denetimsiz veri madenciliği yöntemidir.
- Değişik boyutlardaki veriler üzerinde çalışır.
- Gerekli olan hesaplamalar diğer yöntemlere (yapay sinir ağları, genetik algoritmalar vb.) göre çok daha basittir.

Sepet analizinin başarısız olduğu noktalar:

- Problemin boyutu büyüdükçe, gerekli hesaplamalar üstel olarak artmaktadır.

- Sepet analizinde kullanılacak doğru ürünlerin seçimi. Ürün gruplandırma (süt ürünleri, unlu mamüller vb.) biraz bilgi kaybı getirirse de analizin boyutlarını küçültebilir.
- Kayıtlarda çok az rastlanan ürünleri yok sayar.

6.2. Karar Ağaçları

Tahmin edici ve tanımlayıcı özelliklere sahip olan karar ağaçları, veri madenciliğinde

- Kuruluşlarının ucuz olması,
- Yorumlanmalarının kolay olması,
- Veri tabanı sistemleri ile kolayca entegre edilebilmeleri,
- Güvenilirliklerinin daha iyi olması

nedenleri ile sınıflama modelleri içerisinde en yaygın kullanıma sahiptir.

Karar ağacı temelli analizlerin yaygın olarak kullanıldığı sahalara,

- Belirli bir sınıfın muhtemel üyesi olacak elemanların belirlenmesi,
- Çeşitli vakaların yüksek, orta, düşük risk grupları gibi çeşitli kategorilere ayrılması,
- Gelecekteki olayların tahmin edilebilmesi için kurallar oluşturulması,
- Parametrik modellerin kurulmasında kullanılmak üzere çok miktardaki değişken ve veri kümesinden faydalı olacakların seçilmesi,
- Sadece belirli alt gruplara özgü olan ilişkilerin tanımlanması,
- Kategorilerin birleştirilmesi ve sürekli değişkenlerin kesikliye dönüştürülmesidir.

Karar ağacı temelli tipik uygulamalar ise,

- Hangi demografik grupların mektupla yapılan pazarlama uygulamalarında yüksek cevaplama oranına sahip olduğunun belirlenmesi (Direct Mail),
- Bireylerin kredi geçmişlerini kullanarak kredi kararlarının verilmesi (Credit Scoring),

- Geçmişte işletmeye en faydalı olan bireylerin özelliklerini kullanarak işe alma süreçlerinin belirlenmesi,
- Tıbbi gözlem verilerinden yararlanarak en etkin kararların verilmesi,
- Hangi değişkenlerin satışları etkilediğinin belirlenmesi,
- Üretim verilerini inceleyerek ürün hatalarına yol açan değişkenlerin belirlenmesidir [16].

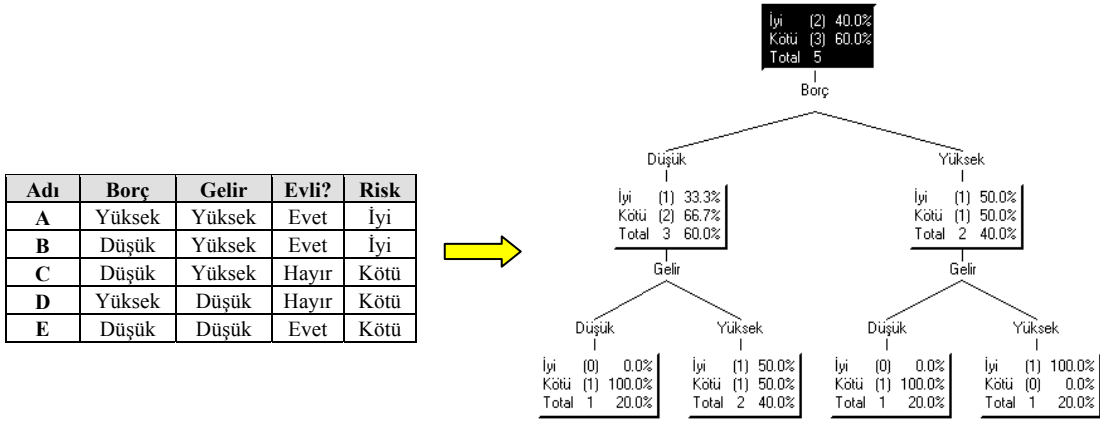
Bazı uygulamalarda, sınıflandırmanın ya da öngörünün doğruluğu önemli olan tek şeydir. Örneğin doğrudan posta ilanları ile iş yapan bir firma, hangi müşterilerin kendilerine gönderilen ilanlara olumlu yanıt vereceğini öngören bir model sahibi olduğunda bu modelin nasıl veya neden çalıştığını sorgulamaz.

Karar ağaçlar, bir dizi soru sorup bunların cevapları doğrultusunda hareket ederek en kısa sürede sonuca gider. Karar ağaçları, sorduğu bir soruya gelen cevap ile soracağı diğer soruları belirler. Eğer sorular iyi seçilmiş olursa, yeni gelen bir kaydın sınıflandırılması işlemi, en az sayıda soru sorarak gerçekleştirilebilir.

Sorulacak sorular ve bu sorulara gelebilecek cevapların yönlendirdiği başka soruların bulunduğu bir ağaç yapısı olarak adlandırılan karar ağaçları ile değerlendirme yaparken, yeni gelen bir kayıt ağacın kökünden giriş yapar. Kökte test edilen bu yeni kayıt, yapılan testin sonucuna göre bir alt düğüme gönderilir. Bu süreç, yeni kayıt herhangi bir yaprak düğüme gelene kadar devam eder. Ağacın belirli bir yaprağına gelen bütün yeni kayıtlar aynı şekilde sınıflandırılırlar. Kökten her bir yaprağa giden sadece tek bir yol vardır. Bu yol, kayıtları sınıflandırmak için kullanılan bir kuralı tanımlamaktadır. Bazı yapraklar aynı sınıflandırmayı yapabilirler fakat her bir yaprak bu sınıflandırmayı farklı nedenlere dayanarak yapar.

Gerçek dünyanın sosyal ve ekonomik olaylarını daha güvenilir bir şekilde gösterebilmek için standart istatistik tekniklerin dışında yeni analiz tekniklerinin geliştirilmesi ile ilgilenen Morgan ve Sonquist tarafından University of Michigan'da 1970'li yılların başlarında kullanıma alınan Automatic Interaction Detector – AID, karar ağacı temelli ilk algoritma ve yazılımdır. AID tekniği en kuvvetli ve en iyi tahmini gerçekleştirebilmek için bağımlı ve bağımsız değişkenler arasındaki mümkün bütün ilişkilerin incelenmesine dayanmaktadır. Şekil 6.1'de görüldüğü gibi en kuvvetli ilişkiye sahip bağımsız değişken bulunduğunda, veri kümesi bu bağımsız

değişken değerlerine göre ikiye ayrılmakta ve süreç mümkün bölünmeler tamamlanıncaya kadar devam etmektedir. Karar ağacı tekniğinin sağladığı kuruluş ve yorumlama kolaylıkları, AID yazılımının başlangıçta istatistikçi ve veri analistleri tarafından büyük coşku ile karşılanmasına neden olmuştur [16].



Şekil 6.1 : Verilerden Karar Ağacına [16]

Ancak AID'in bağımlı ve bağımsız değişkenler arasındaki ilişkilerin tanımlanmasında aşırı saldırgan davrandığı ve bunun sonucunda anlamlı ve anlamsız ilişkileri ayırt edemediği yönünde Einhorn başta olmak üzere bir çok araştırmacı tarafından yayınlar yapılmıştır [16].

İlk temelleri AID yöntemi ile atılan karar ağacı modelleri çeşitli algoritmalar ile sürdürülmüştür. Geliştirilen bu algoritmalar içerisinde CHAID (Chi-Squared Automatic Interaction Detector; G.V. Kass; 1980), C&RT (Classification and Regression Trees; Breiman, Friedman, Olshen ve Stone; 1984), ID3 (Quinlan; 1986), Exhaustive CHAID (Biggs, de Ville ve Suen; 1991), C4.5 (Quinlan; 1993), MARS (Multivariate Adaptive Regression Splines; Friedman), QUEST (Quick, Unbiased, Efficient Statistical Tree; Loh ve Shih, 1997), C5.0 (Quinlan), SLIQ (Supervised Learning in Quest; Mehta, Agarwal ve Rissanen), SPRINT (Scalable Parallelizable Induction of Decision Trees; Shafer, Agrawal ve Mehta) başlıcalarıdır [16].

Bu algoritmalarından CHAID ve C&RT hakkında sırasıyla Bölüm 8.1 ve Bölüm 8.2'de kısa bilgi verilmiş, kullanıldıkları birer uygulamadan da bahsedilmiştir. Ayrıca yine Bölüm 5.1'de bir sınıflandırma örneği, karar ağacı kullanılarak anlatılmıştır.

Karar ağaçlarının güçlü olduğu noktalar şunlardır:

- Üretilen sonuçlar kolayca anlaşılabilir.

- Denetimli öğrenme için kullanılan bir tekniktir.
- Sonuçlar kurallara dönüştürülebilir.
- Çok sayıda işlem yapılmasına gerek duymadan sınıflandırma işlemini gerçekleştirebilir.
- Hem kategorik (nominal / ordinal) hem de sayısal veriler üzerinde işlem yapabilmektedir.
- Karar ağaçları, sınıflandırma ve tahmin problemleri için hangi değişkenlerin daha önemli olduğunu açıkça ifade etmektedir.

Karar ağaçlarının zayıf olduğu nokta ise şöyledir:

- Karar ağaçlarının öngörü için kullanıldığı çalışmalarda, öngörü yapılacak değişkenin sürekli değerler alması durumunda uygun sonuçlar üretilmemektedir.

6.3. Yapay Sinir Ağları (Artificial Neural Networks)

Yapay sinir ağları (YSA), veri madenciliği ve karar destek sistemlerinde önceden kanıtlanmış başarılarından dolayı yaygın olarak kullanılmaktadır. YSA, sınıflandırma, öngörü ve kümeleme modellerinde doğrudan uygulanabilen çok güçlü bir yöntemdir.

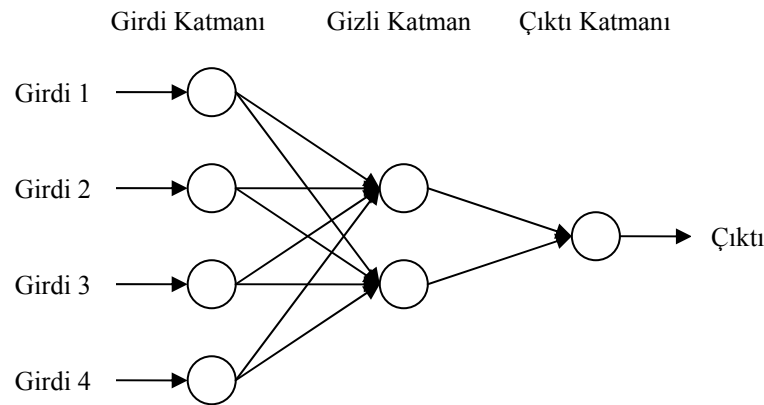
Mali serilerin tahmininden sağlık durumlarının teşhisine, değerli müşterilerin belirlenmesinden kredi kartı sahtekarlıklarının tespitine, el yazısı formlarının değerlendirilmesinden makina arızalanma oranının öngörülmesine kadar birçok alanda uygulanmaktadır [19].

YSA'nın başlangıcından günümüze kadar olan gelişim süreci içinde en iyi bilinen mimarilerinin kısa bir tarihçesi Tablo 6.5'de görülmektedir.

Tablo 6.5: Bilinen YSA Mimarilerinin Tarihsel Gelişimi [20]

Yıl	Ağ Mimarisi	Bulucu Bilim Adanı
1942	Mc Culloh-Pitts Hücresi	Mc Culloh-Pitts
1957	Algılayıcı (Perceptron)	Roserblatt
1960	Madaline	Widrow
1969	Cerebellatron	Albus
1974	Geriye Yayınım (Backpropagation)	Werbos, Parker, Rumelhart
1977	Bir Kutu İçinde Zeka	Anderson
1978	Neocognitron	Fukushima
1978	Adaptif Rezonans Teorisi	Carpenter, Grossberg
1980	Ön Düzenlemeli Harita	Kohonen
1982	Hopfield	Hopfield
1985	İki Yönlü Bileşik Hafıza	Kosko
1985	Boltzman Makinesi	Hinton, Sejnowsky, Szu
1986	Sayıc Yayılım (Counterpropagation)	Hecht-Nielsen
1988	Hücrel Sinir Ağı	Chua, Yang

Sinir ağları, biyolojik nöronlar şeklinde modellenmiş temel birimlerden oluşurlar. Her bir birimin bir çok girdisi bulunmakta ve bu girdiler, bir çıktı değeri oluşturacak şekilde birleştirilmektedirler. Birimler Şekil 6.2'deki gibi birbirlerine bağlanırlar ve bazı birimlerin çıktıları, başka birimlerin çıktıları olabilmektedir.

**Şekil 6.2 :** İleri Beslemeli Sinir Ağı Örneği [4]

Şekil 6.2'de de gözüktüğü gibi ağı oluşturan birimler üç katman şeklinde düzenlenmiştir. İlk katman ağı *girdi katmanı*dır. Girdi katmanındaki her birim tek bir kaynağa bağlı olup, giriş verilerinin 0 ile 1 arasında olması zorunludur.

İkinci katman, ağı ne girdilerine ne de çıktısına bağlı olmaması nedeniyle *gizli katman* olarak isimlendirilir. Gizli katmanın her birimi, girdi katmanındaki tüm birimlere tam bağlıdır. Gizli katman, daha fazla örüntünün tanınmasını mümkün kılmasından dolayı ağı daha güçlü kılar. Bu katmanın büyümesi ağda uymama riskini arttıracığı için genellikle tek bir gizli katman yeterlidir. Oluşturulan ağ yapılarında

gizli katmanın bulunma zorunluđu bulunmamaktadır. Ağ, sadece girdi ve çıktı katmanlarından oluşabilir.

Son katman, *çıktı katmanı*dır. Gizli katmanın olduđu hallerde gizli katmandaki tüm birimlere, olamadığı durumlarda da girdi katmanındaki tüm birimlere tam bağıldır. Çoğunlukla sinir ağı tek bir değeri hesaplar. Bu nedenle bu katman tek bir değeri oluşturur ve ürettiği değeri de 0 ile 1 aralığında kalır. Bazı durumlarda çıktı katmanını birden fazla birimden de oluşabilir.

Yapay sinir ağları, insanların deneyimlerinden bir takım bilgiler çıkartması gibi kendisine verilen örneklerden bir takım bilgiler çıkartma yeteneğine sahiptir. YSA, öncelikle bir veri kümesi üzerinde öğrenme algoritmaları çalıştırılarak eğitilir. Bu eğitim neticesinde yapay sinir ağının içerisindeki bir takım ağırlıklar belirlenir. Bu ağırlıklar kullanılarak yeni gelen veriler işlenir ve bir sonuç üretilir. Yapay sinir ağlarının en olumsuz tarafı ise bu ağırlıkların neden ilgili değeri aldıklarının bilinmemesidir. Çıkan sonucun nedenleri açıklanamamaktadır. Bu olumsuz özellik yüzünden, Amerika Birleşik Devletleri'nde yapay sinir ağlarının, kredi taleplerinin değerlendirilmesinde kullanılması yasaklanmıştır. Çünkü sistem bir kişiye kredi vermeme sebebini açıklayamayacaktır. Kişinin riskli gözükmesinin sebebi ten rengi veya ırkı olabilir. Bu sorunun cevabının alınamayacak oluşu, böyle bir uygulamayı zorunlu kılmıştır. Bu sebeple, yapay sinir ağlarını kullanmak için en iyi yaklaşım, onları içi bilinmeyen bir şekilde çalışan kara kutular olarak düşünmek olacaktır.

6.3.1. Yapay Sinir Ağlarının Temel Özellikleri [21]

YSA'nın hesaplama ve bilgi işleme gücünü, paralel dağılmış yapısından, öğrenme ve genelleme yeteneğinden aldığı söylenebilir. Genelleme, eğitim ya da öğrenme sürecinde karşılaşılmayan girişler için de YSA'nın uygun tepkileri üretmesi olarak tanımlanır. Bu üstün özellikleri, YSA'nın karmaşık problemleri çözebilme yeteneğini gösterir. Günümüzde birçok bilim alanında YSA, aşağıdaki özellikleri nedeniyle etkin olmuş ve uygulama yeri bulmuştur.

- *Doğrusal Olmama*; YSA'nın temel işlem elemanı olan hücre, doğrusal değildir. Dolayısıyla hücrelerin birleşmesinden meydana gelen YSA da doğrusal değildir ve bu özellik bütün ağa yayılmış durumdadır. Bu özelliği ile YSA, doğrusal olmayan karmaşık problemlerin çözümünde en önemli araç olmuştur.

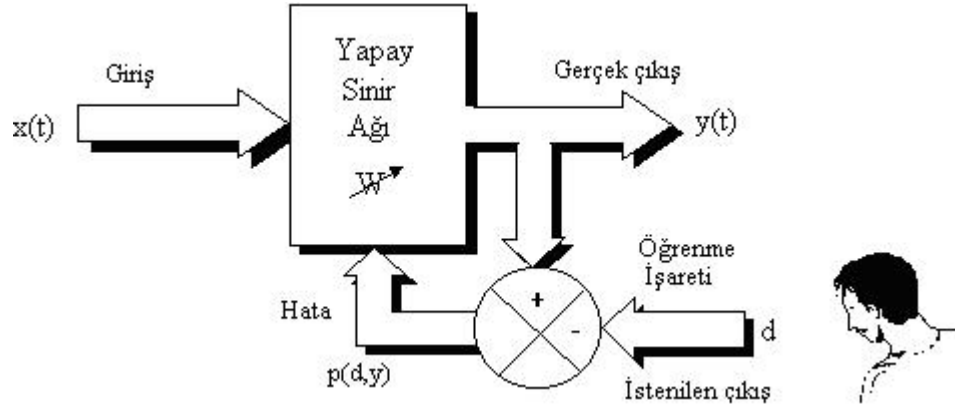
- *Öğrenme*; YSA'nın arzu edilen davranışı gösterebilmesi için amaca uygun olarak ayarlanması gerekir. Bu, hücreler arasında doğru bağlantıların yapılması ve bağlantıların uygun ağırlıklara sahip olması gerektiğini ifade eder. YSA'nın karmaşık yapısı nedeniyle bağlantılar ve ağırlıklar önceden ayarlı olarak verilemez ya da tasarlanamaz. Bu nedenle YSA, istenen davranışı gösterecek şekilde ilgilendiği problemde aldığı eğitim örneklerini kullanarak problemi öğrenmelidir.
- *Genelleme*; YSA, ilgilendiği problemi öğrendikten sonra eğitim sırasında karşılaşmadığı test örnekleri için de arzu edilen tepkiyi üretebilir. Örneğin, karakter tanıma amacıyla eğitilmiş bir YSA, bozuk karakter girişlerinde de doğru karakterleri verebilir ya da bir sistemin eğitilmiş YSA modeli, eğitim sürecinde verilmeyen giriş sinyalleri için de sistemle aynı davranışı gösterebilir.
- *Uyarlanabilirlik*; YSA, ilgilendiği problemdeki değişikliklere göre ağırlıklarını ayarlar. Yani, belirli bir problemi çözmek amacıyla eğitilen YSA, problemdeki değişimlere göre tekrar eğitilebilir ve değişimler devamlı ise gerçek zamanda da eğitime devam edilebilir. Bu özelliği ile YSA, uyarlamalı örnek tanıma, sinyal işleme, sistem tanılama ve denetim gibi alanlarda etkin olarak kullanılır.
- *Hata Toleransı*; YSA, çok sayıda hücrenin çeşitli şekillerde bağlanmasından oluştuğu için paralel dağılmış bir yapıya sahiptir ve ağına sahip olduğu bilgi, ağıdaki bütün bağlantılar üzerine dağılmış durumdadır. Bu nedenle, eğitilmiş bir YSA'nın bazı bağlantılarının hatta bazı hücrelerinin etkisiz hale gelmesi, ağına doğru bilgi üretmesini önemli ölçüde etkilemez. Bu nedenle, geleneksel yöntemlere göre hatayı tolere etme yetenekleri son derece yüksektir.

6.3.2. Öğrenme Şekillerine Göre Yapay Sinir Ağları [20]

Sinir ağlarının en önemli özelliği, öğrenme yeteneğidir. Bir sinir ağına öğrenmenin anlamı, ağına belirli bir probleme ait doğru çıktıları üretmesini sağlayacak optimum ağırlık değerlerinin bulunmasıdır. Öğrenme, ağırlık değerlerinin nasıl değiştirilmesi gerektiğini ifade eden bir öğrenme kuralına dayanır. Bir öğrenme kuralının temel ilkesi ise, benimsenen öğrenme stratejisi ile tanımlanır. Literatürde, üç tip öğrenme stratejisinden söz edilmektedir.

- Denetimli Öğrenme

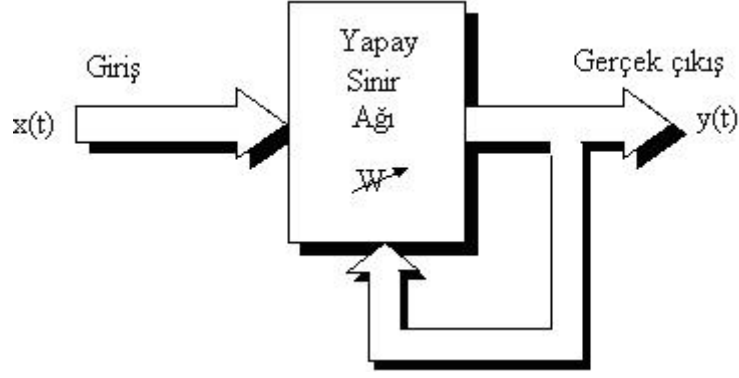
Denetimli öğrenme moduyla, yapay sinir ağının eğitimi için eğitici veriler (eğitim kümesi) kullanılmaktadır. Eğitim kümesi, giriş bilgileri ve istenen (hedef) bilgiler olmak üzere iki ayrı vektör gibi düşünülebilir. Vektörlerin her bir karşılıklı elemanları bir eğitim çiftini oluşturmaktadır. Eğitim kümesi, ağın eğitimine başlamadan önce belirlenmektedir. Ağın eğitimi için, öncelikle bağlantı ağırlıklarına rastgele değerler atanmaktadır. Daha sonra eğitim çiftlerine bağlı olarak bir algoritma dahilinde ağırlıklar yenilenmektedir. İstenilen bilgiler ve ağın çıkışı arasındaki fark (hata) azalınca kadar eğitim sürdürülmektedir. Ağ çıkışındaki hatanın azalması, ağırlıkların kararlılık kazanması demektir. Ağırlıklar istenilen kararlılığa ulaştığında eğitim bitirilmektedir. Şekil 6.3’de denetimli öğrenme yapısı gösterilmiştir.



Şekil 6.3 : Denetimli Öğrenme Yapısı

- Denetimsiz Öğrenme

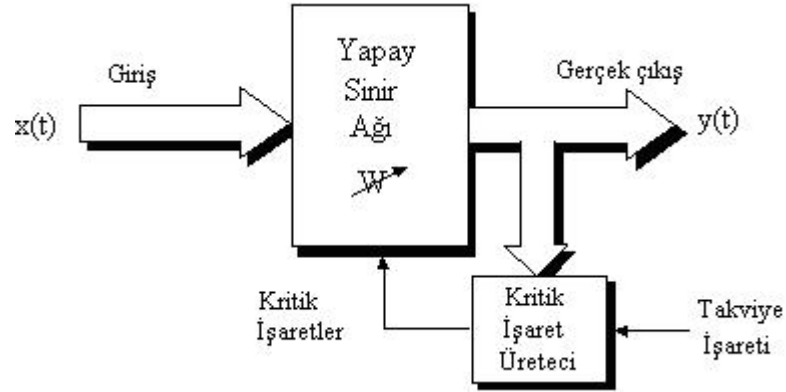
Denetimsiz öğrenme moduna, “kendi kendine öğrenilebilen mod” da denilmektedir. Bu öğrenme modunda eğitim kümesi kullanılmamaktadır. Ağ, birbirine benzer giriş bilgilerini gruplamakta veya giriş bilgisinin hangi gruba ait olduğunu göstermektedir. Ağ eğitimi için sadece giriş bilgileri yeterli olmakta, referans alınacak (eğitici) bilgiye ihtiyaç duyulmamaktadır. Ağın performansını kendiliğinden izlenmesi söz konusudur. Ağ, giriş sinyallerinin yönüne veya düzenine bakmakta ve ağın fonksiyonuna göre ayarlama yapmaktadır. Bu nedenle, denetimsiz öğrenme stratejisini kullanan ağlar, kendi kendine organize olan ağlar olarak adlandırılır. Şekil 6.4’de denetimsiz öğrenme yapısı gösterilmiştir.



Şekil 6.4 : Denetimsiz Öğrenme Yapısı

- Takviyeli Öğrenme

Bu öğrenme kuralı denetimli öğrenmeye yakın bir metottur. Denetimsiz öğrenme algoritması istenilen çıkışın bilinmesine gerek duymaz. Hedef çıktıyı vermek için "öğretmen" yerine, burada YSA'ya bir çıkış verilmemekte fakat elde edilen çıkışın verilen girişe karşılık iyiliğini değerlendiren bir kriter kullanılmaktadır. Şekil 6.5'de takviyeli öğrenme yapısı gösterilmektedir.

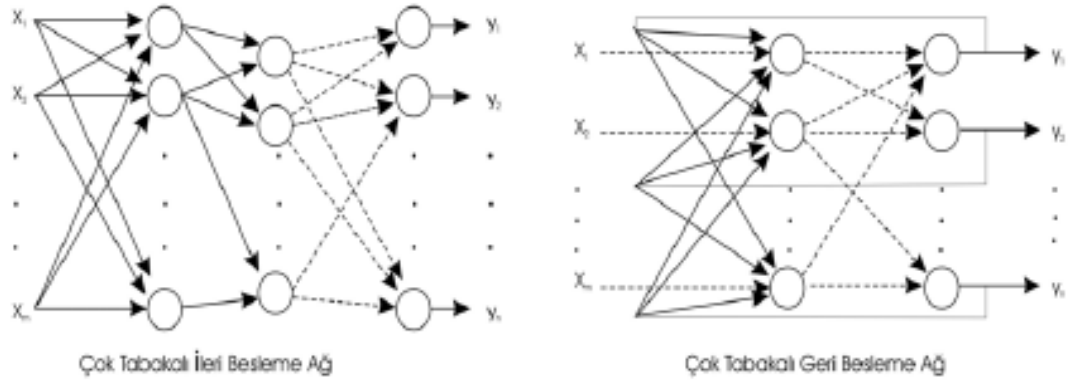


Şekil 6.5 : Takviyeli Öğrenme Yapısı

6.3.3. Ağ Yapılarına Göre Yapay Sinir Ağları [22]

YSA'lar, ağın yapısına göre sınıflandırılabilirler. Bazı ağlar ileri besleme şeklinde yapılandırılırken, bazı ağlar ise geri besleme yapısı içermektedir. İleri besleme sinir ağlarında, işlem elemanları arasındaki bağlantılar bir döngü oluşturmazlar ve bu ağlar girdi veriye genellikle hızlı bir şekilde karşılık üretirler. Geri beslemeli ağlarda (Recurrent Networks) ise bağlantılar döngü içerirler ve hatta her seferinde yeni veri kullanabilmektedirler. Bu ağlar, döngü sebebiyle girdinin karşılığını yavaş bir şekilde oluştururlar. Bu yüzden, bu tür ağların eğitime süreci daha uzun olmaktadır. Ayrıca, hem ileri besleme hem de geri yayılma olarak tanımlanabilecek ağ yapıları da

mevcuttur. Şekil 6.6’da, çok tabakalı ileri besleme ağ yapısı ile birlikte çok tabakalı geri besleme ağ yapısı örneklenmektedir.



Şekil 6.6 : İleri Beslemeli ve Geri Beslemeli Ağ Yapıları

6.3.4. Yapay Sinir Ağlarının Kuvvetli ve Zayıf Yönleri

Yapay sinir ağlarının veri madenciliği açısından kuvvetli yönleri şunlardır:

- Çok geniş bir yelpazedeki problemlerde kullanılabilir.
- Çok karmaşık durumlarda dahi iyi sonuçlar üretir.
- Hem sayısal hem de kategorik veriler üzerinde işlem yapabilir.

Bütün bu olumlu özelliklerine rağmen yapay sinir ağlarının olumsuz yönleri de vardır:

- Girdi verilerinin 0 ile 1 arasında değerler alması zorunludur.
- Ürettikleri sonuçların nedenleri hakkında açıklama yapamazlar.
- Varılan sonucun olası en iyi sonuç olduğunun garantisi yoktur.
- Kullanılması zordur ve uzmanlık gerektirir.

6.4. Genetik Algoritmalar

Genetik algoritmalar evrimsel hesaplama (evolutionary computing) metodlarının örnekleridir ve optimizasyon tipli algoritmalar [3]. Genetik algoritmalar da yapay sinir ağları gibi biyolojik işlemlerden kaynağını almıştır. Yüzyıllar boyu süren adaptasyonlar ve doğal seleksiyon sonucunda çevre koşullarına en fazla uyum sağlayanlar hayatta kalmışlardır. Genetik algoritmalarında benzer bir çalışma biçimi

vardır. Geçtiğimiz yıllar boyunca genetik algoritmalar, veriyi modellemek için yapay sinir ağlarıyla birlikte sıkça kullanılmıştır.

Veri madenciliğinde genetik algoritmalar kümeleme ve öngörme problemleri hatta birliktelik kuralları için kullanılabilirler. Bu teknikler, veriyi tanımlamak için kurulan modellere ait kümeden en uygun (fittest) modelleri bulmak olarak düşünülebilir. Bu yaklaşımda öncelikle bir başlangıç modeli kabul edilir ve bir çok iterasyondan sonra modeller yeni modeller oluşturmak üzere birleştirilir. Bu modeller arasından bir uygunluk fonksiyonu (fitness function) tarafından tanımlanan en iyi model, bir sonraki iterasyon için girdi olarak kullanılır. Algoritmalar modelin nasıl tanımlandığına, modeldeki farklı bireylerin / kromozomların nasıl birleştirildiğine ve uygunluk fonksiyonun nasıl kullanıldığına göre farklılık gösterir [3].

Temel bir genetik öğrenme algoritması şu şekilde ifade edilebilir [14]:

1. n tane elemandan oluşan bir popülasyon, P , seçilir. Buradaki elemanlar genelde kromozom olarak anılır.
2. Belirlenmiş bir bitim koşulu sağlanana kadar
 - a. Geçerli çözümün her elemanını değerlendirmek için bir uygunluk fonksiyonu kullan. Eğer bir eleman uygunluk kriterini geçerse, popülasyonun içinde kalır.
 - b. Popülasyon şu an m adet elemandan oluşmaktadır ($m \leq n$). $(n-m)$ adet yeni eleman üretmek için genetik operatörleri kullan. Yeni elemanları popülasyona ekle.

Veri madenciliği için düşünüldüğünde elemanların / kromozomların değişkenler ve değerlerle tanımlanan örnekler olduğu düşünülür.

En yaygın genetik operatörler çaprazlama (crossover) ve mutasyon (mutation) operatörleridir. Çaprazlama, o an popülasyonda bulunan iki elemanın parçalarını birleştirerek, popülasyon için yeni elemanlar oluşturmaktadır. Çaprazlama için kullanılacak olan elemanlar genelde, popülasyondan atılacak olan elemanlardır. İkinci genetik operatör olan mutasyon, atma işlemi için seçilen elemanlara uygulanır. Mutasyon, bir elemanın içindeki genleri (veya değişkenin değerlerini) rassal olarak değiştirerek uygulanmaktadır. Seçme (Selection) işlemi ise sık kullanılmayan üçüncü genetik operatördür. Seçme işlemi ile, popülasyondan silinen elemanlar uygunluk

testinin yüksek skorla geçen elemanların kopyalarıyla değiştirilirler. Böylelikle popülasyonun genel uygunluk değerinin yükselmesi garantilenir [14].

Genetik algoritmalar açıklanabilir sonuçlar üretirler. Çok değişik tiplerdeki verileri işleme özelliğine sahip olan genetik algoritmalar, optimizasyon amacı ile kullanılabilirler. Ayrıca genetik algoritmalar yapay sinir ağları ile çalışarak başarılı sonuçlar üretmektedirler.

Tüm bu güzel yönlerine rağmen genetik algoritmaların kullanılmalarında bazı sıkıntılar da vardır.

- Genetik algoritmaları anlamak ve son kullanıcıya anlatmak zordur
- Karmaşık sorunların genetik kodlamasını yapmak çok zordur.
- En iyi uygunluk fonksiyonunu belirlemek zordur.
- Çaprazlama ve mutasyon işlemlerinin nasıl yapılacağını belirlemek zordur.
- Optimal sonucun üretildiğine dair bir garanti bulunmamaktadır.

6.5. Yapısal Risk Minimizasyonu [11]

Yapısal Risk Minimizasyonu'nun (Structured Risk Minimization, SRM) temeli Vladimir Vapnik tarafından ispatlanan İstatistik Öğrenme Teorisi'ne (Statistical Learning Theory) dayanmaktadır.

Vladimir Vapnik'in çalışmaları 1970'lerde İstatistik Öğrenme Teorisi'nin yöntemlerini ortaya koyan iki önemli çalışma yayınlaması ile başlamaktadır.

1930'larda Glivenko-Cantelli-Kolmogorov ve Fisher tarafından ileri sürülen iki yaklaşım, veri modellemenin birbirinden çok farklı iki kolda gelişmesine neden olmuştur. Fisher yaklaşımı, genel çıkarsama problemlerini inceleyen teorik istatistik ile özel parametrik modeller kullanan uygulamalı istatistiği ayırmaktadır. Fisher yaklaşımının kalitesi ve özel sonuçları, uygulamalı istatistiğe çok güvenilmesine ve teorik matematikten uzaklaşılmasına neden olmuştur.

1960'larda çok sayıda ve yüksek korelasyonlu değişkenler içeren ilk büyük veri dosyalarının ortaya çıkması ile, geleneksel uygulamalı istatistik yönteminin bu tür verilerde kabul edilebilir modeller oluşturamayacağı, yani "çok boyutluluğun laneti" ortaya çıkmıştır. Başlıca Bileşen Analizi (Principal Component Analysis, PCA,) ve

sinir ağırları gibi o zamanlar ispatlanmamış olan ancak iyi sonuç veren yöntemler ise uygulamalı istatistik çevrelerinde büyük tartışmalara yol açmıştır.

Sinir ağlarının ilk anlaşılabilir sonuçları vermesi (1990) ve “çok boyutluluğun laneti”nden kurtulmanın mümkün olduğunun ispatı için 25 yıl daha geçmesi gerekmiştir. 1995’de Vapnik tarafından ispatlanan teori ise öngörüye yönelik modellemenin tanımını sorgulayarak yeni bir çözüm sistemi yaratmıştır. O zamana kadar kullanılan çözümlerin aksine, bu çözüm tam olarak ispatlanmış bir istatistik teorisine dayanıyordu. Böylece Vapnik’in çalışması Fisher parametrelerini bırakıp, Glivenko-Cantelli-Kolmogorov’un genel yaklaşımları çerçevesinde teorik istatistiğin temellerine geri dönüşü gerçekleştirdi.

- Öğrenmenin Temel Meselesi

Her biri n parametreden ve “iş sorusu” olarak adlandırılan son bir sütundan oluşan satırlarla tanımlanan bir veri takımı düşünelim. Bu satırlar, iş sorusu y olmak üzere, $[x_1, \dots, x_n | y]$ şeklinde gösterebilir.

$X, R^n : X=(x_1, \dots, x_n)$ ’de bir vektör olsun. $R^n \rightarrow R$ (regresyon) veya $R^n \rightarrow [0,1]$ (sınıflandırma) modeli oluşturmak istenirse

- w, R^p ’nin modeli tanımlayan bir parametresi
- $Z_i = (X_i, y)$ mümkün veri değerleri
- $Q(z, w); f(X, w) = y$ olduğunda modelin hata oranı
- $P(z), Z$ verisinin bilinmeyen olasılığı

olmak üzere, sonucu y değeri olan $f(X, w)$ fonksiyonunu hesaplamak için bir model kullanılır.

Amaç, $w: R(w) = \int Q(z, w) dP(z)$ ile gösterilen model riskini en aza indirmektir. Bunu yapabilmek için ise sadece $(z_1, \dots, z_L)$ ile gösterilecek ve bilinmeyen $P(z)$ dağılımına sahip L öğrenme olayı mevcuttur. Dolayısıyla,

$$E(w) = (1/L) \sum \{Q(z_i, w) | i=1, \dots, L\} \quad (6.4)$$

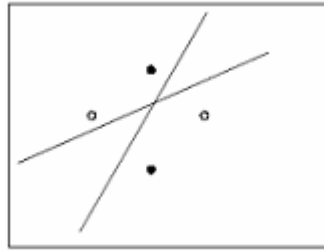
ile gösterilen ampirik risk en aza indirilmeye çalışılmaktadır.

Vapnik’in teorisinin gücü, modelin R riskinin bütün öğrenmenin ampirik riski ile deterministik bir niceliğin toplamıyla arttığını göstermesidir.

L artarken, modelin yeni veri üzerinde hatası öğrenme verisindeki hataya yaklaşıyorsa bu modele “tutarlı model” denir.

$f \in F$ modeli tanımlayan fonksiyon olsun: $Y = f(X, w)$. Vapnik $F : R^n \rightarrow R$ fonksiyon ailesine F ailesinin Vapnik-Chervonenkis, VC, boyutu denen bir h tamsayısı atamaktadır. Bu sayı F ailesinin R^n uzayındaki noktaları ne kadar iyi bölebildiğini göstermektedir. $F : R^n \rightarrow R$ fonksiyon ailesi ve $(x_1, \dots, x_n) \in R^n$ noktalar kümesi ele alınsın. m tane “beyaz” ve $L-m$ tane “siyah” noktadan oluşan L adet noktanın rengi ne olursa olsun (2^L mümkün durum var), “beyaz” noktalarda pozitif, “siyah” noktalarda negatif değer alan bir $f \in F$ fonksiyonu bulunabiliyorsa F fonksiyon ailesi $(x_1, \dots, x_n)$ noktalar kümesini böler denir. Eğer R^n uzayından alınan her h vektörlük alt küme F ailesinden bir fonksiyon ile bölünebiliyorsa ve F fonksiyon ailesinden hiçbir fonksiyon ile bölünemeyen $h+1$ vektörlük en az bir alt küme varsa, F fonksiyon ailesinin VC boyutu h olur.

Örneğin bir doğru, düzlemdeki 4 noktayı her zaman bölemeyebilir. Şekil 6.7’de gösterildiği gibi F bir düzlemdeki doğruların kümesi ise, $h_F = 3$ olur.



Şekil 6.7 : VC Boyutu

Temel Vapnik teoremi şöyledir:

- (X, w) modelinin öğrenmesi yalnız ve yalnız model ailesinin VC boyutu h sonlu ise tutarlıdır.
- $1-q$ olasılığı ile aşağıdaki eşitsizlik sağlanır:

$$R(w) < E(w) + \sqrt{\frac{h(\ln(2L/h) + 1) - \ln(q)}{L}} \quad (6.5)$$

Denklem 6.5 çok önemlidir, çünkü:

- Yeni veriye uygulanan modelin riskinin $1-q$ olasılıkla (risk eşiği, yani $q=0.01$ veya $0,01$) bütün öğrenmenin ampirik riski ile deterministik bir niceliğin toplamı ile artacağını gösterir.
- Problemdaki değişken sayısından bağımsızdır. Bu teorem istatistik modellemeye yeni bir yaklaşım getirmektedir.
- Bilinmeyen istatistik dağılım $P(z)$ 'den bağımsızdır ve $P(z)$ için hiçbir hipoteze gerek kalmaz.
- h/L sıfıra giderken son terim de sıfıra gitmektedir.

Bu denklem, limitin çok yüksek olduğu veya çok parametrelili durumlarda bile h boyutlu bir F ailesinden alınan $f(X,w)$ modelinin yeni veride hata oranının, h boyutu L 'ye göre düşük kaldığı sürece kontrol edilebildiğini göstermektedir. Ayrıca, model milyonlarca değişken içerdiğinde bile h/L oranı düşük kalıyorsa ($1/20$ iyi bir değer sayılır) model yararlı ve tutarlıdır. Ayrıca, modeli kurmak için kullanılan veriyle elde edilen sonuçlar ile kıyaslanabilir sonuçlar verecektir.

- *Yapısal Risk Minimizasyonu (YRM) İlkesi*

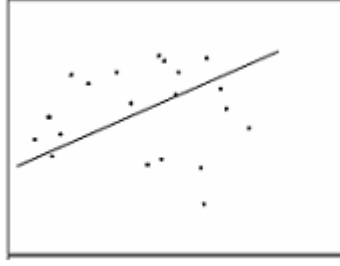
YRM ilkesinin ana fikri $E(w)$ ile ölçülen modelin doğruluğunu ve $\sqrt{(h(\ln(2L/h)+1)-\ln(q))/L}$ teriminin tersi ile ölçülen tutarlılığını saptayarak verilen bir F_m model ailesinden bir $f(X,w)$ modeli seçilmesidir. Bu seçimi yapabilmek için olası model ailelerinden, gitgide daha “zengin” (daha fazla bilgi içeren) olacak şekilde bir dizi yaratılır:

$$h_1 < h_2 < \dots < h_p \text{ olmak üzere } F_1 \subset F_2 \subset \dots \subset F_p$$

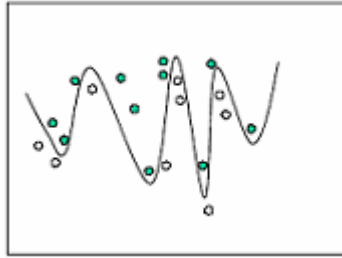
Model aileleri gittikçe “zenginleştiği” için, $p < q$ olmak üzere F_q ailesinin en iyi modeli F_p ailesinin en iyi modelinden daha doğru olacaktır. Ancak $h_p < h_q$ olduğu için daha az tutarlı olacak, yani yeni veride hatası daha fazla olacaktır.

Şekil 6.8 ve Şekil 6.9’da sırasıyla, aynı nokta kümesinin bir doğru ve daha yüksek dereceli bir polinom ile bölünmesi örnekleri gösterilmiştir. Noktalar kümesi doğru ile bölündüğünde modelin kalitesi daha düşük olmasına rağmen yeni veriye daha iyi uyacaktır. Bunun yanında aynı noktalar kümesi daha yüksek dereceli bir polinomla bölündüğünde, modelin kalitesi artacak (eğri her noktaya yakın geçiyor) ancak bu

seferde modelin tutarlılığı / genellenebilirliği azalacaktır. Dolayısıyla da yeni veri üzerinde iyi çalışamayacaktır.



Şekil 6.8 : Bir Noktalar Kümesinin Bir Doğru ile Bölünmesi Örneği



Şekil 6.9 : Noktalar Kümesinin Bir Polinom İle Bölünmesi Örneği

Bir YRM yaklaşımında modelleme;

1. Verinin bilinmeyen istatistik dağılımı üzerine bir hipotez ileri sürmek
2. Yüksek boyutlu verinin çok fazla değişirge ve uzun hesaplama zamanı gerektireceğini kabul etmek veya tutarlılık sorunu olan bazı değişkenleri önceden seçerek değişken sayısını azaltmak
3. Daha doğru bir model bulup geçerli olduğunu ispatlamak

adımlarından oluşan geleneksel modelleme yaklaşımını

1. YRM bakış açısı ile, VC boyutunu kontrol ederek en iyi F model ailesini bulmak
2. Tanım olarak modelin tutarlılığı kontrol altında olduğundan tüm parametrelerle çalışmak
3. Doğruluk ve tutarlılık arasındaki en iyi dengeyi bulmak

adımlarından oluşan süreçle değiştirmektedir.

6.6. Destek Vektör Makinaları (DVM) [23,24]

İstatistikteki, genelleme kuramındaki hesapsal öğrenme ve makina öğrenme kuramlarındaki yeni ilerlemeler, model yapım/öğrenme/sağlamlaştırma sürecinin genel karakterine ve doğasına yeni izleme metodları ve derin özellikler sağlamıştır. Bazı araştırmacılar, istatistiksel ve makina öğrenme modellerin kavramsal olarak tamamıyla farklı olmadıklarına dikkat çekmektedir. Yeni hesapsal ve makina öğrenme yöntemlerinin birçoğu, istatistikte parametre tahmini fikrini genellemektedir. Geçtiğimiz bir kaç yılda bu yeni yöntemlerden en çok ilgi çeken Destek Vektör Makinaları (Support Vector Machines, SVM) olmuştur.

DVM, ilk olarak Vapnik tarafından ortaya atılan yeni bir öğrenme makinasıdır. Hesapsal öğrenme kuramındaki Yapısal Risk Minimizasyonu prensibine dayanmaktadır. Hearst ve diğerleri DVM algoritmasını, öğrenme kuramı ve pratiğin kesişme noktasına yerleştirmişlerdir: “Sinirsel ağların geniş bir sınıfını, radial bazlı fonksiyon (RBF) ağını ve özel durumlar için de polinomsal sınıflandırıcıları içermektedir. Aynı zamanda matematiksel olarak analiz edilebilecek kadar basittir, çünkü non-lineer olarak girdi uzayına bağlı yüksek boyutlu bir nitelik uzayında, lineer bir yönteme benzeyecek biçimde gösterilebilir.” Bu bağlamda DVM’ler, kuram odaklı ve kolay analiz edilebilen geleneksel istatistiki yöntemlerin güçlü yanları ile daha fazla veri odaklı, dağıtımdan muaf ve sağlıklı makina öğrenme yöntemlerini birleştirmek için iyi bir aday olabilir.

Geçtiğimiz birkaç yılda, DVM’lerin farklı safhalarında büyük gelişmeler olmuştur. Bu safhalar kuramsal kavrayış, uygulama ve gerçek hayata uyarlama için algoritmik stratejileri içermektedir. DVM, bioinformatik, belge kategorizasyonu ve görüntü tespitini gibi problemleri de kapsayan geniş problem yelpazesinde mükemmel bir genelleme performansı göstermiştir. Bu uygulama alanları yüksek boyutlu girdi uzayını içermekteydi ve bu iyi performans DVM’nin öğreti kabiliyetinin nitelik uzayının boyutsallığından bağımsız olabilme gerçeğine de bağlıydı.

DVM yaklaşımı son zamanlarda temel olarak zaman serileri tahmini ve sınıflandırma gibi birçok finansal uygulamada da kullanılmıştır. Veri madenciliğindeki uygulamaları üstün genelleme performansından ötürü genellikle sınıflama tekniğinde ortaya çıkmıştır. Elde edilen sonuçlar bu yöntemin sınıflama tekniğinde oldukça başarılı olduğunu göstermiştir.

Denetimli öğrenme yöntemlerinin ana teması gözlemlerden öğrenmedir. X ile gösterilen bir girdi uzayı, $X \subseteq R^n$, Y ile gösterilen bir çıktı uzayı ve S ile gösterilen bir çalışma kümesi vardır. $S = ((x_1, y_1), (x_2, y_2), \dots, (x_l, y_l)) \subseteq (X \times Y)^l$ ve burada l , çalışma setinin boyutudur. Öğrenme için genel varsayım, gizli bir fonksiyonun varlığı $Y = f(X)$ ve sınıflandırmanın görevinin sezgisel (heuristic) bir fonksiyon $h(X)$ oluşturmak olduğudur. Şöyle ki Y 'nin tahmini üzerine $h \rightarrow f$ dir. Çıktı uzayı Y 'nin doğası, öğrenme tipine karar verir. $Y = \{1, -1\}$ ikili (binary) bir sınıflandırma problemine, $Y = \{1, 2, 3, \dots, m\}$ çok-sınıflı bir sınıflandırma problemine ve $Y \subseteq R^n$ bir regresyon problemine yol açar.

DVM, sınıflandırma probleminin Denklem 6.6'da gösterildiği üzere bir optimizasyon problemi gibi sunulduğu maksimal marjın sınıflandırıcı tipine aittir.

$$\min_{w,b} < w, w >$$

$$\text{Şöyle ki } y_i(< w, \Phi(x_i) > + b) \geq 1 \quad (6.6)$$

$$i = 1, \dots, l$$

Vapnik bir DVM'nin örüntü tespiti için çalıştırmanın, sınırlı kısıtlar ve bir lineer eşitlik kısıtıyla birlikte nasıl ikinci dereceden bir optimizasyon problemine yol açtığını Denklem 6.7'de göstermektedir. İkinci dereceden optimizasyon problemi çok iyi anladığımız bir problem tipine aittir. Çalışma örneklerinin sayısı problemin büyüklüğünü belirlediği için standart ikinci dereceden problem çözücüleri kullanmak, hesaplamayı büyük çalışma setleri için kolayca imkansız hale getirecektir. DVM'nin özel özelliklerini kullanarak, DVM'nin içindeki ikinci dereceden programlama problemini çözmek için farklı çözümler öne sürülmüştür. Bu stratejiler; bir kernel fonksiyonunun, $K(x_i, x_j)$, gerekli tüm hesaplamaların girdi uzayında direkt olarak yapılmasına izin vermek için başvurulduğu eğim tırmanma yöntemi (gradient ascent method), irileme ve ufaltma (chunking and decomposition) ve Sıralı Minimal Optimizasyon (Sequential Minimal Optimization) algoritmasını içerir.

$$\max W(\alpha) = \sum_{i=1}^l \alpha_i - \frac{1}{2} \sum_{i,j=1}^l y_i y_j \alpha_i \alpha_j < \Phi(x_i), \Phi(x_j) > \quad (6.7)$$

$$= \sum_{i=1}^l \alpha_i - \frac{1}{2} \sum_{i,j=1}^l y_i y_j \alpha_i \alpha_j K(x_i, x_j)$$

şöyle ki $\sum_{i=1}^l y_i \alpha_i = 0$, $\alpha_i \geq 0$, $i = 1, \dots, l$

Kavramsal olarak kernel fonksiyonları orijinal veriyi daha yüksek bir boyut uzayına taşır ve girdi veri kümesini dönüştürülmüş uzayda lineer olarak ayrılabilir hale getirir. Kernel fonksiyonlarının seçimi büyük ölçüde uygulamaya bağlıdır ve DVM uygulamalarındaki en önemli faktördür.

Denklem 6.7'deki formülasyon sadece sıfırın ampirik hatasına karşılık gelen ayrılabilir durumu dikkate almıştır.

Gürültülü (noisy) verilerde, zor marjın kısıtlamalarını kolaylaştırmak için zayıf değişkenler bazı sınıflandırma hatalarına izin vermek için Denklem 6.8'deki gibi ortaya koyulur. Bu formülasyonda, gürültü seviyesi, $C > 0$, amprik hata ve karışıklık terimi arasındaki feragati belirlemektedir.

$$\min_{w,b,\xi} < w, w > + C \sum_{i=1}^n \xi_i \quad (6.8)$$

$$y_i(< w, \Phi(x_i) > + b) \geq 1 - \xi_i$$

$$\xi_i \geq 0, i = 1, \dots, l$$

Bu genişletilmiş formülasyon Denklem 6.9'da ifade edilen ikili (dual) probleme yol açar.

$$\max W(\alpha) = \sum_{i=1}^l \alpha_i - \frac{1}{2} \sum_{i,j=1}^l y_i y_j \alpha_i \alpha_j < \Phi(x_i), \quad (6.9)$$

$$\Phi(x_j) \geq \sum_{i=1}^l \alpha_i - \frac{1}{2} \sum_{i,j=1}^l y_i y_j \alpha_i \alpha_j K(x_i, x_j)$$

şöyle ki $\sum_{i=1}^l y_i \alpha_i = 0$, $0 \leq \alpha_i C$, $i = 1, \dots, l$

Standard DVM formülasyonu sadece ikili sınıflandırma problemini çözmektedir. Bu nedenle ya çok-sınıflı bir sınıflandırıcı oluşturmak için çeşitli ikili sınıflandırıcılar kullanılmalı ya da bütün sınıfları aynı anda ele alabilmek için orijinal formülasyonda köklü değişiklikler yapılmalıdır.

7. FARKLI ENDÜSTRİLERDEKİ VERİ MADENCİLİĞİ UYGULAMALARI

7.1. Bankacılık ve Finans Endüstrilerinde Veri Madenciliği Uygulamaları

Bankacılık endüstrisi müşterileri hakkında sahip oldukları bilgilerin öneminin farkına varmıştır. Bilgi teknolojisi sadece servis kalitesini geliştirmek için değil bunun dışında rekabet avantajı kazanmak içinde geniş kapsamlı olarak kullanılmaktadır. Bankalar tarafından yıllardır toplanmakta olan muazzam miktardaki veriler, manuel olarak incelenemeyecek bir hal almış ve veri madenciliği teknolojileri için kaçınılmaz bir fırsat oluşturmuştur.

Finansal pazarda ise, yapay sinir ağları gibi öngörüselle modelleme tekniklerinin kullanıldığı analizlerle portföy oluşturma ve iyileştirme, bono fiyatlandırması ve finansal tehlikeleri tahmin gibi konulara çözümler aranmaktadır [25].

Bankacılık ve Finans endüstrilerinde veri madenciliğinin kullanıldığı konulardan bazıları şunlardır:

- Sahtecilik tespiti ve önlenmesi
- Müşteri segmentasyonu
- Sadık müşterilerin belirlenmesi
- Bankadan ayrılacak olan müşterilerin öngörülmesi
- Müşteri karlılığı
- Farklı finansal göstergeler arasındaki saklı korelasyonların bulunması
- Kampanya yönetimi
- Çapraz / Dikey satış
- Ürün yönetimi
- Fiyatlandırma
- Risk yönetimi

Farmer's Group Inc. veri madenciliğini “spor arabası olan bir kişinin yüksek kaza riski yoktur” senaryosunu bulmak için kullanmıştır. Senaryonun şartları, spor arabanın ikinci araba olmasını ve aile arabasının bir station wagon veya sedan olmasını gerektirmektedir [14].

Bank of America ise veri madenciliğini, hangi müşterilerin hangi Bank of America ürünlerini kullandıklarını tespit etmek ve böylece müşteri ihtiyaçları ile örtüşen doğru ürünleri ve servisleri önerebilmek için kullanmaktadır [14].

New York'taki Chase Manhattan Bankası müşterilerini rakiplerine kaybetmeye başlayınca, müşteri hesaplarını analiz etmek ve kendi hesap gereksinimlerinde değişiklikler yapabilmek için veri madenciliği kullanmaya başlamış, bu sayede karlı müşterilerini elinde tutabilmiştir [6].

7.2. Perakende Sektöründe Veri Madenciliği Uygulamaları

Perakende sektörü de veri madenciliği kullanarak, rekabet avantajı kazanmanın mümkün olduğunun farkına varmıştır. Bankacılık sektöründe olduğu gibi yıllar boyunca toplanmış çok büyük miktardaki veriye sahip olan perakende sektöründe, bu verilerle yapılan en bilinen analiz türü sepet analizidir.

Veri madenciliğinin çoğunlukla pazarlama ve müşteri ilişkileri yönetimi konularında kullanıldığı perakende sektöründe, çözüm aranan başlıca konular aşağıda belirtilmiştir.

- Pazar sepet analizi
- Ürünlerin satış eğilimleri
- Müşterilerin alış veriş alışkanlıkları ve tercihleri
- Promosyon şemalarının seçimi
- En iyi stok kararlarını vermek
- Müşterilerin demografik bilgileri arasındaki ilişkiler
- Müşterilerin kampanya veya reklamlara olan cevaplarını öngörme
- Tedarikçilerin teslimat performansları
- Dönemsel farklılıklar

Eskiden sadece ürünlerle ilgili bilgilere ulaşılabilen perakende sektöründe, artık bir çok perakende mağazasının uyguladığı bir başka pazarlama taktiği olan sadakat kartı (loyalty card) kullanımı sayesinde, müşterilerin bilgilerine de ulaşmak mümkün olmuştur. Mağazamızdan sıkça alışveriş yapan müşterileri ödüllendirmek, müşterileri mağazamızdan daha çok ürün almaya teşvik etmekte ve onların diğer mağazalardan alışveriş yapma isteğini azaltmaktadır.

Süper marketlerde de sadakat kartlarını kullanmak mümkündür. Böylelikle marketin “dondurulmuş döner” gibi önerebileceği yeni bir servisi olduğunda, veri madenciliği kullanarak kimlerin bu ürünü alma yönünde eğilim göstereceğini tespit edip, sadece bu kişilere yönelinebilir. Ayrıca bu kartların kullanılmasıyla, müşteri kasaya ödeme yapmak için geldiğinde müşteriyi anında promosyona tabi tutmak mümkün olur. Müşteri kartındaki demografik ve alışveriş bilgilerini daha önce kurmuş olduğunuz modele tanıttığınız takdirde, müşterinize almamış olduğu bir ürünü mağazadan ayrılmadan önce önerebilirsiniz.

7.3. Telekomünikasyon Sektöründe Veri Madenciliği Uygulamaları

Telekomünikasyon sektörü çok büyük miktarda veri üretip saklamaktadır. Bu veri; telekomünikasyon ağları içinden geçen konuşmaları tanımlayan “ ayrıntılı konuşma verisi”ni, şebeke içerisindeki donanım ve yazılım bileşenlerinin durumunu gösteren “şebeke verisi”ni ve müşterileri anlatan “müşteri verisi”ni içermektedir. Veri miktarı çok büyük olduğundan verinin manuel olarak analiz edilmesi çok zordur. Bu kadar büyük hacimli veriyi analiz edip faydalı bilgi elde etme ihtiyacı, bilgi tabanlı uzman sistemlerin gelişimine izin vermiştir. Dolayısıyla telekomünikasyon sektörü veri madenciliği teknolojisini erken benimseyen sektörlerden biri olmuştur. Sektörde veri madenciliği kullanılarak çözüm bulunabilecek konulardan bazıları aşağıda belirtilmiştir.

- Pazar araştırması
- Müşteri segmentasyonu
- Satış gücü optimizasyonu
- Kampanya optimizasyonu
- Çapraz / Dikey satış

- Müşteri sadakat analizi
- Call center optimizasyonu
- Sahtekarlık tespiti ve öngörüsü
- Şebeke planlama
- Şebeke bakımı
- Şebeke kapasite planlaması
- Şebeke hata analizi ve öngörüsü
- Gelir tahmini

Günümüzde AT&T, sahtecilik içeren uluslararası konuşmaları belirleyebilmek için veri madenciliği yoluyla geliştirdiği bir sistem kullanmaktadır [14].

Bunun yanı sıra RightPoint Corporation, özellikle müşteri sadakat analizleri başta olmak üzere telekomünikasyon sektöründeki çeşitli veri madenciliği konularında çalışmalar yapmaktadır [26].

US West Communication ise aile boyutu, ortalama aile bireyi yaşı ve konum gibi özelliklere dayanan müşteri eğilimlerini ve ihtiyaçlarını belirlemek için veri ambarı ve veri madenciliği teknolojilerini kullanmaktadır. Veri madenciliği projesinin sonuçları yeni müşterileri firmaya kazandırmak için kullanılmıştır [14].

7.4. Üretim Sektöründe Veri Madenciliği Uygulamaları

Rekabetin bol olduğu bir ortamda üreticiler artık, düşük fiyatların, kalitenin ve zamanında yapılan teslimatların kendilerini üst seviyelerde tutacağına güvenemezler. Bu ölçütler on yıl önce birer avantajken, günümüzde iş hayatında var olabilmek için sadece gereksinim halini almıştır. Üreticiler yükselen globalleşme ve her zamankinden daha fazla rekabet ile yüzleşmektedirler. Dolayısıyla üreticilerin rekabet avantajı sağlayabilmek için üretim sistemlerinde bilgi yönetim sistemlerini uygulamaları zorunlu bir hal almıştır [26]. Veri madenciliği, değer zincirindeki temel teknolojilerden biridir ve üretim sektöründe aşağıda da belirtilen bir çok uygulama alanı içerir.

- Talep Planlama
- Kalite Geliştirme
- Tedarikçi İlişkileri Yönetimi
- Tedarik Zinciri Analizi
- Değer Zinciri Analizi
- Garanti Analizleri

Ayrıntı vermek gerekirse;

- Malzemelerin seçimi
- Sürecin seçimi
- Süreç kontrolü ve optimizasyonu
- Kalite kontrolü
- Makina arıza zamanı tespiti
- Önleyici bakım

gibi konulardan söz edilebilir.

7.5. Diğer Sektörlerdeki Veri Madenciliği Uygulamaları

Yukarıdaki bölümlerde bahsedilen endüstriler dışında da, veri madenciliği kullanılmakta ve kullanımı giderek artmaktadır. Kamu, sigorta ve sağlık sektöründen spor ve sinema endüstrisine kadar bir çok alanda da veri madenciliği kullanımı söz konusudur. Kamu sektörünü düşünecek olursak, kaçak elektrik veya su kullanımının tesbiti veri madenciliği teknolojisi sayesinde yapılabilir.

Sağlık sektöründe, Vysis ilaç geliştirmek için yürüttüğü protein analizlerinde sinir ağlarını kullanmaktadır. The University of Rochester Cancer Center ise araştırmalarına yardımcı olması için karar ağacı teknolojinin kullanmaktadır [26].

Sinema sektöründe ise Twentieth Century Fox örnek olarak gösterilebilir. Twentieth Century Fox, çeşitli pazarlama bölgelerinde hangi aktörlerin, senaryoların ve filmlerin daha iyi talep gördüğünü anlamak için gişe makbuzlarını analiz etmektedir.

Elde edilen bilgiler ayrıca hangi film fragmanlarının gösterileceğine karar vermek içinde kullanılır [14].

Bir diğ er ilgin  örnekte Amerika Profesyonel Basketbol Ligi NBA'den verilebilir. NBA takımlarından Toronto Raptors'ın yardımcı antren r  Brian James, uygun oyuncu e leşmelerini olu şturabilmek ve en iyi oyunları se ebilmek i in IBM tarafında NBA i in geli tirilen bir veri madenciliđi yazılımını kullanmaktadır [26].

8. LİTERATÜRDE BANKACILIK VE FİNANS SEKTÖRÜNE AİT UYGULAMA ÖRNEKLERİ

8.1. CHAID Algoritmasının Kullanıldığı Bir Kredi Değerlendirme Problemi [16]

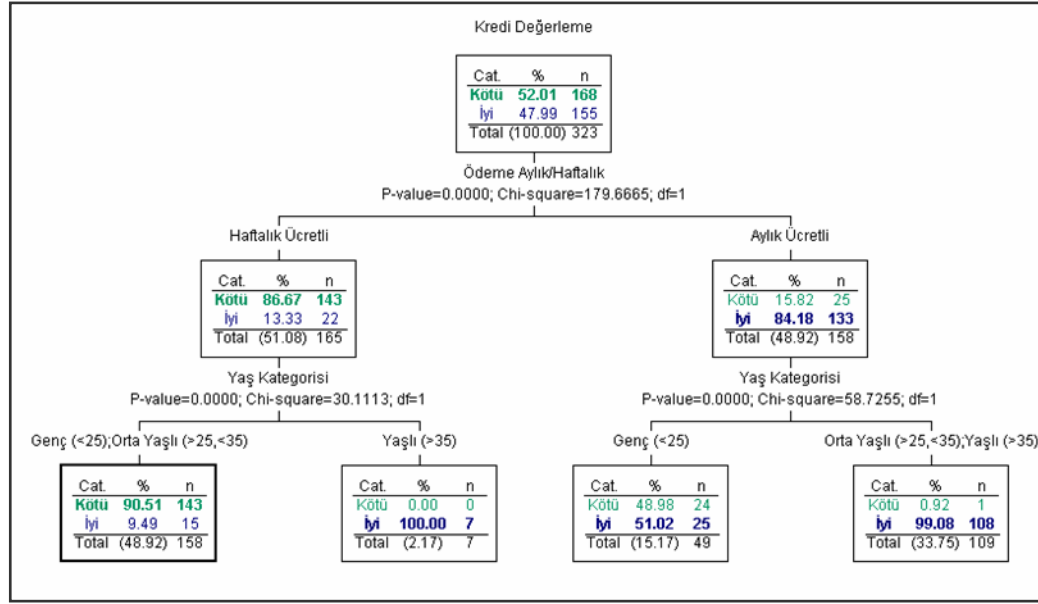
1980 yılında G.V. Kass tarafından geliştirilen CHAID algoritmasında, bağımlı değişkeni en fazla etkileyen bağımsız değişken, bağımlı değişkenin sürekli olması durumunda F testi, kategorik olması durumunda Ki Kare testi kullanılarak belirlenir. Kategorik ve sürekli değişkenler üzerinde çalışabilmesi, ağaçta her düğümü ikiden fazla alt gruba ayırabilmesi gibi nedenlerle günümüzde de tercih edilen bir algoritmadır.

Örnek uygulamada 323 kişiye ilişkin veriler toplanmış ve zaman içerisinde bu kişilerin geri ödemelerini düzenli veya düzensiz yapmalarına bağlı olarak, kredi değerlemeleri iyi ve kötü şeklinde sınıflandırılmıştır. Amaç, belirtilen bu bağımlı değişkeni etkileyen bağımsız değişkenlerin belirlenmesi ve bu çerçevede kredilerini düzenli olarak geri ödeyen müşteri örüntülerinin ortaya çıkartılmasıdır. Veri kümesinde yer alan bağımsız değişkenler Tablo 8.1’de görülmektedir.

Tablo 8.1: Veri Kümesinde Yer Alan Bağımsız Değişkenler

Özellik (Bağımsız Değişken)	Özellik Değerleri
Yaş	Genç, Orta Yaşlı, Yaşlı
Mağaza Kredi Kartı Var mı ?	Evet, Hayır
Ücretini Aldığı Zaman Dilimi	Haftalık, Aylık
İş Sınıfı	Yönetici, Meslek Sahibi, Tecrübeli İşçi, Tecrübesiz İşçi

Örnek uygulama SPSS Answer Tree 2.0 yazılımı ile hazırlanmıştır. Yazılımın çalıştırılması sonucunda elde edilen karar ağacının ilk iki aşaması Şekil 8.1’de görülmektedir.



Şekil 8.1 : CHAID Algoritması Uygulanan Bir Karar Ağacı

CHAID algoritmasına göre yapılan hesaplama sonucunda aylık ücretli olarak çalışan ve genç olmayan kişiler ile, haftalık ücretli olarak çalışan ve yaşlı kişilerin kredi geri ödemelerinde iyi olarak değerlendirilebileceği açıkça görülmektedir.

7.2. C&RT Algoritmasının Kullanıldığı Bir Kredi Kartı Başvurusu Değerlendirme Problemi [16]

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. Ayırım 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.

Örnek uygulamada 925 kredi kartı başvurusuna ilişkin veriler Tablo 8.2'deki özelliklere uygun olarak toplanmış, uzman kişiler tarafından yapılan değerlendirmeler sonucunda ise 925 kredi kartı başvurusunun 725'i kabul edilmiştir. Örnek çalışmanın amacı, kredi kartı değerlendirmesine ilişkin olarak banka uzmanlarının karar örüntülerinin belirlenmesidir. Tablo 8.2'de nominal, ordinal ve sürekli tipteki değişkenler sırası ile N, O, ve S harfleri ile gösterilmiştir.

Tablo 8.2: Kredi Kartı Başvurusuna İlişki Veriler

Özellik (Bağımsız Değişken)	Tipi	Özellik Değerleri
Medeni Durumu	N	Bekar, Evli, Dul, Boşanmış
Önceki İşi	N	Yönetici, Memur, Esnaf, Öğrenci, Emekli, İşçi, İşsiz ...
Şimdiki İşi	N	Yönetici, Memur, Esnaf, Öğrenci, Emekli, İşçi, İşsiz ...
Kredi Kartı Promosyonuna Cevap Verdi mi ?	N	Evet, Hayır
Çalıştığı Yere Ortak mı ?	N	Evet, Hayır
Mağaza Kredi Kartı Var mı ?	N	Evet, Hayır
Başka Bankadan Kredi Kartı Var mı ?	N	Evet, Hayır
Tasarruf Hesabı Var mı ?	N	Evet, Hayır
Şimdiki İşinde Çalışma Süresi (Ay)	S	
Kaç Yıldır Banka Müşterisi ?	S	

Her bir düğümün her aşamada ikiye ayrıldığı C&RT algoritmasında, ilk aşamada en kuvvetli ayırım kriteri olarak müşterinin tasarruf hesabı gini indeksi kullanılarak bulunmuştur.

$p(j | t)$, t. düğümde j sınıfının nispi frekansı olmak üzere, t. düğümün gini indeks değeri,

$$g(t) = 1 - \sum_j p^2(j | t) \quad (8.1)$$

eşitliği ile hesaplanmaktadır. Bir düğümde kategoriler arasında eşit dağılım söz konusu olduğunda, bağımlı değişken için kategori sayısı k olmak üzere gini indeks değeri $1 - (1/k)$ işlemi sonucunda elde edilecek maksimum değere erişecektir. Bir düğümdeki bütün vakaların aynı kategoriye ait olması durumunda gini indeks değeri 0 olacaktır.

Gini kriter fonksiyonu $\Phi(s,t)$; p_L sol, p_R sağ alt düğüme gönderilen vakaların oranı olmak üzere,

$$\Phi(s,t) = g(t) - p_L g(t_L) - p_R g(t_R) \quad (8.2)$$

eşitliği ile hesaplanmaktadır. Bu eşitlikte $\Phi(s,t)$ değerini maksimize edecek s değerinin seçilmesi amaçlanmakta, t düğümünde bütün vakaların katılımıyla hesaplanan bu değer, C&RT ağacında ilerleme (improvement) kavramı ile ifade edilmektedir.

C&RT algoritması sonucunda kurulan modelin, sınıflara ayırmadaki doğruluk derecesi Tablo 8.3’de sunulan matriste görülmektedir. Bu matrise göre fiilen kabul edilen 725 kredi kartı başvurusundan 699’u, fiilen kabul edilmeyen 200 başvurunun 128’i doğru olarak sınıflandırılmıştır. Bu durumda kurulan modelin doğruluk derecesi % 89.4 ($=1 - 0.106$) olacaktır.

Tablo 8.3: Risk Matrisi

		Fiili		
		Evet	Hayır	Toplam
Tahmini	Evet	699	72	771
	Hayır	26	128	154
	Toplam	725	200	925

Risk Tahmini	0.10594
Risk Tahmininin Standart Hatası	0.0101194

Tablo 8.4’de sunulan kazançlar matrisinde *Kredi Kartı Alsın mı ?* isimli hedef değişkenin *Hayır* sınıfı ile ilgili sonuçları yer almaktadır.

Tablo 8.4: Kazançlar Matrisi

Düğüm No	30	18	29	21	3	27
Düğüm: n	91	19	23	16	5	19
Düğüm: %	9.84	2.05	2.49	1.73	0.54	2.05
Cevap: n	84	15	16	10	3	8
Cevap: %	42	7.5	8	5	1.5	4
Kazanç %	92.3077	78.9474	69.5652	62.5	60	42.1053
İndeks %	426.923	365.132	321.739	289.063	277.5	194.737

30 nolu düğümde *Başvuru Sahibinin Mağaza Kredi Kartı Var mı ?* isimli bağımsız değişken değerinin *Hayır* olarak belirtildiği sınıf yer almaktadır. Bu düğümde 84

cevap *Hayır*, 7 cevap *Evet* olmak üzere toplam 91 vaka bulunmaktadır. Düğüm:% değeri 91/925; cevap:% değeri 84/200; kazanç % değeri 84/91; Indeks % değeri 92.3/21.62 işlemleri sonucunda elde edilmektedir. Indeks % değeri bu düğümün genel toplam içerisinde 4.26 kat daha fazla anlamlı olduğunu göstermektedir.

Örneğe ilişkin karar kuralı ise,

if (tasarruf hesabı var mı ? = ??? or tasarruf hesabı var mı ? = diğer banka) and (önceki kredi kartı promosyonuna = ??? or önceki kredi kartı promosyonuna = hayır) and başka kredi kartı var mı ? = hayır and (mağaza kredi kartı var mı ? = hayır or mağaza kredi kartı var mı ? = ???)

then

node=30

prediction='Hayır'

probability=0.923

olacaktır. Bu cümlede belirtilen şartların gerçekleşmesi durumunda % 92.3 olasılıkla kredi kartı başvurusunun red edileceği görülmektedir.

8.3. Destek Vektör Makinaları ve Sinir Ağları Kullanımı ile Kredi Değerlendirme Analizi [24]

Bu çalışma için biri Birleşik Devletler bir diğeri de Tayvan pazarından olmak üzere iki adet kredi değerlendirme veri kümesi hazırlanmıştır.

- *Tayvan Veri Kümesi*

Bu veri kümesini oluşturmak için, Securities and Futures Institute (SFI)'den alınan finansal veri ile Taiwan Ratings Corporation (TRC)'den alınan değerlendirme bilgileri kullanılmıştır. Bu iki kurumdan gelen veriler eşleştirilip gerekli filtreleme işlemi yapıldıktan sonra, banka kredi değerlendirmesi ve 21 adet finansal değişken ile 74 durumdan oluşan bir veri kümesi oluşturulmuştur. Bu veri kümesi, 1998'den 2002 yılına kadar 25 enstitüyü kapsamaktadır. Veri kümesinde 5 adet değerlendirme kategorisi (twAAA, twAA, twA, twBBB ve twBB) oluşmuştur.

- *Birleşik Devletler Veri Kümesi*

Standard and Poor's Compustat'ın (S&P) veri kümesinden, Tayvan veri kümesiyle karşılaştırılabilir bir Birleşik Devletler değerlendirme veri kümesi oluşturulmuştur. Veri kümesi, 1991 yılından 2000 yılına kadar ki periyotta finansal değişkenler ve değerlendirmeleri kapsamaktadır. Veri kümesinin son hali, 36 ticari banka için 10 yıllık veriye ait 265 durumu içermektedir. Veri kümesinde 5 adet değerlendirme kategorisi (AA, A, BBB, BB, B) oluşmuştur.

İki veri kümesinin kredi değerlendirme kategorilerine ait dağılımlar Tablo 8.5'de gösterilmiştir.

Tablo 8.5: Kredi Değerlendirme Kategorilerine Ait Dağılımlar

Tayvan Verisi		Birleşik Devletler Verisi	
twAAA	8	AA	20
twA	11	A	181
twA	31	BBB	56
twBBB	23	BB	7
twBB	1	B	1
Toplam	74	Toplam	265

Tayvan veri kümesindeki elde edilen finansal veriler Tablo 8.6'da listelenmiştir. Bu değişkenler, SFI veri tabanındaki mevcut finansal oranları ve literatürde kredi değerlendirmesi için sıkça kullanılan iki denge ölçümünü içermektedir. Bunlar aktif toplamı (total assets) ve pasif toplamıdır (total liabilities). İlk yedi değişken, önceki kredi değerlendirme tahmini çalışmalarında sıkça kullanılmış finansal değişkenlerdir. Diğer bazı finansal oranlar çoğunlukla Birleşik Devletler'de kullanılmamaktadır. Bu sebeple kısa tanımları verilmiştir.

Ayrı değerlendirme sınıfları arasındaki farklılıkların, her bir finansal değişkende anlamlı olup olmadığını test etmek için Tayvan veri kümesinde ANOVA (Analysis of Variance) testi uygulanmıştır. Eğer fark anlamlı değil ise (yüksek p -değeri), finansal verinin kredi değerlendirme kararı için bilgi verici olmadığı düşünülür. Tablo 8.6 her değişkene ait p -değerini göstermektedir. Bu p -değerleri, farkın anlamlı olup olmadığı hakkında bilgi sağlar.

Tablo 8.6: Veri Kümesinde Kullanılan Finansal Oranlar

	Finansal oran adı / tanımı	ANOVA gruplar arası <i>p</i> -değeri
X1	Aktif toplamı	0.00
X2	Pasif toplamı	0.00
X3	Uzun vadeli borçlar / Toplam sermaye yatırımları	0.12
X4	Borç oranı	0.00
X5	Cari oran	0.36
X6	Faiz ve vergi öncesi kar (FVOK) / Faiz	0.00
X7	Faaliyet kar marjı (FKM)	0.00
X8	(Öz sermaye + Uzun vadeli borçlar) / Duran varlıklar	0.00
X9	Çabuk oran	0.37
X10	Aktif karlılığı	0.01
X11	Özkaynak karlılığı	0.04
X12	Faaliyet geliri / Alınmış sermaye	0.00
X13	Vergi öncesi net gelir / Alınmış sermaye	0.00
X14	Net kar marjı	0.00
X15	Hisse başına kar	0.00
X16	Brüt kar marjı	0.02
X17	Faaliyet dışı gelir / Satış	0.81
X18	Vergi öncesi net gelir / Satış	0.00
X19	Faaliyetlerden sağlanan nakit / Cari borçlar	0.84
X20	(Faaliyetlerden sağlanan nakit / (Sermaye harcamaları + Envanterdeki yükseliş + Peşin ödenen kar)) son 5 yıl içinde	0.64
X21	(Faaliyetlerden sağlanan nakit - Peşin ödenen kar) / (Duran varlıklar + Diğer varlıklar + Çalışma sermayeleri)	0.08

Tayvan veri kümesine ANOVA testi uygulandıktan sonra, yüksek *p*-değerine sahip 5 adet oran (X5, X9, X17, X19 ve X20) elenmiştir. Bu işlemten sonra, Tayvan veri kümesinin son halinde 14 oran ve iki denge ölçütü bulunmaktadır. İki pazarın daha iyi karşılaştırılması için araştırmacılar, Birleşik Devletler pazarında da benzer değişkenler kullanmaya çalışmışlardır. Birleşik Devletler veri kümesinde 2 oran (X6 ve X21) mevcut değildir. Bu yüzden Birleşik Devletler veri kümesi 12 mevcut orandan ve 2 adet denge ölçütünden oluşmaktadır.

Çalışma sırasında değişken seçme işlemi yüzünden değerli bilgilerin kaybolmadığından emin olmak için, orjinal veri kümeleri üzerinde farklı öngörü modelleri ile ayrıca çalışılmıştır. Yüksek *p*-değerine sahip değişkenler eklendiğinde, DVM ve sinir ağları (SA) modellerinin öngörü doğruluklarında gerileme olduğu görülmüştür.

8.3.1. Kurulan Modeller ve Öngörü Doğruluk Dereceleri

Eldeki iki veri kümesi kullanılarak, 4 adet model oluşturulmuştur. Her bir pazar için, sıkça kullanılan finansal verilerle basit bir model ve mevcut tüm verilerin kullanıldığı karmaşık ikinci bir model oluşturulmuştur. Yaratılan bu modeller şöyledir:

- Tayvan I: Değerlendirme = $f(X1, X2, X3, X4, X6, X7)$
- Tayvan II: Değerlendirme = $f(X1, X2, X3, X4, X6, X7, X8, X10, X11, X12, X13, X14, X15, X16, X18, X21)$
- Birleşmiş Devletler I: Değerlendirme = $f(X1, X2, X3, X4, X7)$
- Birleşmiş Devletler II: Değerlendirme = $f(X1, X2, X3, X4, X7, X8, X10, X11, X12, X13, X14, X15, X16, X17)$

Dört modelin hepsinde kredi değerlendirmesini öngörmek için geri yayılım sinir ağları (Backpropagation neural networks) ve destek vektör makinaları kullanılmıştır. Öngörü performansını değerlendirmek için 10-katlı çapraz geçerlilik testi uygulanmıştır. Bazı kredi değerlendirme sınıflarının her iki veri kümesinde de az sayıda veriye sahip olmasından dolayı, öngörü performanslarına erişebilmek için birini dışarda bırakmalı çapraz geçerlilik (leave-one-out cross validation) testi de uygulanmıştır. SA için çapraz geçerlilik testi uygulanırken, verinin %10'u geçerlilik kümesi olarak kullanılmıştır. Tablo 8.7'de her iki çapraz geçerlilik testi kullanılarak, dört modelin öngörü doğruluk dereceleri gösterilmiştir. Karşılaştırma olanağı sağlamak için, literatürde oldukça iyi performansa ulaşmış bir regresyon modeline ait doğruluk dereceleri de Tablo 8.7'de ayrıca belirtilmiştir. Burada kullanılan lojistik regresyon (logistic regression) modelidir.

Tablo 8.7: Doğruluk Dereceleri

	10-katlı çapraz geçerlilik			Birini dışarda bırakmalı çapraz geçerlilik		
	LojR (%)	DVM (%)	SA (%)	LojR (%)	DVM (%)	SA (%)
Tayvan I	72.97	79.73	75.68	75.68	79.73	74.32
Tayvan II	70.27	77.03	75.68	70.27	75.68	74.32
BD I	76.98	78.87	80.00	75.09	80.38	80.75
BD II	75.47	80.00	79.25	75.47	80.00	75.68

Analizler sonrasında özetlenebilecek gözlemler; test edilen dört modelin üçünde DVM'lerin en iyi performansa ulaştığı, DVM ve SA modellerininin lojistik regresyon modelinden sürekli olarak üstün oldukları ve her iki çapraz geçerlilik testinin karşılaştırılabilir doğruluk dereceleri elde ettiğidir.

Bu çalışmada elde edilen en yüksek öngörü doğruluk dereceleri Tayvan veri kümesi için %79.73, Birleşik Devletler veri kümesi için %80.75'dir. Bunun yanında, destek vektör makinalarının kredi değerlendirme öngörü doğruluk derecelerini bir parça geliştirdiği görülmüştür. Bulunan sonuçlar ayrıca göstermektedir ki; literatürde sıkça kullanılan finansal verilerden oluşan küçük veri kümelerini kullanan modeller, büyük veri kümelerini kullanan modellere göre karşılaştırılabilir, hatta bazı durumlarda daha iyi sonuçlar üretmişlerdir. Bu olay daha önceki çalışmalarda tanımlanan finansal değişkenlerin kredi değerlendirme kararı için en uygun bilgileri içerdiklerini doğrulamaktadır.

9. TÜRKİYE’DE KREDİ KARTI KULLANIMI VE KREDİ KARTI DOLANDIRICILIK TİPLERİ

9.1. Türkiye’de Kredi Kartı Kullanımı [27]

Banka kartı (Debit Card), doğrudan kart hamilinin mevduat hesabına bağlıdır. Bu hesaba şifre aracılığı ile ulaşarak, Otomatik Vezne Makinaları (ATM) ve Satış Noktası Terminalleri'nden (POS) mal ve hizmet alma, para çekme ve sunulan diğer bankacılık hizmetlerini yerine getirmek amacıyla banka müşterilerine verilen karttır.

Kredi kartı ise, bankalar ve çıkartmaya yetkili kuruluşların müşterilerine belirli limitler dahilinde açtıkları kredilerle, nakit kullanmaksızın mal ve hizmet alabilmeleri, nakit kredi çekebilmeleri için verdikleri ödeme aracıdır.

Yaklaşık 40 yıldır hayatımızda olan kredi kartları, günlük hayatımızın vazgeçilmezi haline gelmiş durumdadır. Başlangıçta 10 bini geçmeyen kredi kartı sayısı Ağustos 2006 itibariyle 31 milyonu aşmıştır. Alışverişi kolaylaştırmasının yanı sıra ekonomiye sağladığı katkılarla da finans dünyasının lokomotif sektörlerinden biri haline gelmiştir.

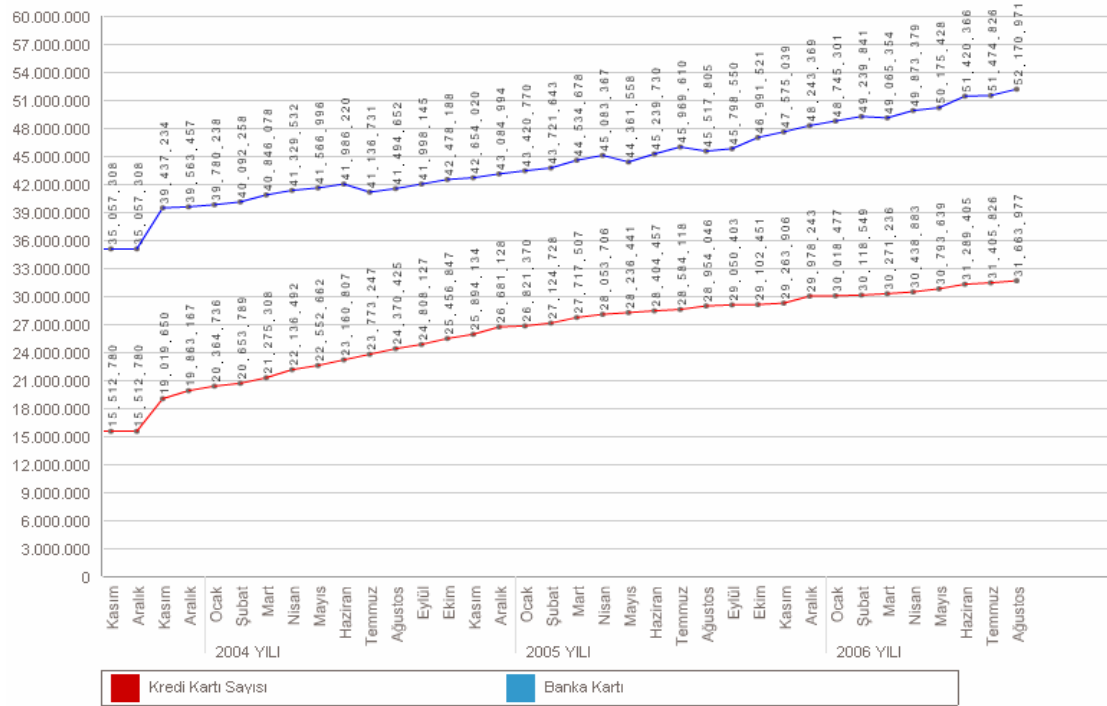
Diners Club 1968 yılında Türkiye’de çıkartılan ilk kredi kartıdır. Diners Club’tan sonra Türk Ekspres Havacılık ve Turizm Limited şirketi “American Express” kartları ile piyasaya giriş yapmıştır. Bu iki kart 1975 yılına kadar rakipsiz olarak faaliyetini sürdürürken, 1980’den başlayarak da Master-Card ve Visa markalı bankalara ait kredi kartları piyasaya girmiştir.

Diners Club kartları, “Charge Card” yani son ödeme tarihi itibariyle bakiyenin tümünün ödenmesi gereken kartlar niteliğindedir. Bu kartlarda limit tahsisi nakit bloke tutar karşılığı olarak veriliyordu. Prestij unsuru olarak algılandığı için, son derece itibarlı ve varlıklı kişilere verilirdi ve bu nedenle kullanıcı sayısı son derece düşüktü. 1975 yılına kadar Diners Club’ın sayısı 10,000’i geçmedi. Bu kartların kullanım alanları Türkiye ile sınırlıydı ve bugün kullandığımız POS terminalleri yerine Imprinter adı verilen mekanik cihazlar vasıtası ile işlem gerçekleştiriliyordu.

Provizyon ise telefonla alınmaktaydı. Kartların kullanılabildiği işyerleri de son derece sınırlıydı.

1980’li yılların başında Türkiye, nüfus sayısı açısından Avrupa’daki en büyük nüfusların başında gelmekteydi. Bir başka deyişle potansiyeli yüksek bir pazar konumundaydı. Yatırımları çok düşük ve gelişmekte olan bir ülkeydi. Bu nedenle Türkiye’de faaliyet gösteren bankalar da müşterilerine kredi kartı vermeye başladılar. Bu dönemde gold, classic gibi farklı tipte ürünler de ortaya çıkmaya başladı. 1984 yılında Visa’nın Türkiye Ofisi’ni açması ivmeyi hızlandırdı. 1990’da 13 özel ve kamu bankasının ortaklığı ile farklı bankalara ait kredi kartları arasında otorizasyon ve takas işlemlerini gerçekleştirmek amacıyla Bankalararası Kart Merkezi A.Ş. (BKM) kuruldu. Aynı yıl kredi kartı ile yapılan harcamalarda puan uygulaması başladı. 1993 yılında ilk elektronik POS terminali kullanıma girdi. 1993 yılında ise Europay/MasterCard’ın Türkiye Ofisi’ni açmasıyla kart sayısı artışı hız kazandı. Türkiye’de taksitli kredi kartlarının kullanımına ise 1999 yılında başlandı. Ayrıca katalog ve puan uygulamaları farklı bir boyut kazanarak, kazanılan puanların anlaşmalı işyerlerinde nakit gibi harcanabilmesi gibi özellikler pazara sunuldu.

Ağustos 2006 itibariyle Türkiye’de toplam 31,7 milyon adet kredi kartına karşılık 15,5-16 milyon adet kredi kartı kullanıcısı olduğu varsayılmaktadır. Kredi kartı ve kart müşterisi sayılarındaki artış, özellikle 1999 yılı başı itibariyle son derece ciddi bir ivme kazanmıştır. Bundaki başlıca sebep, kredi kartları ile birlikte müşterilere sunulan hizmetlerdeki çeşitliliğin artmasıdır. Ayrıca bankalar açısından da hazine operasyonlarının eski karlılığını devam ettirmemesi ve bankaların bireysel bankacılık alanında (özellikle de kredi kartları) yapmış oldukları yatırımlar da, bu sonucun oluşmasında etkili olmuştur. Hala gelişmiş batılı ülkelerin nüfuslarına oranla pazarda bulunan kredi kartı sayıları baz alındığında, Türkiye’deki kredi kartı sayısının daha da artması gerektiği ortaya çıkmaktadır. Şu andaki konumu ile de Türkiye Avrupa’daki üçüncü en büyük pazar olma özelliğine sahiptir. Türkiye’de son yıllardaki kredi kartı ve banka kartı sayıları ile ilgili bilgiler Şekil 9.1’de gösterilmiştir.



Şekil 9.1 : Son Yıllara Ait Kredi Kartı ve Banka Kartı Sayıları

BKM tarafından hazırlanan ve Tablo 9.1’de gösterilen rapora göre 2006 yılı ilk dokuz ayı sonunda, kredi kartı ile yapılan işlemlerin adedi bir önceki yılın aynı dönemine göre yaklaşık %4 artış göstererek 1 milyar 5 milyona, tutar ise %25 artarak 78 milyar 972 milyon YTL’ye ulaşmıştır. 2006 yılı üçüncü dönemde yapılan ve toplamda 28 milyar 800 milyon YTL olarak gerçekleşen işlemlerin 26 milyar 95 milyon YTL’si alışveriş, 2 milyar 704 milyon YTL’si ise nakit çekim işlemleri olarak değerlendirilmiştir.

BKM’nin 2006 yılının ikinci dönemi için yaptığı açıklamaya göre, kredi kartlarıyla gerçekleştirilen harcamalarda ilk sırayı %16.05’lik payla market ve alışveriş merkezleri almıştır. Market ve alışveriş merkezlerinde gerçekleştirilen işlemleri %15.7 ile benzin istasyonları, %9.6 ile giyim ve aksesuar alışverişleri izlemiştir. Telekomünikasyon sektörü %6.8 pay ile dördüncü sıraya yerleşirken, elektirik-elektronik eşya ve bilgisayar alışverişleri ise %6.7 ile beşinci sırada yer almıştır.

İnternet üzerinden yapılan alışverişler incelendiğinde ise 2006 yılının ilk altı aylık döneminin Türkiye’de sanal poslar üzerinden gerçekleştirilen e-ticaret işlemlerinin adedinin, bir önceki yılın aynı dönemine göre %12’lik artışla 8 milyon 582 bin adede yükseldiği, işlem hacminin de %70’lik artışla 957 milyon 680 bin YTL’ye ulaştığı saptanmıştır.

Tablo 9.1: Yerli ve Yabancı Kredi Kartlarının Yurtiçi Kullanımı

Dönem	İşlem Adedi			İşlem Tutarı(Milyon YTL)		
	Alışveriş	Nakit Çekme	Toplam	Alışveriş	Nakit Çekme	Toplam
OCAK	95,730,796	4,759,479	100,490,275	5,283.37	574.43	5,857.80
ŞUBAT	88,658,019	4,571,566	93,229,585	5,082.47	566.49	5,648.96
MART	102,790,271	5,087,390	107,877,661	6,061.72	633.69	6,695.42
1. DÖNEM	287,179,086	14,418,435	301,597,521	16,427.57	1,774.62	18,202.19
NİSAN	97,583,062	4,970,532	102,553,594	5,894.33	666.3	6,560.63
MAYIS	111,397,604	5,662,818	117,060,422	6,792.63	743.86	7,536.50
HAZİRAN	107,177,726	5,306,514	112,484,240	6,745.91	735.49	7,481.40
2. DÖNEM	316,158,392	15,939,864	332,098,256	19,432.88	2,145.65	21,578.54
TEMMUZ	107,715,741	5,214,481	112,930,222	6,933.19	764.22	7,697.41
AĞUSTOS	105,478,224	5,625,831	111,104,055	6,823.76	831.49	7,655.26
EYLÜL	107,599,389	5,123,061	112,722,450	7,087.06	766.65	7,853.72
3. DÖNEM	320,793,354	15,963,373	336,756,727	20,844.02	2,362.38	23,206.40
EKİM	107,887,943	4,838,007	112,725,950	7,139.93	725.48	7,865.41
KASIM	100,259,964	4,574,149	104,834,113	6,769.98	686	7,455.99
ARALIK	108,721,964	5,090,053	113,812,017	7,427.42	758.04	8,185.47
4. DÖNEM	316,869,871	14,502,209	331,372,080	21,337.34	2,169.53	23,506.88
2005 YILI	1,241,000,703	60,823,881	1,301,824,584	78,041.83	8,452.19	86,494.02
OCAK	104,289,274	4,424,487	108,713,761	6,684.31	684.79	7,369.10
ŞUBAT	95,841,800	4,662,458	100,504,258	6,434.21	693.87	7,128.09
MART	110,269,108	5,027,558	115,296,666	7,573.94	764.35	8,338.29
1. DÖNEM	310,400,182	14,114,503	324,514,685	20,692.47	2,143.02	22,835.49
NİSAN	107,776,790	4,816,203	112,592,993	7,702.02	742.45	8,444.47
MAYIS	114,320,833	5,462,625	119,783,458	8,491.35	860.48	9,351.83
HAZİRAN	115,628,194	5,139,327	120,767,521	8,712.32	827.18	9,539.51
2. DÖNEM	337,725,817	15,418,155	353,143,972	24,905.70	2,430.12	27,335.83
TEMMUZ	112,795,805	5,177,734	117,973,539	8,702.06	875.54	9,577.61
AĞUSTOS	101,283,898	5,494,602	106,778,500	8,774.19	927.05	9,701.25
EYLÜL	98,216,274	5,151,357	103,367,631	8,619.44	902.37	9,521.82
3. DÖNEM	312,295,977	15,823,693	328,119,670	26,095.71	2,704.98	28,800.69

9.2. Kredi Kartı Dolandırıcılık Tipleri [28]

Kredi kartlarının sayısının giderek artması ve aynı doğrultuda sürekli büyüme gösterip çok yüksek meblağlara ulaşan işlem hacmi, beraberinde bu pazardan haksız kazanç sağlamak isteyen dolandırıcıları da ortaya çıkarmıştır. Aşağıda belirtilen alt başlıklarda, kredi kartları ile ilgili çeşitli dolandırıcılık tipleri anlatılmıştır.

9.2.1. Kayıp/Çalıntı Kart Kullanımı

Kart hamilinin kartını kaybetmesi veya çaldırması halinde kartın kötü niyetli kişiler tarafından kullanılmasıdır. Kart hamilleri kartlarının kayıp/çalıntı olduğunu Banka'ya bildirmeleri durumunda, bildirim sonrası işlemlerden sorumlu olmazlar.

Bankalar kart hamillerinin, kredi kartları hareketlerini takip ederek kayıp/çalıntı kart vakalarını en erken zamanda tespit etmeyi hedeflerler. Kart hamillerinin harcama alışkanlıkları dışındaki işlemlerde, kart hamilleri Bankalar tarafından aranarak işlem için teyit talep edilir.

Harcamaların yapıldığı işyerleri, yasal olarak kimlik ve imza kontrol etmek zorunda olduğundan, kimlik kontrol edilmediğinin tespit edildiği durumlarda üye işyerleri söz konusu işlemlerden sorumlu olabilir. Son dönemlerde sadece kayıp/çalıntı kartların kullandırılması amacıyla açılmış paravan işyerleri kurulduğu da gözlenen dolandırıcılıklar arasındadır.

9.2.2. Ele Geçmeyen Kartlar

Kartlar posta, özel kargo şirketleri ve Banka şubesi aracılığıyla kart hamillerine ulaştırılırken kötü niyetli kişilerce ele geçirilebilmekte ve kullanılabilmektedir.

Bankalar kartlardaki ilk kullanımları takip ederek, kartıyla ilk kez yüksek meblağlı işlem yapan kart hamillerinden harcama teyiti ve kartın ellerine ulaşip ulaşmadığı hakkında bilgi talep ederler. Özel kargo şirketi üzerinden harcama yapılan kartın sorumluluğu ilgili kurumdadır.

9.2.3. Sahte Başvuru

Dolandırıcılar, sahte kimlik hazırlayarak, çalıntı kimlik kullanarak ya da 3. şahıslara ait ele geçirdikleri kimlik fotokopileri ve paravan şirket bilgileri ile Bankalara kredi kartı başvurusunda bulunurlar.

Daha önceleri genellikle sahte nüfus cüzdanı hazırlayan dolandırıcıların, son dönemlerde mağdur şahıslardan nüfus cüzdanı fotokopisi elde ederek başvuru yaptıkları tespit edilmektedir. Dolandırıcılar, yurt dışına işçi olarak gönderme bahanesiyle ve gazetelere verdikleri iş ilanlarına başvuran mağdur şahıslardan başvuru formundaki özlük bilgileri ile birlikte kimlik fotokopisini talep ederek kimlik elde etmektedirler.

Banka tarafından tahsis edilen kartları teslim alır almaz tüm limitini bitiren dolandırıcılar, kart teslim adresi olarak kullanılan paravan şirketleride kapatırlar.

Harcamalar sonrasında Banka tarafından şahıslara ulaşılması mümkün olmaz. Ödenmeyen kartta oluşan borç Banka zararını oluşturmakla birlikte, adına kart çıkartılan mağdur şahısların olumsuz kredibilite sicilleri oluşur.

9.2.4. Posta, İnternet ve Telefon ile Yapılan Mal Sipariş Dolandırıcılığı

Kredi kartı numarası kullanılarak posta, internet ve telefon ile yapılan siparişlerde kart hamili ve kart fiziksel olarak satıcı firmanın karşısında bulunmaz. Bu nedenle bu tip işlemler riskli kabul edilmektedir. Başkasına ait geçerli bir kart numarası ile önceden ayarlanmış bir adrese mal gönderilerek dolandırıcılık gerçekleştirilebilmektedir.

Bu yollarla yapılan dolandırıcılıkların önlenmesi için kart hamillerinin, kart bilgilerini (kart numarası, son kullanma tarihi, güvenlik numarası) tanınmayan firmalara vermemeleri önem taşımaktadır.

9.2.5 ATM Dolandırıcılıkları

Dolandırıcının, kart sahibinin şifresini işlem sırasında gözleyerek ya da başka yöntemlerle ele geçirmesi ve daha sonra kartı değiştirme, çalma veya ATM'in giriş haznesine önceden çeşitli cisimler yerleştirilerek kartı ele geçirmesi ve kullanması ile gerçekleştirilen dolandırıcılık türüdür.

ATM üzerinden kart kopyalama vakalarında son zamanlarda yaşanan ATM dolandırıcılık çeşitlerindendir. Kart okuyucu haznesine, kart kopyalama cihazı yerleştiren dolandırıcılar, ATM klavyesini gören bir alana da ATM'in parçasıymış gibi mikrokamera yerleştirmektedir. Böylece işlem yapan bir kart hamilinin, kopyalama cihazı kart bilgilerini, kamera ise şifre giriş anını görüntüleyerek şifre bilgisini ele geçirmektedirler.

Kart hamillerinin, ATM'de işlem sırasında yaşanabilecek herhangi bir sorunda 3. şahıslardan gelen yardım tekliflerini reddederek, Bankalarını aramaları gerekir. Bankalar hiçbir durumda kart hamilinden şifre bilgisini talep etmezler. ATM telefonlarından ya da 3. şahıslar tarafından Bankayla görüştüklerini belirterek verdikleri telefonlardan şifre bilgisi talep ediliyorsa yanıt verilmemelidir.

9.2.6. Sahte Kart

Sahte kart, gerçeğine benzer şekilde sahte olarak üretilmiş kart kavramını açıklamak için kullanılır.

Üretiliş itibariyle çeşitli şekillerde sahte kartlar vardır. Öncelikle gerçek kart hamillerinin kart bilgileri, kötü niyetli işyerleri tarafından encoder denilen kart kopyalama cihazından geçirilerek kopyalanmaktadır. Söz konusu kopyalama işlemi özellikle restoranlar gibi, işlem anında kartın kart hamilinden uzaklaştığı işyeri tiplerinde görülür. Dolandırıcılar lüks restoranlardaki garson, kasiyer gibi görevlilerle anlaşarak gold, platinum kart gibi yüksek limitli kartların kopyalanması işlemini gerçekleştirirler.

En çok kullanılan yöntem gerçek bir kartın manyetik şeridinin kopyalanarak sahte bir kartın manyetik şeridine aktarımıdır. Farklı bir yöntem, gerçek bir kartın manyetik şerit bilgilerinin daha önce kayıp/çalıntı olmuş başka bir gerçek kartın manyetik şeridine aktarılmasıdır.

Sahte kart dolandırıcılıklarının önlenmesi için kart hamillerinin işlem anında kartlarını göz önünden ayırmamaları önem arz etmektedir. Bununla birlikte Bankalar, kart hamillerinin harcama alışkanlıkları dışındaki yüksek meblağlı işlemler için kart hamillerinden onay talep ederek sahte kart dolandırıcılıklarını en erken zamanda tespit etmeyi hedeflerler.

9.2.7. Nakit Temin Etmeye Yönelik Kullanım

Kart hamilinin kredi kartını kendi, ortağı olduğu ya da yakını olduğu firmada kullanmasıdır. Böylece kart hamili Bankaya hiçbir faiz ödmeden çok düşük maliyetle nakit para kullanmaktadır. Bu kullanım şekli uluslararası kredi kartı kullanım kurallarına da aykırı olup, kart hamilinin nakit sıkıntısı içerisinde olduğunu da göstermektedir.

Özellikle faiz oranlarının yükseldiği kriz dönemlerinde, kart hamillerinin kartlarıyla nakit temin etmeye yönelik işlemler yaparak söz konusu nakitleri Bankalardan faiz kazanmak amacıyla kullandıkları gözlenmiştir.

Bankalar nakit temin etmeye yönelik işlem tespit ettiklerinde, öncelikle kart hamillerini uyarmakta, aynı işlemin tekrarı durumunda kart iptaline gitmektedirler.

10. UYGULAMA: XYZ BANK'TA KREDİ KARTI BAŞVURU AŞAMASINDA VERİ MADENCİLİĞİ İLE SAHTEKARLIK TESPİTİ

10.1. Giriş

Veri madenciliği yöntemlerinin, çeşitli karar verme aşamalarında kullanımı kuramsal açıdan mümkündür. Veri madenciliğinde amaç, çok büyük miktardaki veriden manuel olarak çıkartılması zor veya oldukça zahmetli olan değerli bilginin otomatik olarak elde edilmesidir.

Her sektörün kendine özgü sorunlarına veya her şirketin kendi iç disiplinine göre öncelikli olarak çözüm bulmak istenilen konular rahatlıkla değişkenlik gösterebilir. Sonuçta her şirket, kendisini kara geçirecek yolda farklı taktikler kullanacak ve farklı adımlarla yoluna devam edecektir. XYZ Bank eski yöneticilerinden Kunter Kutluay, kendisiyle yaptığım görüşmede bir bankanın kara geçmesi için gerekli olan durumu Denklem 10.1'deki gibi formülize etmiştir [29].

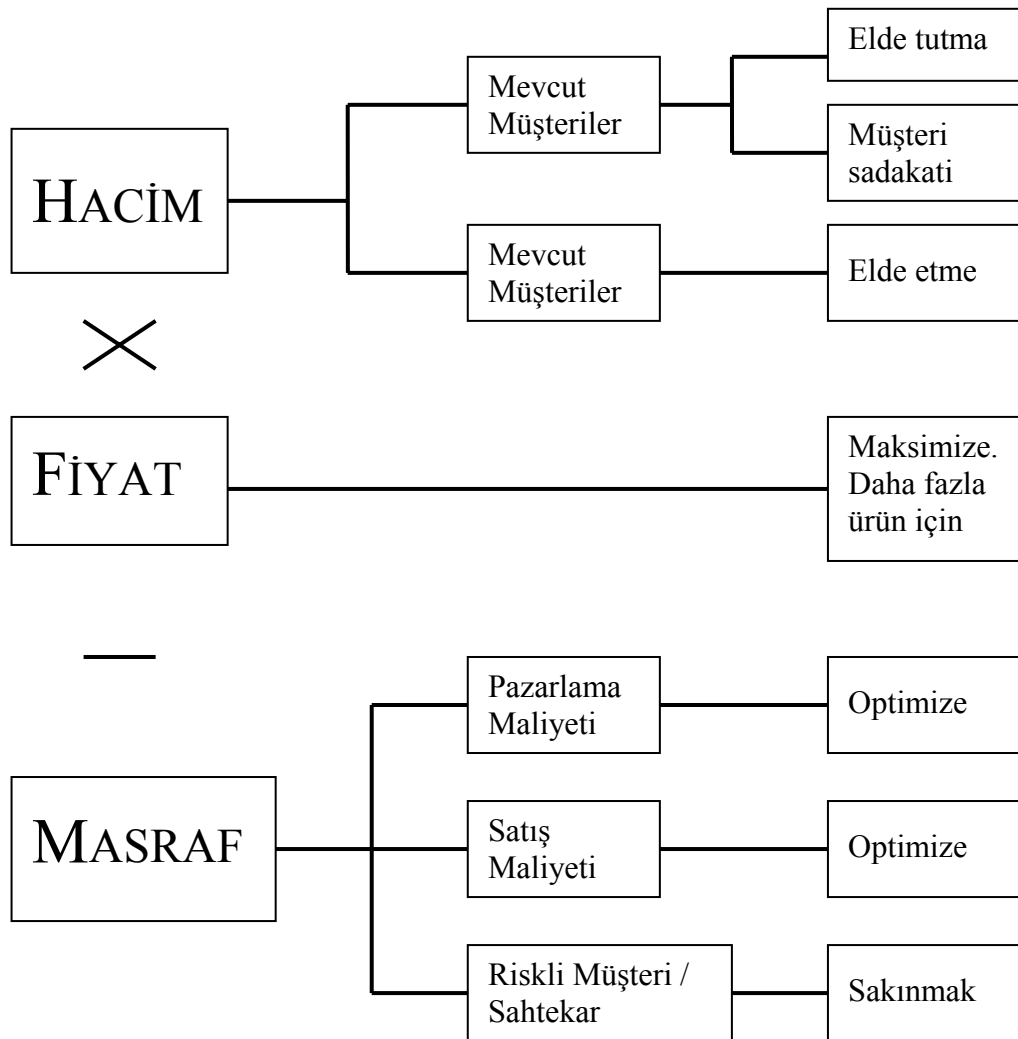
$$Kar = (Hacim \times Fiyat) - Masraf \quad (10.1)$$

Denklem 10.1'den de anlaşıldığı gibi bir Banka'nın kara geçebilmesi için hacim ve fiyatın artırılması, masrafların ise azaltılması gerekir. Burada hacim ile anlatılmak istenen Banka'ya ait müşteri sayısıdır. Dolayısıyla hacmi artırabilmek için Banka'nın öncelikli olarak mevcut müşterilerini elinde tutması gerekir. Bu amaç doğrultusunda sistemli analizler yapılarak, müşterilerin memnuniyet seviyeleri belirlenmeli ve Banka'dan ayrılma ihtimali olan müşterileri bu fıkırdan vazgeçirecek Müşteri İlişkileri Yönetimi (Customer Relationship Management, CRM) çalışmaları yapılmalıdır. Hacmi artırmak için önemli bir diğer husus da yeni müşterileri elde edebilmektir. Mevcut müşteri profilini öğrenmek ve benzer nitelikteki olası müşterileri Banka'ya kazandırabilmek için yine sistemli analizler yapmak şarttır. Yeni bir müşteri elde etmenin, müşteriyi elde tutmaktan çok daha maliyetli olduğu unutulmamalı ve mevcut müşteriye çok önem verilmelidir.

Fiyatlandırma da bir Banka'yı kara geçirecek önemli kalemlerden biridir. Mevcut ürünleri müşterilerin talep gösterebileceği maksimum ücretlerle piyasaya sunabilmek, Banka'yı kara geçirebilecek faaliyetlerden biri olacaktır. Yapılacak analizlerle, müşterilerin hangi ürünleri hangi fiyat aralıklarında satın aldıkları belirlenmeli ve daha fazla ürün için maksimum fiyat oluşturulmalıdır.

Bir firmanın masraflarını azaltması ise kara geçmek için şüphesiz etkin bir yoldur. Bir banka için yoğun olarak masraf yapılan pazarlama ve satış maliyetlerinde yapılacak tasarruf, bu amaç doğrultusunda faydalı olacaktır. Buna ek olarak bankaları maddi olarak büyük zarara uğratan riskli müşterilerden ve sahtekarlardan sakınmak da, çözüm aranan ilk konular arasındadır.

Şekil 10.1'de de gösterilen bu kalemlerle ilgili analizlerde veri madenciliği teknolojileri kullanılarak etkin sonuçlara ulaşılabilir.



Şekil 10.1 : Bankalarda Kara Geçmek İçin Gerekli İşlemler

10.2. XYZ Bank

XYZ Bank, 1960'lı yıllarda kurulmuştur. 1990'lı yılların ortalarında yerli bir Holding'in XYZ Bank'ın ana sermayedarı olmasıyla birlikte bankanın sermaye tabanını güçlendirme, verimliliğini ve piyasa etkinliğini artırma süreci hızlanmıştır. 1999 yılından bu yana ticari ve bireysel bankacılıkta lider bir konumda yer almak üzere hızlı bir büyüme stratejisi benimseyen XYZ Bank, bu yönde çalışmalarını sürdürmektedir. 2005 yılında ise , Avrupa'nın önde gelen bir finans grubu, XYZ Bank'ın ana hissedarlarının elinde bulunan % 89.3 oranındaki hisseleri satın almıştır.

Bugün yeni yapılanmasında kurumsal bankacılık için oluşturulan beş özel şube dışında tüm segmentlere hizmet veren aktif 159 şube, 254 ATM, 94 kiosk, 65.000 POS terminali, 6 web portalı, İnternet bankacılığı ve çağrı merkezinden oluşan gelişmiş bir dağıtım kanalına sahip olan XYZ Bank, Türkiye'nin lider finansal kurumlarından biridir.

Alternatif dağıtım kanallarına ve kredi kartlarına önem veren XYZ Bank'ın 700,000 civarında kredi kartı müşterisi bulunmaktadır.

10.3. XYZ Bank'ın Problemi

Türkiye, kredi kartı sayısı bakımından Avrupa'nın üçüncü büyük pazarı konumundadır. Bankacılık sektöründe, özellikle de kredi kartlarında yaşanan büyük rekabet bankaları, müşteri tabanlarını büyütme ve daha riskli segmentler hedeflemeye yöneltilmektedir. Bankalar daha fazla müşteri kazanma çabalarını sürdürdükçe de, kredi kartı başvuru sayısı giderek artmaktadır. Bütün bunların sonucunda hem başvuru hem de işlem aşamasındaki sahtekarlıkların sayısı hızlı bir artış göstermektedir.

XYZ Bank Kredi Kartı Güvenlik Birimi, olası sahtekarlıkları başvurusu onaylanmadan önce tespit etmek istemektedir.

10.4. Sahtecilik Tespiti

Zaman ve kaynak kısıtları nedeniyle, günümüzde pek çok kurumun sadece bilinen sahtecilik yöntemlerine yoğunlaştıkları görülmektedir. Veri madenciliği, bu engeli aşmaya ve basit sahtecilik kurallarından daha fazlası ile uğraşmaya yardımcı

olmaktadır. Günümüzde gelişmiş yazılımlar kullanılarak hem bilinen hem de bilinmeyen sahtecilik kuralları tespit edilebilir.

İki çeşit sahtecilik vardır. Bunlar:

- 1) *Bilinen*: Daha önce tespit edilerek tanımlanmış belli sahtecilikler bu sınıfa girer. Bu çeşit sahteciliklerin yakalanması daha kolaydır. Dolayısıyla çoğu sahtecilik projesinde bu sınıfa giren durumlara ağırlık verilmektedir.
- 2) *Bilinmeyen*: Daha önce tespit edilmemiş sahtecilikler bu sınıfı oluşturur. Yakalanması zor olan sahtecilik çeşididir. Elle tespit ve otomatik kurallar, en çok kullanılan sahtecilik tespit yöntemleridir.

Sahtecilik tespiti için elle tespit, otomatik tespit ve öngörüşel modelleme teknikleri kullanılmaktadır.

Elle yapılan incelemeler sahtecilik yakalanmasında en duyarlı sonuçları verir. Ancak incelenmesi gereken büyük veri miktarları böyle bir yöntemin kullanılmasını engellemektedir. Otomatik kurallar ise uzman bilgisinin kısmi kodlamasıdır. Bilinen sahtecilik tipleri için iyi sonuçlar vermelerine rağmen iki kısıtlamaları vardır:

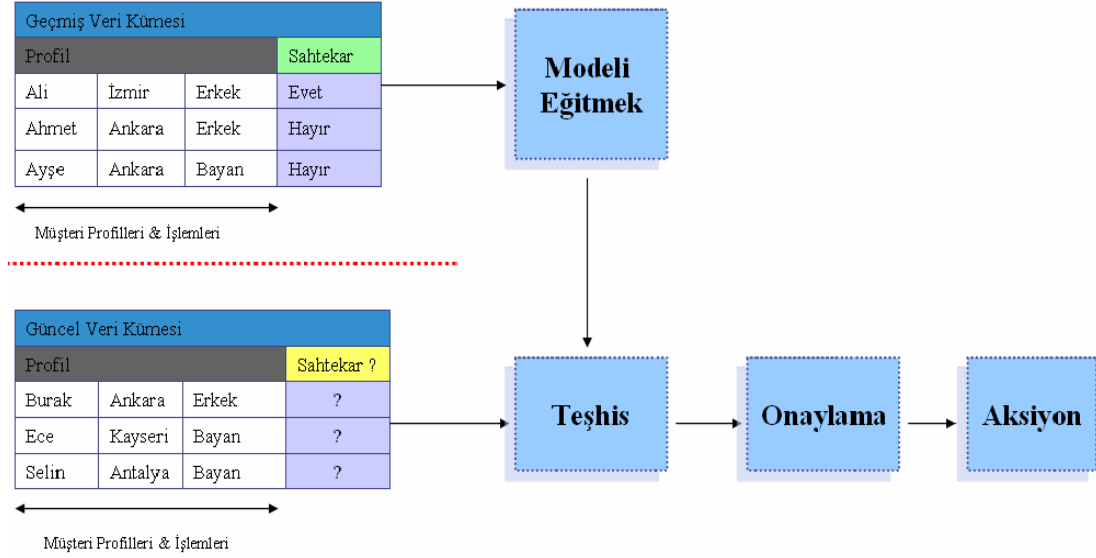
- 1) Otomatik kuralları güncel tutmak zordur çünkü yeni kuralların elle eklenmesi gerekmektedir.
- 2) Önceden tanımlanmış kuralları kullanarak yeni sahtecilik tiplerini yakalamak zordur.

Öngörüşel modelleme, varolan sahtecilik tespit yöntemlerinin bu kısıtlarını ortadan kaldırmaktadır. Öngörüşel modelleme ile bilinen sahtecilik durumları otomatik olarak yakalanırken, bilinmeyen tipleri tespit edebilmek için de sıradışı değerler bulunur. Etkin bir sahtecilik tespit yöntemi elle tespit, otomatik kurallar, bilinen sahtecilik tiplerinin profili ve bilinmeyen tiplerin ortaya çıkarılması adımlarının bir birleşimi olmalıdır.

10.4.1. Bilinen Sahtecilikler İçin Süreç

Bu yöntem bilinen sahtecilik durumlarını da içeren bir öğrenme veri kümesi olan durumlarda kullanılır. Ödemelerini düzgün yapan müşterileri ve sahtecilik yaptığı tespit edilmiş olan müşterileri birlikte içeren bir veri kümesi bu duruma iyi bir örnektir. Şekil 10.2’de belirtilen süreçte de gözüktüğü gibi, öncelikle sahtecilik

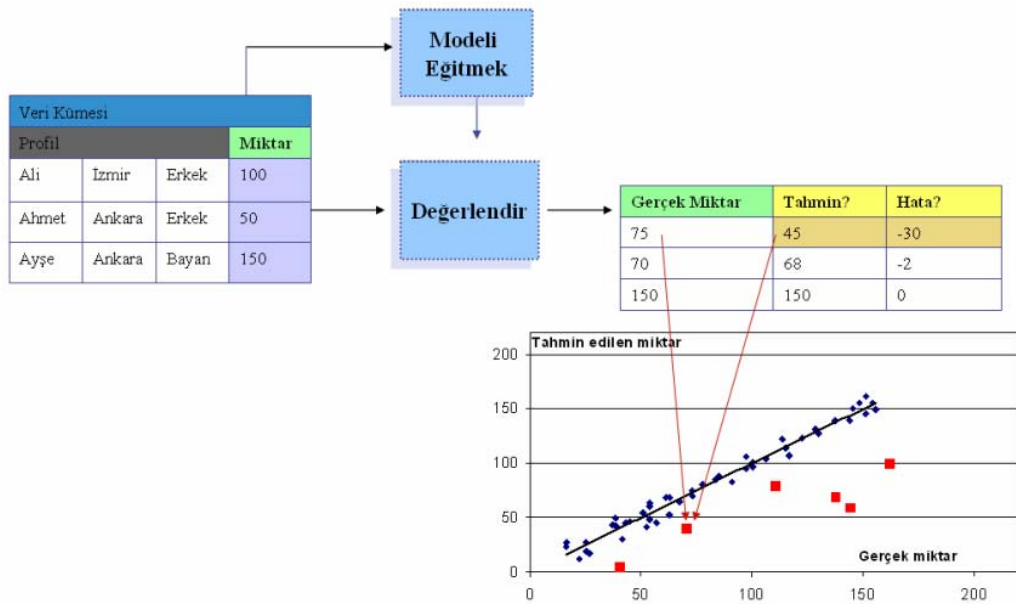
yapan müşterilerin özellikleri analiz edilir ve bu model diğer müşterilere uygulanarak mümkün sahtecilik durumları tespit edilir.



Şekil 10.2 : Bilinen Sahtecilikler İçin Süreç [30]

10.4.2. Bilinmeyen Sahtecilikler İçin Süreç

Bu yöntem, sahtecilik kaynakları bilinmediğinde ve bulmanın kolay olmadığı durumlarda kullanılır. Sıradışı değerler tespit edilerek incelenir. Böylece bilinmeyen sahtecilik durumlarının yakalanması kolaylaşır. Bu duruma örnek olarak, kurumun tüm müşterilerinin fatura bilgilerinin analizi sonucunda ortaya çıkan sıradışı değerler için daha yakın bir inceleme yapılması verilebilir.



Şekil 10.3 : Bilinmeyen Sahtecilikler İçin Süreç [30]

10.5. Proje Konusu

Uygulamamın konusu; bireysel müşterilerin kredi kartı başvurularını, müşterilerin sahteci olmasında etkili müşteri bilgilerini dikkate alarak değerlendiren bir sistem kurmak ve XYZ Bank'ı zarar uğratabilecek sahtecileri tespit etmektir.

2004 yılının Nisan ayında başlanan projede, oluşturulan model 20 Nisan 2004 tarihinde uygulamaya alınmıştır. Projede, Yapısal Risk Mizimizasyonu algoritmasını kullanan KXEN Analytic Framework adlı veri madenciliği yazılımı kullanılmıştır.

10.5.1. Projenin Amacı

Gerçekleştirdiğim projenin amaçları

- Müşterinin sahteci olma riskini doğru bir şekilde belirlemek,
- Mevcut çalışan kapasitesini artırmadan, tespit edilen sahte başvuru sayısını artırmak,
- XYZ Bank'ın sahteciler nedeniyle uğradığı maddi zararı azaltmak,
- Başvuru değerlendirme çalışma gruplarına karar vermede yardımcı olmak,
- Başvuru değerlendirmeyi olabildiğince gerçekçi koşullara dayandırarak, insiyatif kullanımını azaltmaktır.

10.6. XYZ Bank'daki Mevcut Durum [31,32]

Sahte başvurular (SB), Güvenlik Birimi Ön İzleme Servisi tarafından takip edilmektedir. Olası sahtekarlıkları başvurusu onaylanmadan önce tespit etmek isteyen XYZ Bank Kredi Kartı Güvenlik Birimi, proje öncesinde başvuruları elle değerlendirmekteydi. Bu yöntem, toplam başvuruların yalnız %14'ünün kontrol edilebilmesine imkan veriyordu. Ayrıca, sahtekarlığı tespit etmek için başvuru sırasında verilen demografik bilgilere ve güvenlik ekibinin tecrübesine dayanıldığı için, kart sayısının artmasıyla birlikte, sahtekarlık tespitinin maliyetinde de hızlı bir artış görüldü.

Proje öncesi Ön İzleme Servisi'nin performansına ait bilgiler Tablo 10.1'de belirtilmiştir.

Tablo 10.1: Proje Öncesi Ön İzleme Servisi Performansı

	Günlük Ortalama İncelenen	Günlük Ortalama Sahte Başvuru	Günlük Ortalama Bağlantılı	Günlük Ortalama Kurtarılan (YTL)
Ön İzleme	517	7	1	14,312.5

Ön İzleme Servisi, Proje öncesi günlük olarak ortalama 517 başvuruyu inceleyebilmektedir. Günlük olarak belirlenen ortalama sahte başvuru sayısı ise 7 ve bu sahtecilerle bağlantılı olduğu tespit edilen ortalama kişi sayısı da 1 ile sınırlı kalmaktaydı. Ortalama olarak baktığımızda ise günlük 14,312.5 YTL kurtarılmaktaydı.

Bağlantılı kişi tespiti, sahteci olduğu belirlenen bir kişinin bilgilerinin veri ambarında sorgulanmasıyla yapılmaktadır. Genellikle telefon ve iş yeri bilgileri sorgulanır. Örneğin sahteci bir kişinin verdiği iş telefonu numarasını veren başka kişiler de var ise, bu kişiler detaylı incelemeye alınır.

XYZ Bank'ın proje öncesi başvuru değerlendirme süreci Sekil 10.4'de gösterilmiştir. Bu şekilde de görüldüğü gibi gelen başvurular öncelikle İstihbarat Bölümü tarafından incelenmektedir. Bu bölümde başvuru ilk olarak havuz kontrolüne tabi tutulur. BKM, TCMB (Türkiye Cumhuriyeti Merkez Bankası) listelerinde ve SABAS (Sahte Bilgi/Belge/Beyan/Başvuru Alarm Sistemi) sisteminde otomatik olarak değerlendirilen başvuruların, buralardaki bir kayıtlarla eşleşip eşleşmediği kontrol edilir. Bir eşleşme tespit edilemeyen başvurular kredibilite durumunu öğrenmek için KKB'ye (Kredi Kayıt Bürosu) gönderilir. Olumlu rapor alınanlar bir sonraki aşamaya geçer ve bu kişiler için telefon ile iş yeri bilgilerini doğrulama çalışması yapılır. Bahsedilen aşamaların herhangi birinden olumsuz yanıt alan başvuru reddedilir. Reddedilen başvurular arasında sahteci olduğu tespit edilenler ise Sahte Başvuru Havuzuna yollanır.

XYZ Bank'ta gelen başvuruların %58'i İstihbarat Bölümü tarafından reddedilmekte, %42'si ise onaylanarak Tahsis kısmına geçmektedir. Tahsis aşamasında kartlar basıma gider.

Kartlar basımdayken, bir sonraki aşama olarak ön izleme raporları alınır. Manuel olarak yapılan bu işlemde, kartı basıma giden onaylanmış başvurular için XYZ Bank tarafından riskli olarak belirlenmiş bazı ölçütlere göre raporlar çekilir ve inceleme

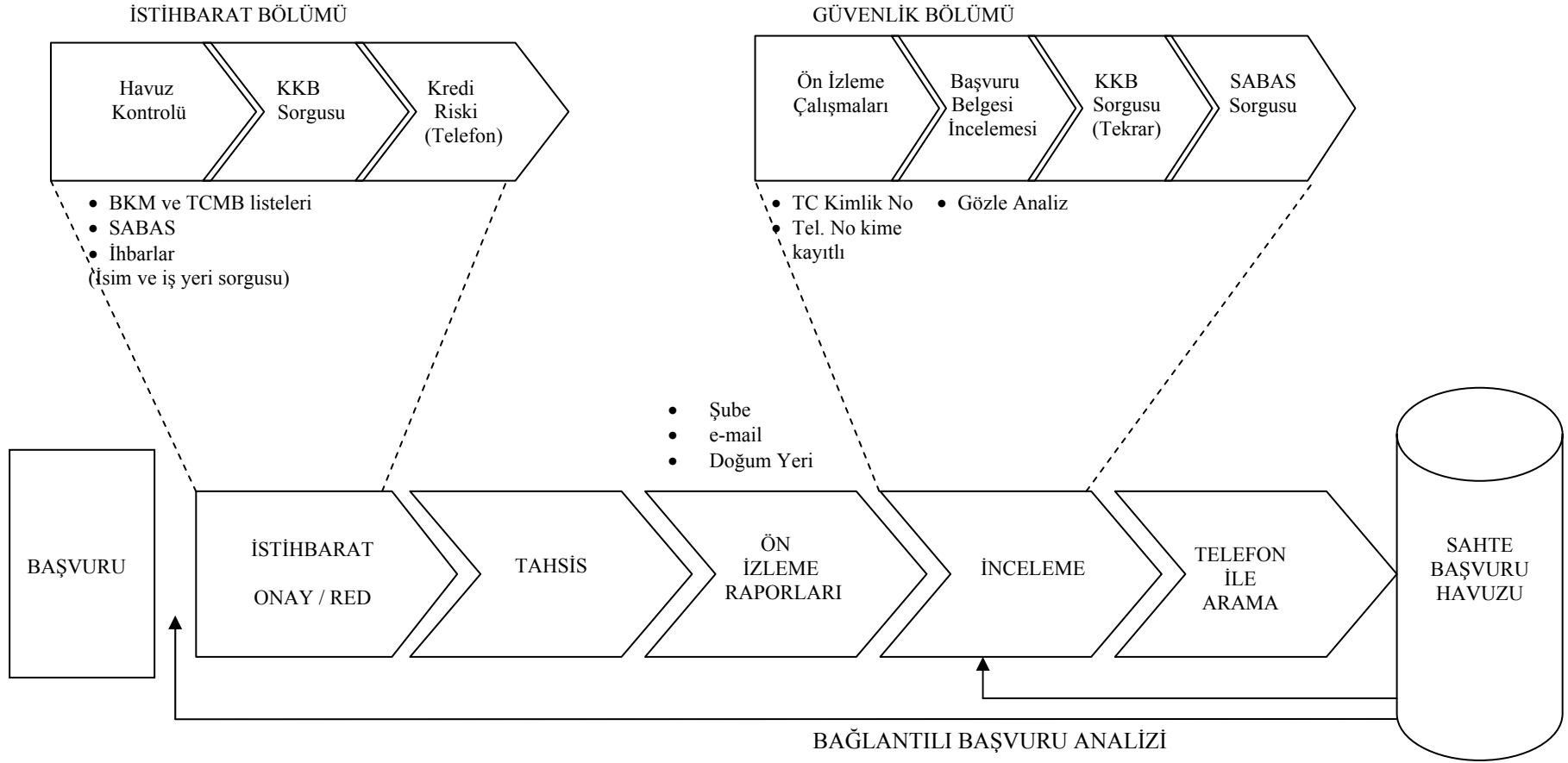
yapılır. Bu ölçütler; başvuru yapılan şube, e-mail hesabı ve doğum yeri bilgileridir. Bu aşamada tüm başvuruların sadece %14'ü kontrol edilebilmektedir. Bu da günde yaklaşık olarak 1000 başvuruya denk gelir.

Tüm başvuruların elde kalan bu %14'lük kısmı ise bir sonraki aşamada Güvenlik Bölümü'ne yollanır ancak zaman kısıtlaması yüzünden Güvenlik Bölümü'nde gelen belgelerin yarısı yani tüm başvuruların %7'si incelenebilmektedir. Bu da günde yaklaşık olarak 500 kişinin Güvenlik Bölümü tarafında incelenebildiği anlamına gelmektedir.

Güvenlik Bölümü öncelikle ön izleme çalışmaları ile çeşitli ekran kontrolleri yapmaktadır. Bu kontrollerin arasında, "tckimlik.nvi.gov.tr" web adresinden yapılan T.C. kimlik numarası kontrolü ve "ttrehber.gov.tr" adresinden yapılan sabit telefon numarası kontrolü yer almaktadır. Bu kontrollerden olumlu yanıt alınan başvurulara ait başvuru belgelerinin incelenmesi ise ikinci aşamadır. Gözle yapılan bu analiz bittikten sonra gerekli duyulursa tekrar bir KKB ve detaylı SABAS sorgusu yapılmaktadır. Güvenlik Bölümü'ndeki bahsedilen aşamaların herhangi birinden olumsuz yanıt alan başvuru Sahte Başvuru Havuzuna yollanır.

Güvenlik Bölümü'nden de sorunsuz geçen başvurular için son kontrol, başvuru sahibinin ibraz ettiği telefon numaralarından ev, cep ve iş telefonu sırasında uygun olarak aranarak güvenlik teyidinden geçirilmesine yönelik kontroldür. XYZ Bank'ta tüm başvuruların sadece %1.5'i aranabilmektedir. Bu da günde yaklaşık 100 aramaya denk gelmektedir.

Başvuru değerlendirme sürecinin herhangi bir aşamasında belirlenen sahteciler, İstihbarat ve İnceleme kısmında tekrar kullanılmak üzere Sahte Başvuru Havuzu'nda toplanırlar. Bağlantılı başvuru analizi yapılırken bu havuzdaki sahteciler kullanılır.

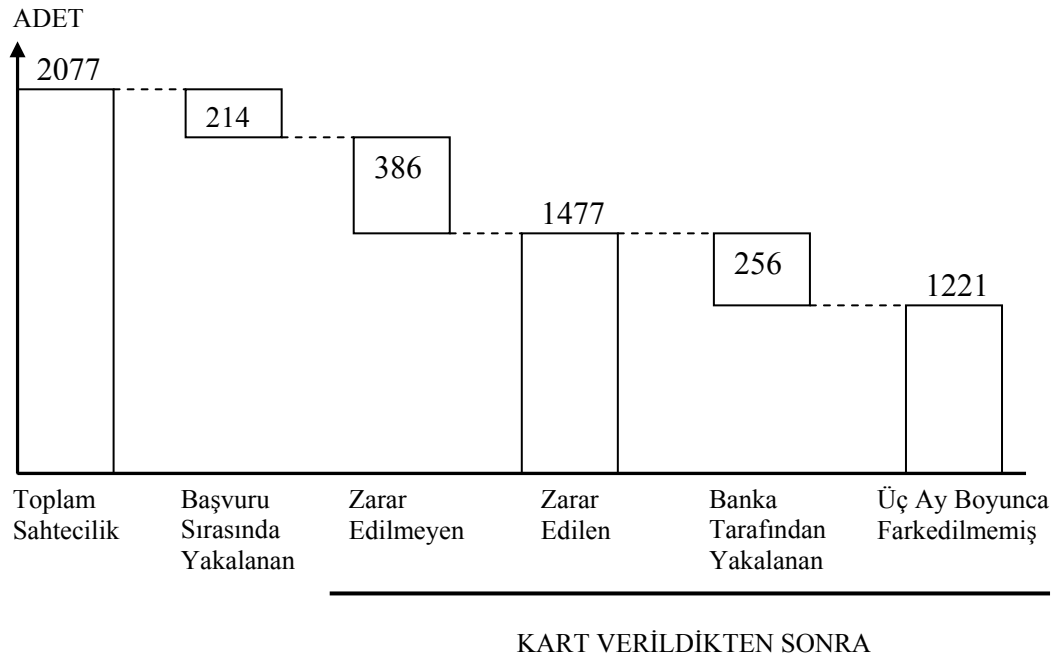


Şekil 10.4 : XYZ Bank Proje Öncesi Başvuru Değerlendirme Süreci

XYZ Bank'ın Ekim 2003 ve Ocak 2004 tarihleri arasındaki üç aylık döneme ait sahtecilik vakalarını incelediğimde ulaştığım bilgiler Şekil 10.5'de gösterilmiştir. Bu dönemde toplam 2077 adet sahtecilik vakası tespit edilmiştir. Bunlardan sadece 214 tanesi henüz başvuru aşamasındayken yakalanabilmiştir. Geriye kalan 1863 başvuru sahibi ise kartlarını almışlardır. Bu 1863 kişinin 386 tanesinden zarar edilmezken, 1477 kişi XYZ Bank'ı zarara uğratmıştır. Zarara sebebiyet veren 1477 kişinin ise 256 tanesi XYZ Bank tarafından yakalanırken, 1221 kişi fark edilmemiştir. Bu 1221 kişi, limitlerinin %90'ını kullanmalarına rağmen ilk 3 ay içerisinde hiç ödeme yapmadan kanuni takibe düşmüş olan kişilerdir.

XYZ Bank, kendi hesaplamalarına göre sahteci bir kişiden ortalama 3,000 YTL zarar etmektedir. Bu bilgiye göre, sadece 3 aylık dönemde XYZ Bank 1477 kişiden toplamda yaklaşık 4,500,000 YTL gibi büyük bir zarara uğramıştır.

Fark edilemeyen sahtecilik vakalarının, fark edilenler veya yakalananlardan daha fazla olması, yapılacak modelleme çalışmasında bu kısma da kesinlikle önem verilmesi gerektiğini ortaya çıkarmaktadır.



Şekil 10.5 : Üç Aylık Döneme Ait Sahtecilik Vakaları

10.7. Verilerin Hazırlanması

Doğru ve iyi hazırlanmış veri kümesiyle yapılan bir çalışma bizi etkin modelleme sonuçlarına götüreceğinden dolayı veri hazırlığı çok önemlidir. Modelin kurulması

sırasında ortaya çıkacak sorunlar, bu aşamaya sık sık geri dönülmesine verilerin yeniden düzenlenmesine neden olacaktır. Projede bilinen sahtecilik durumları için gerekli süreç takip edilmiştir. Bu sebeple Bölüm 10.4.1’de belirtildiği gibi, modeli eğitmek için bir geçmiş veri kümesi, modelin uygulanması için de ayrı bir güncel veri kümesi oluşturulmuştur.

Veri tablosunun her bir satırı ayrı bir müşteriyi, her bir sütunu da müşterilere ait ayrı bir özelliği belirtmektedir. Tablonun en son sütununda ise hedef değişken yer almaktadır. Projede oluşturulan tablolarda, hedef değişkene “durum” adı verilmiş ve “gerçek” ile “sahte” olarak kodlanmıştır. Durum değişkeninin de değeri “sahte” olanlar sahte başvuruları, değeri “gerçek” olanlar ise sahte olmayan başvuruları ifade etmektedir. Geçmiş veri kümesinde hedef değişken değerlerinin dolu olması gerekmektedir. Aksi taktirde projede kullanılan yazılım hata verecektir.

XYZ Bank Güvenlik Ekibi tarafından proje öncesinde tespit edilmiş sahte başvuru sayısının oldukça az oluşu, modelin sahtekar profilini sahtekar olmayanlardan ayırmasını güçleştiriyordu. Bunun önüne geçebilmek ve daha etkin bir model kurabilmek için daha çok sahtekarlık vakası örneğine ihtiyaç vardı. Sahte başvuru havuzunu büyütmek amacıyla, potansiyel sahte başvuru (PSB) olarak nitelendirilen ve XYZ Bank’ı büyük miktarda zarara uğratan kişiler de proje kapsamında sahtekar olarak nitelendirildi. Mevcut durum analizin de gözüken bu kişiler hatırlanacağı gibi, kredi kartlarını alıp kullanmaya başlamış ve limitlerinin en az %90’ını kullanmalarına rağmen üç ay boyunca hiçbir ödeme yapmadan kanuni takibe düşmüş kişilerdi. Banka yetkilileriyle yapılan görüşmelerde, bu kişilerin belli bir kısmının sahteci olmayabileceği, ancak böyle bir uygulamanın bankanın gelecekte oluşabilecek zarar riskinin önüne geçme yolunda yardımcı olacak bir uygulama olduğu kabul ettirildi.

Projede kullanılan yazılım değişkenleri saklama şekillerine göre beşe, değerlerinin türüne göre ise üçe ayırmaktadır. Saklama şekilleri şöyledir:

- Dizi (String): Değişken karakter dizileri içeriyorsa seçilir. İsim, adres, e-mail adresi vb. değişkenleri örnek olarak verebiliriz.
- Sayı (Number): Hesaplanabilir sayılar içeren değişkenlere atanmalıdır. Telefon numarası veya hesap numarası gibi değişkenler bu kategoride yer almamalıdır.

- Tamsayı (Integer)
- Tarih (Date): Değişken sadece tarih bilgisi içeriyorsa seçilir.
- Tarih ve zaman (Datetime): Değişken hem tarih hem de zaman bilgisi içeriyorsa seçilir. Örnek: 2004/12/22 24:44:33

Değişkene ait değerleri ise aşağıdaki üç gruba göre ayırmak gerekmektedir.

- Sürekli (Continuous): Ortalama, varyans gibi hesaplamalar yapabileceğimiz sayısal değişkenler için seçilir.
- Nominal: Değişkenin bir birinden farklı değerlere sahip olduğu ve bu değerlerin bir birine karşı herhangi bir üstünlüğünün olmadığı durumlarda seçilir. Hem sayısal hem de kategorik değerlere uygulanabilir. En iyi örneklerden biri Medeni Durum değişkenidir. Burada yer alabilecek değerler “Evli”, “Bekar”, “Dul” ve “Boşanmış” olabilir. Bu değerlerin hepsi bir birinden farklıdır ve bir birlerine karşı bir üstünlükleri yoktur. Evli olmanın dul olmaya göre artı bir etkisi olmayacaktır. Bu sebeple, bu değişkenin nominal olarak belirlenmesi gerekir.
- Sırasal (Ordinal): Burada farklı değerlerin bir birine karşı olan üstünlüğü söz konusudur. Okulda alınan notlar basit bir örnek olabilir. 0’dan 100’e kadar olan not sisteminde 65 puan 45 puandan iyi bir not olacaktır. Dolayısıyla böyle bir değişken nominal değerlendirilmelidir. Yine firma politikasına göre müşterilerin firmalarla olan çalışma süreleri ordinal değerlendirilebilir. Eğer sizinle uzun süredir çalışan müşterinizin, sizinle daha az çalışan müşteriye göre daha değerli olduğunu düşünüyorsanız, böyle bir seçim yapılabilir.

Proje sırasında, değişkenler bu yönleriyle de tanınıp, analiz edilmiş ve modellemeden önce bunlar yazılım da girdi olarak belirtilmiştir.

Sonuç olarak öğrenme veri kümesini hazırlanırken, XYZ Bank’ın veri ambarındaki son 1 yıla ait tüm başvuru bilgileri çekilmiş ve her müşteri tek bir satırda ifade edilecek şekilde tablo oluşturulmuştur. Tablonun sonuna da hedef değişken yaratılıp, sahteciler “sahte”, sahteci olmayanlar da “gerçek” olarak işaretlenmişlerdir. Güncel veri kümesi kurulurken de aynı yollar izlenmiş, yalnızca hedef değişken değerleri boş bırakılmıştır.

Başvuru sırasında müşteri hakkında elde edilen bilgiler genellikle, müşterilerin kimlik, adres ve çalışma bilgilerini içeren genel bilgilerdir. Bunlara ek olarak Banka'ya özel şube ve kampanya gibi bilgilere de ulaşılır.

Bu aşama da elde edilen değişkenler, Tablo 10.2'de gösterilmiştir.

Tablo 10.2: Analiz Öncesi Elde Edilen Değişkenler

Değişken Adı	Saklama Tipi	Değer Tipi	Açıklama
Güncelleme Tarihi	date	continuous	
Basvuru No	integer	nominal	
Ad	string	nominal	
Soyad	string	nominal	
Cinsiyet	string	nominal	
Medeni Durum	integer	nominal	0: Bilinmiyor, 1: Evli, 2: Bekar, 3: Dul / Boşanmış
Kişisel Vergi Numarası	integer	nominal	
TC Kimlik Numarası	integer	nominal	
Öğrenim Durumu	integer	ordinal	0:Bilinmiyor, 1:İlkokul, 2:Ortaokul, 3:Lise, 4:Yüksekokul, 5:Üniversite, 6:Lisansüstü
Meslek Kodu	integer	nominal	43 gruba ayrılıp, kodlanmıştır.
Çalışma Şekli	integer	nominal	10 gruba ayrılıp, kodlanmıştır.
İşyeri Adı Unvanı	string	nominal	
İş Adres1	string	nominal	
İş Adres2	string	nominal	
İş Adres İlçe Semt	string	nominal	
İş Adres İl Kodu	integer	nominal	99: Bilinmeyen. 81 ilin plaka numaraları
İş Ülke Kodu	integer	nominal	
İş Posta Kodu	integer	nominal	
İş Telefonu	integer	nominal	
İş Telefonu Dahili	integer	nominal	
Çalışma Süresi Yıl	integer	continuous	
Çalışma Süresi Ay	integer	continuous	
Ev Adres1	string	nominal	
Ev Adres2	string	nominal	
Ev Adresi İlçe Semt	string	nominal	
Ev Adresi İl Kodu	integer	nominal	99: Bilinmeyen. 81 ilin plaka numaraları
Ev Ülke Kodu	integer	nominal	
Ev Posta Kodu	integer	nominal	
Ev Telefonu	integer	nominal	
Cep Telefonu	integer	nominal	
Email	string	nominal	
Ev Durumu	integer	nominal	0:Bilinmiyor, 1:Kendi, 2:Kira, 3:Lojman,4: Aile Ferdi, 5:Diğer
Araba Durumu	integer	nominal	0:Bilinmiyor, 1:Var, 2:Yok
Aylık Kisisel Net Gelir	number	continuous	
İmza Tarihi	date	continuous	

Tablo 10.2: Devamı

Değişken Adı	Saklama Tipi	Değer Tipi	Açıklama
Kart Tipi	integer	nominal	Verilebilecek Kart Tipleri 0:Bilinmiyor, 1:Visa, 2:Master, 3: Visa_Gold, 4:Master_Gold
Ödeme Dönemi	integer	nominal	0:Bilinmiyor, 1:İlk Dönem, 2:İkinci Dönem
Kart Adres	integer	nominal	0:Bilinmiyor, 1:Ev, 2:İş
Ekstre Adres	integer	nominal	0:Bilinmiyor, 1:Ev, 2:İş
Ek Kart Ad	string	nominal	
Ek Kart Soyad	string	nominal	
Ek Kart Cinsiyet	string	nominal	
Ek Kart Başvuru Tarihi	date	continuous	
Ek Kart Doğum Tarihi	date	continuous	
Kampanya Kodu 1	string	nominal	Kişinin daha önce katıldığı kampanya bilgisi.
Kampanya Kodu 2	string	nominal	Kişinin daha önce katıldığı kampanya bilgisi.
Kaynak Kodu	integer	nominal	0:Bilinmiyor, 1:Şube, 2:Dergi, 3:Mağaza, 4:Direk Satış
Direk Satış Sicil No	integer	nominal	Kredi kartını satan memurun sicil numarası
Şube Kodu	integer	nominal	
Sube Gorus Formu	string	nominal	E: Evet, H:Hayır
Sisteme Giriş Tarihi	datetime	continuous	
Doğum Yeri Ülke Kodu	integer	nominal	
Doğum Yeri İl	integer	nominal	99: Bilinmiyor. 81 ilin plaka numaraları
Doğum Yeri İlçe	string	nominal	
Doğum Tarihi	date	continuous	
Uyruk	integer	nominal	0:Bilinmiyor, 1:TC, 2:Diğer
Baba Adı	string	nominal	
Anne Adı	string	nominal	
Kimlik Tipi	integer	nominal	0:Bilinmiyor, 1:Nüfus Cüzdanı, 2:Pasaport, 3:Ehliyet, 4:Kurum Kimliği
Kimlik Seri No	string	nominal	
Kimlik Sıra No	integer	continuous	
Nufusa Kayıtlı İl	integer	nominal	81 ilin plaka numaraları
Müşteri No	integer	continuous	
Pasaport No	integer	continuous	
Red Sayısı	integer	continuous	Kişinin daha önce kaç kez reddedildiği bilgisi.
İdeal Kart Taksitli Limit	number	continuous	Taksitli limiti
Kart Ad Soyad	string	nominal	Kredi kartı üzerinde yazan ad soyad bilgisi
TRL Hesap Sube	integer	nominal	KişininTürk Lirası hesabının hangi şubede bulunduğu bilgisi
durum	string	nominal	Sahte, Gerçek

10.7.1. Eksik Değerler (Missing Values)

Elde edilen veri kümesinde bazı değişkenler eksik değerler barındırmaktaydı. Kullanılan yazılım bu sorunla başa çıkabildiği için bu eksik değerlere sahip değişkenler de analize sokulmuştur. Analiz sonuçlandığında ise bu değişkenlerdeki eksik bilgilerin de bir önemi olduğu bir nevi ispatlandı. Örneğin kişilerin Çalışma Şekli bilgisi eksik ise, bu o kişinin sahtekar olma özelliğini azaltıcı bir faktör olarak gözüktü. Bunun sebebi de, sahtekarların kredi kartını alabilmek için kendilerini çalışıyor olarak göstermeleriydi.

Şu unutulmamalıdır ki, aslında sahte başvuruların tespiti sırasında kullanılan veri kümesinde bir kişiye ait eksik değerlerin oluşu, bu kişiyi daha risksiz kılmaktadır. Bunun sebebi, sahtekarların genelde herhangi bir sorun çıkmaması için tüm bilgilerini eksiksiz olarak bankaya sunmasıdır. Dolayısıyla eksik değerli değişkenlerin analize sokulmasının, modelleme sonuçlarına katkısı olabileceği bir gerçektir.

Kullanılan yazılım, eğitim veri kümesini 3 alt kümeye bölerek çalışmaktadır. Bu alt kümeler; tahmin (estimation), doğrulama (validation) ve test alt kümeleri olup, işlevleri Bölüm 10.8 Modelin Kurulması kısmında anlatılacaktır. Yazılım eksik değerleri şu şekilde analize katmaktadır.

Nominal ve ordinal değişkenler için tahmin alt kümesinde eksik değerlere rastlanmışsa KxMissing adlı yeni bir değer oluşturulur. Güncel veride yeni müşterinin bu değişkene ait değeri eksikse, bu değer yerine KxMissing değeri koyularak işlem yapılır. Tahmin alt kümesinde eksik değerlere rastlanmadığı durumlarda değerlendirilecek yeni müşteri ilgili değişkende eksik değere sahipse, bu eksik değer tahmin alt kümesinde en sık kullanılan (frekansı en büyük olan) değerle doldurulur.

Sürekli değişkenler için ise tahmin alt kümesinde eksik değerlere rastlanmışsa yine KxMissing adlı yeni bir değer oluşturulur. Eğer güncel veride yeni müşterinin bu değişkene ait değeri eksikse, bu değer yerine KxMissing değeri koyularak işlem yapılır. Tahmin alt kümesinde eksik değerlere rastlanmadığı durumlarda ise değerlendirilecek yeni müşteri ilgili değişkende eksik değere sahipse, bu eksik değer tahmin kümesinde ilgili değişkene ait değerlerin ortalama değeriyle doldurulur.

10.7.2. Verileri Dönüştürme

Mevcut bilgileri en etkin şekilde kullanmak, modelleme sırasındaki temel amaçlardan biri olmalıdır. Bazı durumlarda değişken veya değişken değerleri üzerinde yapacağınız işlemler, ilgili değişkenin analiz için daha faydalı hale gelmesine sebep olabilir. Bu nedenle proje sırasında eldeki bilgilerin daha faydalı olabilecek şekil de nasıl kullanılabilecekleri düşünülmüş, bazı değişkenler ve değerleri üzerinde değişiklikler yapılmıştır. Bu değişkenlerden bazıları ve yapılan işlemler aşağıda belirtilmiştir.

- Çalışma Süresi:

Oluşturulan ilk veri kümesinde, çalışma süresi ile ilgili “Çalışma Süresi Yıl” ve “Çalışma Süresi Ay” olmak üzere iki adet değişken bulunmaktaydı. Eğer bir başvuru sahibi mevcut işinde 7 yıl 5 aydır çalıştığını beyan etmişse, XYZ Bank sistemine bu bilgiyi “Çalışma Süresi Yıl” değişkenine 7 ve “Çalışma Süresi Ay” değişkenine 5 yazarak işlemekteydi. Modelleme sırasında böyle bir ayırımın gerekli olmayacağı düşünülerek, daha etkin ve yorumlanabilir sonuç alabilmek adına “Çalışma Süresi Toplam Ay” adlı yeni bir değişken oluşturuldu. Daha sonra bu değişkenin altına gelecek değerlerle ilgili sorgu cümlesi (query) yazıldı ve her başvuru sahibinin “Çalışma Süresi Yıl” değişkeninde sahip olduğu değer 12 ile çarpılıp, “Çalışma Süresi Ay” değişkenindeki değerle toplanarak oluşturulmuş yeni değişkenin altındaki ilgili satırlara işlendi. Yani 7 yıl 5 aydır çalışmakta olan bir kişi, toplamda 89 aydır çalışıyor olarak tek bir değişkenle gösterilmeye başlandı.

Çalışma süresi için yapılan bu değişikliğin, hem baştaki mevcut iki adet değişken arasındaki korelasyonu ortadan kaldırmak hem de değişken sayısını azaltarak modelleme süresini kısaltmak adına analize faydası olmuştur.

- İş ve Ev Telefonları:

Telefon bilgileri bu tarz tablolar hazırlanırken karşılaşılan en sorunlu bilgilerden biridir. Öncelikli olarak genellikle tek bir formatta yazılmamış oluşu bir veri kalitesi problemine yol açar ve bunun düzeltilmesi gerekir.

Ülkemizde telefon numaraları 10 basamaklı bir sayı olarak karşımıza çıkmaktadır. İstatistikçiler bu sayıları analiz etmek oldukça zor olduğundan, ya bu değişkenleri analize sokmamışlar ya da telefon bilgilerini “Var” veya “Yok” olarak kodlayarak

analize dahil etmişlerdir. Bu tarz değişkenlerin analizinin oldukça zor oluşunun sebebi istatistikçiler tarafından kullanılan klasik algoritmalardır.

XYZ Bank projesi için kullanılan yazılımın dayandığı Yapısal Risk Minimizasyonu algoritmasının bir özelliği, çok basamaklı sayıları nominal olarak algılayıp, her farklı sayıyı ayrı ayrı değerlendirmeye alabilme kabiliyetidir. Kullanılan yazılımın bu özelliği bilindiğinden, telefon numaralarından anlamlı bir bilgi elde edilebileceği düşünülerek bazı dönüştürmeler yapılmıştır.

Bilinmektedir ki ülkemizdeki en sorunlu veri posta kodu bilgisidir. Çoğu vatandaş hala daha posta kodunu bilmemektedir. Bundan dolayı da başvuru formlarında bu bilgiyi genellikle ya boş bırakmakta ya da yanlış doldurmaktadırlar. Posta kodu bilgisi bu kadar sorunlu bir bilgi olmasına karşı analizler için bir o kadar da önemlidir. Çünkü bu bilginin doğruluğu sizin diğer adres satırlarını analize sokmanızı gerektirmez. Doğru ev posta kodu bilgisiyle kişinin ev adresi hakkında, doğru iş posta kodu bilgisiyle de kişinin iş adresi hakkında kesin bilgilere ulaşabilirsiniz.

Telefon numaraları düşünüldüğünde de içlerinde böyle bir bilginin saklı olduğunu düşünmek yanlış olmaz. Şöyle ki, 212 272 15 XX numaralı bir telefon düşünelim. Buradaki 212 alan kodu bize, telefonun İstanbul Avrupa yakasına ait olduğunu göstermektedir. Devamındaki 272 rakamı ise Avrupa yakasındaki Gayrettepe mevkinde bulunan konutlarda kullanılmaktadır. Bir başka ifadeyle, bu mevkideki 10,000 adet telefon 272 rakamıyla başlamaktadır. Dolayısıyla bu telefon numarasını 212272 şekline dönüştürüp kullandığımızda, İstanbul Avrupa yakası Gayrettepe mevkindeki 10,000 hanelik bir semte ait bilgiye ulaşmış oluruz ve bu bilgiyi buralara ait bir posta kodu olarak kullanabiliriz.

Projede de analiz öncesi bu tarz bir fayda elde edilecek olan tüm ev ve iş telefon numaraları, ilgili sorgu cümleleri yazılarak dönüştürme işlemine tabi tutulmuş ve oluşan yeni değerler veri tablosuna eklenen “Ev Tel Alan Santral” ve “İş Tel Alan Santral” adlı iki yeni değişkenin altında yazdırılmıştır.

- Cep Telefonu Bilgisi:

Cep telefonu numarası da ev ve iş telefonları gibi analizi zor olan bir değişkendir. Bu bilgiyi analize sokarken “Var” veya “Yok” diye kodlamak yerine, numaraların ilk üç rakamı yani GSM operatör numaraları (ÖR: 532, 542, 555, 535 vb.) alınarak analize

sokulmuştur. Bu 3 haneli rakamlar ilgili sorgu cümlesi yazılarak yeni oluşturulan “Cep Tel Alan” değişkeninin altına yazdırılmıştır.

- Doğum Tarihi Bilgisi:

“Doğum Tarihi” değişkenine de bir dönüştürme işlemi uygulanmıştır. Tarih formatında olduğunda analize bir fayda sağlamayan bu değişken, müşterinin başvuru yaptığında kaç yaşında olduğunu gösteren bir başka değişkene dönüştürülmüştür. Bu değerler yeni oluşturulan “Başvuru Yaşı” adlı değişkenin altına eklenmiştir.

- E-mail Bilgisi:

“E-mail” değişkeni de analize anlam katması amacıyla dönüştürme işlemi uygulanan değişkenlerden biridir. Burada kişinin tüm e-mail adresini kullanmak yerine, kişinin sadece e-mail hesabı bilgisi kullanılmıştır. Şöyle ki eğer bir kişinin e-mail adresi abc123@hotmail.com ise, bu adresin sadece “hotmail.com” kısmı alınmıştır. İlgili sorgu cümlesi yazıldıktan sonra tüm bu değerler, yeni oluşturulan “e-mail hesap” değişkeninin altına yazdırılmıştır.

10.7.3. Değişken Seçimi

Başvuru numarası, müşteri numarası gibi anlamlı olmayan ve diğer değişkenlerin modeldeki etkilerini azalmasına da neden olabilecek değişkenler modele sokulmamıştır. Bunun ek olarak tarih bilgileri ve üzerinde değişiklik yapılarak yeni değişkenler üretilen diğer değişkenlerde modellemeye yer almamıştır. Modele sokulmayan bu değişkenler Tablo 10.3’de belirtilmiştir.

Tablo 10.3: Modellemeye Sokulmayan Değişkenler

Değişken Adı	
Güncelleme Tarihi	Ev Adres2
Basvuru No	Ev Telefonu
Kişisel Vergi Numarası	Cep Telefonu
TC Kimlik Numarası	Email
İşyeri Adı Unvanı	İmza Tarihi
İş Adres1	Ek Kart Başvuru Tarihi
İş Adres2	Ek Kart Doğum Tarihi
İş Telefonu	Sisteme Giriş Tarihi
İş Telefonu Dahili	Kimlik Seri No
Çalışma Süresi Yıl	Kimlik Sıra No
Çalışma Süresi Ay	Musteri No
Ev Adres1	Pasaport No

10.8. Modelin Kurulması

Proje konusu bir sınıflandırma problemidir. Dolayısıyla kullanılan yazılımda ilgili fonksiyon kullanılarak bir sınıflandırma modeli oluşturulmuştur. Yazılım kendi içinde birden fazla model oluşturduktan sonra, bu modeller içerisinde en yüksek kalite ve en yüksek tutarlılık arasındaki en iyi uyumu yakalayan tek bir modeli kullanıcıya sunmaktadır.

Modelleme sırasında yazılım öncelikli olarak geçmiş veri kümesini (öğrenme veri kümesini) üç ayrı alt kümeye böler. Bu alt kümelerin modellemedeki rolleri Tablo 10.4’de belirtilmiştir.

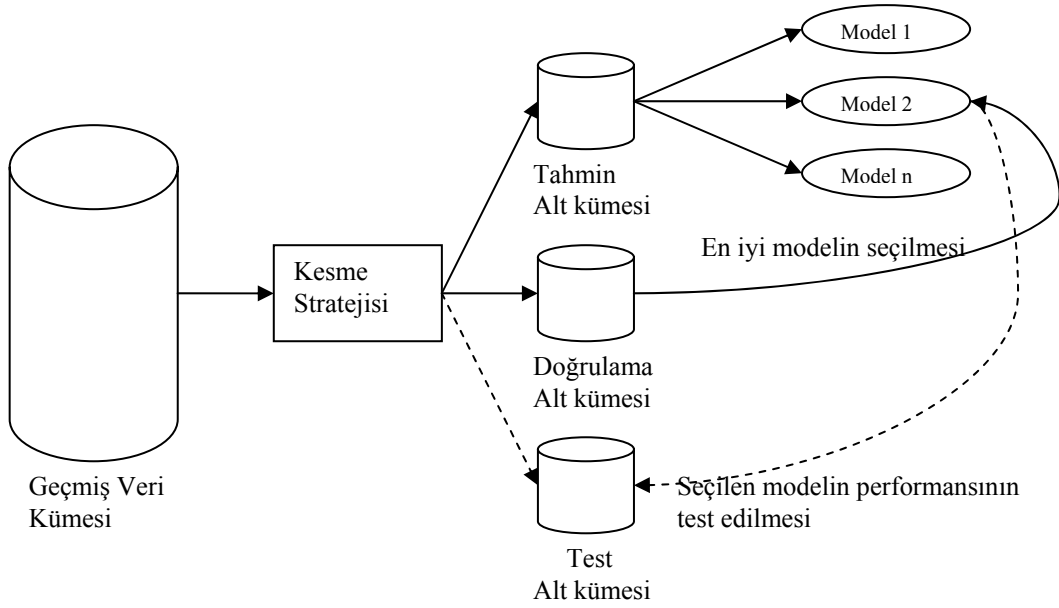
Tablo 10.4: Üç Alt Kümenin Modellemedeki Roller [33]

Veri Kümesi	Kullanım amacı
Tahmin (Estimation)	Farklı modeller oluşturmak için kullanılır. Bu aşamada oluşturulan modeller kuramsaldır (hypothetical).
Doğrulama (Validation)	Tahmin alt kümesi kullanılarak oluşturulan modeller arasından en iyi olanı seçmek için kullanılır. Bu model, en yüksek kalite ile en yüksek tutarlılık arasında en iyi uyumu sağlayan modeldir.
Test	Seçilen modelin performansını yeni bir veri kümesi üzerinde doğrulamak için kullanılır.

Yazılım bu üç alt kümeyi oluştururken bir kesme stratejisi (cutting strategy) kullanmaktadır. Seçilen strateji, oluşturulan alt kümelere toplam verinin alt kümelere hangi oranda dağıtılacağını belirler. Projede, rasgele (random) kesme stratejisi kullanılmıştır. Bu sayede toplam verinin %60’ı tahmin alt kümesine, %20’si doğrulama alt kümesine ve geriye kalan son %20’lik kısım da test alt kümesine paylaştırılmıştır.

Modeli oluşturmak için kullandığım veri tablosu, %1.6’sı sahtecilik vakası olan yaklaşık 850,000 kayıttan ve 51 adet değişkenden oluşmaktadır.

Yazılımın model oluşturma süreci Şekil 10.6’da gösterilmiştir.



Şekil 10.6 : Model Oluşturma

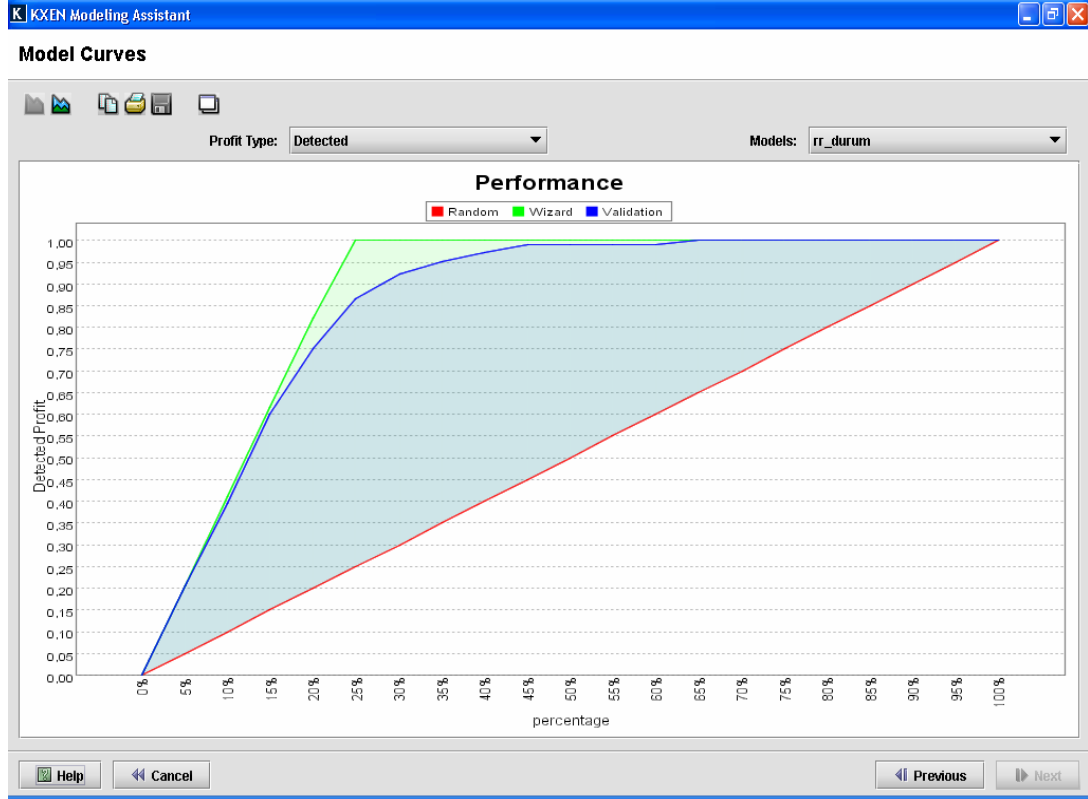
10.9. Kurulan Modelin Değerlendirilmesi

Kurulan modeldeki değerlendirme faktörlerinden en önemlisi modelin kalitesidir. Bir modelin kalitesini, oluşturulacak performans eğrileri ile ölçmek mümkündür.

Performans eğrisi çizilirken, öncelikle genele uygun bir örneklem alınır. Daha sonra model bu örnekleme uygulanır ve her bir kişinin aldığı skor belirlenir. Bir sonraki aşamada ise kişiler aldıkları skora göre büyükten küçüğe doğru sıralanırlar. Son olarak örnekler %5'lik dilimlere ayrılır ve her dilime giren “aranan” örneklerin yüzdesi işaretlenir. XY grafiğinde işaretlenen bu noktalardan geçen eğri, performans eğrisi olarak ifade edilir.

Projede kullanılan yazılımla elde edilen örnek bir performans eğrisi grafiği Şekil 10.7’de gösterilmiştir. Bu grafik XYZ Bank tarafından proje öncesinde yazılımın kabiliyetlerini görmek için hazırlanan veri kümesiyle oluşturulmuş modele aittir. Bu veri kümesi, 1018 tanesi sahte başvuru olan toplam 4107 başvuruyu içermektedir. Bunun dışında XYZ Bank tarafından hazırlanan 40 adet değişken veri kümesinde yer almaktadır. Projede yapılabilecek benzer şekilde bazı, değişkenler için (Ev Telefonu, Cep Telefonu, İş Telefonu, Çalışma Süresi, Başvuru Yaşı) daha önce Bölüm 10.7.2’de belirtilen ilgili dönüşümler yapılarak, yeni değişkenler oluşturulmuş ve veri kümesine eklenmiştir. Modellemeye ise bu değişkenler arasından 30 tanesi

sokulmuştur. Hedef değişken bu çalışmada da yine “gerçek” ve “sahte” değerlerine sahip olan “durum” adlı değişkendir.



Şekil 10.7 : Performans Eğrisi

Şekil 10.7’de %5’lik dilimlere ayrılmış x eksenin toplam popülasyonu gösterirken, y eksenini ise arananların yani sahtecilerin yüzdesini gösterir. Grafikte gözüken üç eğriden en altta olanı rasgele (random) eğrisidir ve popülasyonun belli bir yüzdesinin rasgele incelendiğinde yakalanacak sahtecilerin yüzdesini gösterir. Şöyle ki, rasgele bir seçimle popülasyonun %10’unu incelenirse toplam sahtecilerin de %10’unun veya popülasyonun %35’i incelenirse toplam sahtecilerin de %35’inin yakalanacağı istatistiksel bir gerçektir.

Grafiğin en üstünde yer alan eğri ise kahin (wizard) eğrisidir. Kahin, herşeyi bilen yani kimin sahteci olduğunu kimin gerçek olduğunu hatasız söyleyebilir. Dolayısıyla kahin eğrisi en iyi tahmin durumunu gösterir. Bu sebeple popülasyonun %25’i incelendiğinde tüm sahteciler kahin tarafından hatasız olarak bulunacak ve daha sonra yapılacak bir incelemenin herhangi bir faydası olmayacaktır.

Grafiğin ortasında yer alan doğrulama (validation) eğrisi ise oluşturulan modelin eğrisidir. Model kahin gibi davranmasa da, rasgele bir seçime göre fayda

sağlamaktadır. Şöyle ki tüm popülasyonun %20'si rasgele bir seçimle incelendiğinde tüm sahtecilerin %20'si tespit edilirken, kahin kişi sahtecilerin %82'sini yakalayabilmektedir. Oluştulan modelin sahteci tespit oranı ise %75 olup, rasgele performansın çok üstünde bir değere ulaştığı gözükmemektedir. Dolayısıyla modeli bu grafiğe göre değerlendirirken, doğrulama eğrisinin kahin eğrisine yakın oluşu modelin kaliteli bir model olduğunu gösterecektir.

Proje de kurduğum model için de benzer bir performans eğrisi elde edilmiştir. Modelin performansı rasgele performansın oldukça üstünde değerler alırken, kahin performansına da çok uzak kalmamıştır. Modelin performans eğrisine göre popülasyonun %5'i incelendiğinde tüm sahtekarların %62.85'ini, popülasyonun %12.5'i incelendiğinde ise tüm sahtekarların %72.80'ini tespit etmek mümkündür.

Proje için kullanılan yazılım, performans grafiğinden elde edilen sayısal bir değeri de kullanıcıya sunmaktadır. KI olarak gösterilen bu değer bilgi katsayı olarak adlandırılır ve eldeki değişkenlerle hedef değişkenin ne kadar iyi açıklanabileceğini sayısal olarak gösterir. KI değeri hesaplanırken, performans eğrilerinin altına kalan alanlar kullanılır. İlgili formül Denklem 10.2'de verilmiştir.

$$KI = \frac{\text{Rasgele eğrisi ile doğrulama eğrisi arasında kalan alan}}{\text{Rasgele eğrisi ile kahin eğrisi arasında kalan alan}} \quad (10.2)$$

KI değerleri 0 ile 1 arasında yer alır. Sıfırdan farklı her değer modelin rasgele seçimden daha başarılı olduğunu gösterir. Herhangi bir kısıtlaması olmamakla beraber, mümkün olduğu kadar büyük bir KI değerine sahip olunmaya çalışılır. Düşük olan KI değerini artırmak için değişken sayısını çoğaltmak gerekir.

Proje de kurduğum model için KI değeri olarak 0.8142 değeri elde edilmiştir. Bu değer oluşturulan modeldeki değişkenlerin hedef değişkeni açıklamak için yetkin olduklarını, modelin kaliteli bir model olduğunu ve rasgele seçime göre oldukça başarılı olduğunu göstermiştir.

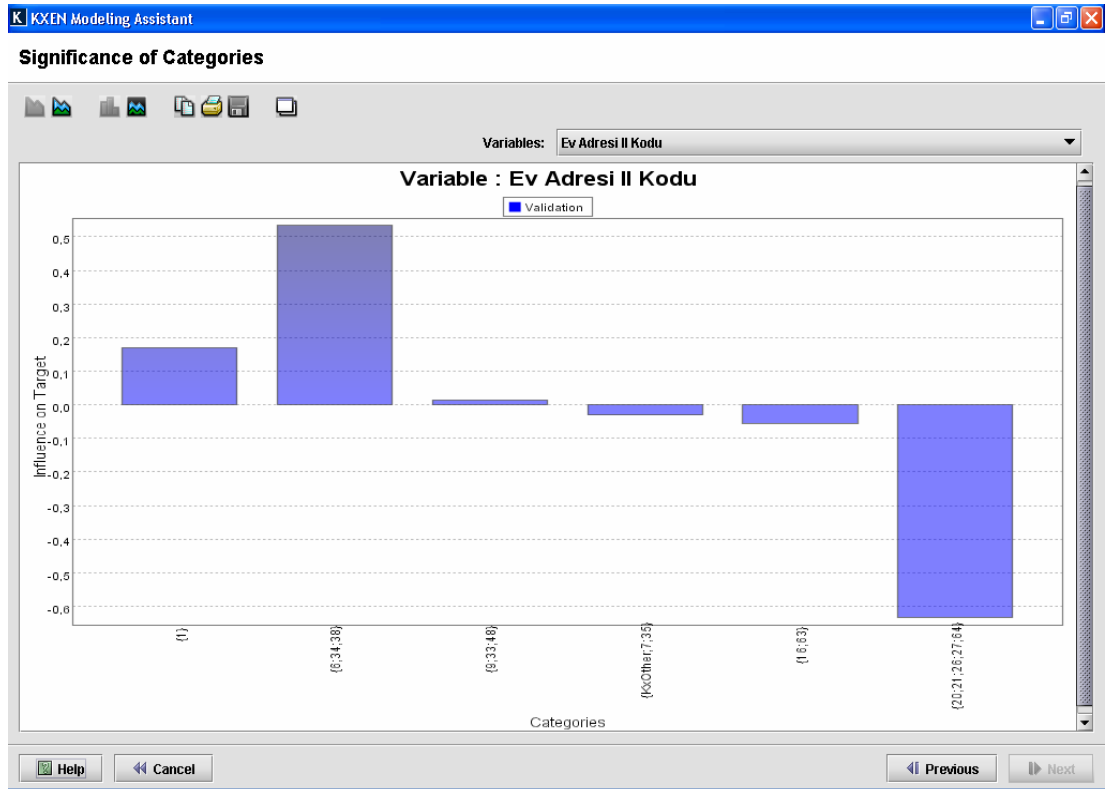
Oluşturulan modelin değerlendirilmesindeki bir diğer önemli kriter modelin genellenebilirlik, bir başka deyişle yeni veri üzerinde de başarılı bir şekilde çalışabilme özelliğidir. Eğer model yeni veri kümesi üzerinde başarısız olursa, yüksek KI değerine sahip olması bir anlam ifade etmeyecektir. Kullanılan yazılım bu özelliği de bir gösterge ile sunmaktadır. KR olarak gösterilen ve sağlamlık (robustness) katsayısı olarak adlandırılan bu değer de 0 ile 1 arasında değer alır. Bu

Net Gelir” deęiřkenleri takip etmektedir. Yazılımın sunduęu bu grafikte, modellemeye katılan tm deęiřkenlerin modele olan etkilerini grmek mmkndr.

Projede kurduęum modelde ise nem sırasına gre deęiřkenler řu řırayla çıkmıřtır.

1. Aylık Kiřisel Net Gelir
2. Cep Tel Alan
3. Meslek Kodu
4. Bařvuru Yařı
5. Ev Adresi İlçe Semt

Bir deęiřkenin modele olan etkisini bilmek nemlidir ama deęiřkenin bu etkiyi hangi deęerleri sayesinde saęladığını ęrenmekte, modeli daha anlařılabilir kılmak iin kesinlikle faydalı olacaktır. XYZ Bank ile yapılan deneme alıřmasından “Ev Adres İl Kodu” deęiřkenin hedefe olan etkisi iin elde edilen bilgi řekil 10.9’da gsterilmiřtir.

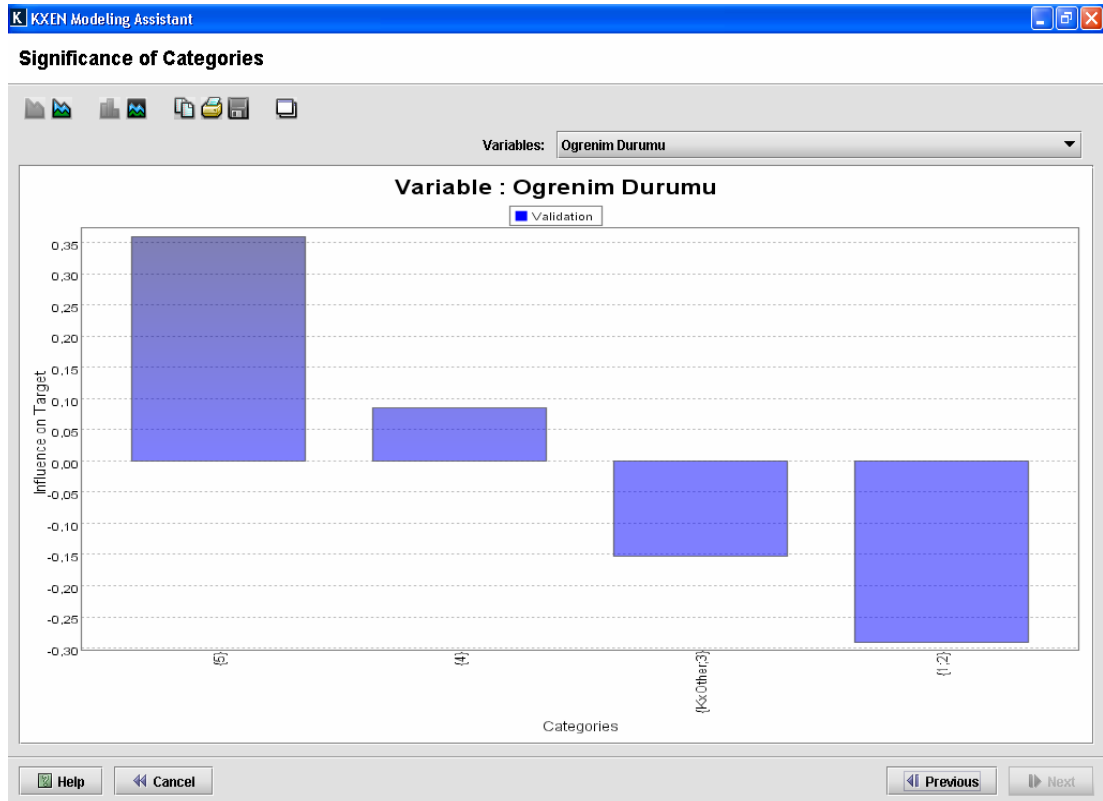


řekil 10.9 : Ev Adresi İl Kodu Deęiřkeninin Hedefe Olan Etkisi

Bu tarz grafiklerde x eksenı deęiřkene ait deęerleri (kategorileri) gsterirken, y eksenı bu deęerlerin hedef deęiřken zerindeki pozitif veya negatif etkilerini

göstermektedir. Grafiğe bakıldığında Adana ilinden oturuyor olmak, bir kişinin sahteci olma riskini artıran en önemli bilgi olarak karşımıza çıkmıştır. Bunun dışında kişi Ankara, İstanbul veya Kayseri illerinden birinde oturuyor ise, yine sahteci olma yönünde pozitif bir eğilime sahip olduğu gözükmemektedir. Bu üç ili içinde barındıran kategorinin Adana'dan daha fazla etkiye sahipmiş gibi gözükmesinin nedeni, bu grubu ait kişilerin sayısının veri kümesinde daha çok yer almasından kaynaklanmaktadır. Sahteci olma yönünde en risksiz olan şehirler ise grafiğin en sağındaki sütundan da gözüktüğü gibi Denizli, Diyarbakır, Eskişehir, Gaziantep ve Uşak şehirleridir.

Benzer bir diğer grafikte Şekil 10.10'da "Öğrenim Durumu" değişkeni için sunulmuştur. Bu grafikte ise görülmektedir ki, kişinin üniversite mezunu oluşu bu değişkene göre sahteci oluşunu en pozitif şekilde etkileyen değerdir. Kişinin yüksekokul mezunu oluşu da sahteci olma yönünde risk taşımaktadır. İlkokul veya ortaokul mezunu olmak ise sahtecilik yönünde en risksiz grup olarak karşımıza çıkmaktadır.



Şekil 10.10 : Öğrenim Durumu Değişkeninin Hedefe Olan Etkisi

Proje de elde ettiğim ilginç bilgilerden bazıları şunlardır:

- “Cep Tel Alan”değişkeni incelendiğinde sahteci olma konusunda en riskli grubun faturasız (kontrollü) hat sahibi oldukları, en risksiz grubun ise faturalı hat sahipleri olduğu tespit edilmiştir.
- “Ev Adres İl Kodu” değişkeni incelendiğinde sahteci olma konusunda en riskli illerin Adana, İstanbul, İzmir ve Antalya olduğu gözükmiştir. En risksiz iller arasında ise Bursa, Eskişehir, Gaziantep ve Zonguldak yer almaktadır.
- “Aylık Kişisel Net Gelir” değişkeni incelendiğinde sahtecilerin gelirlerini yüksek beyan ettikleri görülmüştür.
- “Öğrenim Durumu” değişkeni incelendiğinde üniversite ve yüksekokul mezunu olmanın, sahteci olma yönünde en riskli grup olduğu görülmüştür.
- “Ev Adres İlçe Semt” değişkeni incelendiğinde ise İstanbul’da en riskli semtler olarak Avcılar, Bahçelievler, Bakırköy, Merter, İkitelli ve Güneşli, en risksiz semtler olarak ise Bostancı, Etiler, Kadıköy, Levent, Moda ve Nişantaşı gözükmiştir.

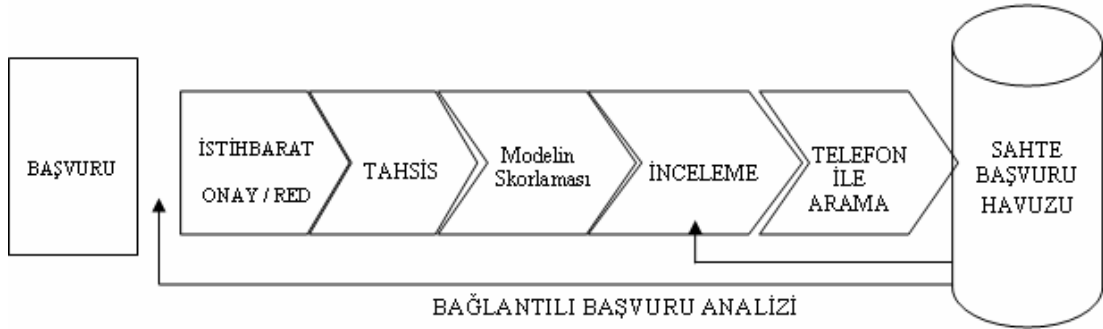
Tespit edilen sahte başvurular incelendiğinde, sahte başvuruda bulunan kişilerle ilgili aşağıdaki sonuçlara ulaşılmıştır:

- %48’i Üniversite Mezunu %20’si mühendistir.
- %94’ü Cep Telefonu kullanmaktadır.
- %55’i İstanbul’dan başvuru yapmaktadır.
- %71’inin kendine ait evi, %66’sının ise arabası vardır.
- %93’ü Nüfus Cüzdanı ile başvuru yapmaktadır.
- %34’ünün geliri 2000– 4000 YTL arasındadır.
- %50’sinin işyerinde çalışma süresi 24 – 60 Ay’dır.
- Sahte başvuruların %20’si Direkt Satış % 80’i Şube kanalı ile yapılmıştır.

Kurulan modelin değerini belirlenmesinde kullanılan bir diğer ölçü, model tarafından önerilen uygulamadan elde edilecek kazancın bu uygulamanın gerçekleştirilmesi için katlanılacak maliyete bölünmesi ile elde edilecek olan yatırımın geri dönüş (Return On Investment) oranıdır. XYZ Bank tarafından, projeye yapılan yatırımın 4 gün içerisinde karşılandığı beyan edilmiştir.

10.10. Modelin Kullanılması

Projede oluşturulan modelin aldığı KI, KR değerleri ve modelin anlaşılabilirliği değerlendirilip, doğrudan bir uygulama olarak kullanılmasına karar verilmiştir. Model uygulamaya alınmasıyla mevcut sahte başvuru değerlendirme süreci Şekil 10.11'deki halini almıştır.



Şekil 10.11 : XYZ Bank Proje Sonrası Başvuru Değerlendirme Süreci

Şekil 10.11'de de gözüktüğü gibi başvuru değerlendirme sürecinde ön izleme raporlarının yerini artık oluşturulan model almıştır. Güvenlik Birimi tarafından günlük olarak onaylanan kartlar üzerinden “sahte başvuru modeli” çalıştırılır. Söz konusu model tüm başvuruları, sahte başvuru riskini göz önüne alarak 1 ile 0 arasında skorlar. 1'e yakın skoru olan başvuruların sahte başvuru olma ihtimali yüksek, 0'a yakın başvuruların sahte başvuru olma ihtimali düşüktür. Başvurular aldıkları skora göre büyükten küçüğe sıralanarak, bir Excel dosyası şeklinde kullanıcıya sunulur. Daha sonra bu liste ilgili çalışanlara paylaşılır ve yüksek skora sahip müşterilerden başlanarak değerlendirme yapılır.

10.11. Modelin İzlenmesi

Zaman içinde oluşabilecek değişiklere bağlı olarak kurulan modellerin sürekli olarak izlenmesi ve gerekiyorsa yeniden düzenlenmesini gerektirecektir. Sahtecilik tespiti

sürekli olarak takip edilmesi gereken bir konudur çünkü siz sahtekarları yakaladıkça onlar taktiklerini değiştirip karşınıza çıkmaya devam ederler. Sahtecilik trendlerini takip edebilmek için modelin sürekli olarak güncel tutulması gerekir.

XYZ Bank'ta her ay sonu, o ay içinde gelen gerçek ve sahte başvuru bilgileri geçmiş veri kümesine eklenerek bu yeni veri kümesi üzerinden güncel bir model kurulmaktadır.

10.12. Sonuçların Değerlendirmesi

Ön İzleme Servisi'nin proje öncesi ve sonrası Tablo 10.5'de karşılaştırmalı olarak gösterilmiştir. Projenin performansta sağladığı yüzdesel artışlar ise Tablo 10.6'da sunulmuştur.

Tablo 10.5: Proje Öncesi ve Sonrası Ön İzleme Servisi Performansı

	Günlük Ortalama İncelenen	Günlük Ortalama Sahte Başvuru	Günlük Ortalama Bağlantılı	Günlük Ortalama Kurtarılan (YTL)
Proje Öncesi Ön İzleme	517	7	1	14,312.5
Proje Sonrası Ön İzleme	597	23	9	65,380.1

Proje sonrası günlük ortalama incelenen başvuru adedi %15.5'lik bir artış göstererek 517'den 597 adede ulaşmıştır. Bununla beraber çalışan kapasitesi artırılmadan, günlük ortalama sahte başvuru tespit adedi %228.6'lık bir artışla 7'den 23'e, günlük ortalama bağlantılı sahte başvuru tespit adedi ise %800'lük bir artışla 1'den 9 adede yükselmiştir.

Sahtecilik tespitindeki bu artış doğal olarak kurtarılan para tutarına da yansımış ve günlük ortalama kurtarılan para tutarı %356.8'lik artışla 14,312.5 YTL'den 65,380.1 YTL'ye yükselmiştir.

Tablo 10.6: Proje Sonrası Performansta Gözüken Yüzdesel Artış

İncelenen Artış (%)	Sahte Başvuru Tespit Artış (%)	Bağlantılı Sahte Başvuru Artış (%)	Kurtarılan Artış (%)
15.5	228.6	800	356.8

Projenin finansal faydalarının dışında XYZ Bank’a sağladığı diğer katma değerler şunlardır:

- Sahte başvuruların kart teslim edilmeden önce tespiti ile birlikte Mali Şube Ekipleri ile organize edilen operasyonlar sonucu 2004 yılı içinde 43 vakada yakalanan 100 dolandırıcı cezaevine yollanmıştır.
- Modelin istenilen sıklıkta güncellenebilmekte ve bu sayede sahte başvuru trendinin takip edilebilmektedir.
- Kullanım kolaylığı ve rakamsal sonuçlar itibarıyla yakalanan başarı ile birlikte çalışan motivasyonu artmıştır.
- Etkin kullanım sonucu, projenin toplam maliyeti kısa sürede amorti edilmiştir.

Belirlenen sahte başvurulara ait istatistiksel bilgiler Tablo 10.7’de gösterilmiştir.

Tablo 10.7: Sahte Başvuru Adet Bazında Analiz

Aylar	Başvuru	Onaylanan Kart	SB Tespit Adet	Potansiyel SB Tespit Adet	SB+PSB Tespit Adet	SB/(SB+PSB) (%)	Onaylanan /Başvuru (%)	(SB+PSB)/Başvuru (%)	(SB+PSB)/Onaylanan Kart (%)
Eyl.03	67,185	26,592	234	61	295	79.3	39.6	0.4	1.1
Eki.03	85,476	36,396	203	40	243	83.5	42.6	0.3	0.7
Kas.03	76,186	32,576	208	59	267	77.9	42.8	0.4	0.8
Ara.03	116,678	50,063	257	120	377	68.2	42.9	0.3	0.8
Oca.04	104,285	36,575	378	80	458	82.5	35.1	0.4	1.3
Şub.04	120,070	41,853	464	75	539	86.1	34.9	0.4	1.3
Mar.04	186,652	72,995	564	102	666	84.7	39.1	0.4	0.9
Nis.04	156,668	100,930	351	117	468	75.0	64.4	0.3	0.5
May.04	145,394	59,963	619	109	728	85.0	41.2	0.5	1.2
Haz.04	148,606	86,125	707	51	758	93.3	58.0	0.5	0.9
Tem.04	155,103	62,388	795	88	883	90.0	40.2	0.6	1.4
Ağu.04	102,921	62,344	638	41	679	94.0	60.6	0.7	1.1
Eyl.04	49,075	22,383	844	13	857	98.5	45.6	1.7	3.8
Eki.04	22,014	2,407	351	5	356	98.6	10.9	1.6	14.8

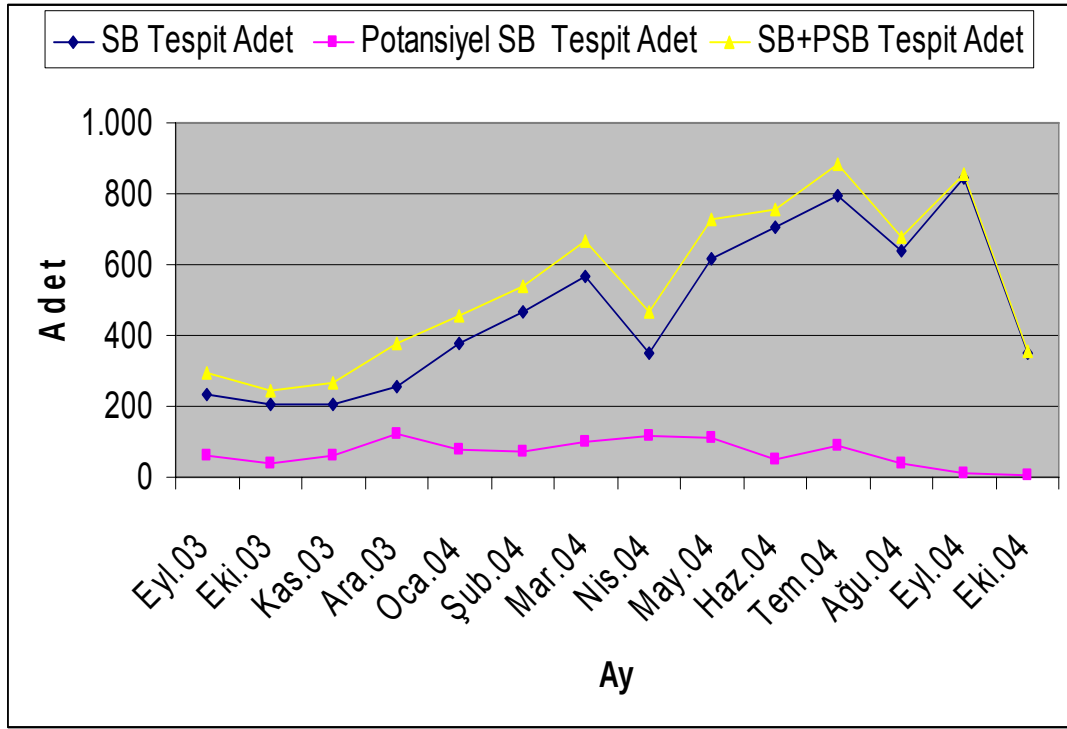
Tablo 10.7 incelendiğinde görülmektedir ki, sahte başvuru (SB) adetleri Ekim 2003’de bir düşüş yaşamış, daha sonra ise Nisan 2004’e kadar artış göstermiştir. Mart 2004’de 564 olan SB tespit adedi, Nisan 2004’te onaylanan kart sayısı (100,930 adet) önceki aylardan çok daha fazla olmasına rağmen 351’e gerilemiştir. Bunda mevcut inceleme yönteminin yetersizliğinin ve çalışan personel performansının da etkisi vardır.

20 Nisan 2004’de kullanıma alınan modelin faydalarını takip eden aylarda daha iyi gözükmetedir. Mayıs 2004’de yakalanan SB adedi Eylül 2003’den beri ulaşılan en yüksek sayıdır. Bu sayı Ağustos 2004’e kadar artış göstermiştir. Ağustos ayındaki gerileme ise bu ay sonunda yapılan model güncellemesiyle aşılmış ve 2004 yılı Eylül ayında tespit edilen SB adedi 844’e çıkmıştır.

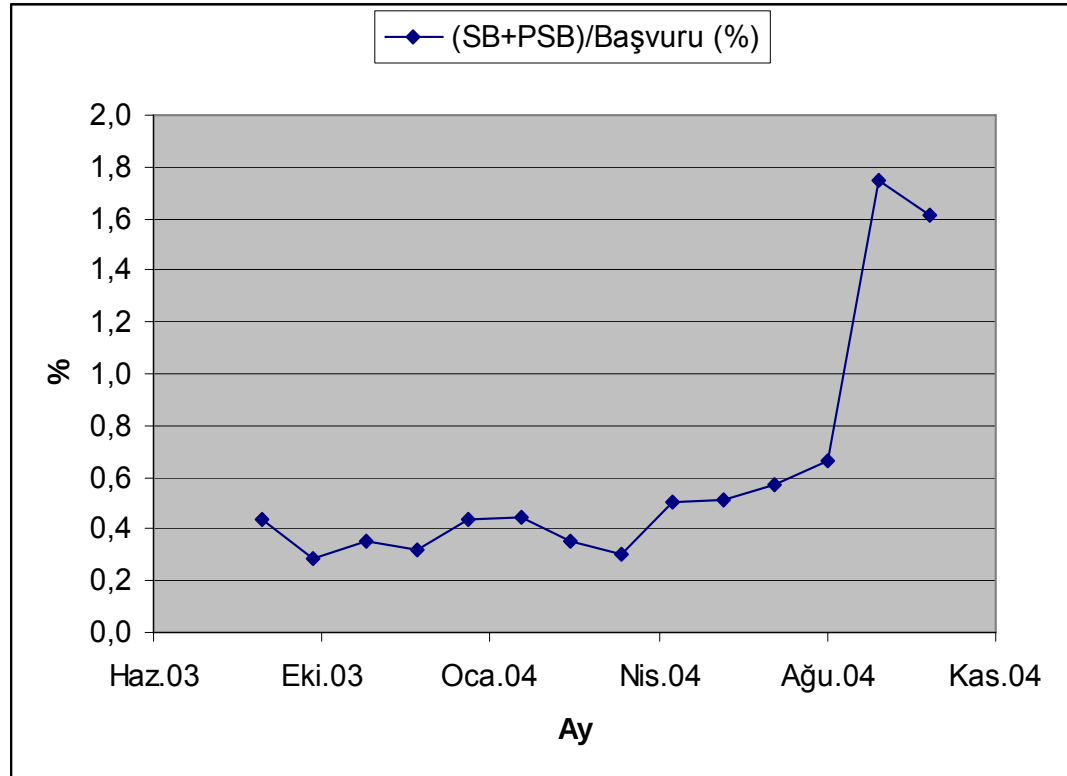
XYZ Bank, Eylül 2004’de kendi içinde gelişen olaylara mütakip, sahtecilik tespit projesinden bağımsız olarak aldığı genel bir kararla pazarlama faaliyetlerini durdurduğu için bu aydan başlayarak başvuru adetlerinde belirgin bir azalış gözükmiştir. Bunun sonucunda İstihbarat Bölümü başvuruları daha dikkatli inceleme fırsatı bulmuş ve onaylanan kart adetlerinde ciddi bir azalma olmuştur. Bu sebeple, Eylül 2004’de tespit edilen 844 adet SB ile 13 adet potansiyel sahte başvurunun (PSB) ayrı bir önemi ortaya çıkmıştır. Çünkü şöyle bir gerçek ortaya çıkmıştır ki, Eylül 2004’de onaylanan başvuruların %3.8’i sahte başvurudur. Bu oran Ekim 2004’te 2,407 adet başvuru içerisinde toplamda (SB + PSB) 356 adet sahte başvuru yakalayarak %14.8’e çıkmıştır.

Tablo 10.7’den gözüken bir başka bilgi de, tespit edilen PSB sayısının giderek azalarak, Eylül 2004’de 5 adede kadar düşmesidir. Bu olay tespit edilen sahte başvuruların (SB) toplam sahte başvurular (SB + PSB) içerisindeki oranını da %98.6’ya çıkarmıştır.

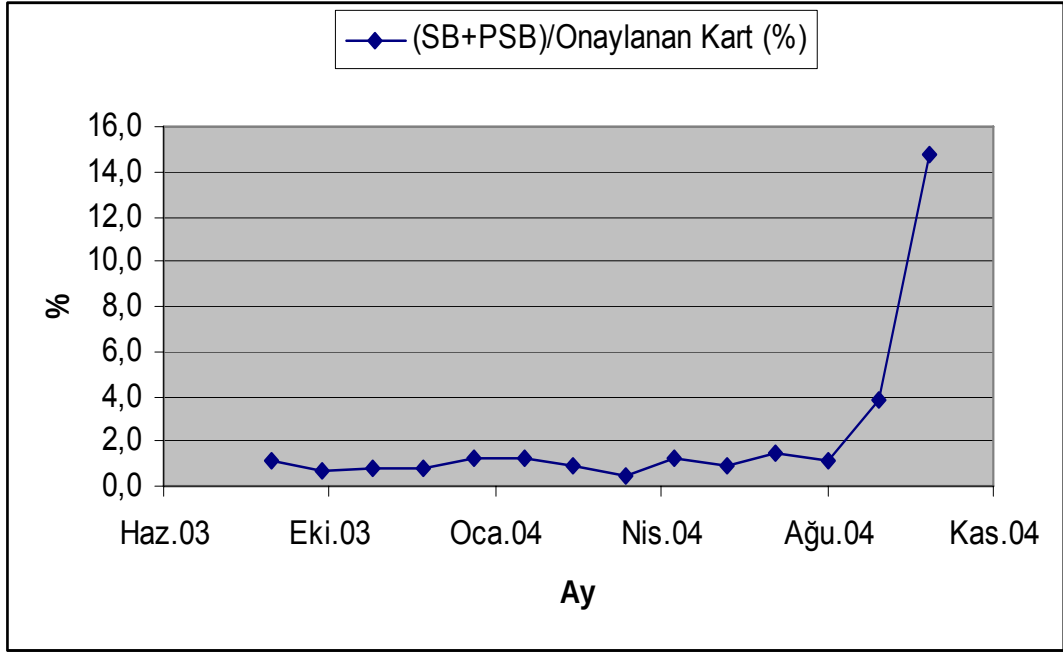
Tablo 10.7’deki bilgiler kullanılarak elde edilen bazı grafikler Şekil 10.12, Şekil 10.13 ve Şekil 10.14’te gösterilmiştir.



Şekil 10.12 : Sahte Başvuru Tespit Adetleri



Şekil 10.13 : Tespit Edilen Toplam Sahte Başvuruların Başvurulara Oranı



Şekil 10.14 : Tespit Edilen Toplam Sahte Başvuruların Onaylanan Kartlara Oranı

11. SONUÇ

Yapılan bu çalışmada veri madenciliği kavramı, teknikleri ve uygulama alanları incelenmiş, gerçekleştirilen bir veri madenciliği projesiyle de bir bankanın ne kadar büyük faydalar elde edebileceği detaylı bir şekilde gösterilmiştir.

Elde edilen başarılı sonuçlar göz önüne alındığında firmaların, ellerinde bulunan çok büyük boyutlardaki veriyi kolaylıkla analiz etmelerini sağlayan ve bu şekilde kendilerini sektörlerinde daha iyi pozisyonlara taşıyabilecek bilgilere ulaşmalarına fırsat tanıyan teknolojileri kullanmaları gerektiği gözükmektedir. Her 20 ayda dünyadaki veri miktarının 2 katına çıktığı tahmin edilen günümüzde, gelişen teknolojilerin sunduğu analiz yöntemlerini ve araçlarını kullanmak, gelecekte elinde veri depolayan her firmanın benimsemek zorunda kalacağı bir yoldur.

Bankaların ortak problemi olan sahte başvuruları, şüphesiz her banka farklı metotlarla önlenmeye çalışmaktadır. Ancak veri madenciliği gibi teknolojileri kullandıklarında daha iyi ve daha hızlı sonuç alabilecekleri de bir gerçektir. XYZ Bank'ta elde edilen başarılı sonuçlar bunu kanıtlamaktadır.

Literatürde sahtecilik tespiti üzerine yapılan çalışmalar, müşterilerin kullanım aşamasında yapmış oldukları sahtekarlıkları tespit etmek üzerine odaklanmış durumdadır. Bankacılık sektörü için kredi kartı kullanımındaki sahteciliklerin tespiti, telekomünikasyon sektörü için ise yapılan konuşmalardaki sahteciliklerin tespiti üzerine yapılan çalışmalara rastlanmaktadır. Ancak sahtekar müşterinin kullanım aşamasına geçmeden yakalanmasına yönelik bir veri madenciliği çalışması bulunmamaktadır. Yaptığım çalışma, literatürdeki bu açığı kapatmaya hizmet etmektedir.

Bankaları büyük zararlara uğratan sahtekarların henüz başvuru aşamasındayken yakalanıp, doğabilecek zararlara ait belli bir kısmın müşteriyle teması kurulduğu ilk aşamada önlenebileceği fikri yapmış olduğum bu çalışmayla ispatlanmıştır.

XYZ Bank'ta tespit edilen sahte başvuru adedini 7'den 23'e çıkaran bu proje, günlük kurtarılan ortalama para tutarını da %356.8'lik artışla 14,312.5 YTL'den 65,380.1 YTL'ye yükseltmiştir.

Proje ile birlikte giderek azalan potansiyel sahte başvuru adedi ve proje hayata geçtikten sonra tespit edilen sahte başvuruların aynı ay içerisindeki tüm sahte başvuruların %98.6'sını oluşturur hale gelmesi, kullanım aşamasını beklemeden yapılacak bu tür analizlerin ne kadar faydalı olabileceğinin bir başka kanıtıdır.

Bu çalışma aynı zamanda, veri madenciliği projelerinde analize sokulabilecek bazı değişkenlerin kullanımına da yenilikler getirmektedir. İstatistikçiler tarafından klasik teknolojiler kullanılarak analiz edilmeleri zor olduğu için ya değerlendirmeye alınmayan ya da sadece “Var” veya “Yok” şeklinde kodlanarak analize sokulabilen telefon numarası bilgilerinin, gelişen teknolojiler sayesinde modelleme çalışmalarında kullanılabileceği yine bu projede gösterilmiştir. Projede, ev/iş telefon numaralarının ilk 6 hanesi seçilmiş ve kişilerin lokasyonlarını belirten bir bilgi olarak kullanılmıştır. Telefon numaraları, posta kodu bilgisinin eksik olduğu başvurular için anlamlı bir bilgi olarak analiz sürecinde yer almışlardır.

Eksik bilgilere sahip kayıtların modelleme sürecine sokulmaması fikrinin yanlış olduğu ise, bu çalışmayla ortaya çıkartılan bir diğer konudur. Kayıtların eksik değerler sahip olmalarına rağmen modellemeye katıldıklarında fayda sağladıkları gözükmiştir. Özellikle sahte başvuru sahiplerinin, başvuru sırasında eksik bilgi vermemesi, sahtecileri gerçek müşterilerde ayıran bir özellik olarak karşımıza çıkmıştır.

Başvuru sırasında tüm sahtecilerin yakalanamadığı bilinmektedir. Bu kişilerin takibine devam edilip kullanım aşamasında yakalanması, bu tez çalışmasının bir sonraki adımıdır. Müşteri işlem bilgilerini de analize sokarak, kredi kartı kullanım aşaması için de bir veri madenciliği modeli oluşturulabilir. Kişinin davranış bilgisi özlük bilgisine göre çok daha anlamlıdır. Davranış bilgilerinin katıldığı modellerle projenin ikinci safhası oluşturulup, daha başarılı sonuçlara imza atmak mümkündür. Oluşturulan bu modelle sahtecilerin kart kullanım alışkanlıkları anlaşılmalı, sahtecilerin ilk olarak nerelerde ve ne kadar tutarda harcama yapabilecekleri belirlenmeye çalışılabilir. Böyle bir çalışma için birliktelik kuralları ve öngörüsül modellemeyi birlikte kullanmak doğru olacaktır.

KAYNAKLAR

- [1] **Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P. and Uthurusamy, R.,** 1996. Advances in Knowledge Discovery and Data Mining, MIT Press, Cambridge.
- [2] **Akpınar, H.,** 2004. Business Intelligence & Data Mining, Dönence Basın ve Yayın Hizmetleri, İstanbul.
- [3] **Dunham, M.H.,** 2003. Data Mining Introductory and Advanced Topics, Pearson Education Inc., New Jersey.
- [4] **Han, J. and Kamber, M.,** 2001. Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, San Francisco.
- [5] **Cabena, P., Hadjinian, P., Stadler, R., Verhees, J. and Kamber, M.,** 1998. Discovering Data Mining: From Concept to Implementation, Prentice Hall, New Jersey.
- [6] **Fabris, P.,** 1998. Advanced Navigation, *CIO*, May 15.
- [7] **Chung, H. and Gray, M.,** 1999. Special Section: Data Mining, *Journal of Management Information Systems*, **16(1)**.
- [8] **Hui, S.C. and Jha, G.,** 2000. Data Mining for Customer Service Support, *Information & Management*, **38(1)**, 1-13.
- [9] **Berry, M.J.A. and Linoff, G.S.,** 2000. Mastering Data Mining, John Wiley & Sons, New York.
- [10] **Chopoorian, J.A., Witherell, R., Khalil, O.E.M and Ahmed, M.,** 2001. Mind Your Business by Mining Your Data, *SAM Advanced Management Journal*, **66(2)**.
- [11] **Bera, M.,** 2001. The New Approaches to Predictive Modeling With A Very Great Number Of Variables, KXEN Inc.
- [12] **Wirth, R. and Hipp, J.,** 2001. CRISP-DM: Towards a Standard Process Model for Data Mining.
<http://www-db.informatik.uni-tuebingen.de/forschung/papers>

- [13] **Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. and Wirth, R.**, 1996. CRISP-DM 1.0: Step-by-Step Data Mining Guide, <http://www.crisp-dm.org/CRISPWP-0800.pdf>
- [14] **Roiger, R.J. and Geatz, M.W.**, 2003. Data Mining: A Tutorial-Based Primer, Pearson Education Inc.,USA.
- [15] **Giudici, P.**, 2003. Applied Data Mining: Statistical Methods for Business and Industry, John Wiley & Sons Ltd, England.
- [16] **Akpınar, H.**, 2000. Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği, *İ.Ü. İşletme Fakültesi Dergisi*, **29(1)**, 1-22.
- [17] **Berry, M.J.A. and Linoff, G.**, 1997. Data Mining Techniques For Marketing, Sales and Customer Support, John Wiley & Sons, New York.
- [18] **KXEN**, 1999. KXEN Association Rules User Guide, USA.
- [19] **Berson, M.A. and Smith, S.**, 1997. Data Warehousing, Data Mining and OLAP, McGraw-Hill, New York
- [20] <http://www.backpropagation.netfirms.com>
- [21] **Ergezer, H., Dikmen, M. ve Özdemir, E.**, 2003. Yapay Sinir Ağları ve Tanıma Sistemleri, *PİVOLKA*, **2(6)**, 14-17.
- [22] **Yurtoğlu, H.**, 2005. Yapay Sinir Ağları Metodolojisi İle Öngörü Modellemesi: Bazı Makroekonomik Değişkenler İçin Türkiye Örneği, *Uzmanlık Tezi*, Devlet Planlama Teşkilatı, Ankara. <http://ekutup.dpt.gov.tr/ekonomi/tahmin/yurtoglu/ysa.pdf>
- [23] **Jack, L.B. and Nandi, A.K.**, 2004. Fault Detection Using Support Vector Machines and Artificial Neural Networks, Augmented By Genetic Algorithms, *Mechanical Systems and Signal Processing*, **16(2-3)**, 373-390.
- [24] **Hsieh, N.**, 2004. An Integrated Data Mining and Behavioral Scoring Model for Analyzing Bank Customers, *Expert Systems with Applications*, **27**, 623-633.
- [25] **Dahlan, N., Ramayah, T. and Hoe K.A.**, 2002. Data Mining in the Banking Industry: An Exploratory Study, *International Conference on Internet Economy And Business*, Kuala Lumpur, 17-18th September 2002.
- [26] **Groth, R.**, 1999. Data Mining: Building Competitive Advantage, Prentice Hall, New Jersey.
- [27] <http://www.bkm.com.tr>
- [28] **Kayalık, B.**, 2005. Kişisel görüşme.

- [29] **Kutluay, K.**, 2005. Kişisel görüşme.
- [30] <http://www.kratis.com/tr/fraud.asp>
- [31] **Gültürk, A.**, 2005. Kişisel görüşme.
- [32] **Usluel, D.**, 2005. Kişisel görüşme.
- [33] **KXEN**, 1999. KXEN Analytic Framework User Guide, USA.

EK A :KXEN ANALYTIC FRAMEWORK PAZAR SEPETİ ANALİZİ SONUÇLARI

Tablo A.1: KXEN Analytic Framework Pazar Sepeti Analizi Sonuçları

Rules		Confidence	KI	Rule Support	Consequent Support	Antecedent Support
Rule4	MANDIRA -> SÜTLÜK	79,40%	0,290808	38,04%	65,72%	47,91%
Rule89	SÜTLÜK -> MANDIRA	57,88%	0,262507	38,04%	47,91%	65,72%
Rule1	MANAV -> SÜTLÜK	72,24%	0,182215	45,51%	65,72%	63%
Rule26	MANDIRA -> MANAV	71,85%	0,18186	34,43%	63%	47,91%
Rule23	SÜTLÜK -> MANAV	69,25%	0,1761	45,51%	63%	65,72%
Rule92	UNLU MAMULLER -> MANDIRA	56,65%	0,171912	27,83%	47,91%	49,13%
Rule70	MANDIRA -> UNLU MAMULLER	58,08%	0,171665	27,83%	49,13%	47,91%
Rule49	SOĞUK İÇECEKLER -> BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER	70,24%	0,171233	26,06%	59,09%	37,11%
Rule90	MANAV -> MANDIRA	54,64%	0,169861	34,43%	47,91%	63%
Rule214	ZÜCCACİYE/HEDİYELİK EŞYA/MUHTELİF EV EŞYALARI -> TEMİZLİK ÜRÜNLERİ	52,59%	0,163707	10,36%	33,96%	19,70%
Rule182	SÜTLÜK & MANAV & BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER -> SOĞUK İÇECEKLER	50,07%	0,16266	14,66%	37,11%	29,28%
Rule213	KOZMETİK VE KİŞİSEL BAKIM -> TEMİZLİK ÜRÜNLERİ	51,92%	0,159537	10,34%	33,96%	19,91%
Rule95	ŞEKER VE ŞEKERLİ MAMÜLLER -> MANDIRA	63,29%	0,156336	16,06%	47,91%	25,38%
Rule5	SOĞUK İÇECEKLER -> SÜTLÜK	75,17%	0,155632	27,89%	65,72%	37,11%
Rule186	SÜTLÜK & BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER & UNLU MAMULLER -> SOĞUK İÇECEKLER	51,98%	0,154402	12,60%	37,11%	24,23%
Rule7	ŞEKER VE ŞEKERLİ MAMÜLLER -> SÜTLÜK	79,30%	0,152927	20,13%	65,72%	25,38%
Rule187	SÜTLÜK & BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER & MANDIRA -> SOĞUK İÇECEKLER	51,33%	0,151547	12,77%	37,11%	24,87%

Tablo A.1: Devamı

Rules		Confidence	KI	Rule Support	Consequent Support	Antecedent Support
Rule119	BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER & TEMİZLİK ÜRÜNLERİ -> SOĞUK İÇECEKLER	52,79%	0,15138	11,89%	37,11%	22,53%
Rule289	MANDIRA & KATKI MADDELERİ -> ŞEKER VE ŞEKERLİ MAMÜLLER	50,66%	0,151327	5,74%	25,38%	11,34%
Rule3	UNLU MAMULLER -> SÜTLÜK	72,50%	0,147758	35,62%	65,72%	49,13%
Rule112	SÜTLÜK & TEMİZLİK ÜRÜNLERİ -> SOĞUK İÇECEKLER	50,51%	0,143474	12,62%	37,11%	24,98%
Rule2	BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER -> SÜTLÜK	71,05%	0,139649	41,98%	65,72%	59,09%
Rule67	SÜTLÜK -> UNLU MAMULLER	54,19%	0,133187	35,62%	49,13%	65,72%
Rule10	KURU GIDALAR -> SÜTLÜK	80,14%	0,133058	16,66%	65,72%	20,79%
Rule318	MANDIRA & SOĞUK İÇECEKLER & TEMİZLİK ÜRÜNLERİ -> ŞEKER VE ŞEKERLİ MAMÜLLER	50,87%	0,131702	4,98%	25,38%	9,79%
Rule196	MANAV & BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER & UNLU MAMULLER -> SOĞUK İÇECEKLER	50,97%	0,130193	11,17%	37,11%	21,91%
Rule45	SÜTLÜK -> BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER	63,87%	0,130136	41,98%	59,09%	65,72%
Rule260	SÜTLÜK & MANDIRA & SOĞUK İÇECEKLER -> TEMİZLİK ÜRÜNLERİ	50,49%	0,127991	8,77%	33,96%	17,36%
Rule197	MANAV & BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER & MANDIRA -> SOĞUK İÇECEKLER	50,53%	0,127384	11,19%	37,11%	22,15%
Rule217	MANDIRA & ŞEKER VE ŞEKERLİ MAMÜLLER -> TEMİZLİK ÜRÜNLERİ	51,66%	0,126771	8,30%	33,96%	16,06%
Rule94	TEMİZLİK ÜRÜNLERİ -> MANDIRA	57,08%	0,124709	19,38%	47,91%	33,96%
Rule304	SÜTLÜK & MANDIRA & KONSERVE GIDALAR -> ŞEKER VE ŞEKERLİ MAMÜLLER	51,79%	0,124295	4,62%	25,38%	8,91%
Rule96	SICAK İÇECEKLER -> MANDIRA	61,52%	0,123849	13,98%	47,91%	22,72%
Rule274	SÜTLÜK & MANAV & BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER & SOĞUK İÇECEKLER -> TEMİZLİK ÜRÜNLERİ	52,69%	0,122418	7,72%	33,96%	14,66%
Rule292	TEMİZLİK ÜRÜNLERİ & KATKI MADDELERİ -> ŞEKER VE ŞEKERLİ MAMÜLLER	52,82%	0,120207	4,38%	25,38%	8,30%
Rule307	SÜTLÜK & TEMİZLİK ÜRÜNLERİ & SICAK İÇECEKLER -> ŞEKER VE ŞEKERLİ MAMÜLLER	52,13%	0,119888	4,43%	25,38%	8,49%
Rule25	UNLU MAMULLER -> MANAV	68,69%	0,119874	33,74%	63%	49,13%

Tablo A.1: Devamı

Rules		Confidence	KI	Rule Support	Consequent Support	Antecedent Support
Rule250	SÜTLÜK & BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER & ŞEKER VE ŞEKERLİ MAMÜLLER -> TEMİZLİK ÜRÜNLERİ	52,19%	0,118336	7,60%	33,96%	14,55%
Rule8	SICAK İÇECEKLER -> SÜTLÜK	77,43%	0,118128	17,60%	65,72%	22,72%
Rule6	TEMİZLİK ÜRÜNLERİ -> SÜTLÜK	73,56%	0,118109	24,98%	65,72%	33,96%
Rule330	SÜTLÜK & BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER & MANDIRA & SICAK İÇECEKLER -> ŞEKER VE ŞEKERLİ MAMÜLLER	51,11%	0,117607	4,43%	25,38%	8,66%
Rule357	SÜTLÜK & TEMİZLİK ÜRÜNLERİ & KONSERVE GIDALAR -> KURU GIDALAR	51,70%	0,117437	3,23%	20,79%	6,26%
Rule291	SOĞUK İÇECEKLER & KATKI MADDELERİ -> ŞEKER VE ŞEKERLİ MAMÜLLER	50,24%	0,116717	4,47%	25,38%	8,89%
Rule305	SÜTLÜK & SOĞUK İÇECEKLER & SICAK İÇECEKLER -> ŞEKER VE ŞEKERLİ MAMÜLLER	50,36%	0,116449	4,45%	25,38%	8,83%
Rule270	BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER & MANDIRA & SOĞUK İÇECEKLER -> TEMİZLİK ÜRÜNLERİ	51,74%	0,116403	7,60%	33,96%	14,68%
Rule120	BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER & ŞEKER VE ŞEKERLİ MAMÜLLER -> SOĞUK İÇECEKLER	52,41%	0,1158	9,26%	37,11%	17,66%
Rule246	SÜTLÜK & MANAV & ŞEKER VE ŞEKERLİ MAMÜLLER -> TEMİZLİK ÜRÜNLERİ	51,78%	0,114272	7,45%	33,96%	14,38%
Rule122	BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER & KÜMES HAYVANLARI -> SOĞUK İÇECEKLER	55,11%	0,114062	8,15%	37,11%	14,79%
Rule319	MANDIRA & SOĞUK İÇECEKLER & SICAK İÇECEKLER -> ŞEKER VE ŞEKERLİ MAMÜLLER	56,23%	0,114011	3,94%	25,38%	7%
Rule303	SÜTLÜK & BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER & KONSERVE GIDALAR -> ŞEKER VE ŞEKERLİ MAMÜLLER	50,12%	0,113113	4,34%	25,38%	8,66%
Rule113	SÜTLÜK & ŞEKER VE ŞEKERLİ MAMÜLLER -> SOĞUK İÇECEKLER	50,21%	0,113025	10,11%	37,11%	20,13%
Rule298	SICAK İÇECEKLER & KATKI MADDELERİ -> ŞEKER VE ŞEKERLİ MAMÜLLER	62,45%	0,112022	3,57%	25,38%	5,72%
Rule223	SOĞUK İÇECEKLER & ŞEKER VE ŞEKERLİ MAMÜLLER -> TEMİZLİK ÜRÜNLERİ	55,23%	0,111832	6,51%	33,96%	11,79%
Rule68	MANAV -> UNLU MAMULLER	53,56%	0,111804	33,74%	49,13%	63%

Tablo A.1: Devamı

Rules		Confidence	KI	Rule Support	Consequent Support	Antecedent Support
Rule137	MANDIRA & TEMİZLİK ÜRÜNLERİ -> SOĞUK İÇECEKLER	50,49%	0,11119	9,79%	37,11%	19,38%
Rule358	SÜTLÜK & ŞEKER VE ŞEKERLİ MAMÜLLER & KONSERVE GIDALAR -> KURU GIDALAR	51,43%	0,11086	3,06%	20,79%	5,96%
Rule208	BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER & UNLU MAMÜLLER & MANDIRA -> SOĞUK İÇECEKLER	50,98%	0,110135	9,45%	37,11%	18,53%
Rule51	ŞEKER VE ŞEKERLİ MAMÜLLER -> BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER	69,57%	0,110116	17,66%	59,09%	25,38%
Rule218	MANDIRA & SICAK İÇECEKLER -> TEMİZLİK ÜRÜNLERİ	51,45%	0,109009	7,19%	33,96%	13,98%
Rule322	MANDIRA & TEMİZLİK ÜRÜNLERİ & SICAK İÇECEKLER -> ŞEKER VE ŞEKERLİ MAMÜLLER	53,85%	0,108074	3,87%	25,38%	7,19%
Rule93	SOĞUK İÇECEKLER -> MANDIRA	55,16%	0,107731	20,47%	47,91%	37,11%
Rule13	KATKI MADDELERİ -> SÜTLÜK	78,58%	0,1077	14,83%	65,72%	18,87%
Rule75	KÜMES HAYVANLARI -> UNLU MAMÜLLER	60,82%	0,106715	13,87%	49,13%	22,81%
Rule123	BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER & KURU GIDALAR -> SOĞUK İÇECEKLER	54,48%	0,106109	7,77%	37,11%	14,26%
Rule115	SÜTLÜK & KÜMES HAYVANLARI -> SOĞUK İÇECEKLER	51,43%	0,105112	8,81%	37,11%	17,13%
Rule371	MANDIRA & ŞEKER VE ŞEKERLİ MAMÜLLER & KONSERVE GIDALAR -> KATKI MADDELERİ	50,85%	0,104865	2,55%	18,87%	5,02%
Rule364	MANDIRA & TEMİZLİK ÜRÜNLERİ & KATKI MADDELERİ -> KURU GIDALAR	50,36%	0,104719	2,94%	20,79%	5,83%
Rule132	UNLU MAMÜLLER & TEMİZLİK ÜRÜNLERİ -> SOĞUK İÇECEKLER	50,41%	0,102991	9,11%	37,11%	18,06%
Rule212	SÜTLÜK & MANAV & UNLU MAMÜLLER & MANDIRA -> SOĞUK İÇECEKLER	50,85%	0,102773	8,87%	37,11%	17,45%
Rule50	TEMİZLİK ÜRÜNLERİ -> BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER	66,35%	0,102096	22,53%	59,09%	33,96%
Rule98	KURU GIDALAR -> MANDIRA	60,08%	0,101343	12,49%	47,91%	20,79%
Rule293	TEMİZLİK ÜRÜNLERİ & KONSERVE GIDALAR -> ŞEKER VE ŞEKERLİ MAMÜLLER	50,85%	0,100712	3,81%	25,38%	7,49%
Rule262	MANAV & BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER & ŞEKER VE ŞEKERLİ MAMÜLLER -> TEMİZLİK ÜRÜNLERİ	52,34%	0,100628	6,43%	33,96%	12,28%

Tablo A.1: Devamı

Rules		Confidence	KI	Rule Support	Consequent Support	Antecedent Support
Rule361	MANAV & TEMİZLİK ÜRÜNLERİ & KONSERVE GIDALAR -> KURU GIDALAR	50%	0,100407	2,83%	20,79%	5,66%
Rule248	SÜTLÜK & MANAV & KURU GIDALAR -> TEMİZLİK ÜRÜNLERİ	52,17%	0,099679	6,40%	33,96%	12,28%
Rule313	BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER & MANDIRA & KONSERVE GIDALAR -> ŞEKER VE ŞEKERLİ MAMÜLLER	52,11%	0,099675	3,68%	25,38%	7,06%
Rule365	MANDIRA & TEMİZLİK ÜRÜNLERİ & KONSERVE GIDALAR -> KURU GIDALAR	52,03%	0,099318	2,72%	20,79%	5,23%
Rule219	MANDIRA & KURU GIDALAR -> TEMİZLİK ÜRÜNLERİ	51,79%	0,099304	6,47%	33,96%	12,49%
Rule47	UNLU MAMÜLLER -> BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER	63,97%	0,099212	31,43%	59,09%	49,13%
Rule247	SÜTLÜK & MANAV & SICAK İÇECEKLER -> TEMİZLİK ÜRÜNLERİ	51,25%	0,098588	6,55%	33,96%	12,79%
Rule114	SÜTLÜK & SICAK İÇECEKLER -> SOĞUK İÇECEKLER	50,18%	0,098581	8,83%	37,11%	17,60%
Rule368	SÜTLÜK & MANAV & TEMİZLİK ÜRÜNLERİ & KATKI MADDELERİ -> KURU GIDALAR	50%	0,098142	2,77%	20,79%	5,53%
Rule271	UNLU MAMÜLLER & MANDIRA & SOĞUK İÇECEKLER -> TEMİZLİK ÜRÜNLERİ	51,26%	0,097979	6,51%	33,96%	12,70%
Rule375	SÜTLÜK & MANDIRA & TEMİZLİK ÜRÜNLERİ & KONSERVE GIDALAR -> KATKI MADDELERİ	50,68%	0,097682	2,38%	18,87%	4,70%
Rule315	BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER & TEMİZLİK ÜRÜNLERİ & SICAK İÇECEKLER -> ŞEKER VE ŞEKERLİ MAMÜLLER	50,43%	0,097644	3,72%	25,38%	7,38%
Rule373	SÜTLÜK & MANAV & ŞEKER VE ŞEKERLİ MAMÜLLER & KONSERVE GIDALAR -> KATKI MADDELERİ	50,45%	0,096555	2,36%	18,87%	4,68%
Rule340	MANAV & BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER & MANDIRA & SICAK İÇECEKLER -> ŞEKER VE ŞEKERLİ MAMÜLLER	50,14%	0,096521	3,70%	25,38%	7,38%
Rule69	BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER -> UNLU MAMÜLLER	53,19%	0,095965	31,43%	49,13%	59,09%
Rule102	KONSERVE GIDALAR -> MANDIRA	62,13%	0,095408	10,40%	47,91%	16,74%
Rule252	SÜTLÜK & BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER & KURU GIDALAR -> TEMİZLİK ÜRÜNLERİ	51,97%	0,095364	6,17%	33,96%	11,87%

Tablo A.1: Devamı

Rules		Confidence	KI	Rule Support	Consequent Support	Antecedent Support
Rule306	SÜTLÜK & SOĞUK İÇECEKLER & KONSERVE GIDALAR -> ŞEKER VE ŞEKERLİ MAMÜLLER	51,54%	0,095216	3,55%	25,38%	6,89%
Rule9	KÜMES HAYVANLARI -> SÜTLÜK	75,09%	0,094867	17,13%	65,72%	22,81%
Rule276	SÜTLÜK & MANAV & UNLU MAMULLER & SOĞUK İÇECEKLER -> TEMİZLİK ÜRÜNLERİ	51,30%	0,094631	6,28%	33,96%	12,23%
Rule360	SÜTLÜK & KATKI MADDELERİ & KONSERVE GIDALAR -> KURU GIDALAR	52,61%	0,094571	2,57%	20,79%	4,89%
Rule301	KURU GIDALAR & KONSERVE GIDALAR -> ŞEKER VE ŞEKERLİ MAMÜLLER	54,93%	0,094264	3,32%	25,38%	6,04%
Rule138	MANDIRA & ŞEKER VE ŞEKERLİ MAMÜLLER -> SOĞUK İÇECEKLER	50,73%	0,093764	8,15%	37,11%	16,06%
Rule317	UNLU MAMULLER & SOĞUK İÇECEKLER & SICAK İÇECEKLER -> ŞEKER VE ŞEKERLİ MAMÜLLER	53,36%	0,093642	3,38%	25,38%	6,34%
Rule329	SÜTLÜK & BİSKÜVİ/KEK/ÇİKOLATA/Ş EKER & UNLU MAMULLER & KATKI MADDELERİ -> ŞEKER VE ŞEKERLİ MAMÜLLER	52,27%	0,093038	3,43%	25,38%	6,55%
Rule362	MANAV & ŞEKER VE ŞEKERLİ MAMÜLLER & KONSERVE GIDALAR -> KURU GIDALAR	50,41%	0,09264	2,60%	20,79%	5,15%
Rule372	MANDIRA & KURU GIDALAR & KONSERVE GIDALAR -> KATKI MADDELERİ	52,53%	0,092597	2,21%	18,87%	4,21%
Rule146	TEMİZLİK ÜRÜNLERİ & ŞEKER VE ŞEKERLİ MAMÜLLER -> SOĞUK İÇECEKLER	55,54%	0,092576	6,51%	37,11%	11,72%
Rule121	BİSKÜVİ/KEK/ÇİKOLATA/Ş EKER & SICAK İÇECEKLER -> SOĞUK İÇECEKLER	51,03%	0,092531	7,91%	37,11%	15,51%
Rule116	SÜTLÜK & KURU GIDALAR -> SOĞUK İÇECEKLER	50,06%	0,092497	8,34%	37,11%	16,66%
Rule97	KÜMES HAYVANLARI -> MANDIRA	58,02%	0,092375	13,23%	47,91%	22,81%
Rule101	KATKI MADDELERİ -> MANDIRA	60,09%	0,092071	11,34%	47,91%	18,87%
Rule370	UNLU MAMULLER & ŞEKER VE ŞEKERLİ MAMÜLLER & KONSERVE GIDALAR -> KATKI MADDELERİ	53,12%	0,091391	2,17%	18,87%	4,09%
Rule117	SÜTLÜK & KONSERVE GIDALAR -> SOĞUK İÇECEKLER	53,47%	0,090381	6,89%	37,11%	12,89%
Rule27	SOĞUK İÇECEKLER -> MANAV	68,64%	0,089707	25,47%	63%	37,11%

Tablo A.1: Devamı

Rules		Confidence	KI	Rule Support	Consequent Support	Antecedent Support
Rule251	SÜTLÜK & BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER & SICAK İÇECEKLER -> TEMİZLİK ÜRÜNLERİ	50,17%	0,08953	6,21%	33,96%	12,38%
Rule300	KURU GIDALAR & KATKI MADDELERİ -> ŞEKER VE ŞEKERLİ MAMÜLLER	50,47%	0,089348	3,40%	25,38%	6,74%
Rule302	KATKI MADDELERİ & KONSERVE GIDALAR -> ŞEKER VE ŞEKERLİ MAMÜLLER	55,26%	0,089287	3,13%	25,38%	5,66%
Rule324	MANDIRA & TEMİZLİK ÜRÜNLERİ & ZÜCCACIYE/HEDİYELİK EŞYA/MUHTELİF EV EŞYALARI -> ŞEKER VE ŞEKERLİ MAMÜLLER	50,97%	0,089097	3,36%	25,38%	6,60%
Rule128	BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER & KONSERVE GIDALAR -> SOĞUK İÇECEKLER	56,14%	0,089021	6,13%	37,11%	10,91%
Rule354	UNLU MAMÜLLER & MANDIRA & SOĞUK İÇECEKLER & ŞEKER VE ŞEKERLİ MAMÜLLER -> SICAK İÇECEKLER	51,79%	0,088408	2,77%	22,72%	5,34%
Rule220	MANDIRA & KATKI MADDELERİ -> TEMİZLİK ÜRÜNLERİ	51,41%	0,088239	5,83%	33,96%	11,34%
Rule295	SICAK İÇECEKLER & KURU GIDALAR -> ŞEKER VE ŞEKERLİ MAMÜLLER	51,50%	0,087741	3,28%	25,38%	6,36%
Rule342	BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER & UNLU MAMÜLLER & MANDIRA & SICAK İÇECEKLER -> ŞEKER VE ŞEKERLİ MAMÜLLER	50,81%	0,087706	3,32%	25,38%	6,53%
Rule287	SÜTLÜK & YAĞLAR -> ŞEKER VE ŞEKERLİ MAMÜLLER	51,69%	0,087473	3,26%	25,38%	6,30%
Rule321	MANDIRA & SOĞUK İÇECEKLER & KONSERVE GIDALAR -> ŞEKER VE ŞEKERLİ MAMÜLLER	54,51%	0,08704	3,09%	25,38%	5,66%
Rule374	SÜTLÜK & MANAV & KURU GIDALAR & KONSERVE GIDALAR -> KATKI MADDELERİ	50,50%	0,087038	2,13%	18,87%	4,21%
Rule299	SICAK İÇECEKLER & KONSERVE GIDALAR -> ŞEKER VE ŞEKERLİ MAMÜLLER	57,20%	0,086858	2,96%	25,38%	5,17%
Rule369	BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER & MANDIRA & TEMİZLİK ÜRÜNLERİ & KOZMETİK VE KİŞİSEL BAKIM -> ZÜCCACIYE/HEDİYELİK EŞYA/MUHTELİF EV EŞYALARI	50%	0,086384	2,26%	19,70%	4,51%
Rule52	SICAK İÇECEKLER -> BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER	68,26%	0,086227	15,51%	59,09%	22,72%
Rule133	UNLU MAMÜLLER & ŞEKER VE ŞEKERLİ MAMÜLLER -> SOĞUK İÇECEKLER	51,27%	0,086155	7,28%	37,11%	14,19%

Tablo A.1: Devamı

Rules		Confidence	KI	Rule Support	Consequent Support	Antecedent Support
Rule345	SÜTLÜK & BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER & UNLU MAMÜLLER & SOĞUK İÇECEKLER & TEMİZLİK ÜRÜNLERİ -> ŞEKER VE ŞEKERLİ MAMÜLLER	51,54%	0,086081	3,21%	25,38%	6,23%
Rule326	SOĞUK İÇECEKLER & TEMİZLİK ÜRÜNLERİ & SICAK İÇECEKLER -> ŞEKER VE ŞEKERLİ MAMÜLLER	55,51%	0,085968	3%	25,38%	5,40%
Rule347	SÜTLÜK & ŞEKER VE ŞEKERLİ MAMÜLLER & KOZMETİK VE KİŞİSEL BAKIM -> SICAK İÇECEKLER	51,42%	0,085874	2,70%	22,72%	5,26%
Rule225	SOĞUK İÇECEKLER & KURU GIDALAR -> TEMİZLİK ÜRÜNLERİ	53,85%	0,085854	5,21%	33,96%	9,68%
Rule263	MANAV & BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER & SICAK İÇECEKLER -> TEMİZLİK ÜRÜNLERİ	51,90%	0,085266	5,53%	33,96%	10,66%
Rule31	KÜMES HAYVANLARI -> MANAV	71,64%	0,084559	16,34%	63%	22,81%
Rule297	SICAK İÇECEKLER & ZÜCCACIYE/HEDİYELİK EŞYA/MUHTELİF EV EŞYALARI -> ŞEKER VE ŞEKERLİ MAMÜLLER	52,73%	0,084474	3,09%	25,38%	5,85%
Rule281	MANAV & BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER & UNLU MAMÜLLER & SOĞUK İÇECEKLER -> TEMİZLİK ÜRÜNLERİ	50,86%	0,084175	5,68%	33,96%	11,17%
Rule37	KONSERVE GIDALAR -> MANAV	74,71%	0,084148	12,51%	63%	16,74%
Rule15	KONSERVE GIDALAR -> SÜTLÜK	77%	0,083827	12,89%	65,72%	16,74%
Rule224	SOĞUK İÇECEKLER & SICAK İÇECEKLER -> TEMİZLİK ÜRÜNLERİ	51,84%	0,083117	5,40%	33,96%	10,43%
Rule349	MANDIRA & ŞEKER VE ŞEKERLİ MAMÜLLER & ZÜCCACIYE/HEDİYELİK EŞYA/MUHTELİF EV EŞYALARI -> SICAK İÇECEKLER	53,36%	0,082789	2,53%	22,72%	4,74%
Rule71	SOĞUK İÇECEKLER -> UNLU MAMÜLLER	54,70%	0,08276	20,30%	49,13%	37,11%
Rule366	MANDIRA & KATKI MADDELERİ & KONSERVE GIDALAR -> KURU GIDALAR	53,89%	0,082543	2,21%	20,79%	4,11%
Rule140	MANDIRA & KÜMES HAYVANLARI -> SOĞUK İÇECEKLER	51,61%	0,082233	6,83%	37,11%	13,23%
Rule367	SOĞUK İÇECEKLER & TEMİZLİK ÜRÜNLERİ & KONSERVE GIDALAR -> KURU GIDALAR	52,76%	0,082223	2,23%	20,79%	4,23%

Tablo A.1: Devamı

Rules		Confidence	KI	Rule Support	Consequent Support	Antecedent Support
Rule314	BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER & SOĞUK İÇECEKLER & KONSERVE GIDALAR -> ŞEKER VE ŞEKERLİ MAMÜLLER	50,69%	0,08189	3,11%	25,38%	6,13%
Rule126	BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER & KATKI MADDELERİ -> SOĞUK İÇECEKLER	52,09%	0,081808	6,64%	37,11%	12,74%
Rule337	SÜTLÜK & MANDIRA & TEMİZLİK ÜRÜNLERİ & KOZMETİK VE KİŞİSEL BAKIM -> ŞEKER VE ŞEKERLİ MAMÜLLER	53,26%	0,081725	2,96%	25,38%	5,55%
Rule54	KURU GIDALAR -> BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER	68,58%	0,081621	14,26%	59,09%	20,79%
Rule264	MANAV & BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER & KURU GIDALAR -> TEMİZLİK ÜRÜNLERİ	51,65%	0,081559	5,34%	33,96%	10,34%
Rule359	SÜTLÜK & SICAK İÇECEKLER & KONSERVE GIDALAR -> KURU GIDALAR	51,21%	0,081367	2,26%	20,79%	4,40%
Rule32	KURU GIDALAR -> MANAV	72,06%	0,080771	14,98%	63%	20,79%
Rule142	MANDIRA & KATKI MADDELERİ -> SOĞUK İÇECEKLER	53,66%	0,080432	6,09%	37,11%	11,34%
Rule141	MANDIRA & KURU GIDALAR -> SOĞUK İÇECEKLER	52,13%	0,080398	6,51%	37,11%	12,49%
Rule336	SÜTLÜK & MANDIRA & TEMİZLİK ÜRÜNLERİ & KURU GIDALAR -> ŞEKER VE ŞEKERLİ MAMÜLLER	50,90%	0,079963	3,02%	25,38%	5,94%
Rule229	ŞEKER VE ŞEKERLİ MAMÜLLER & SICAK İÇECEKLER -> TEMİZLİK ÜRÜNLERİ	53,20%	0,079946	4,96%	33,96%	9,32%
Rule12	ZÜCCACİYE/HEDİYELİK EŞYA/MUHTELİF EV EŞYALARI -> SÜTLÜK	74,84%	0,079714	14,74%	65,72%	19,70%
Rule363	MANAV & KATKI MADDELERİ & KONSERVE GIDALAR -> KURU GIDALAR	50%	0,079269	2,23%	20,79%	4,47%
Rule356	MANDIRA & SOĞUK İÇECEKLER & TEMİZLİK ÜRÜNLERİ & ŞEKER VE ŞEKERLİ MAMÜLLER -> SICAK İÇECEKLER	50,43%	0,078549	2,51%	22,72%	4,98%
Rule290	MANDIRA & YAĞLAR -> ŞEKER VE ŞEKERLİ MAMÜLLER	53,20%	0,078122	2,83%	25,38%	5,32%
Rule296	SICAK İÇECEKLER & KOZMETİK VE KİŞİSEL BAKIM -> ŞEKER VE ŞEKERLİ MAMÜLLER	50%	0,077984	3%	25,38%	6%
Rule333	SÜTLÜK & BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER & TEMİZLİK ÜRÜNLERİ & KOZMETİK VE KİŞİSEL BAKIM -> ŞEKER VE ŞEKERLİ MAMÜLLER	51,71%	0,077785	2,89%	25,38%	5,60%

Tablo A.1: Devamı

Rules		Confidence	KI	Rule Support	Consequent Support	Antecedent Support
Rule348	MANDIRA & ŞEKER VE ŞEKERLİ MAMÜLLER & KOZMETİK VE KİŞİSEL BAKIM -> SICAK İÇECEKLER	53,88%	0,077776	2,36%	22,72%	4,38%
Rule139	MANDIRA & SICAK İÇECEKLER -> SOĞUK İÇECEKLER	50,08%	0,077686	7%	37,11%	13,98%
Rule323	MANDIRA & TEMİZLİK ÜRÜNLERİ & KÜMES HAYVANLARI -> ŞEKER VE ŞEKERLİ MAMÜLLER	50%	0,077431	2,98%	25,38%	5,96%
Rule255	SÜTLÜK & UNLU MAMÜLLER & SICAK İÇECEKLER -> TEMİZLİK ÜRÜNLERİ	50,40%	0,077085	5,30%	33,96%	10,51%
Rule143	MANDIRA & KONSERVE GIDALAR -> SOĞUK İÇECEKLER	54,40%	0,077083	5,66%	37,11%	10,40%
Rule231	ŞEKER VE ŞEKERLİ MAMÜLLER & KURU GIDALAR -> TEMİZLİK ÜRÜNLERİ	56,74%	0,076953	4,30%	33,96%	7,57%
Rule74	SICAK İÇECEKLER -> UNLU MAMULLER	57,58%	0,076889	13,09%	49,13%	22,72%
Rule35	KATKI MADDELERİ -> MANAV	72,49%	0,076846	13,68%	63%	18,87%
Rule256	SÜTLÜK & UNLU MAMULLER & KURU GIDALAR -> TEMİZLİK ÜRÜNLERİ	51,29%	0,076618	5,09%	33,96%	9,91%
Rule232	ŞEKER VE ŞEKERLİ MAMÜLLER & KATKI MADDELERİ -> TEMİZLİK ÜRÜNLERİ	55,83%	0,07656	4,38%	33,96%	7,85%
Rule253	SÜTLÜK & BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER & KATKI MADDELERİ -> TEMİZLİK ÜRÜNLERİ	50,41%	0,076476	5,26%	33,96%	10,43%
Rule184	SÜTLÜK & MANAV & KATKI MADDELERİ -> SOĞUK İÇECEKLER	53,04%	0,076418	5,94%	37,11%	11,19%
Rule100	ZÜCCACİYE/HEDİYELİK EŞYA/MUHTELİF EV EŞYALARI -> MANDIRA	57,56%	0,076139	11,34%	47,91%	19,70%
Rule125	BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER & ZÜCCACİYE/HEDİYELİK EŞYA/MUHTELİF EV EŞYALARI -> SOĞUK İÇECEKLER	51%	0,076	6,51%	37,11%	12,77%
Rule221	MANDIRA & KONSERVE GIDALAR -> TEMİZLİK ÜRÜNLERİ	50,31%	0,075849	5,23%	33,96%	10,40%
Rule350	TEMİZLİK ÜRÜNLERİ & ŞEKER VE ŞEKERLİ MAMÜLLER & KOZMETİK VE KİŞİSEL BAKIM -> SICAK İÇECEKLER	54%	0,075793	2,30%	22,72%	4,26%
Rule265	MANAV & UNLU MAMULLER & ŞEKER VE ŞEKERLİ MAMÜLLER -> TEMİZLİK ÜRÜNLERİ	50,63%	0,075598	5,15%	33,96%	10,17%

Tablo A.1: Devamı

Rules		Confidence	KI	Rule Support	Consequent Support	Antecedent Support
Rule320	MANDIRA & SOĞUK İÇECEKLER & ZÜCCACİYE/HEDİYELİK EŞYA/MUHTELİF EV EŞYALARI -> ŞEKER VE ŞEKERLİ MAMÜLLER	50,76%	0,075253	2,85%	25,38%	5,62%
Rule124	BİSKÜVİ/KEK/ÇİKOLATA/Ş EKER & KOZMETİK VE KİŞİSEL BAKIM -> SOĞUK İÇECEKLER	50,93%	0,074721	6,43%	37,11%	12,62%
Rule215	YAĞLAR -> TEMİZLİK ÜRÜNLERİ	55,11%	0,074645	4,36%	33,96%	7,91%
Rule28	TEMİZLİK ÜRÜNLERİ -> MANAV	68,11%	0,074409	23,13%	63%	33,96%
Rule267	BİSKÜVİ/KEK/ÇİKOLATA/Ş EKER & UNLU MAMÜLLER & ŞEKER VE ŞEKERLİ MAMÜLLER -> TEMİZLİK ÜRÜNLERİ	50,31%	0,074023	5,11%	33,96%	10,15%
Rule226	SOĞUK İÇECEKLER & KATKI MADDELERİ -> TEMİZLİK ÜRÜNLERİ	52,39%	0,073107	4,66%	33,96%	8,89%
Rule149	TEMİZLİK ÜRÜNLERİ & KURU GIDALAR -> SOĞUK İÇECEKLER	55,06%	0,072823	5,21%	37,11%	9,47%
Rule257	SÜTLÜK & UNLU MAMÜLLER & KATKI MADDELERİ -> TEMİZLİK ÜRÜNLERİ	52,40%	0,072803	4,64%	33,96%	8,85%
Rule134	UNLU MAMÜLLER & KURU GIDALAR -> SOĞUK İÇECEKLER	51,34%	0,072783	6,13%	37,11%	11,94%
Rule266	MANAV & UNLU MAMÜLLER & KURU GIDALAR -> TEMİZLİK ÜRÜNLERİ	51,89%	0,072123	4,68%	33,96%	9,02%
Rule332	SÜTLÜK & BİSKÜVİ/KEK/ÇİKOLATA/Ş EKER & TEMİZLİK ÜRÜNLERİ & KÜMES HAYVANLARI -> ŞEKER VE ŞEKERLİ MAMÜLLER	50,79%	0,071935	2,72%	25,38%	5,36%
Rule294	TEMİZLİK ÜRÜNLERİ & YAĞLAR -> ŞEKER VE ŞEKERLİ MAMÜLLER	56,59%	0,071856	2,47%	25,38%	4,36%
Rule103	KIRMIZI ET -> MANDIRA	65,23%	0,071728	6,74%	47,91%	10,34%
Rule48	MANDIRA -> BİSKÜVİ/KEK/ÇİKOLATA/Ş EKER	62,70%	0,071645	30,04%	59,09%	47,91%
Rule308	SÜTLÜK & KÜMES HAYVANLARI & KATKI MADDELERİ -> ŞEKER VE ŞEKERLİ MAMÜLLER	52,14%	0,070327	2,60%	25,38%	4,98%
Rule288	BİSKÜVİ/KEK/ÇİKOLATA/Ş EKER & YAĞLAR -> ŞEKER VE ŞEKERLİ MAMÜLLER	50,82%	0,069722	2,64%	25,38%	5,19%
Rule351	TEMİZLİK ÜRÜNLERİ & ŞEKER VE ŞEKERLİ MAMÜLLER & ZÜCCACİYE/HEDİYELİK EŞYA/MUHTELİF EV EŞYALARI -> SICAK İÇECEKLER	51,24%	0,06946	2,19%	22,72%	4,28%

Tablo A.1: Devamı

Rules		Confidence	KI	Rule Support	Consequent Support	Antecedent Support
Rule328	SOĞUK İÇECEKLER & TEMİZLİK ÜRÜNLERİ & KURU GIDALAR -> ŞEKER VE ŞEKERLİ MAMÜLLER	50,61%	0,069437	2,64%	25,38%	5,21%
Rule91	BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER -> MANDIRA	50,85%	0,0694	30,04%	47,91%	59,09%
Rule277	SÜTLÜK & MANAV & MANDIRA & KÜMES HAYVANLARI -> TEMİZLİK ÜRÜNLERİ	51,44%	0,069312	4,57%	33,96%	8,89%
Rule352	TEMİZLİK ÜRÜNLERİ & ŞEKER VE ŞEKERLİ MAMÜLLER & KATKI MADDELERİ -> SICAK İÇECEKLER	50,49%	0,069295	2,21%	22,72%	4,38%
Rule76	KURU GIDALAR -> UNLU MAMÜLLER	57,42%	0,068977	11,94%	49,13%	20,79%
Rule73	ŞEKER VE ŞEKERLİ MAMÜLLER -> UNLU MAMÜLLER	55,91%	0,068878	14,19%	49,13%	25,38%
Rule344	BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER & MANDIRA & TEMİZLİK ÜRÜNLERİ & KOZMETİK VE KİŞİSEL BAKIM -> ŞEKER VE ŞEKERLİ MAMÜLLER	54,25%	0,068737	2,45%	25,38%	4,51%
Rule327	SOĞUK İÇECEKLER & TEMİZLİK ÜRÜNLERİ & KÜMES HAYVANLARI -> ŞEKER VE ŞEKERLİ MAMÜLLER	52,19%	0,068668	2,53%	25,38%	4,85%
Rule234	ŞEKER VE ŞEKERLİ MAMÜLLER & KONSERVE GIDALAR -> TEMİZLİK ÜRÜNLERİ	57,01%	0,068663	3,81%	33,96%	6,68%
Rule355	UNLU MAMÜLLER & SOĞUK İÇECEKLER & TEMİZLİK ÜRÜNLERİ & ŞEKER VE ŞEKERLİ MAMÜLLER -> SICAK İÇECEKLER	51,53%	0,068413	2,15%	22,72%	4,17%
Rule183	SÜTLÜK & MANAV & ZÜCCACIYE/HEDİYELİK EŞYA/MUHTELİF EV EŞYALARI -> SOĞUK İÇECEKLER	51,89%	0,067789	5,55%	37,11%	10,70%
Rule151	TEMİZLİK ÜRÜNLERİ & KATKI MADDELERİ -> SOĞUK İÇECEKLER	56,15%	0,067725	4,66%	37,11%	8,30%
Rule147	TEMİZLİK ÜRÜNLERİ & SICAK İÇECEKLER -> SOĞUK İÇECEKLER	52,37%	0,067496	5,40%	37,11%	10,32%
Rule353	SÜTLÜK & MANAV & ŞEKER VE ŞEKERLİ MAMÜLLER & ZÜCCACIYE/HEDİYELİK EŞYA/MUHTELİF EV EŞYALARI -> SICAK İÇECEKLER	51,28%	0,067477	2,13%	22,72%	4,15%
Rule309	SÜTLÜK & KÜMES HAYVANLARI & KONSERVE GIDALAR -> ŞEKER VE ŞEKERLİ MAMÜLLER	54,07%	0,067345	2,40%	25,38%	4,45%
Rule236	SICAK İÇECEKLER & KURU GIDALAR -> TEMİZLİK ÜRÜNLERİ	57,53%	0,066855	3,66%	33,96%	6,36%

Tablo A.1: Devamı

Rules		Confidence	KI	Rule Support	Consequent Support	Antecedent Support
Rule335	SÜTLÜK & MANDIRA & SOĞUK İÇECEKLER & KOZMETİK VE KİŞİSEL BAKIM -> ŞEKER VE ŞEKERLİ MAMÜLLER	52,02%	0,066724	2,47%	25,38%	4,74%
Rule159	ŞEKER VE ŞEKERLİ MAMÜLLER & KATKI MADDELERİ -> SOĞUK İÇECEKLER	56,91%	0,066624	4,47%	37,11%	7,85%
Rule312	SÜTLÜK & ZÜCCACIYE/HEDİYELİK EŞYA/MUHTELİF EV EŞYALARI & KATKI MADDELERİ -> ŞEKER VE ŞEKERLİ MAMÜLLER	51,79%	0,066438	2,47%	25,38%	4,77%
Rule227	SOĞUK İÇECEKLER & KONSERVE GIDALAR -> TEMİZLİK ÜRÜNLERİ	52,37%	0,066375	4,23%	33,96%	8,09%
Rule254	SÜTLÜK & BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER & KONSERVE GIDALAR -> TEMİZLİK ÜRÜNLERİ	51,11%	0,066215	4,43%	33,96%	8,66%
Rule154	ŞEKER VE ŞEKERLİ MAMÜLLER & SICAK İÇECEKLER -> SOĞUK İÇECEKLER	53,65%	0,066074	5%	37,11%	9,32%
Rule57	KATKI MADDELERİ -> BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER	67,53%	0,065934	12,74%	59,09%	18,87%
Rule338	SÜTLÜK & SOĞUK İÇECEKLER & TEMİZLİK ÜRÜNLERİ & KOZMETİK VE KİŞİSEL BAKIM -> ŞEKER VE ŞEKERLİ MAMÜLLER	56,38%	0,06547	2,26%	25,38%	4%
Rule148	TEMİZLİK ÜRÜNLERİ & KÜMES HAYVANLARI -> SOĞUK İÇECEKLER	54,16%	0,065443	4,85%	37,11%	8,96%
Rule268	BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER & UNLU MAMÜLLER & KURU GIDALAR -> TEMİZLİK ÜRÜNLERİ	50,86%	0,065266	4,40%	33,96%	8,66%
Rule99	KOZMETİK VE KİŞİSEL BAKIM -> MANDIRA	56,09%	0,065234	11,17%	47,91%	19,91%
Rule244	KURU GIDALAR & KONSERVE GIDALAR -> TEMİZLİK ÜRÜNLERİ	58,10%	0,065046	3,51%	33,96%	6,04%
Rule242	KURU GIDALAR & KATKI MADDELERİ -> TEMİZLİK ÜRÜNLERİ	55,52%	0,06485	3,74%	33,96%	6,74%
Rule193	SÜTLÜK & MANDIRA & ZÜCCACIYE/HEDİYELİK EŞYA/MUHTELİF EV EŞYALARI -> SOĞUK İÇECEKLER	52,78%	0,064176	5,04%	37,11%	9,55%
Rule346	SÜTLÜK & BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER & MANDIRA & SOĞUK İÇECEKLER & KÜMES HAYVANLARI -> ŞEKER VE ŞEKERLİ MAMÜLLER	51,13%	0,063924	2,40%	25,38%	4,70%

Tablo A.1: Devamı

Rules		Confidence	KI	Rule Support	Consequent Support	Antecedent Support
Rule311	SÜTLÜK & KURU GIDALAR & ZÜCCACİYE/HEDİYELİK EŞYA/MUHTELİF EV EŞYALARI -> ŞEKER VE ŞEKERLİ MAMÜLLER	50,22%	0,063889	2,45%	25,38%	4,87%
Rule339	SÜTLÜK & SOĞUK İÇECEKLER & TEMİZLİK ÜRÜNLERİ & ZÜCCACİYE/HEDİYELİK EŞYA/MUHTELİF EV EŞYALARI -> ŞEKER VE ŞEKERLİ MAMÜLLER	51,63%	0,063388	2,36%	25,38%	4,57%
Rule275	SÜTLÜK & MANAV & BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER & KÜMES HAYVANLARI -> TEMİZLİK ÜRÜNLERİ	50,12%	0,063029	4,38%	33,96%	8,74%
Rule258	SÜTLÜK & UNLU MAMULLER & KONSERVE GIDALAR -> TEMİZLİK ÜRÜNLERİ	51,89%	0,062955	4,09%	33,96%	7,87%
Rule341	MANAV & UNLU MAMULLER & TEMİZLİK ÜRÜNLERİ & SICAK İÇECEKLER -> ŞEKER VE ŞEKERLİ MAMÜLLER	50,22%	0,062783	2,40%	25,38%	4,79%
Rule127	BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER & SİGARA -> SOĞUK İÇECEKLER	51,24%	0,062364	5,28%	37,11%	10,30%
Rule152	TEMİZLİK ÜRÜNLERİ & KONSERVE GIDALAR -> SOĞUK İÇECEKLER	56,53%	0,062346	4,23%	37,11%	7,49%
Rule24	BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER -> MANAV	65,43%	0,061603	38,66%	63%	59,09%
Rule105	YAĞLAR -> MANDIRA	67,20%	0,061176	5,32%	47,91%	7,91%
Rule334	SÜTLÜK & UNLU MAMULLER & TEMİZLİK ÜRÜNLERİ & KOZMETİK VE KİŞİSEL BAKIM -> ŞEKER VE ŞEKERLİ MAMÜLLER	50,23%	0,061124	2,34%	25,38%	4,66%
Rule279	SÜTLÜK & BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER & MANDIRA & KÜMES HAYVANLARI -> TEMİZLİK ÜRÜNLERİ	50,79%	0,060378	4,09%	33,96%	8,04%
Rule38	KIRMIZI ET -> MANAV	76,54%	0,060078	7,91%	63%	10,34%
Rule331	SÜTLÜK & BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER & MANDIRA & SİGARA -> ŞEKER VE ŞEKERLİ MAMÜLLER	50%	0,059733	2,30%	25,38%	4,60%
Rule150	TEMİZLİK ÜRÜNLERİ & ZÜCCACİYE/HEDİYELİK EŞYA/MUHTELİF EV EŞYALARI -> SOĞUK İÇECEKLER	50,51%	0,059526	5,23%	37,11%	10,36%
Rule135	UNLU MAMULLER & KONSERVE GIDALAR -> SOĞUK İÇECEKLER	51,55%	0,059514	4,96%	37,11%	9,62%
Rule46	MANAV -> BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER	61,36%	0,059399	38,66%	59,09%	63%

Tablo A.1: Devamı

Rules		Confidence	KI	Rule Support	Consequent Support	Antecedent Support
Rule230	ŞEKER VE ŞEKERLİ MAMÜLLER & KÜMES HAYVANLARI -> TEMİZLİK ÜRÜNLERİ	51,72%	0,058658	3,83%	33,96%	7,40%
Rule30	SICAK İÇECEKLER -> MANAV	69,01%	0,058563	15,68%	63%	22,72%
Rule269	BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER & UNLU MAMÜLLER & KATKI MADDELERİ -> TEMİZLİK ÜRÜNLERİ	51,12%	0,058283	3,89%	33,96%	7,62%
Rule79	KATKI MADDELERİ -> UNLU MAMÜLLER	56,82%	0,058092	10,72%	49,13%	18,87%
Rule310	SÜTLÜK & KURU GIDALAR & KOZMETİK VE KİŞİSEL BAKIM -> ŞEKER VE ŞEKERLİ MAMÜLLER	50%	0,058073	2,23%	25,38%	4,47%
Rule169	KÜMES HAYVANLARI & KURU GIDALAR -> SOĞUK İÇECEKLER	58,14%	0,057719	3,72%	37,11%	6,40%
Rule278	SÜTLÜK & MANAV & SOĞUK İÇECEKLER & KÜMES HAYVANLARI -> TEMİZLİK ÜRÜNLERİ	53,55%	0,057618	3,53%	33,96%	6,60%
Rule316	BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER & KÜMES HAYVANLARI & KATKI MADDELERİ -> ŞEKER VE ŞEKERLİ MAMÜLLER	50,49%	0,057537	2,19%	25,38%	4,34%
Rule163	SICAK İÇECEKLER & KURU GIDALAR -> SOĞUK İÇECEKLER	58,19%	0,057484	3,70%	37,11%	6,36%
Rule156	ŞEKER VE ŞEKERLİ MAMÜLLER & KURU GIDALAR -> SOĞUK İÇECEKLER	54,78%	0,057346	4,15%	37,11%	7,57%
Rule155	ŞEKER VE ŞEKERLİ MAMÜLLER & KÜMES HAYVANLARI -> SOĞUK İÇECEKLER	55,17%	0,057318	4,09%	37,11%	7,40%
Rule325	MANDIRA & SICAK İÇECEKLER & KÜMES HAYVANLARI -> ŞEKER VE ŞEKERLİ MAMÜLLER	51,27%	0,057287	2,15%	25,38%	4,19%
Rule190	SÜTLÜK & UNLU MAMÜLLER & KATKI MADDELERİ -> SOĞUK İÇECEKLER	52,16%	0,057106	4,62%	37,11%	8,85%
Rule343	BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER & MANDIRA & SOĞUK İÇECEKLER & KOZMETİK VE KİŞİSEL BAKIM -> ŞEKER VE ŞEKERLİ MAMÜLLER	51,28%	0,056734	2,13%	25,38%	4,15%
Rule11	KOZMETİK VE KİŞİSEL BAKIM -> SÜTLÜK	72,12%	0,056506	14,36%	65,72%	19,91%
Rule204	MANAV & MANDIRA & ZÜCCACIYE/HEDİYELİK EŞYA/MUHTELİF EV EŞYALARI -> SOĞUK İÇECEKLER	52,49%	0,056372	4,49%	37,11%	8,55%
Rule198	MANAV & UNLU MAMÜLLER & SICAK İÇECEKLER -> SOĞUK İÇECEKLER	50,78%	0,056206	4,87%	37,11%	9,60%

Tablo A.1: Devamı

Rules		Confidence	KI	Rule Support	Consequent Support	Antecedent Support
Rule192	SÜTLÜK & MANDIRA & KOZMETİK VE KİŞİSEL BAKIM -> SOĞUK İÇECEKLER	51,26%	0,056148	4,74%	37,11%	9,26%
Rule161	ŞEKER VE ŞEKERLİ MAMÜLLER & KONSERVE GIDALAR -> SOĞUK İÇECEKLER	56,69%	0,056056	3,79%	37,11%	6,68%
Rule201	MANAV & UNLU MAMULLER & KATKI MADDELERİ -> SOĞUK İÇECEKLER	52,99%	0,055741	4,34%	37,11%	8,19%
Rule81	KONSERVE GIDALAR -> UNLU MAMULLER	57,43%	0,055647	9,62%	49,13%	16,74%
Rule72	TEMİZLİK ÜRÜNLERİ -> UNLU MAMULLER	53,20%	0,05527	18,06%	49,13%	33,96%
Rule29	ŞEKER VE ŞEKERLİ MAMÜLLER -> MANAV	68,06%	0,05514	17,28%	63%	25,38%
Rule239	KÜMES HAYVANLARI & KURU GIDALAR -> TEMİZLİK ÜRÜNLERİ	53,16%	0,054825	3,40%	33,96%	6,40%
Rule280	SÜTLÜK & BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER & SOĞUK İÇECEKLER & KÜMES HAYVANLARI -> TEMİZLİK ÜRÜNLERİ	51,70%	0,054379	3,55%	33,96%	6,87%
Rule53	KÜMES HAYVANLARI -> BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER	64,83%	0,054222	14,79%	59,09%	22,81%
Rule273	MANDIRA & SOĞUK İÇECEKLER & KÜMES HAYVANLARI -> TEMİZLİK ÜRÜNLERİ	51,71%	0,054075	3,53%	33,96%	6,83%
Rule200	MANAV & UNLU MAMULLER & ZÜCCACİYE/HEDİYELİK EŞYA/MUHTELİF EV EŞYALARI -> SOĞUK İÇECEKLER	53,44%	0,054067	4,13%	37,11%	7,72%
Rule245	KATKI MADDELERİ & KONSERVE GIDALAR -> TEMİZLİK ÜRÜNLERİ	55,26%	0,053768	3,13%	33,96%	5,66%
Rule189	SÜTLÜK & UNLU MAMULLER & ZÜCCACİYE/HEDİYELİK EŞYA/MUHTELİF EV EŞYALARI -> SOĞUK İÇECEKLER	52,03%	0,053608	4,36%	37,11%	8,38%
Rule283	MANAV & BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER & SOĞUK İÇECEKLER & KÜMES HAYVANLARI -> TEMİZLİK ÜRÜNLERİ	53,90%	0,053357	3,23%	33,96%	6%
Rule237	SICAK İÇECEKLER & KATKI MADDELERİ -> TEMİZLİK ÜRÜNLERİ	54,28%	0,051852	3,11%	33,96%	5,72%
Rule179	ZÜCCACİYE/HEDİYELİK EŞYA/MUHTELİF EV EŞYALARI & KATKI MADDELERİ -> SOĞUK İÇECEKLER	58,82%	0,050488	3,19%	37,11%	5,43%
Rule171	KÜMES HAYVANLARI & KATKI MADDELERİ -> SOĞUK İÇECEKLER	56,74%	0,050471	3,40%	37,11%	6%

Tablo A.1: Devamı

Rules		Confidence	KI	Rule Support	Consequent Support	Antecedent Support
Rule158	ŞEKER VE ŞEKERLİ MAMÜLLER & ZÜCCACİYE/HEDİYELİK EŞYA/MUHTELİF EV EŞYALARI -> SOĞUK İÇECEKLER	54,81%	0,050351	3,64%	37,11%	6,64%
Rule188	SÜTLÜK & UNLU MAMULLER & KOZMETİK VE KİŞİSEL BAKIM -> SOĞUK İÇECEKLER	50,88%	0,049858	4,30%	37,11%	8,45%
Rule16	KIRMIZI ET -> SÜTLÜK	76,54%	0,049664	7,91%	65,72%	10,34%
Rule82	KIRMIZI ET -> UNLU MAMULLER	61,11%	0,049581	6,32%	49,13%	10,34%
Rule203	MANAV & MANDIRA & KOZMETİK VE KİŞİSEL BAKIM -> SOĞUK İÇECEKLER	50,75%	0,04952	4,30%	37,11%	8,47%
Rule166	SICAK İÇECEKLER & KATKI MADDELERİ -> SOĞUK İÇECEKLER	57,25%	0,049399	3,28%	37,11%	5,72%
Rule176	KURU GIDALAR & KONSERVE GIDALAR -> SOĞUK İÇECEKLER	55,99%	0,048883	3,38%	37,11%	6,04%
Rule238	SICAK İÇECEKLER & KONSERVE GIDALAR -> TEMİZLİK ÜRÜNLERİ	55,14%	0,048844	2,85%	33,96%	5,17%
Rule18	YAĞLAR -> SÜTLÜK	79,57%	0,048648	6,30%	65,72%	7,91%
Rule168	SICAK İÇECEKLER & KONSERVE GIDALAR -> SOĞUK İÇECEKLER	58,85%	0,048166	3,04%	37,11%	5,17%
Rule33	KOZMETİK VE KİŞİSEL BAKIM -> MANAV	68,59%	0,047756	13,66%	63%	19,91%
Rule240	KÜMES HAYVANLARI & KATKI MADDELERİ -> TEMİZLİK ÜRÜNLERİ	51,77%	0,047664	3,11%	33,96%	6%
Rule42	BAKLİYATLAR -> MANAV	80,33%	0,047464	5,13%	63%	6,38%
Rule191	SÜTLÜK & UNLU MAMULLER & SİGARA -> SOĞUK İÇECEKLER	52,02%	0,047054	3,83%	37,11%	7,36%
Rule199	MANAV & UNLU MAMULLER & KOZMETİK VE KİŞİSEL BAKIM -> SOĞUK İÇECEKLER	51,39%	0,046876	3,94%	37,11%	7,66%
Rule144	MANDIRA & KIRMIZI ET -> SOĞUK İÇECEKLER	53,31%	0,046836	3,60%	37,11%	6,74%
Rule210	UNLU MAMULLER & MANDIRA & ZÜCCACİYE/HEDİYELİK EŞYA/MUHTELİF EV EŞYALARI -> SOĞUK İÇECEKLER	52,74%	0,046762	3,68%	37,11%	6,98%
Rule56	ZÜCCACİYE/HEDİYELİK EŞYA/MUHTELİF EV EŞYALARI -> BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER	64,79%	0,046534	12,77%	59,09%	19,70%
Rule181	KATKI MADDELERİ & KONSERVE GIDALAR -> SOĞUK İÇECEKLER	56,02%	0,045855	3,17%	37,11%	5,66%
Rule110	YAĞLAR -> SOĞUK İÇECEKLER	50,54%	0,045552	4%	37,11%	7,91%

Tablo A.1: Devamı

Rules		Confidence	KI	Rule Support	Consequent Support	Antecedent Support
Rule175	KURU GIDALAR & KATKI MADDELERİ -> SOĞUK İÇECEKLER	52,68%	0,045013	3,55%	37,11%	6,74%
Rule19	KURUYEMİŞLER -> SÜTLÜK	78,59%	0,044844	6,17%	65,72%	7,85%
Rule194	SÜTLÜK & MANDIRA & SİGARA -> SOĞUK İÇECEKLER	51,63%	0,044629	3,70%	37,11%	7,17%
Rule129	BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER & KIRMIZI ET -> SOĞUK İÇECEKLER	53,31%	0,044617	3,43%	37,11%	6,43%
Rule172	KÜMES HAYVANLARI & KONSERVE GIDALAR -> SOĞUK İÇECEKLER	56,57%	0,044548	3,02%	37,11%	5,34%
Rule235	SICAK İÇECEKLER & KÜMES HAYVANLARI -> TEMİZLİK ÜRÜNLERİ	50,53%	0,044496	3,04%	33,96%	6,02%
Rule20	BAKLİYATLAR -> SÜTLÜK	81,33%	0,044229	5,19%	65,72%	6,38%
Rule162	SICAK İÇECEKLER & KÜMES HAYVANLARI -> SOĞUK İÇECEKLER	54,06%	0,043751	3,26%	37,11%	6,02%
Rule241	KÜMES HAYVANLARI & KONSERVE GIDALAR -> TEMİZLİK ÜRÜNLERİ	52,19%	0,04342	2,79%	33,96%	5,34%
Rule80	SİGARA -> UNLU MAMULLER	55,25%	0,043189	9,74%	49,13%	17,64%
Rule157	ŞEKER VE ŞEKERLİ MAMÜLLER & KOZMETİK VE KİŞİSEL BAKIM -> SOĞUK İÇECEKLER	53,47%	0,042971	3,28%	37,11%	6,13%
Rule174	KURU GIDALAR & ZÜCCACİYE/HEDİYELİK EŞYA/MUHTELİF EV EŞYALARI -> SOĞUK İÇECEKLER	55%	0,042415	3,04%	37,11%	5,53%
Rule228	SOĞUK İÇECEKLER & KIRMIZI ET -> TEMİZLİK ÜRÜNLERİ	53,51%	0,042292	2,60%	33,96%	4,85%
Rule59	KONSERVE GIDALAR -> BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER	65,18%	0,042246	10,91%	59,09%	16,74%
Rule153	TEMİZLİK ÜRÜNLERİ & KIRMIZI ET -> SOĞUK İÇECEKLER	59,80%	0,042214	2,60%	37,11%	4,34%
Rule107	BAKLİYATLAR -> MANDIRA	64,33%	0,041992	4,11%	47,91%	6,38%
Rule130	BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER & KURUYEMİŞLER -> SOĞUK İÇECEKLER	54,94%	0,041136	2,96%	37,11%	5,38%
Rule165	SICAK İÇECEKLER & ZÜCCACİYE/HEDİYELİK EŞYA/MUHTELİF EV EŞYALARI -> SOĞUK İÇECEKLER	53,45%	0,040987	3,13%	37,11%	5,85%
Rule261	SÜTLÜK & MANDIRA & KIRMIZI ET -> TEMİZLİK ÜRÜNLERİ	50,19%	0,040808	2,83%	33,96%	5,64%
Rule170	KÜMES HAYVANLARI & KOZMETİK VE KİŞİSEL BAKIM -> SOĞUK İÇECEKLER	55,02%	0,040666	2,91%	37,11%	5,30%

Tablo A.1: Devamı

Rules		Confidence	KI	Rule Support	Consequent Support	Antecedent Support
Rule286	MANAV & BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER & UNLU MAMULLER & MANDIRA & KÜMES HAYVANLARI -> TEMİZLİK ÜRÜNLERİ	51,90%	0,040341	2,62%	33,96%	5,04%
Rule285	BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER & UNLU MAMULLER & SOĞUK İÇECEKLER & KÜMES HAYVANLARI -> TEMİZLİK ÜRÜNLERİ	51%	0,04027	2,70%	33,96%	5,30%
Rule77	KOZMETİK VE KİŞİSEL BAKIM -> UNLU MAMULLER	54,17%	0,040153	10,79%	49,13%	19,91%
Rule205	MANAV & MANDIRA & SİGARA -> SOĞUK İÇECEKLER	51,48%	0,039955	3,34%	37,11%	6,49%
Rule136	UNLU MAMULLER & KIRMIZI ET -> SOĞUK İÇECEKLER	51,85%	0,039927	3,28%	37,11%	6,32%
Rule83	SU ÜRÜNLERİ -> UNLU MAMULLER	61,60%	0,039818	4,91%	49,13%	7,98%
Rule209	UNLU MAMULLER & MANDIRA & KOZMETİK VE KİŞİSEL BAKIM -> SOĞUK İÇECEKLER	50,30%	0,039703	3,53%	37,11%	7,02%
Rule160	ŞEKER VE ŞEKERLİ MAMÜLLER & SİGARA -> SOĞUK İÇECEKLER	55%	0,039152	2,81%	37,11%	5,11%
Rule284	MANAV & UNLU MAMULLER & SOĞUK İÇECEKLER & KÜMES HAYVANLARI -> TEMİZLİK ÜRÜNLERİ	51,05%	0,038443	2,57%	33,96%	5,04%
Rule202	MANAV & UNLU MAMULLER & SİGARA -> SOĞUK İÇECEKLER	50,15%	0,038424	3,45%	37,11%	6,87%
Rule34	ZÜCCACIYE/HEDİYELİK EŞYA/MUHTELİF EV EŞYALARI -> MANAV	67,49%	0,037989	13,30%	63%	19,70%
Rule164	SICAK İÇECEKLER & KOZMETİK VE KİŞİSEL BAKIM -> SOĞUK İÇECEKLER	51,77%	0,037708	3,11%	37,11%	6%
Rule282	MANAV & BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER & UNLU MAMULLER & KONSERVE GIDALAR -> TEMİZLİK ÜRÜNLERİ	50,64%	0,03719	2,53%	33,96%	5%
Rule106	KURUYEMİŞLER -> MANDIRA	59,62%	0,036825	4,68%	47,91%	7,85%
Rule118	SÜTLÜK & KURUYEMİŞLER -> SOĞUK İÇECEKLER	51,03%	0,036825	3,15%	37,11%	6,17%
Rule233	ŞEKER VE ŞEKERLİ MAMÜLLER & SİGARA -> TEMİZLİK ÜRÜNLERİ	50%	0,036528	2,55%	33,96%	5,11%
Rule39	SU ÜRÜNLERİ -> MANAV	73,60%	0,036283	5,87%	63%	7,98%
Rule272	UNLU MAMULLER & MANDIRA & KIRMIZI ET -> TEMİZLİK ÜRÜNLERİ	51,90%	0,035757	2,32%	33,96%	4,47%

Tablo A.1: Devamı

Rules		Confidence	KI	Rule Support	Consequent Support	Antecedent Support
Rule177	KOZMETİK VE KİŞİSEL BAKIM & KATKI MADDELERİ -> SOĞUK İÇECEKLER	55,71%	0,035626	2,49%	37,11%	4,47%
Rule259	SÜTLÜK & UNLU MAMULLER & KIRMIZI ET -> TEMİZLİK ÜRÜNLERİ	50%	0,035615	2,49%	33,96%	4,98%
Rule55	KOZMETİK VE KİŞİSEL BAKIM -> BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER	63,35%	0,035173	12,62%	59,09%	19,91%
Rule180	ZÜCCACIYE/HEDİYELİK EŞYA/MUHTELİF EV EŞYALARI & KONSERVE GIDALAR -> SOĞUK İÇECEKLER	55,24%	0,034714	2,47%	37,11%	4,47%
Rule222	MANDIRA & BAKLIYATLAR -> TEMİZLİK ÜRÜNLERİ	52,85%	0,034593	2,17%	33,96%	4,11%
Rule211	UNLU MAMULLER & MANDIRA & SİGARA -> SOĞUK İÇECEKLER	51,53%	0,034445	2,87%	37,11%	5,57%
Rule78	ZÜCCACIYE/HEDİYELİK EŞYA/MUHTELİF EV EŞYALARI -> UNLU MAMULLER	53,46%	0,034119	10,53%	49,13%	19,70%
Rule243	KURU GIDALAR & SİGARA -> TEMİZLİK ÜRÜNLERİ	52,91%	0,033984	2,13%	33,96%	4,02%
Rule178	KOZMETİK VE KİŞİSEL BAKIM & KONSERVE GIDALAR -> SOĞUK İÇECEKLER	56,99%	0,033716	2,26%	37,11%	3,96%
Rule167	SICAK İÇECEKLER & SİGARA -> SOĞUK İÇECEKLER	52,86%	0,03261	2,55%	37,11%	4,83%
Rule41	KURUYEMİŞLER -> MANAV	72,63%	0,032431	5,70%	63%	7,85%
Rule173	KURU GIDALAR & KOZMETİK VE KİŞİSEL BAKIM -> SOĞUK İÇECEKLER	51,42%	0,032226	2,70%	37,11%	5,26%
Rule216	BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER & BAKLIYATLAR -> TEMİZLİK ÜRÜNLERİ	50,24%	0,031676	2,19%	33,96%	4,36%
Rule40	YAĞLAR -> MANAV	72,31%	0,031618	5,72%	63%	7,91%
Rule63	KURUYEMİŞLER -> BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER	68,56%	0,030783	5,38%	59,09%	7,85%
Rule249	SÜTLÜK & MANAV & BAKLIYATLAR -> TEMİZLİK ÜRÜNLERİ	50%	0,03044	2,13%	33,96%	4,26%
Rule145	MANDIRA & KURUYEMİŞLER -> SOĞUK İÇECEKLER	52,27%	0,030419	2,45%	37,11%	4,68%
Rule85	KURUYEMİŞLER -> UNLU MAMULLER	58,81%	0,030408	4,62%	49,13%	7,85%
Rule87	BAHARATLAR -> UNLU MAMULLER	63,76%	0,028517	3,11%	49,13%	4,87%
Rule108	BAHARATLAR -> MANDIRA	62,45%	0,028368	3,04%	47,91%	4,87%
Rule131	BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER & BAKLIYATLAR -> SOĞUK İÇECEKLER	52,20%	0,0282	2,28%	37,11%	4,36%

Tablo A.1: Devamı

Rules		Confidence	KI	Rule Support	Consequent Support	Antecedent Support
Rule195	SÜTLÜK & KOZMETİK VE KİŞİSEL BAKIM & ZÜCCACİYE/HEDİYELİK EŞYA/MUHTELİF EV EŞYALARI -> SOĞUK İÇECEKLER	50,92%	0,027449	2,36%	37,11%	4,64%
Rule206	MANAV & TEMİZLİK ÜRÜNLERİ & SİGARA -> SOĞUK İÇECEKLER	51,46%	0,02695	2,26%	37,11%	4,38%
Rule86	BAKLİYATLAR -> UNLU MAMULLER	59,67%	0,026916	3,81%	49,13%	6,38%
Rule21	BAHARATLAR -> SÜTLÜK	78,17%	0,026911	3,81%	65,72%	4,87%
Rule111	DONDURULMUŞ GIDALAR -> SOĞUK İÇECEKLER	52,36%	0,026555	2,13%	37,11%	4,06%
Rule207	MANAV & KOZMETİK VE KİŞİSEL BAKIM & ZÜCCACİYE/HEDİYELİK EŞYA/MUHTELİF EV EŞYALARI -> SOĞUK İÇECEKLER	51,23%	0,026142	2,21%	37,11%	4,32%
Rule104	SU ÜRÜNLERİ -> MANDIRA	56%	0,025849	4,47%	47,91%	7,98%
Rule84	YAĞLAR -> UNLU MAMULLER	57,26%	0,025748	4,53%	49,13%	7,91%
Rule185	SÜTLÜK & MANAV & BAKLİYATLAR -> SOĞUK İÇECEKLER	50,50%	0,024422	2,15%	37,11%	4,26%
Rule64	BAKLİYATLAR -> BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER	68,33%	0,024419	4,36%	59,09%	6,38%
Rule109	DONDURULMUŞ GIDALAR -> MANDIRA	62,30%	0,02343	2,53%	47,91%	4,06%
Rule62	YAĞLAR -> BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER	65,59%	0,021302	5,19%	59,09%	7,91%
Rule43	BAHARATLAR -> MANAV	72,93%	0,020747	3,55%	63%	4,87%
Rule65	BAHARATLAR -> BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER	68,56%	0,019094	3,34%	59,09%	4,87%
Rule66	DONDURULMUŞ GIDALAR -> BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER	70,16%	0,018612	2,85%	59,09%	4,06%
Rule22	DONDURULMUŞ GIDALAR -> SÜTLÜK	75,39%	0,017443	3,06%	65,72%	4,06%
Rule36	SİGARA -> MANAV	65,26%	0,017096	11,51%	63%	17,64%
Rule14	SİGARA -> SÜTLÜK	67,67%	0,015256	11,94%	65,72%	17,64%
Rule44	DONDURULMUŞ GIDALAR -> MANAV	71,73%	0,015216	2,91%	63%	4,06%
Rule61	SU ÜRÜNLERİ -> BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER	63,20%	0,013581	5,04%	59,09%	7,98%
Rule60	KIRMIZI ET -> BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER	62,14%	0,013067	6,43%	59,09%	10,34%
Rule17	SU ÜRÜNLERİ -> SÜTLÜK	69,07%	0,011841	5,51%	65,72%	7,98%

Tablo A.1: Devamı

Rules		Confidence	KI	Rule Support	Consequent Support	Antecedent Support
Rule88	DONDURULMUŞ GIDALAR -> UNLU MAMULLER	54,97%	0,009506	2,23%	49,13%	4,06%
Rule58	SİGARA -> BİSKÜVİ/KEK/ÇİKOLATA/ŞEKER	58,38%	-0,00512	10,30%	59,09%	17,64%

ÖZGEÇMİŞ

Mustafa Aykut GÖRAL, 30/08/1979 tarihinde Ankara’da doğmuştur. Orta ve lise eğitimini TED Ankara Koleji’nde tamamladıktan sonra 1998 yılında İstanbul Üniversitesi Makina Mühendisliği Bölümü’nde lisans eğitimine başlamış ve 2002 yılında mezun olmuştur. Aynı yıl içinde İstanbul Teknik Üniversitesi Endüstri Mühendisliği Yüksek Lisans Programı’na başlamıştır. 2004 yılında iş hayatına giriş yapmış ve 2006 yılına kadar Kratis Bilgi Teknolojileri A.Ş.’de Satış Mühendisi olarak çalışmıştır. Çalıştığı süre içerisinde Veri Madenciliği konusunda çeşitli projelerde yer almıştır.