VERİ MADENCİLİĞİNDE CHAID ALGORİTMASININ SOSYAL GÜVENLİK KURUMU VERİ TABANINA UYGULANMASI

BETÜL DEMİREL

YÜKSEK LİSANS TEZİ
İSTATİSTİK

GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

MAYIS 2010
ANKARA

Betül DEMİREL tarafından hazırlanan "VERİ MADENCİLİĞİNDE CHAID ALGORİTMASININ SOSYAL GÜVENLİK KURUMU VERİ TABANINA UYGULANMASI" adlı bu tezin Yüksek Lisans tezi olarak uygun olduğunu onaylarım.

Yrd. Doç. Dr. Necla GÜNDÜZ TEKİN ……..
Tez Danışmanı, İstatistik Anabilim Dalı

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

Prof. Dr. Cevriye GENCER ……..
Endüstri Mühendisliği, Gazi Üniversitesi
Yrd. Doç. Dr. Necla GÜNDÜZ TEKİN ……..
İstatistik, Gazi Üniversitesi
Yrd. Doç. Dr. Jale BALIBEYOĞLU ……..
İstatistik, Gazi Üniversitesi

Tarih: 24 / 05 / 2010

Bu tez ile G.Ü. Fen Bilimleri Enstitüsü Yönetim Kurulu Yüksek Lisans derecesini onamıştır.

Prof. Dr. Bilal TOKLU
……..……
Fen Bilimleri Enstitüsü Müdürü

TEZ BİLDİRİMİ

Tez içindeki bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edilerek sunulduğunu, ayrıca tez yazım kurallarına uygun olarak hazırlanan bu çalışmada bana ait olmayan her türlü ifade ve bilginin kaynağına eksiksiz atıf yapıldığını bildiririm.

Betül DEMİREL

VERİ MADENCİLİĞİNDE CHAID ALGORİTMASININ SOSYAL GÜVENLİK KURUMU VERİ TABANINA UYGULANMASI
(Yüksek Lisans Tezi)

Betül DEMİREL

GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
Mayıs 2010

ÖZET

"Veri Madenciliğinde CHAID Algoritmasının Sosyal Güvenlik Kurumu Veri Tabanına Uygulanması" isimli çalışmamızda, öncelikle verinin tanımından başlanılarak, veri tabanı, veri tabanı yönetim sistemleri, veri madenciliğinin uygulanabilmesi için gerekli olan veri ambarı mimarisi, verilerin ön işlemlerden geçirilmesi ve veri madenciliği tekniklerinden bahsedilerek, sınıflandırmada karar ağaçları yönteminde kullanılan CHAID (Otomatik Ki-Kare Etkileşim Belirleyicisi) algoritması uygulaması ile ülkemizde Sosyal Güvenlik Kurumu ile ilişkilendirilmiş işyerlerinde yapılandırma kanunundan yararlanan işyerlerinin profilleri belirlenmeye çalışılmış ve daha sonra çıkarılacak olan kanunların bu yapılan çalışmada öngörülen bilgiler ışığında karar destek amaçlı olarak katkıda bulunması amaçlanmıştır.

Bilim Kodu : **205.1.066**
Anahtar Kelimeler : **Veri madenciliği, sınıflandırma, karar ağaçları, CHAID**
Sayfa Adedi : **129**
Tez Yöneticisi : **Yrd. Doç. Dr. Necla GÜNDÜZ TEKİN**

DATA MINING CHAID ALGORITHM SOCIAL SECURITY ADMINISTRATION DATABASE IMPLEMENTATION

(M.Sc. Thesis)

Betül DEMİREL

GAZİ UNIVERSITY
INSTITUTE OF SCIENCE AND TECHNOLOGY
May 2010

ABSTRACT

"Data Mining CHAID Algorithm Social Security Administration Database Implementation" named in our study, first data definition was started, the data base, data base management systems, data mining for the implementation of the necessary data warehouse architecture, data pre-processing of the spent and data mining techniques mentioned by the classification decision trees in the method used CHAID (Chi-Squared Automatic Interaction Detector) algorithm implementation in our country with the Social Security Administration has been associated with workplace configuration laws that benefit from workplace profiles was determine and then to be issued the law that the study provided information in the light of the decision to support the contribution is intended.

Science Code : **205.1.066**
Key Words : **Data mining, classification, decision trees, CHAID**
Page Number : **129**
Adviser : **Assist. Prof. Dr. Necla GÜNDÜZ TEKİN**

TEŞEKKÜR

Çalışmalarım boyunca değerli yardım ve katkılarıyla beni yönlendiren Hocam Yrd. Doç. Dr. Necla GÜNDÜZ TEKİN'e, "Sosyal Güvenlik Kurumu Aktüerya ve Fon Yönetimi Daire Başkanlığı" biriminde görev yapan Dr.Rasim KÖSELERLİ'ye, manevi destekleriyle beni hiçbir zaman yalnız bırakmayan çok değerli ablam Birsemin ALTINKAYNAK JURGENS ve sevgili eşim Suat DEMİREL ile en değerli varlığım oğlum Batuhan DEMİREL'e teşekkürü bir borç bilirim.

Sevgili Annem'e ve Babam'a

İÇİNDEKİLER

Sayfa

Sayfa

Sayfa

ÇİZELGELERİN LİSTESİ

Çizelge **Sayfa**

Çizelge **Sayfa**

ŞEKİLLERİN LİSTESİ

Şekil **Sayfa**

Şekil **Sayfa**

SİMGELER VE KISALTMALAR

Bu çalışmada kullanılmış bazı simgeler ve kısaltmalar, açıklamaları ile birlikte aşağıda sunulmuştur.

Simgeler	**Açıklama**
x	Bağımsız/açıklayıcı değişken
y	Bağımlı/hedef değişken
c	Sütun değişkeninin düzey sayısı
r	Satır değişkeninin düzey sayısı
s.d.	Serbestlik derecesi
α	Anlamlılık düzeyi
G_{ij}	Gözlenen değer
B_{ij}	Beklenen değer
p	Önem istatistiği
χ^2	Ki-kare test istatistiği
B	Bonferroni çarpanı
$n_{.j}$	j.düzeydeki birimlerin sayısı
$n_{i.}$	i.düzeydeki birimlerin sayısı

Kısaltmalar	Açıklama
AI	Artificial Intelligence (Yapay Zekâ)
ABD	Amerika Birleşik Devletleri
BK	Bağ-Kur
CHAID	Chi-Squared Automatic Interaction Detector (Otomatik Ki-kare Etkileşim Belirleyicisi)
ES	Emekli Sandığı
IPC	İdari Para Cezası
OLTP	Online Transaction Processing (Çevrimiçi İşlem İşleme)
OLAP	Online Analytical Processing (Çevrimiçi Analitik İşleme)
RDBMS	Relational Database Management System (İlişkisel Veri Tabanı Yönetim Sistemi)
SGK	Sosyal Güvenlik Kurumu
SSK	Sosyal Sigortalar Kurumu

TÜİK	Türkiye İstatistik Kurumu
VT	Veri Tabanı
VTYS	Veri Tabanı Yönetim Sistemi
VA	Veri Ambarı
VM	Veri Madenciliği
VTŞ	Veri Tabanı Şeması
VTBK	Veri Tabanında Bilgi Keşfi
YSA	Yapay Sinir Ağları

1. GİRİŞ

Dünya var olduğundan beri yaşamla ilgili birçok veriyi bünyesinde barındırmaktadır. İnsanlık tarihinin başlangıcından itibaren insanoğlu, yaşamında önemli bir yere sahip olan nesneleri resmetmekle ilk verileri oluşturmaya başlamıştır. Günümüzde, bilimsel araştırmalar için yapılan kazı çalışmalarında ortaya çıkan veriler halen birçok bilinmezin var olduğunu ortaya koymakta olup, bilim çevreleri gelişmiş teknolojik aletlerle bu bilinmezleri bilgiye çevirme çalışmalarını halen devam ettirmektedirler.

Gerek resim yolu ile gerekse yazı ile ve günümüzde ise gelişmiş elektronik sistemlerle kayıt altına alınan bu veri yığınlarından bilginin çıkartılması bir süreç gerektirmekte ve bu sürecin tamamı "veri tabanlarında bilgi keşfi" olarak adlandırılmaktadır. Bu sürecin en önemli adımı ise "veri madenciliği"dir. Süreç sonunda varolan eldeki veriler ışığı altında bazı bağıntılar, örüntüler[1] veya kurallar elde edilmesiyle, geleceğe yönelik tahminlerin yapılması veya kararların alınması sağlanacaktır.

Bu çalışmada; verinin tanımından başlanarak, bilginin ortaya çıkarılması süreci ve bu sürecin en önemli adımı olan veri madenciliği ile veri madenciliğinde kullanılan teknikler ayrıntılı olarak anlatılmıştır. Yine aynı şekilde, veri madenciliği tekniklerini uygulayabilmek için yapılması gerekli olan verilerin ön işlemden geçirilmesi yöntemlerinden de bahsedilmiştir.

Tezin son kısmında ise, veri madenciliğinde karar ağacı oluşturmak için geliştirilen algoritmalardan biri olan CHAID (Chi-Squared Automatic Interaction Detector/Otomatik Ki-kare Etkileşim Belirleyicisi) algoritması, Sosyal Güvenlik Kurumu (SGK) veri tabanına uygulanmış ve elde edilen sonuçlar ortaya konularak yorumlanmıştır.

---

[1] İlgilenilen varlıkla ilgili gözlenebilir veya ölçülebilir bilgilere verilen isim

2. VERİ TANIMI,VERİ AMBARI, VERİ ÖN İŞLEME, VERİ MADENCİLİĞİ

2.1. Veri Kavramı ve Klasik Dosya Yapıları

Bilgiyi elde etmeye yarayan işlenmemiş ham malzemeye "veri" denilmektedir. Bu veriler işlenerek bilgi elde edilir. Veriyi bilgiye çevirme süreci ise "veri analizi" olarak adlandırılır [27]. Bilgi ise, şimdi bilinen ve gelecek zamanda verilecek olan kararlar için var olan gerçek bir değerdir ve anlamlı biçimde derlenen ve birleştirilen verilerden oluşur [66].

Bilgisayar teknolojisindeki büyük gelişme ile beraber hızlı bir şekilde veri toplanmaya, birikmeye başlamış, kurumların bu verileri saklama ve işleme tekniklerinin yetersiz kalması neticesinde bilgiye ulaşmak da zorlaşmıştır. Bu verilerin kurum ve kuruluşlarda normal günlük işlemlerde kullanılmasının yanı sıra, üst düzey yöneticilerinin karar vermek amacıyla talepleri de dikkate alınırsa, verinin öneminin yanında veriden bilgiye ulaşmanın da ne denli önemli olduğu ortaya çıkmaktadır.

Veri saklama birimlerinde depolanan veri topluluklarına "dosya" (file/kütük) denir. Günümüzde, veri toplulukları bilgisayar içinde sabit disk üzerinde dosyalarda saklanmaktadır [52]. Dosyalar da kendi içlerinde alanlara bölünmüştür. Örnek olarak bir sınıftaki öğrenci listesini ele alalım. Bu liste çok sayıda veri içerebilir. Bu listedeki her bir öğrenci bilgisi bir mantıksal kayıt oluşturur. Her kayıt farklı bilgiler içerebilir: Öğrencinin adı, soyadı, öğrenci numarası, doğduğu yer, baba adı gibi. Öğrenciye ait bu bilgilerin her birine "alan" denilmektedir ve bu alanlara veri girişi ile kayıt oluşur. Her kayıt birbiriyle ilişkili alanlardan oluşur [52].

Sabit disk üzerinde depolanan ve birbirleri ile ilişkilendirilen veri dosyalarına erişim, oluşturulan dosyanın tipine (sıralı/dizinli) bağlıdır. Üç tip dosya vardır [52].

- *Sıralı Dosyalar (Ardışık Dosyalar)*

İçerdiği kayıtlara birinci kayıttan başlayarak sırayla erişim yapmak üzere tasarlanmış dosyalardır. Bu tür dosyaların kayıtlarına ardışık olarak erişilebilmesine karşılık kayıtlar fiziksel olarak ardışık/sıralı olmayabilir. Bu dosyalarda kayıtlara erişim, tüm kayıtların teker teker taranarak istenilen kayda ulaşılması şeklindedir. Eğer bir kayda ulaşmak için taradığımız dosyanın hacmi çok büyük ise bu bize çok fazla zaman kaybettirir ve bu kayda ulaşmak gereksiz yere dosyadaki bütün kayıtları boş yere taramamıza neden olur.

- *İndeksli Dosyalar (Dizinli Dosyalar – İndeks Sıralı Dosyalar)*

Bu dosyalama sisteminde veri dosyasından ayrı olarak bir indeks dosyası oluşturulmaktadır. Kayıtlarda tekilliği sağlayan alan (anahtar alan) üzerinde bir dizin oluşturulur ve dizin dosyası olarak kaydedilir. Herhangi bir kayda ulaşılmak istendiğinde dizin dosyasına gidilerek bu kaydın adresi alınır ve tüm kayıtların bulunduğu diğer dosyada bu adrese karşılık gelen kayda doğrudan erişim yapılır.

- *Hesaba Dayalı Dosyalar*

Doğrudan erişim dosyası olarak bilinen hesaba dayalı dosyalarda, indeksli dosyalar gibi ayrı bir indeksin tutulmasına gerek olmadan, dosyanın herhangi bir kaydına doğrudan doğruya erişebilmek için bir hesaplama algoritması kullanılır.

Bu devasa veri yığınlarının, yukarıda adı geçen klasik dosyalama sistemleri ile kayıt altına alındığı ve verilerin sürekli güncel ve değişken olduğu düşünülürse, dosyalama sistemlerinin bu tür verileri saklamaya yetersiz kaldığı açıktır. Zaman içerisinde kurum ve kuruluşlardaki veri artışı nedeniyle artık bu dosyalardan veriye erişmek birtakım sorunları da beraberinde getirmiştir. Veri artışı ile birlikte birçok dosya oluşmuş ve farklı kullanıcılar tarafından bu dosyalara aynı anda erişim sorunu,

verileri işleme sorunlarını da beraberinde getirmiştir. Bu nedenle klasik dosya sistemlerinden veri tabanlarına geçme gereği ortaya çıkmıştır.

2.2. Veri Tabanı

Kişinin doğumundan hayatının sonuna kadar ki yaşam sürecinde kendisi ile ilgili birçok bilgi ve belge kayıt altına alınmaktadır. Doğum anından itibaren hastanelerde kayıt altına alınan doğum bilgileri, nüfus bilgileri, okul kayıtları, evlilik kayıtları, elde edilen taşınır ve taşınmaz mallara ilişkin kayıtlar, iş hayatındaki çalışma süreleri ile ilgili kayıtlar, banka kayıtları gibi şahsi kayıtlar ile birlikte kişinin yaşamı boyunca günlük hayatta yapmış olduğu faaliyetler ile ilgili marketlerden ya da büyük alışveriş merkezlerinden yapılan alışverişler, yolda yürürken kameraya çekilen görüntüler ve benzeri kayıtların dosyalama sistemleri ile sağlıklı bir şekilde tutulmasının ne denli zor olduğu tek bir kişi bazında düşünülürse, ülkedeki tüm insanların da bu tür kayıtlarının tutulduğu göz önüne alındığında devasa bir veri yığınının oluştuğu görülmektedir.

Klasik dosya sistemlerinin, verilerin saklanması, depolanması ve erişiminde yetersiz kalması nedeniyle veri tabanı kavramı ortaya çıkmıştır. Veri tabanlarında veriler dosyalarda değil tablolarda tutulur.

Birbiriyle ilişkisi olan verilerin tutulduğu, kullanım amacına uygun olarak düzenlenmiş veriler topluluğunun mantıksal ve fiziksel olarak tanımların olduğu bilgi depoları "veri tabanı" olarak tanımlanmaktadır [66]. Örneğin; SGK'da Türkiye genelinde il müdürlüklerinde sürekli olarak bilgi girişi olmakta ve bunlar SGK Başkanlığı Bilgi İşlem Merkezinde toplanarak veri tabanı oluşturmaktadır. SGK'na tabii kişilerin sayısı göz önüne alınacak olursa, günlük veri giriş ve çıkışının ne denli yoğun olduğu kolayca düşünülebilir.

Veri tabanları da işlemsel veri tabanları ve ilişkisel veri tabanları olarak ikiye ayrılmaktadır [66].

- *İşlemsel Veri Tabanları*

İşlemsel veri tabanları, birim bazında süreç ve işlemleri hızlı ve etkin bir şekilde gerçekleştirmek amacıyla tasarlanmaktadırlar.

Bu veri tabanları hareketli ve dinamiktir, yani günlük veri giriş ve çıkışı yapılarak veriler üzerinde sürekli oynamalar olduğu gibi veri tabanı sürekli değişikliğe uğramaktadır. Hastanelerde, okullarda, işletmelerde vb. gibi kullanılan veri tabanları işlemsel veri tabanlarına örnek olarak gösterilebilir.

- *İlişkisel Veri Tabanları*

İlişkisel veri tabanı, ayrı tablolara yerleştirilmiş verilerin belirli alanlarına göre ilişkilendirilerek, ilişkisel modele göre düzenlenen veri tabanlarına verilen isimdir.

İlişkisel veri tabanlarında, tüm veriler tablolar içerisinde saklanır. Tablolar satır ve sütunlardan oluşur. Tablolardaki her satır bir kaydı, her sütun ise bir ismi temsil ettiği gibi bu tablolarda türetilen nitelikler de olabilir. Örneğin, tablo sütununda doğum tarihi varsa buradan yaş değişkeni elde edilerek yeni bir nitelik tabloya sütun olarak eklenebilir.

2.2.1. Veri tabanı şeması

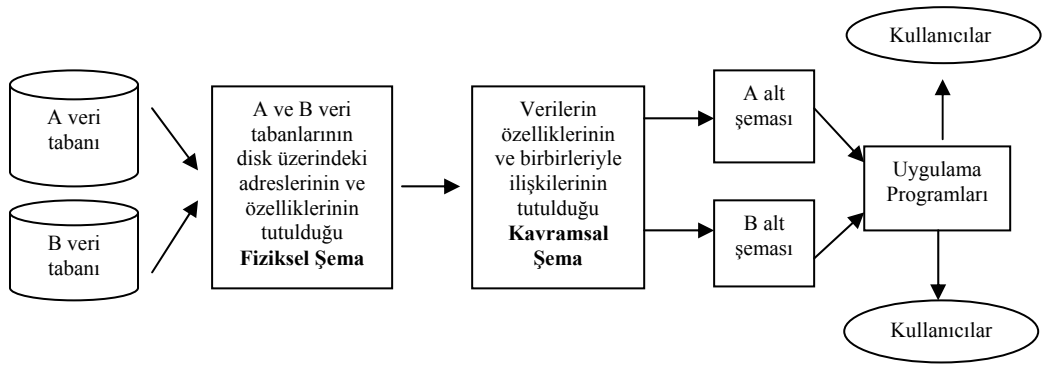
Tablolar ve bu tablolardaki nitelikler veri tabanı şemasını (VTŞ) oluşturur. VTŞ, Şekil 2.1'de görüldüğü üzere iki ana grupta toplanır [52].

- Fiziksel Şema
- Kavramsal Şema

Fiziksel şema; veri tabanının fiziksel çevresi ile ilgili tanımları içerir. Veri tabanının disk üzerindeki adresi ve özellikleri ile ilgili bilgiler fiziksel şemayı oluşturur.

Kavramsal şema; veri tabanındaki tabloların alanları, veri tipleri, veriler arasındaki ilişkiler vb. bilgilerin tutulduğu şemadır.

Veri tabanları, kullanıcıların ilgilendiği alanlara yönelik olarak alt şemalara ayrılır ve ilgilenilen alt şema kullanılarak veri tabanına erişim sağlanır. Uygulama programları, veri tabanının her noktasına erişemez sadece ilgili alt şemaya erişim yapılır. Örneğin, bir kurumun veri tabanında sigortalama–tahsis–sağlık–personel tabloları olsun. Sigortalama uygulayıcıların sadece sigortalama ile ilgili alt şema ile ilgilenmesi, diğer tablolara erişmesi gereksizdir.



Şekil 2.1. Veri tabanı şeması

2.2.2. Veri tabanı yönetim sistemleri

Veri tabanı yönetim sistemleri (VTYS) ile ilgili birçok tanımlama yapılmıştır.

Veri tabanlarını çeşitli yazılım uygulamalarıyla yöneten sisteme "Veri Tabanı Yönetim Sistemi" denilmektedir [51].

Bir diğer tanımlama; "yeni bir veri tabanı oluşturmak, veri tabanını düzenlemek, geliştirmek ve bakımını yapmak gibi çeşitli karmaşık işlemlerin gerçekleştirildiği birden fazla programdan oluşmuş bir yazılım sistemidir" şeklindedir [66].

Y. Özkan'a göre ise VTYS, "veri kümelerinin düzenli biçimde tutulduğu ve bu verinin çeşitli yazılımlar aracılığıyla yönetildiği bir ortam" olarak düşünülebilir [51].

VTYS, kullanıcı ile veri tabanı arasında bir arabirim oluştur ve veri tabanına her türlü erişimi sağlar [66]. VTYS'nde verinin girişi ve depolanması, veriye erişen uygulama programlarından tamamen bağımsızdır. Klasik dosya sistemlerinde ise, dosyaların kayıt desenlerindeki herhangi bir değişiklik bütün sistemi etkiler, uygulama programlarının yeniden düzenlenmesine neden olur [51].

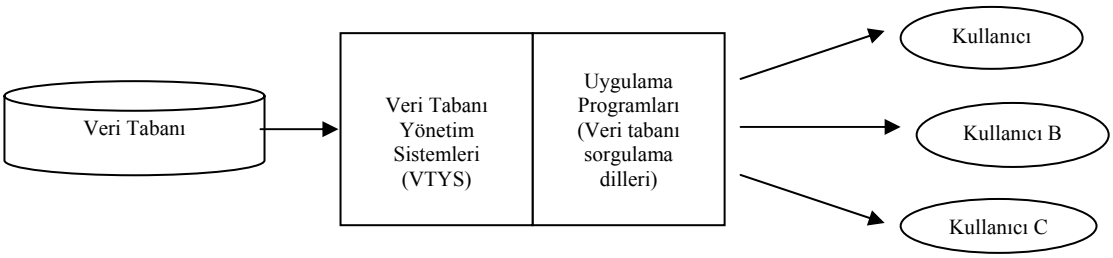
VTYS yazılımları, günümüzde özellikle kamu kurumlarında işletim sistemlerinden sonra en çok kullanılan yazılım olma özelliğine sahiptir.

VTYS ile ilgili yazılımlar; Sybase, Oracle, MySQL, Firebird, PostgreSQL, Berkeley DB, IBM DB2, FileMaker, Microsoft Accsess, Microsoft SQL Server, SQLite, 1C, Microsoft Visual Fox Pro olarak sayılabilir. Veri tabanlarını yöneten yazılımların bir kısmı sadece ilişkisel veri tabanlarını yöneten yazılımlar olma özelliğine sahiptir.

VTYS sorgulamak için de SQL, PL/SQL, TCL gibi veri tabanı dilleri geliştirilmiştir.

Yapısal Sorgulama Dili olan SQL (Structured Query Language–SQL), veri tabanı yönetim sistemlerinin hepsinde çalışabilen ortak bir sorgulama dili olma özelliğine sahiptir.

Veri tabanlarındaki bilgiler uç kullanıcılar tarafından Şekil 2.2.'de gösterildiği üzere, yukarıda sayılan sorgulama dilleri kullanılarak görüntülenebilir, güncellenebilir, silinebilir veya yazdırılabilir. Hatta bu bilgiler kullanılarak yeni bilgiler elde edilebilir [51].



Şekil 2.2. Veri tabanı ile kullanıcı arasındaki ilişki

VTYS'nin sağladığı yararlar arasında en önemlileri aşağıda maddeler halinde özetlenmiştir [51].

- *Verinin tekrarlanmasını önler,*

Klasik dosya sistemlerinin kullanıldığı uygulamalarda, her bir sistem için veriler ayrı ayrı tutulur. Uygulamalar alt sistemlere bölünmüştür. Örneğin; il kodlarının tutulduğu bir tablo, hem pazarlama alt sisteminde hem de personel bilgilerinin tutulduğu bir başka alt sistemde kullanılıyor olabilir. İşte bunun gibi birçok alt uygulamada kullanılan bu tür tablolar verilerin gereksiz yere birçok alt sistemde tekrarlanmasına neden olmaktadır. VTYS'lerinde ise, veri kaynağı tek olarak tasarlanır ve böylece veri tekrarları önlenmiş olur.

- *Verinin tutarlı olmasını sağlar,*

Verinin tutarlı olması, verinin doğruluğunu ifade etmektedir. VTYS'leri veri bütünlüğünü sağlamak için, uygulamalar aracılığıyla birtakım kısıtlamalar getirerek verileri tutarlı olmaya zorlar. Mesela, sigortalı bilgilerinin tutulduğu bir tabloda doğum yeri alanına bir kısıtlama getirilerek bu alana kullanıcı 100 değerini girdiğinde kullanıcının bu veriyi girmesi engellenerek, kullanıcının tutarlı ve düzgün bir veri girişi yapması denetlenmiş olacaktır.

- *Aynı andaki erişimlerde tutarsızlıkların ortaya çıkmasını önler,*

Veri tabanları aynı anda birçok kişi tarafından sorgulandığından, yapılan sorgulamalarda ortaya çıkabilecek sorunları VTYS otomatik olarak çözer. Şöyle ki; bir mamul stokunda 100 adet rulman[2] olsun. İki farklı kullanıcıdan biri 50 adet diğeri ise 55 adet stoktan rulman çıkışı yapmaya kalkarsa VTYS, önce 50 âdetin çıkışını

[2] Rulman ya da yuvarlanma elemanı, rulmanlı yatakların iç ve dış bilezikleri arasında yuvarlanarak en az sürtünme ve kayıpla iş yapmasını sağlayan makine elemanlardır

yapar, elde kalan stok yeterli ise diğer mamulünde çıkışına izin verir, yeterli değilse çıkış yaptırmaz.

- *Verilerin güvenliğini sağlar.*

Veri tabanlarında veriler bir bütün olarak depolanmaktadır [67]. VTYS, veri tabanındaki verilere ulaşımın kullanıcılara veri tabanı üzerinde verilen yetkilerle sınırlanmasını sağlayarak, her kullanıcının veri tabanındaki bütün bilgileri görmesini engeller.

2.2.3. Veri modelleri

VTYS'nin temeli veri modeline dayanır. Veriyi mantıksal düzeyde düzenlemek için kullanılan kavramlar, yapılar ve işlemler topluluğuna "veri modeli" adı verilir [51].

Veri modelleri dört ana grupta toplanmıştır.

- Hiyerarşik (Sıradüzensel) Veri Modeli
- İlişkisel Veri Modeli
- Ağ (Network) Veri Modeli
- Nesneye Yönelik Veri Modeli

En yaygın kullanılan veri modeli, verilerdeki karmaşıklığı basite indirgeyen ilişkisel veri modelidir. Bu modelde, veri tabanındaki ilişkiler tablolar şeklinde ifade edilir. Tablolardaki sütunlar nitelikleri, satırlar ise bu niteliklerin değerlerini ifade eder.

Veri tabanını oluşturan tabloların özelliklerinden bahsedecek olursak;

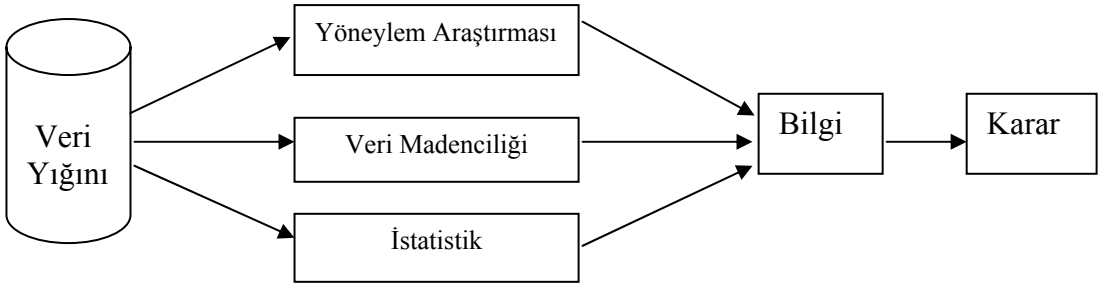
- Tablolar sütunlardan (kolonlardan) oluşur.
- Her bir sütunun ayrı bir değeri vardır.
- Her satır birbirinden farklıdır.

- Satırların ve sütunların sırası önemsizdir.
- Veri tabanındaki tablolar birbirlerine primary key olarak adlandırılan anahtarlarla birbirlerine bağlanırlar. Tablolar arasındaki ilişki bu anahtar sayesinde kurulur.

2.2.4. Veri tabanında bilgi keşif sürecinin aşamaları

Geleneksel sorgu veya raporlama araçlarının veri yığınları karşısında yetersiz kalması, Veri Tabanlarında Bilgi Keşfi (VTBK) adı altında, sürekli ve yeni arayışlara neden olmaktadır. Büyük miktardaki verilerin veri tabanlarında tutuldukları bilindiğine göre bu verilerin veri madenciliği teknikleri ile işlenmesine 'Veri Tabanında Bilgi Keşfi' denir [40].

VTBK aslında Şekil 2.3.'de görüldüğü üzere birçok yöntemle yapılmakta olup, veri madenciliği de VTBK sürecinin bir alt tekniği olarak adlandırılabilir.



Şekil 2.3. Veri tabanında bilgi keşfi yöntemleri

Büyük veri tabanlarında ilginç ve değerli olan bilgiyi algılamak ve erişmek oldukça zordur. Bu değerli ve önceden kestirilemeyen bilgiye, belirli metotlar uygulayarak erişmek için bilgi keşfi süreçlerinin veri tabanında tutulan verilere sıra ile uygulanması gerekir.

VTBK, Şekil 2.4.'de [18] görüldüğü üzere aşağıdaki adımlardan oluşur [42]:

- Uygulama alanının incelenmesi,

  (Konuyla ilgili bilgi ve uygulama amaçların belirlenmesi aşamasıdır.)

- Amaca uygun veri kümesi yaratma,

  (İlgilendiğimiz konu ile ilgili verilerin veritabanından çekilip, birleştirilmesi aşamasıdır.)

- Verilerin temizlenmesi,

- Verilerin birleştirilmesi,

- Verilerin seçilmesi,

- Verilerin dönüşümü,

- Veri madenciliği tekniği seçme,

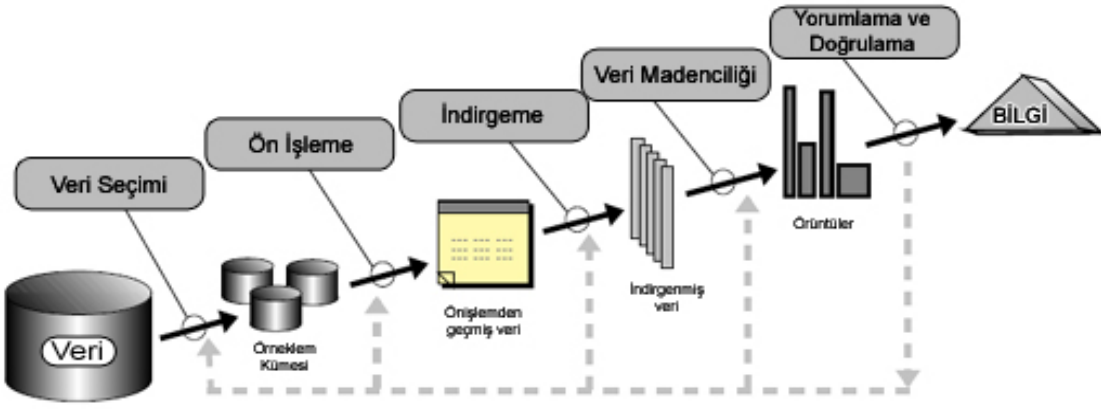
  (Bu aşama da ise uygulanacak olan veri madenciliği teknikleri seçilir.)

- Veri madenciliği algoritması seçme,

  (Bu aşamada seçilen tekniğe göre uygun olan algoritma veriler üzerinde denenir.)

- Örüntülerin değerlendirilmesi,

- Öz bilginin sunumu ve yorumlanması.



Şekil 2.4 Veri tabanında bilgi keşif süreci aşamaları

2.3. Veri Ambarı

2.3.1. Veri ambarının tanımı ve oluşturulma amacı

Kurumlarda günlük işlemlerden üretilen veri sadece işlemsel görevlerde kullanılmakta, saklanan verinin sürekli olarak değişikliğe uğraması ve farklı coğrafi konumlarda bulunan veri tabanlarının yapılan sorgulamalarla bir araya getirilerek verilerden bilgiye ulaşılmasının zorluğu "veri ambarı" (VA) (Database Warehouse) kavramının ortaya çıkmasına neden olmuştur.

Bu kavramın ortaya çıkması ile birlikte Çevrimiçi İşlem İşleme (OLTP) (Online Transaction Processing) veri tabanı (Bkz.2.3.2) sistemine sahip olan birçok kurum da VA teknolojisini uygulayarak, depoladıkları verileri anlamlı bilgiye dönüştürme imkânına kavuşma olanağı elde etmişlerdir.

VA, hemen hemen aynı anlam içinde çeşitli şekillerde tanımlanmıştır.

- VA, zaman içerisinde olabildiğince birikmiş verilerin oluşturduğu bir veri yığınıdır; bir işletmenin sahip olduğu verilerin karar destek amacıyla kullanılmasına olanak sağlar [51].

- VA, ilişkili verilerin sorgulanabilindiği ve analizlerinin yapılabildiği bir depodur [66].

- Kaba bir tanımla, VA "işletimsel sistemlerin çıktısı olan verilerin yönetildiği bilgi ortamıdır" denilebilir [18].

- Belirli bir döneme ait, yapılacak çalışmaya göre konu odaklı olarak düzenlenmiş, birleştirilmiş ve sabitlenmiş işletmelere ait veri tabanlarına veri ambarları denilir [59].

- VA, başlangıçta farklı kaynaklardan gelen verinin üzerinde daha etkili ve daha kolay sorguların yapılmasını sağlayan bütünleştirilmiş bilgi deposudur [66].

Veri Ambarlarının kurulum maliyetinin pahalı olmasına karşın, kurumların ileriye dönük kararlar almasında oldukça etkili sistemler olduğu bilinmektedir.

Erişilebilirlik, zamanlılık, inanılırlık ve anlaşılırlık, tasarım ve kullanım esnekliği gibi kalite faktörleri veri ambarlamanın başarısında çok önemli rol oynar [34].

Veri Ambarları; sağlık, coğrafi bilişim sistemleri, işletmelerin pazarlama bölümlerinde daha birçok sektörde oldukça yaygın bir kullanım alanına sahip, daha etkili sorgulamalar yapılmasına olanak tanıyan, yurt içinde ve yurt dışında pek çok uygulama alanı bulmuş olan sistemlerdir. Örneğin, Maliye Bakanlığı Gelir İdaresi Başkanlığı vergi kayıp ve kaçaklarını önlemek amacıyla, kurmuş olduğu VA teknolojisi ile kurumların otomasyon sistemlerine (online haberleşme) bağlanarak, Telekom, Bankalar, Emniyet, Tapu ve Nüfus İdarelerinden anlık bilgi akışı ile mükelleflerin her türlü parasal ve mal alım-satım hareketlerini vergisel durumu ile birlikte izleyebilme olanağı elde etmiştir.

Veri Ambarlarının oluşturulma amaçları;

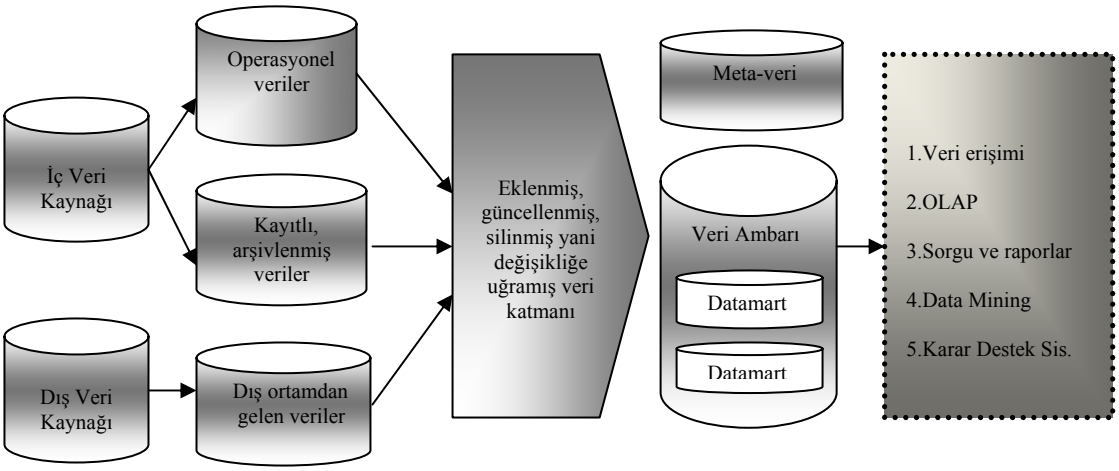
- Operasyonel veriler üzerinden birçok kullanıcı tarafından sorgulamalar yapıldığında operasyonel sistemin yavaşlamasını önlemek dolayısıyla vatandaşa daha iyi hizmet sunmak,

- Verilerin uğradığı değişimlerin tarihçesini tutarak ileride yeniden kullanım amaçlı arşivlemek,

- Kurumların karar destek amaçlı olarak ileriye dönük kararlar alarak kurumların strateji belirlemesine olanak tanımak

şeklinde ifade edilebilir.

Şekil 2.5.'den de anlaşılacağı üzere, farklı coğrafi konumlarda bulunan veri tabanlarındaki verilerin bir araya getirilmesi ile veri ambarları oluşur [30].



Şekil 2.5. Veri ambarı akış mimarisi

VA'nın en önemli bileşenlerinden biri de meta-veri'lerdir. Meta-veri, verilerin tanımlandığı kısım olup, veri hakkında veri anlamına gelmektedir [66].

Veri ambarlarının alt kümesi olarak ifade edilen datamartlar, boyutları 1–10 GB arasında değişen küçük veri ambarlarıdır. Datamart oluşturmaktaki anahtar farklılık, önceden tanımlanmış belli bir ihtiyaca yönelik bir grup bilgi ve seçilmiş veri yapılandırılmasıdır. Datamart yapılandırılması ile ilgili veriye kolay erişim vurgulanmaktadır [15]. Veri ambarlarının içerdiği verinin miktarının fazlalığı ile birlikte, bu verileri sorgulayan kullanıcıların görmemesi gereken ya da kendi bölümlerini ilgilendirmeyen verileri sorgulamaları pek mantıklı olmadığından VA'nda konuya yönelik datamartlar oluşturulur ve kullanıcılar ilgilendikleri verinin bulunduğu datamarttan sorgulama yaparlar.

2.3.2. Veri ambarlarının modellenmesi ve yapıları

Kurumlardaki veri tabanlarının da kurulduğu iki tür sistem mevcuttur.

- *Operasyonel Sistemler*

OLTP sistemler (Online Transaction Processing–Çevrimiçi İşlem İşleme) olarak da adlandırılmaktadır. Bu sistemin temeli işlemsel veri tabanlarına dayanmakta olup, veri tabanı sürekli olarak değişikliğe uğramaktadır.

Günlük yapılan işleri ve işlemleri gerçekleştirmek, sonuçları saklamak bu sistemlerin görevidir. Bu tür sistemlerde erişilebilirlik ana amaçtır; veriye en kısa sürede ulaşmak ve işlemleri en kısa sürede sona erdirmek hedeflenir. Bu nedenden dolayı canlı sistemler çevrimiçi çalışma prensibi ile tasarlanırlar [29].

- *Karar Destek Sistemleri*

Günümüzde firmalar arasındaki rekabet ortamı yeni teknolojiler kullanarak ileriye dönük, hızlı, etkili ve doğru kararlar alınmasını gerektirmiştir. Karar Destek Sistemleri bu şekilde ortaya çıkmıştır.

Bu sistemlerin kurulu olduğu veri tabanı yapısı; yöneticilerin programlanamayan türden karar verme işlemlerine yardımcı olmak üzere geliştirilir. Yöneticinin herhangibir anda, daha önceden öngörülmemiş bir bilgiye aniden gereksinimi olabilir. Böyle bir durumda, hemen yanıt verebilecek bir sistemin varlığı gerekecektir. Karar destek sistemleri bu gibi durumlar için tasarlanır. Tasarlanan bu yapı üzerinden üst yöneticiler ihtiyaçlarını OLAP (Online Analytics Processing–Çevrimiçi Analitik İşleme) adı verilen bir teknoloji ile karşılarlar [51].

Veri ambarları üzerinde, çeşitli taktik ve stratejik konular hakkında karar vermeye yardımcı olacak veri analizi ve sorgulama işlemlerine OLAP denilir [59].

OLAP, son kullanıcıların her gün ihtiyaç duydukları rapor ve analizleri karar vericilere çok boyutlu ve hızlı bir şekilde sunan özel bir teknolojidir [66]. Bu sistemde, verinin uğradığı değişikliğin zaman boyutu tutulduğundan veriler sürekli olarak depolanır, dolayısıyla geçmişe yönelik çıkarsamalara da olanak sağlayan bir teknoloji olması nedeni ile veri ambarları ve datamartlarda da etkin bir şekilde kullanılmaktadır.

2.3.3. Veri ambarı ile veri tabanı arasındaki farklar

- VA, aynı veya farklı bölgelerde bulunan birden çok veri tabanının bütünleştirilmiş halidir.

- Veri ambarları kullanıcıları analistler, yöneticiler ve bu verileri kullanan personel olduğundan kullanıcı sayısı azdır. Veri tabanı kullanıcıları ise kasiyer, sistem uzmanı, veri tabanı yöneticisi vb. gibi olduğundan kullanıcı sayısı daha fazla olmaktadır.

- Veri ambarlarında verilerin uğradığı değişiklikler her akşam sisteme yansıtılarak, eski verilerin tarihçesi tutulmasına karşın veri tabanlarında verilerin uğradığı değişikliklerin kaydı tutulmamaktadır.

- Veri ambarları yıldız, konu tabanlı vb. gibi mimarilerle inşa edilirken, veri tabanları varlık ilişkisel model, nesne tabanlı gibi uygulamaya yönelik mimarilerle yapılandırılırlar [59].

- Veri ambarları kurumların veri tabanlarında kapladıkları alan bakımından “Tera Byte”lar seviyesinde iken, veri tabanları “Giga Byte”lar seviyesinde bir büyüklüğe sahiptir [59].

2.3.4. Veri ambarı ile OLTP sistemler arasındaki farklar

- OLTP sistemler veri tabanı ve süreç tasarımının her ikisiyle, VA ise veri modelleme ve veri tabanı tasarımıyla ilgilenir [51].

- OLTP sistemlerde her türlü veri olmasına karşın, Karar Destek Sistemleri için kurulan veri ambarları ise, karar destek sürecinde kullanılmayacak olan veriyi bünyesinde barındırmaz.

- OLTP sistemler ile veri ambarları farklı fiziksel ortamlarda yaratılırlar. OLTP sistemlerin verilerin işlem gördüğü ortamlar olması nedeni ile veri ambarlarını da aynı fiziksel ortama koymanın günlük işlemlerin hızını yavaşlatacağı düşünülerek, sadece sorgulamaların yapıldığı veri ambarlarının OLTP sistemlerden ayrı bir fiziksel ortama kurulması uygun görülmektedir.

- OLTP sistemleri kuruluşları çalıştıran, veri ambarları ise kuruluşlara yol gösteren bir oluşumdur [51].

- OLTP sistemlerde yapılan sorgulamalarla elde edilen veri daha sonra değişikliğe uğramışsa, değişikliğe uğramadan önceki durumu sistemde tutulmaz, dolayısıyla bu tür sistemlerde çekilen veri o andaki veriyi yansıtmaktadır. Veri Ambarlarında ise verinin uğradığı her türlü değişiklik, değişilikliğin yapıldığı tarih, zaman bilgisi ve değişikliği yapan kullanıcının kodu ile birlikte kayıt altına alınır. Kısaca, veri ambarları OLTP sistemlerin belirli dönemlerdeki anlık görüntülerinden oluşur diyebiliriz.

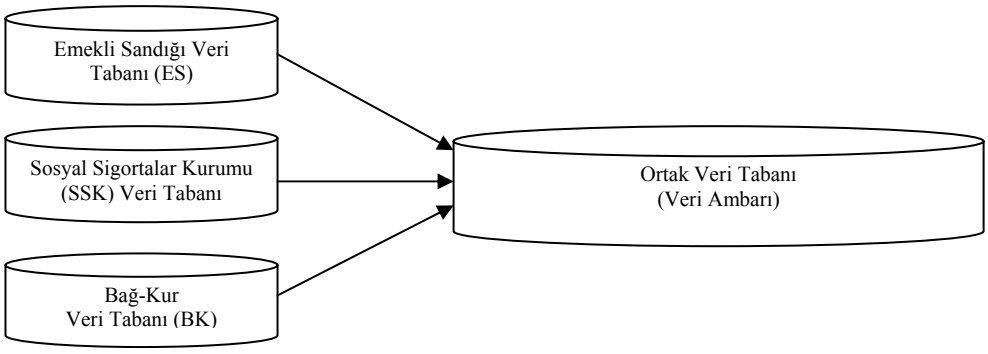
- OLTP sistemlerin zaman boyutu özelliği olmamasına karşın, veri ambarları geçmiş verilerin de kayıtlarının tutulduğu bir yapıda tasarlandıklarından zaman boyutu özelliğini içermektedir.

2.3.5. Veri ambarının özellikleri

Veri ambarlarının özelliklerinden bahsedecek olursak;

- *Entegre Olma,*

Veri ambarları, Şekil 2.6.'da gösterildiği gibi farklı veri tabanlarının entegrasyonu sonucunda oluşan bütünleştirilmiş veri tabanlarıdır. Farklı veri tabanlarındaki veriler bütünleştirilerek ortak bir veri tabanı haline getirilirken, aynı niteliğe sahip verilerin kodlamaları da aynı olmalıdır. Örnek olarak; Emekli Sandığı (ES) veri tabanında cinsiyet alanı Erkek='E' ve Kadın='K' olarak, Sosyal Sigortalar Kurumu (SSK) veri tabanında ise Erkek=0 ve Kadın=1 olarak kodlanmış ise, VA oluşturulurken bu kodlamaların aynı yapıya dönüştürülerek farklılıkların teke indirilmesi ön koşuldur.



Şekil 2.6. Farklı veri tabanlarının birleşimi ile oluşturulan yeni veri tabanı

- *Nesne Yönelimlilik,*

Veri ambarları konuya odaklı tasarlanır ve karar almada kullanılmayacak olan verileri bünyelerinde barındırmazlar.

- *Zaman Değişimlilik,*

Veri ambarlarında zaman boyutu kavramı bulunduğundan, her kaydın değişikliğe uğradığı zaman sistemde kayıt altına alınır. Veri ambarlarının bu özelliği, geçmiş

verilerin kullanılarak bu verilerden elde edilen bilginin geleceğe dönük çıkarsamalarda kullanılmasına olanak sağlar.

- *Verilerin tarihçeleri tutulur,*

OLTP sistemlerdeki verilerin uğradıkları değişimlerin (update, insert, delete gibi.) tarihçeleri veri ambarlarında tutulur. Veri ambarlarında veriler sadece okunabilir, silinemez ve güncelleştirilemez.

- *Veri ambarları statiktir,*

VA'na veri giriş ve çıkışı yapılmaz, veriler sadece okunabilir.

- *Veri Ambarları birçok datamart'tan oluşur.*

Veri Ambarları konu odaklı olup, her konu bir datamartı temsil eder. Örneklendirecek olursak; bir kullanıcı satış verileri ile ilgileniyorsa satış verilerinin olduğu datamarttan sorgusunu yapar, VA'nın tamamı ile ilgilenmez. Zaten her kullanıcının da her bilgiye ulaşması istenen bir durum değildir.

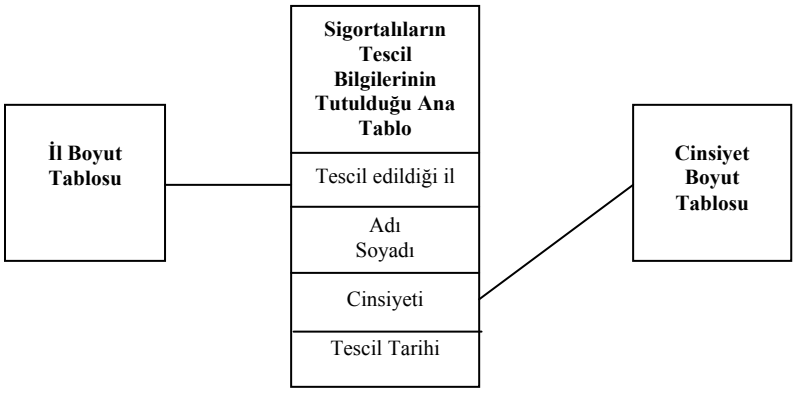
2.3.6. Veri ambarında kullanılan modeller

Veri Ambarları, kurumlardaki günlük işlemleri uygulamak için değil, elde edilen veriyi çok daha hızlı bir biçimde çözümlemeye ve karar vermeye yönelik olarak kurulurlar. Karar verici, bir konuda bilgiye gereksinim duyduğunda bu bilgiyi hareketli olan giriş ve çıkışın çok olduğu veri tabanı sisteminden alamaz, çünkü bu fazla karmaşık bir sorgu gerektirir. İşte veri ambarları bu karmaşık sorgulamaların yapıldığı bir mimariye sahiptir. Veri ambarları genellikle, bir ana tablo (fact tablosu) ve etrafında da bu ana tabloya bağlı boyut tablolarından (dimension tablosu) oluşur.

Veri Ambarları üç değişik şema kullanırlar [59].

- *Yıldız Şema (Star Schema)*

Yıldız şema türünde, Şekil 2.7.'den de görüleceği üzere ortada bir ana tablo ve etrafında da boyut tabloları bulunur.



Şekil 2.7. Yıldız şema modeli (çok boyutlu model)

- *Kar tanesi Şeması (Snowflake Schema)*

Kar tanesi şema türünde ise, yıldız şemadan farklı olarak boyut tabloları normalize edilmiş diğer boyut tablolarına bağlanmıştır. Kar tanesi şeması daha karmaşık sorgular gerektirdiği için sorgu performansı azdır. Bu nedenle VA tasarımında kar tanesi şeması, yıldız şeması kadar yaygın olarak kullanılmamaktadır [19].

- *Ana Tablo Şeması*

Ana tablo şema türünde ise, birden fazla ana tablo bulunur ve boyut tabloları bu ana tablolara bağlanır. Yani birden fazla yıldız şema iç içe girmiş görünümündedir.

2.4. Veri Ön İşleme Teknikleri

Bilişim sektöründeki gelişmeler insan yaşamına birçok açıdan kolaylık getirmesine karşın bir o kadar zorlukları da beraberinde getirmiştir. 20–30 yıl öncesine kadar yüz binlerle ifade edilen veri hacimleri bugün milyon hatta milyar seviyesine ulaşmıştır. Veri hacimlerinin çok büyük olmasından dolayı klasik istatistiksel yöntemler bu verilerden bilgi çıkarmak için yeterli gelmemeye başlamış, verileri analiz edebilmek için yeni tekniklere ihtiyaç duyulmuş ve bu ihtiyaç da veri madenciliği kavramının ortaya çıkmasına sebep olmuştur.

Büyük hacimli veriler üzerinden bir takım analizler yaparak sonuçlar çıkarmak çok sağlıklı sonuçlar vermeyebilir, çünkü veri tabanındaki veriler eksik, hatalı, uç değerler içeren, tekrarlı ve kayıp verilerden oluşabilir. Bu nedenle öncelikle veri tabanındaki verilerin düzeltilmesi gerekir ki doğru kararlar verilebilsin. Veri Madenciliği yöntemlerini uygulayabilmek için de, yeterli ve nitelikli verilere ihtiyaç vardır. Verinin kaliteli olması, bu veriden çıkarılacak bilginin de kaliteli olması anlamına gelir. Bu nedenle veri madenciliği uygulamasına geçmeden önce verilerin bir takım önişlemlerden geçirilmesi gerekmektedir.

Veri ön işleme süreçleri; veri temizleme, veri birleştirme, veri dönüştürme ve veri indirgeme aşamalarından oluşur. Verileri önişlemden geçirebilmek için öncelikle verilerin bir araya toplanması yani farklı veri tabanlarında tutulan verilerin bir araya getirilmesi, kâğıt ortamında tutulan verilerin bilgisayara girilmesi, kayıt edilmemiş değişkenler varsa bu değişkenlerin belirlenmesi ve gürültülü verilerin de tespit edilmesi gerekmektedir. Dolayısıyla öncelikle elimizdeki verileri inceleyerek acaba bu verilere veri madenciliği (VM) tekniklerini uygulayabilir miyiz? sorusunu cevaplamak zorundayız. Ve şu soruları sormalıyız [49];

- VM'nin uygulanabilmesi için elimizde yeterince veri var mı?

- Veriler elde edilebilir mi? Elde edilebilir ise, VM için uygun formlara dönüştürülmeleri, kâğıt ortamında tutulan verilerin de bilgisayar ortamına aktarılması gerekmektedir.

- Veriler, VM tekniklerini uygulayacağımız değişkenleri kapsıyor mu? kapsamıyorsa bu değişkenlerin belirlenerek, verilerle birlikte bilgisayar ortamına aktarılması gerekmektedir.

- Veriler gürültülü mü? Burada gürültüden kasıt; yanlış girilmiş, yanlış ölçülmüş, diğer verilere göre çok uç bir değeri olan aykırı/uç verilerdir. Dolayısıyla hatalı veriler bizi hatalı sonuçlara götüreceğinden mutlaka verilerdeki bu gürültünün de temizlenmesi gerekmektedir.

Yukarıda bahsedilen soruların cevaplarını verebiliyorsak, bir sonraki aşama verileri önişlemden geçirmek olmalıdır. Veri ön işleme teknikleri aşağıda maddeler halinde özetlenmektedir.

2.4.1. Veri temizleme

Veri temizleme; eksik verilerin tamamlanması, aykırı değerlerin tespit edilmesi amacıyla gürültünün düzeltilmesi ve verilerdeki tutarsızlıkların giderilmesi gibi işlemleri gerektirmektedir [49].

Veri tabanında bazı değişkenlerin değeri yok ise eksik değer problemi vardır. Belli bir değeri eksik değerin yerine koyma sürecine "yaklaşık değer verme" denilmektedir. Örneğin; veri tabanında kayıtlı olan sigortalıların bir kısmının medeni hali kodlanmışken, bu bilgi bazı sigortalılarda girilmemiş olabilir. Herhangi bir değişkene ilişkin eksik (kayıp) değerlerin olması durumunda izlenecek birkaç yol vardır;

- Eksik değer içeren kayıtlar atılabilir. Fakat bu çok iyi bir yöntem olarak tercih edilmez, çünkü kayıt sayısı fazla ise veriden elde edeceğimiz bilgiyi

kalitesizleştirir. Bunun için kayıp olan veri sayısı toplam kayıtlı olan veri sayısına oranlanır, bu oran sonuçları etkilemeyecek kadar küçükse bu yöntem kullanılır ama genelde tercih edilen bir yöntem değildir.

- Aynı sınıfa ait değişkenin ortalaması eksik değerlerin yerine kullanılabilir [50].

- Var olan verilere bakılarak en uygun değer kullanılabilir.

- Değişkenin ortalaması eksik değerin yerine kullanılabilir.

- Kayıp olan veriler elle teker teker doldurulabilir. Fakat bu yöntemde eğer veri tabanı küçükse ve kayıp olan veriler elde edilebiliyorsa uygulanmalıdır, yoksa zaman kaybından başka bir şey elde edilemez.

- Tüm kayıp olan, eksik olan verilere aynı değer girilebilir. Örneğin ücret bilgisi eksik olan kayıtlar $+\infty$, medeni hali boş olan kayıtlar da A harfi ile kodlanabilir, fakat bu yöntemde elde edilecek olan bilgiyi kalitesizleştirebilir veya bize yanlış bir bilgi verebilir.

- Regresyon yöntemi kullanılarak eldeki eksik olmayan tüm verilere bir regresyon denklemi ve regresyon katsayıları elde edilerek, kayıp olan veriler tahmin edilebilir. Regresyon dışında zaman serileri analizi, bayesyen sınıflandırma, karar ağaçları, maksimum beklenti gibi VM'nde kullanılan diğer yöntem ve teknikler de kayıp verilerin tahmin edilmesinde kullanılabilir [49].

Kayıp veriler dışında, gürültülü verilerin tespiti içinde kullanılan yöntemlerden birkaçı kümeleme analizi, histogram ve regresyon yöntemi olarak sayılabilir.

2.4.2. Veri birleştirme

VM'nin uygulanabilmesi için farklı veri tabanlarında tutulan verilerin bir araya getirilmesi gerekmektedir. Fakat farklı veri tabanlarında tutulan veriler

birleştirildiğinde "şema birleştirme hataları" olarak adlandırılan bir takım hatalar meydana gelmektedir. Bu tip hatalardan kaçınmak için de meta-veriler kullanılmaktadır. Veri ambarları da genellikle meta-veri temeline dayanarak hazırlanırlar. Meta veri, veriye ilişkin veri demektir [49].

Veri tabanlarının birleştirilmesi ile birlikte farklı tablolardan gelen aynı niteliğe sahip değişkenlerin fazlalıkları VA'nın gereksiz yere büyümesine sebep olmaktadır. Örneğin; veri tabanının birinde tüketici_ıd olarak geçen kolon adı diğer veri tabanında tüketici_numarası olarak adlandırılmış olabilir. Dolayısıyla yapılan birleştirme işlemi ile birlikte aynı değişken çoklanmış olmakta ve şema birleştirme hatası meydana gelmektedir. Yapılması gereken, yukarıda bahsedilen tüketici_ıd ile tüketici numarası arasındaki korelasyon katsayısının hesaplanarak, eğer yüksek bir katsayı elde edilirse değişkenlerden birinin veri tabanından çıkartılması şeklinde olmalıdır.

Veri tabanlarının birleştirilmesi ile birlikte meydana gelen bir diğer sorun da, veri değerlerindeki tutarsızlıklardır. Farklı kaynaklardan gelen değişkenlerin sahip oldukları değerler farklı olabilir. Farklılıktan kasıt; kodlama, ölçekleme gibi sorunlardır ki bu sorunlar da dışarıdan müdahale ile düzeltilmeye çalışılmalıdır.

Ayrıca farklı veri kaynakları birleştirildiğinde gereksiz, kullanılmayacağı düşünülen değişkenler de modele alınmayabilir. Örneğin; tescil tablosunda kayıtları tutulan sigortalıların telefon numarası, kimlik numaraları vb. gibi.

2.4.3. Veri dönüştürme

VM'nde birçok algoritma kullanılmakta olup, bu algoritmaların bazıları kesikli verilerle bazıları ise sürekli verilerle çalışmaktadır. Dolayısıyla kullanacağımız algoritmaya göre bu verilerin bir dönüşüme tabi tutulması gerekmektedir. Örneğin; Yapay Sinir Ağları (YSA), 0–1 aralığındaki verilerle çalışan bir algoritma olup, bu algoritmanın uygulanabilmesi için öncelikle verilerin 0–1 aralığı dönüşüm işleminin yapılarak, VM algoritmalarını uygulayabileceğimiz formlara dönüştürülmesi

sağlanmalıdır [56]. Bu verileri uygun formlara dönüştürmek için de farklı yöntemler geliştirilmiştir.

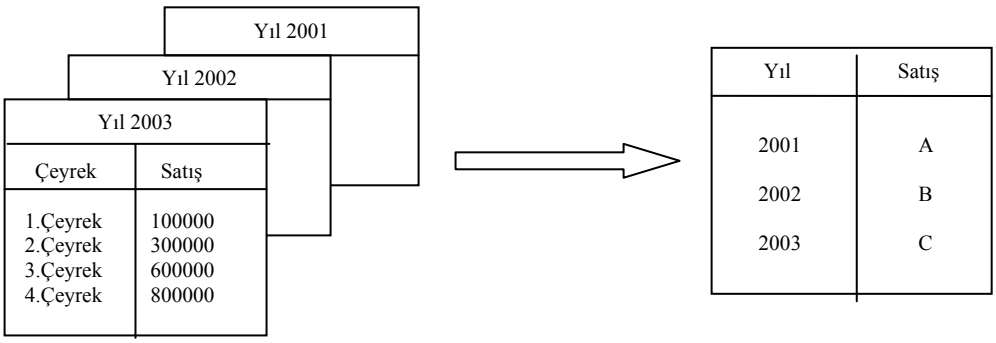
2.4.4. Veri indirgeme

Veri indirgeme, daha küçük hacimli bir veri kümesi elde etmek için uygulanan bir yöntemdir. Her zaman çok büyük veri kümeleri ile çalışmak hem zaman yönünden hem de maliyet yönünden sıkıntılı olduğundan, bu veriyi en iyi şekilde temsil edebilecek bir örneklem ile çalışmak daha doğru bir yöntem olmuştur. Veri indirgeme yöntemleri aşağıda açıklanmıştır [49].

- *Veri birleştirme veya veri küpü*

Veri birleşmeyi bir örnekle açıklayacak olursak; sigortalıların 2000–2002 yılları arası ödedikleri primlerin tutulduğu, çeyrek dönemlerden oluşan bir tabloya sahip olduğumuzu varsayalım.

Şekil 2.8'den de anlaşılacağı üzere bir bilgi kaybı olmaksızın, veriler birleştirilerek veri hacmi azaltılmış oldu.



Şekil 2.8. Veri birleştirme

Veri küpü yöntemi ise; çok değişkenli birleştirilmiş bilginin saklandığı küplerdir.

- *Boyut indirgeme*

VM uygulanacak verilerde bazen çok gereksiz hatta kullanmayacağımız değişkenlere ait veriler tutulmaktadır. Telefon numaraları, posta kodları gibi. Bu veriler elde edilecek olan bilgiyi kalitesizleştirdiği gibi veri tabanı hacminin gereksiz yere büyümesine neden olmaktadır. Bu nedenle bu tür verileri içeren bütün değişkenler veritabanından çıkartılmalıdır. Bunun için de iki yöntem geliştirilmiş olup ilki; ileri yönlü sezgisel seçim, diğeri ise geri yönlü sezgisel seçim olarak adlandırılır.

İleri yönlü sezgisel seçimde; orijinal değişkeni en iyi şekilde temsil edecek olan değişkenler belirlenir. Ardından her bir değişkenin bu kümeye dâhil edilip edilmeyeceğine sezgisel olarak karar verilir.

Geri yönlü sezgisel seçimde ise; bütün değişkenler ele alınarak, daha sonra gereksiz bulunan değişkenler sezgisel olarak bu kümeden atılır.

- *Veri sıkıştırma*

Veri sıkıştırma yönteminde ise; orijinal verileri en iyi şekilde temsil edecek olan veriler veri şifreleme veya veri dönüşümü yöntemi ile sıkıştırılarak veri indirgenir. Eğer veri sıkıştırma işleminde bir bilgi kaybı varsa buna "kayıplı (lossy) sıkıştırma", bir bilgi kaybı oluşmuyorsa buna da "kayıpsız (losless) sıkıştırma" adı verilir.

- *Temel Bileşenler Analizi*

Bu analiz yönteminde ise, p adet değişken orijinal değişkeni en iyi şekilde azda olsa bir bilgi kaybıyla temsil edebilecek k adet ($k<=p$) yeni değişkene indirgenir. Bilgi kaybının olmadığı algoritmalar da vardır fakat bu algoritmalar bir takım kısıtlamalar içerdiklerinden genellikle tercih edilmezler.

- *Kesikli hale getirme*

VM algoritmalarının bazıları sadece kategorik verilerle çalıştığından dolayı sürekli olan verilerin kesikli hale getirilmesi gerekmektedir.

Örneğin; yaş sürekli bir değişken olup, bu değişkenin içerdiği verileri 1–10, 11–21, 22–41, 41+ şeklinde kategorik hale getirebiliriz. Bu yapılan işlem ile detay bilgiler kaybolsa bile genelleştirilmiş veriler daha kolay yorumlanabilecektir.

2.5. Veri Madenciliği

2.5.1. Veri madenciliği tanımı ve amacı

Günlük yaşamda bilgisayarların hayatımıza girmesi ile birlikte yapılan her işlem sayısal ortamda kayıt altına alınmaktadır. Örneğin market alışverişlerinde alınan veya iade edilen her bir ürünün manyetik ortamda yapılan giriş çıkış işlemi, benzer hastanelerdeki hasta kayıtları, sinema ve devlet dairelerindeki kayıtlar, yollarda bulunan kameraların kayıtları, yapılan telefon görüşmeleri gibi kısacası her yerde, her yapılan işlem bir veri topluluğu meydana getirmekte ve verinin olduğu yerde de veri tabanı meydana gelmektedir.

Gelişen teknoloji ile birlikte, veriye erişim kolay olmasına karşın milyon hatta milyar seviyesinde veri kaydının olduğu bir veri tabanında bu verilerden bir takım çıkarsamalar yaparak bilgiye ulaşmak zorlaşmış ve klasik istatistiksel yöntemler bu verilerden bilgi çıkarmak için yeterli gelmemeye başladığından, verileri analiz edebilmek için yeni tekniklere ihtiyaç duyulmuştur. 2001 yılında en büyük işleme sahip veri tabanları 318 tera byte iken, bu rakam 2003 yılında 1029 tera byte'lara çıkarak, iki yıl içerisinde müthiş bir veri artışı olmuş ve halen olmaya da devam etmektedir [10]. Bütün bu veri yığınları içerisinde altın değerinde keşfedilmeyi bekleyen bilgiler bulunmaktadır.

Kurumların kadrolarında görev yapan üst yöneticilerin, bu dev boyuttaki veri yığınları içerisinden bilgiye ulaşarak ileriye dönük, kurumun iyileştirilmesine yönelik kararlar almaları bilgiye erişemedikleri sürece imkânsız hale gelmektedir. Bu verilerin üst yönetime bilgi olarak dönmesi için, bu verilerin uygun yazılımlar aracılığı ile bir takım işlemlerden geçirilerek bilgiye dönüştürülmesi ve üst yönetime sunulması gerekir. Bu bilgilere klasik istatistiksel yöntemlerle erişmek de artık çok zorlaştığından, yeni bir takım tekniklere ihtiyaç duyulmuş ve VM kavramı ortaya çıkmıştır.

Diğer alternatif VM isimleri; Veri Tabanlarında Bilginin Keşfi, Bilgi Çıkarımı, Veri ve Örüntü Analizi, Veri Arkeolojisi, Veri Eşleme olarak sayılabilir [42].

VM, insan merkezlidir ve bazen insan bilgisayar ara yüzü birleştirilir [9]. Veriler arasındaki ilişkiyi, kuralları ve özellikleri belirlemekten bilgisayar sorumludur [41, 17].

VM tanımlamalarından bazıları aşağıda ifade edilmiştir:

- VM, veri yığınları içerisinden anlamlı ve işe yarar bilgi çıkarmayı sağlayan disiplinler arası bir yaklaşımdır [66].

- VM daha önceden bilinmeyen, geçerli ve uygulanabilir bilgilerin geniş veri tabanlarından elde edilmesi ve bu bilgilerin işletme kararları verirken kullanılmasıdır [59].

- VM, eldeki verilerden üstü kapalı, çok net olmayan, önceden bilinmeyen ancak potansiyel olarak kullanışlı bilginin çıkarılmasıdır. Bu da; kümeleme, veri özetleme, değişikliklerin analizi, saptamaların tespiti gibi belirli sayıda teknik yaklaşımları içerir [41].

- VM, en basit tanımı ile çok büyük miktardaki ham veriler içinden amaca uygun modellerin ortaya çıkarılması işlemidir. Başka bir tabirle karmaşık ve düzensiz veriler içindeki modellerin ortaya çıkarıp bunlara karar verme ve eylem planını gerçekleştirmek için kullanma sürecidir [1].

- VM, veri içerisindeki gizli bilgilerin açığa çıkarılması ve verinin karar destek tabanlı bilgiye dönüştürülmesi sürecidir [1, 74].

- VM, büyük miktarda veri içinden gelecekle ilgili tahmin yapmamızı sağlayacak bağıntı ve kuralların bilgisayar programları kullanarak aranmasıdır [12].

- VM, bir kurumda üretilen tüm verilerin belirli yöntemler kullanılarak var olan ya da gelecekte ortaya çıkabilecek gizli bilgiyi su yüzüne çıkarma süreci olarak değerlendirilebilir [41].

- VM, verilerin içerisindeki desenlerin, ilişkilerin, değişimlerin, düzensizliklerin, kuralların ve istatistiksel olarak önemli olan yapıların yarı otomatik olarak keşfedilmesidir [41].

- Gartner Group tarafından yapılan bir diğer tanımda ise VM, istatistik ve matematik tekniklerle birlikte örüntü tanıma teknolojilerini kullanarak, depolama ortamlarında saklanmış bulunan veri yığınlarının elenmesi ile anlamlı yeni korelasyon, örüntü ve eğilimlerin keşfedilmesi sürecidir [59, 13, 55].

- VM'ni, bilgisayar teknolojilerinin sağlamış olduğu çok hızlı veri işleme ve yüksek hacimde veri depolama imkanları yardımıyla ve farklı disiplinlerin (yapay zeka, makine öğrenmesi, uzman sistemler, veri tabanı teknolojileri, paralel bilgi işleme, dağıtık veri işleme, görselleştirme, optimizasyon, veri ambarcılığı, istatistik …) katkısıyla sağlanan araçlarla, sahip olunan çok büyük hacimlerdeki veriden, karar vericinin etkin ve daha fazla bilgiye dayalı karar vermesinde kullanabilmesi amacıyla önceden bilinmeyen, gizli, örtük, klasik metotlarla

ortaya çıkarılması güç, faydalı, ilginç, anlaşılabilir; ilişki, örüntü, bağıntı veya trendlerin otomatik veya yarı otomatik bir şekilde ortaya çıkarılması olarak tanımlamak mümkündür [66].

- VM, istatistiksel analiz tekniklerinin ve yapay zekâ algoritmalarının bir arada kullanılarak çok sayıda durum ve değişkenden oluşan veri yığınları içerisindeki gizli bilgilerin açığa çıkarılması ve verinin karar destek tabanlı bilgiye dönüştürülmesi sürecidir [4].

- VM, geçmiş verilerden gelecek tahminleri yapma işlemidir.

- VM, bir istatistiksel yöntemler serisidir [41].

- VM, birçok disiplini birlikte içerir. Bu disiplinler kümesi; veritabanı teknolojileri, istatistik, makine öğrenme, görselleşme ve diğer disiplinlerden oluşmaktadır.

Bu tanımlara dayalı olarak VM'nin amacını iki madde ile özetleyebiliriz;

1. Veriler arasındaki fark edilmemiş ilişkileri, desenleri ortaya çıkarmak,

2. Kolaylıkla mantıksal kurallara ya da görsel sunumlara çevrilebilecek nitel modelleri elde etmek.

Burada dikkat edilmesi gereken nokta VM'nde elde edilecek bilginin önceden bilinemez, tahmin edilemez olmasıdır. Tahmin edilebilen, beklenen sonuçlar için VM algoritmaları kullanılmamaktadır.

VM'nin üç farklı bakış açısı vardır [65].

- Veri Tabanı Bakış Açısı,
- Makine Öğrenim Bakış Açısı,
- İstatistiksel Bakış Açısı.

VM gürültülü verilerle çalışamamasına rağmen Makine Öğrenim Bakış Açısı tekniği gürültülü verilerle çalışabilmekte ve anlamlı sonuçlar üretebilmektedir.

VM daha çok istatistikçiler, veri analizcileri, bilişim sistemleri ile ilgilenen topluluklar tarafından kullanılmaktadır. VM, sadece verileri bilgisayar ortamına girerek sonuç üreten bir bilgisayar yazılımı değildir. VM yapabilmesi için, öncelikle VM'nin uygulanacağı alanın ve amacın belirlenip, ona göre VM algoritmalarını kullanarak modelleme yapmak gerekmektedir.

2.5.2. Literatür

Sektörel anlamda gerek üretim aşamasında gerekse sunulan hizmetlerin risk faktörleri ile hizmetin kalitesi açısından kullanılabilirliğinin en yüksek seviyede tutulmasını amaçlayan VM, ülkemizde yakın bir geçmişe sahip olmasına karşın kurumlar tarafından büyük oranda uygulama alanı bulmuştur.

VM teknikleri kullanılarak uluslararası yapılan çalışmalardan bazıları aşağıda özetlenmiştir:

Rakowski ve arkadaşları (1998), Amerika'da (ABD) yaşayan 50–75 yaş aralığında örnekleme alınan 1727 kadın üzerinde yaptıkları çalışmada, bu yaş aralığındaki kadınların mammografi yaptırma oranlarını çoklu lojistik regresyon ve karar ağacı ile incelemişlerdir [57].

Bijmolt ve arkadaşları (1998), 5–8 yaş arasındaki 153 hollandalı çocuk ve aileleri ile yaptıkları çalışmada, çocukların televizyon reklâmlarını algılamalarına etki eden faktörleri karar ağacı ile incelemişlerdir [22].

Lopez ve arkadaşları (1999), Arjantin'in Buenos Aires şehrinin alt eyaletlerindeki sosyoekonomik olarak orta sınıftan örnekleme alınan 10–89 yaş aralığındaki 360 kişi ile yaptıkları çalışmada, Hepatit A antikorunun yaygınlığı ile ilişki olan faktörleri CHAID algoritmasını kullanarak incelemişlerdir [46].

Ho ve arkadaşları (2004), demografik, çevresel ve genetik faktörlerin kanser hastalığının gelişimine etkilerini karar ağacı ve lojistik regresyon ile incelemişlerdir [37].

Bayam ve arkadaşları (2005), deneyimli sürücüler ile trafik kazaları arasındaki ilişkiyi karar ağacı ile incelemişlerdir [21].

Diepen ve arkadaşları (2006), CHAID algoritmasından tahmin edilen sonuçların güven sınırlarını hesaplamak için önerdikleri yöntemle, benzetim tekniğiyle elde edilmiş veriler kullanmışlardır [28].

Chan ve arkadaşları (2006), ortopedik engellilerin istihdam sonucunu etkileyen faktörleri CHAID algoritması ile incelemişlerdir [26].

McCarty ve Hastak (2007), iki farklı veri seti kullanarak doğrudan pazarlama bölümlemesi için RFM (Recency, Frequency and Monetary Value/Yenilik, Frekans ve Parasal Değer), CHAID ve Lojistik Regresyon modellerinin karşılaştırmasını yapmışlardır [47].

Lin (2007), mobil telekomünikasyon sektöründe müşterilerin detaylı arama kayıt verileri ile kümeleme analizi yaparak, müşterilerin detaylı arama kayıtlarından müşteri profillerini çıkartmaya çalışmıştır [45].

Sullivan ve arkadaşları (2008), koruyucu aileye verilen farklı yerleşim yeri ve yaş gruplarındaki örnekleme alınan 911 çocuğun, bu ailelerin yanında kalış sürelerine göre fiziksel ve ruhsal ihtiyaçları tanısını karar ağaçlarını kullanarak incelemişlerdir [60].

Seibt ve arkadaşları (2009), 25–60 yaş aralığında 100 öğretmen ve 60 ofis çalışanı olmak üzere toplam 160 kadın örneklem üzerinden yapılan anket çalışması ile çalışanların fiziksel ve ruhsal durumlarının iş ortamları ile olan bağlantısını öngören çalışmayı CHAID karar ağacını kullanarak incelemişlerdir [62].

Horner ve arkadaşları (2010), 1493 ilkokul öğrencisi üzerinde yaptıkları çalışmada öğretmen ve yöneticilerin disiplin kararı almalarını etkileyen öğrenci davranışlarını, öğrencilerin sosyal statüleri, ırk ve cinsiyetlerini esas alarak, regresyon modeli ve karar ağacı kullanarak incelemişlerdir [38].

Hollanda'da bulunan Ignatius Hastanesi; tedavi sürelerinin, belirli bir sürede tedaviye ihtiyacı olan kişi sayısının ve her bir hasta için tedavi süresinin tahmin edilmesi yönünde uygulamalar geliştirerek, yapılan analizler sonucunda elde ettiği bilgi ile hastanenin kadro ve kaynak ihtiyaçlarının doğru belirlenmesini sağlamış ve geçmiş hasta verilerinden elde ettiği bilgi ile kalp hastalıklarında bypass ameliyatlarının riskini minimuma indirmeyi başarmıştır [1].

ABD'ndeki HSBC bankası, VM tekniklerini kullanarak yaptığı çalışmada müşteri ihtiyaçlarını ve davranışlarını tespit etmiş ve doğru müşteriye doğru önerilerle giderek pazarlama maliyetlerinde % 30'luk bir azalma sağlarken, satışlarını % 50 arttırmıştır [1].

Portekiz'de bulunan Banco Espirito Santo bankası, VM çözümleri sayesinde kendisi ile çalışmayı bırakmaya meyilli müşteri profilini tanımlayabilmiştir. Dolayısıyla eldeki müşteriyi tutmaya ve yeni müşteriler edinmeye yönelik modeller geliştirerek, müşteri kaybı yüzdesini azaltırken, karını arttırmıştır [1].

ABD'nin en büyük kablosuz iletişim sağlayıcısı Verizon, kaybetme olasılığı yüksek olan müşterilerini ve müşteri kaybına neden olan faktörleri belirleme amaçlı bir VM çalışması yapmıştır. Bu çalışma ile birlikte müşteriler davranışlarına göre sınıflandırılmış, müşteri kaybına neden olan faktörler belirlenmiş ve müşterilere yönelik yeni pazarlama stratejileri geliştirilmiştir [1].

WEKA, RAPİD-MİNER, SAS, SPSS gibi birçok paket programda VM konusunda çalışmalar yapılmaktadır [59]. SPSS'in önderliğinde yapılan diğer VM çalışmalarından bazıları aşağıda ifade edilmiştir [7]:

İngiltere'de finansal hizmetler ve bankacılık alanında detaylı iş olanakları sağlayan Lloyds TSB bankası, İngiltere'nin lider finansal hizmetler kuruluşlarından birisidir. Kredi kartlarındaki hilekârlık artışı yüzünden Lloyds TSB, yüksek maddi kayıplara uğramış, hilekârlık tespiti yapacak olan bir grup kurarak, gerçek ve hileli harcama arasındaki ayrımı tespit eden bir model oluşturmuş ve yüksek olan maddi kayıplarını en aza indirmeyi başarmıştır.

Benzer şekilde Provident Finans, İngiltere'nin ev kredileri konusunda önde gelen finans şirketlerinden birisidir. Bu şirket de Lloyds TSB şirketi gibi hilekârlık nedeni ile ciddi maddi kayıplara uğramış ve aynı şekilde VM tekniklerini kullanarak hilekârlık yüzdesini aşağı çekmeyi başarmıştır.

İspanya'da bulunan Winterthur Insurance adlı sigorta şirketi, "kimler bizimle çalışmayı bırakabilir?" sorusuna cevap aramayı VM teknikleri ile bulmuştur.

Belçika'da bulunan bir sigorta şirketi olan Corona Direct; araba, eşya sigortası ve kiralama işlemleri gibi hizmetler sunmakta iken, VM tekniklerini kullanarak pazarlama kampanyalarında başarılı olmuş bir sigorta şirketidir.

ABD'nde bulunan Highmark Blue Cross Blue Shield sigorta şirketi, müşteri hizmetlerini geliştirmek için yine VM tekniklerini kullanmıştır.

İngiltere'de otomobil yedek parçaları satan bir marketler zinciri olarak faaliyet gösteren Halfords, müşterilerin satın alma davranışlarını analiz etmek için VM tekniklerini kullanmıştır.

Japonya'da 40 mağazası bulunan ve ülkenin en büyük kişisel bilgisayar ve yazılım perakendecisi olan Sofmap'ın yöneticileri, müşterilerinin birçoğunun donanım ve yazılım ürünleri satın alma kararında zorlandığını, bunun online (çevrimiçi) satışları engellediğini düşünerek, VM teknikleri ile müşteri profiline uygun satışlar için bir altyapı oluşturmuşlardır.

Almanya'nın en büyük ofis ürünleri üreticisi olan Herlitz AG, müşterilerinin satın alma alışkanlıklarını analiz etmek için VM tekniklerini kullanmıştır.

İngiltere'nin en büyük içecek üreticilerinden biri olan Withbread ile dünyada çok sayıda şarap kulübü olan ve şarap satışı yapan Direct Wines da VM tekniklerini kullanan üreticilerdendir.

ABD, kamu ve savunma sektöründe de yaygın olarak VM tekniklerini kullanmaktadır [7]:

- ABD Gelirler İdaresi Kurumu, düşük vergi bildiriminde bulunanların geri ödemelerinin tespit edilmesinde geçmiş ödeme kayıtlarını kullanarak bir model oluşturmuş ve eksik bildirimde bulunan kişilerin profilleri VM teknikleri kullanılarak belirlenmiştir.

- ABD hükümeti istihbarat teşkilatında, güvenlik ile ilgili her gün çok fazla veri toplanmakta ve toplanan verilerin analiz edilerek mevcut saldırıların belirlenmesi teşkilatta görev yapan insanların zekâlarına bağımlı idi. Teşkilatın şüpheli olan network faaliyetlerini belirlemesine yardımcı olacak bir sisteme ihtiyacı vardı ve bu ihtiyaç VM tekniklerini kullanılarak giderilmiştir.

Yine ABD'ndeki Sağlık Sigortası Hizmetleri Merkezi sponsorluğundaki Peer Review Organizasyonları ise; gereksiz yere kabul edilen, tahliye edilen, aynı gün tekrar yatan hasta kayıtlarını veya yanlış tanı kodlarını, ödeme hatalarını tespit etmek için VM tekniklerini kullanmışlardır.

West Midlands Polis Departmanı çözülmemiş kriminal davalarda, izleri ve eğilimleri bulmak için çabuk ve kolay bir yöntem geliştirmeyi amaçlıyordu. Azalan kaynaklar, yetersiz ipuçları ve eskiyen davalar suçlularla mücadeleyi zorlaştırmakta idi. Bu nedenle bilinen suç işlemelerle çözülememiş davaların eşleştirilmesi ve tekrar eden suçluları izleme ve yakalama için VM tekniklerini kullanmışlardır.

Michigan'da bulunan mobilya üreticisi Haworth Inc. ile motosiklet üreticisi Yamaha Motor Europe N.V da hedef kitlelerini belirlemek için, VM tekniklerini kullanan diğer firmalardır.

İtalya'da bulunan Boehringer Ingelheim ilaç firması ise, eczanelerden oluşan müşteri tabanını sınıflandırmak için VM tekniklerini kullanmıştır.

İngiltere'de bulunan St. George Hastanesi, yapılan çalışmalarla hastanede yoğun bakım ünitelerinden çıkartıldıktan sonra ölen hastaların % 39'unun, eğer 48 saat daha yoğun bakımda kalırlarsa, ölüm riskinin kalmayacağını tespit etti. Bilhassa İngiltere'de önceki yıllarda yaşanan grip salgını, yaşlı ve sağlıksız insanların bu krizden etkilenme oranındaki yükseklik, yoğun bakım ünitelerindeki yatak sayısının azlığını önemli bir konu olarak gündeme getirmiştir. Bu sonuçlar, St. Thomas hastanesinde görevli Kathleen Daly ve St. George hastanesinde görevli Rene Chang tarafından hazırlanan bir çalışma ile ortaya konulmuştur. Bu çalışma 1989–1998 yılları arasında, hastanelerin acil bölümlerine başvuran 14 000 hastanın verileri göz önüne alınarak yapılmış ve Mayıs 2001 tarihinde yayınlanmıştır. Kullanılan hipotez, başvuran hastaların üçte birinin, daha acil bölüme başvurdukları an diğerlerine göre 'riskli' olduklarının tespit edilebilir olduğudur. Bu çalışmanın sonucunda da riskli bölümde gösterilen hastaların dörtte birinin öldüğü, buna karşılık risksiz bölümde görülenlerin sadece % 4'nün öldüğü şeklinde bulunmuştur. Ayrıca hastanelerdeki yoğun bakım yatak sayısının % 16 oranında arttırılmasına ihtiyaç duyulduğu belirlenmiştir.

Enerji sektörünün önde gelen firmaları olan; Edf Energy, Almanya'daki Hamburgische Electricitats-Werke AG, Enel Gas ve İtalya'daki Kuwait Petroleum Italia şirketleri de VM tekniklerini kullanarak satış stratejilerini belirleyen diğer firmalardır.

Fortis, bankacılık ve sigorta konusunda çalışan uluslararası bir finansal hizmetler kuruluşu olup, VM teknikleri ile birçok projelerle müşterilerine daha etkin hizmet sağlayan kuruluşlardan biri olmuştur.

Ülkemizde de VM teknikleri kullanılarak farklı alanlarda yapılan çalışmalardan bazıları aşağıda özetlenmektedir:

Garanti Bankası, müşterilerine sunduğu hizmetleri daha iyi bir noktaya taşımak amacıyla VM tekniklerini kullanmıştır [1].

Türkiye İstatistik Kurumu (TÜİK), "2003 Yılı Hane Halkı Bütçe Anketi" çalışmasını VM tekniklerini kullanarak yapmışlardır [1].

Ulaş (2001), Gima Türk A.Ş. marketler zincirine bağlı bir marketin 2000 yılına ait Haziran, Temmuz ve Ağustos aylarında yapılan alışverişlere ilişkin verilere VM tekniklerini uygulamış ve en çok birlikte tüketilen ürünlerin analizini yapmıştır [73].

Yağız (2003), isnat[3] edilen çocuk suçları konusunu VM tekniklerinden CHAID algoritması ile incelemiştir [75].

Doğan ve Özdamar (2003), ailelerin çocuk isteğine etki eden faktörlere ulaşmada bağımsız değişkenlerin birleşmiş kategorilerini ve alt gruplarını VM tekniklerinden karar ağaçları algoritmalarından CHAID ile tahmin etmişlerdir [31].

Erdoğan (2004), Maltepe Üniversitesi öğrenci işleri veri tabanını kullanarak öğrencilere ilişkin veriler üzerinde VM tekniklerini uygulayarak, benzer özelliklere sahip olan öğrenci örüntüleri elde etmistir [33].

Sarıkan (2005), eczane-reçete-hastane üçgeninde VM tekniklerini kullanarak hilekârlık tespiti çalışması yapmıştır [58].

Dolgun (2006), Adana ilinde Groseri Market'in Mersin Pozcu Şubesine ait 2 Ocak 2006–9 Ocak 2006 tarihleri arasında yer alan 8 günlük veriler üzerinde VM

---

[3] Bir düşünceyi, bir konuyu bir kişi veya sebebe dayandırma, yükleme, atfetme

tekniklerini uygulayarak, ürünler arasındaki birliktelik kurallarını belirlemeye çalışmıştır [29].

Türe ve arkadaşları (2006), tümevarım tekniği kullanarak sağlık durumu, beslenme ve bazı diğer faktörlerin okul başarısızlığına etkisini karar ağaçları ile araştırmışlardır [71].

Özçakır ve Çamurcu (2007), bir firmanın pastane satış verileri üzerinde VM tekniklerini uygulayarak ürünlerin birliktelik kurallarını belirlemeye çalışmışlardır [54].

Kılıç ve arkadaşları (2007), osteoporoz (kemik erimesi) riskini VM tekniklerini kullanılarak belirlemeye çalışmışlardır [41].

Gürüler ve arkadaşları (2007), Muğla Üniversitesi İktisadi ve İdari Bilimler Fakültesi 1995 yılı ve sonrası öğrenci verileri üzerinde VM tekniklerini uygulayarak, öğrenci profilini belirlemeye çalışmışlardır [35].

Türe ve arkadaşları (2007), Ocak 1999–Şubat 2003 yılları arasında koroner arter hastalığından şüphelenilen geriye dönük 1381 hasta verileri üzerinde, koroner arter hastalığı gelişimini etkileyen faktörleri sınıflandırma önemine göre inceleyerek bu hastalığı etkileyen en önemli değişkenlerin sırasıyla; cinsiyet, yaş, diyabet, hiperkolesterolemi, ailede koroner arter hastalığı olma ve sigara kullanımı olduğunu tespit etmişlerdir [72].

Timör ve Şimşek (2008), ülkemizde perakende sektöründe faaliyet gösteren büyük bir market zincirine ait verileri kullanarak, müşterilerin satın alma davranışlarını etkileyen faktörleri karar ağaçları ile belirlemeye çalışmışlardır [68].

Şen (2008), İstanbul Güngören semtinde Gün-Bak toptan satış firmasının 5 Nisan 2008–21 Nisan 2008 tarihleri arasındaki veritabanı kayıtları üzerinde VM

tekniklerini uygulayarak, satılan ürünler arasında birliktelik kurallarını belirlemeye çalışmıştır [64].

Gencer ve arkadaşları (2008), İstanbul yolu üzerindeki bir kozmetik mağazanın ayrılma eğilimi gösteren müşteri kesitini VM teknikleri ile belirlemeye çalışmışlardır [36].

Koyuncugil ve Özgülbaş (2009), karar ağacı ve kümeleme tekniğini kullanarak hastanelerin kapasite kullanım oranlarını etkileyen faktörleri belirlemeye yönelik bir çalışma yapmışlardır [44].

Albayrak ve Yılmaz (2009), 173 işletmenin 2004–2006 yıllarına ait finansal göstergelerine VM tekniklerinden karar ağaçları uygulamış, sanayi ve hizmet sektöründe faaliyet gösteren firmaları birbirinden ayıran en önemli değişkenleri belirlemeye çalışmışlardır [76].

Türe ve arkadaşları (2009), meme kanserli hastalarda yinelemesiz sağkalım süresini etkileyen risk faktörlerinin belirlenmesinde karar ağacı yöntemlerinden C&RT, CHAID, QUEST, C4.5 ve ID3 ile Kaplan-Meirer analizini birlikte kullanmışlardır [70].

2.5.3. Veri madenciliğinin tarihçesi

VM, uzun bir geçmişi olan teknoloji alanının evrim geçirmiş hali olarak tanımlanabilir. Kökeni ilk sayısal bilgisayar olan ENIAC (**E**lectrical **N**umerical **I**ntegrator **A**nd **C**alculator)'a kadar dayanmaktadır. 1946 yılında geliştirilen ve bugün kullandığımız kişisel bilgisayarların atası olan ENIAC, ABD'li bilim adamları John Mauchly ve J. Presper Eckert tarafından, II. Dünya Savaşı sırasında ABD ordusu için geliştirildi. 30 tonluk ağırlığıyla 170 m² 'lik bir alanı kaplayan bu ilk bilgisayarın 64 sene içerisinde geçirmiş olduğu evrimin nihai boyutlarını şu anda masa üstünüzdeki bilgisayara bakarak anlamanız mümkündür.

İlk sayısal bilgisayar olan ENIAC'ın ortaya çıkışı ile birlikte, matematikçiler 1950'li yıllarda mantık ve bilgisayar bilimleri alanlarında çalışarak, yapay zekâ (Artificial İntelligence) ve makine öğrenme (Machine Learning) yaratmışlardır [42].

1960'lı yıllarda istatistikçiler yeni bir algoritma keşfetmişlerdir. Örneğin; regresyon analizi (regression analysis), en büyük olabilirlik kestirim (maximum likelihood estimates), sinir ağları (neural networks) vb. gibi metotlar VM'nin ilk adımlarını oluşturmuştur. Ayrıca veri tabanı sistemlerinin gelişmesi ile birlikte büyük sayıda metin dökümanların saklanması ve bilginin geri kazanılması sağlanmıştır [42].

1970, 1980, 1990'lı yıllarda yeni programlama dilleri ve yeni bilgisayar tekniklerinin geliştirilmesi genetik algoritmalar, "genetic algorithms", "EM algorithms", "K-means clustering" ve "decision tree algorithms" gibi algoritmaları da içermiştir [42].

1990 yılı ile beraber veri tabanında bilgi keşifinin ilk adımları oluşturulmuş ve büyük veri tabanları için VA geliştirilmiştir. Ayrıca aynı zaman içerisinde yeni teknolojilerle beraber VM değiştirilerek yaygın olarak kullanılan standart bir işin parçası olmuştur [42].

VM ile ilgili yapılan ilk uygulamanın ise Pazar Sepet Analizi olduğu bilinmektedir. Bazı kaynaklar VM'ni ilk olarak 1980 yılında Londra'da John Graunt adında bir kişinin Doğal ve Siyasi Gözlemler Mortalite Bonosu çalışması üzerine uyguladığını söylemektedir. John Graunt'un o yıllarda ölüm verileri ile ilgili ayrıntılı bir analiz yaparak, model oluşturup bu konuda çalıştığını ve daha sonra da bu modele şehirlerdeki veba hastalığından ölenlerin verilerini kullanarak tahminlerde bulunduğu söylenmektedir.

2.5.4. Veri madenciliği ile istatistik uygulamaları arasındaki fark

VM, klasik istatistiksel uygulamalara çok benzerdir, ancak klasik istatistiksel uygulamalar yeterince düzenlenmiş ve çoğunlukla özet veriler üzerinde çalıştırılır ve

analiz edilen veri sayısı binler, yüz binler iken VM'nde bu sayı milyon hatta milyarlar seviyesinde olmaktadır. Dolayısıyla değişken sayısı da çok fazla olduğundan klasik istatistiksel yöntemler, bu verileri analiz etmeye yeterli gelmemeye ve yeni tekniklere ihtiyaç duyulmaya başlanmıştır.

VM ile istatistik uygulamaları arasındaki en önemli iki farktan birincisi; istatistik çok büyük veriler üzerinden analiz yapmaz, halbuki VM terabaytlarla ölçülen milyarlarca veri üzerinden hesaplamalar yaparak ortaya çıkmamış, gizli kalmış bilgileri, desenleri açığa çıkarmaya çalışır.

İkinci fark ise; istatistikte veriler akıldaki bazı sorular için toplanır, anketler yapılır ve bu sorulara yanıt bulmak için veriler analiz edilir. VM ise, geçmiş verileri kullanarak geleceğe yönelik hiç akla dahi gelmeyecek örüntüleri, gizli kalmış bilgileri keşfetmeye yöneliktir.

VM, klasik istatistik temel olmak üzere yapay zeka, makine öğrenimi ve örüntü tanıma gibi pek çok disiplinin kesişimi olan bir alandır. Her ne kadar klasik istatistik temeline dayansa da, Moss and Atre (2003) istatistiksel analizler ve VM'nin karşılaştırması ve farklılaştığı noktaları Çizelge 2.1.'de vermişlerdir [39].

Çizelge 2.1. Veri madenciliği ile istatistiksel analiz arasındaki fark

İstatistiksel Analiz	**Veri Madenciliği**
İstatistikçiler genellikle bir hipotez ile başlarlar.	Veri madenciliği hipoteze gerek duymaz.
İstatistikçiler hipotezlerini eşleştirmek için kendi eşitliklerini geliştirmek zorundadırlar.	Veri madenciliği algoritmaları eşitlikleri otomatik olarak geliştirir.
İstatistiksel analizler sadece sayısal verileri kullanır.	Veri madenciliği farklı tiplerde data kullanır (örneğin metin, ses) sadece sayısal veriyi değil.
İstatistikçiler kirli veriyi analizleri sırasında bulur ve filtre ederler.	Veri madenciliği temiz veriye dayanır.

2.5.5. Veri madenciliği ile veri ambarı arasındaki fark

Veri Ambarları, farklı veri tabanlarında tutulan verilerin bir araya getirilmesi, toplanması ile oluşturulan bütünleştirilmiş veri depolarıdır. VA kurumlarda var olan, depolanan verileri yönetmek için, VM ise daha çok karar destek sistemleri için verilerden anlamlı bilgiler çıkartarak bu bilgileri üst yönetime hızlı ve doğru bir biçimde sunmak amacı ile kurulurlar.

Veri Ambarlarında veriler içerisinde çok sayıda eksik, hatalı, gürültülü veri mevcuttur ve bu veriler üzerinde kullanıcılar sorgular yaparak raporlar oluşturabilir. VM ise bu tutarsız, hatalı, gürültülü verilerle çalışamadığından verilerin mutlaka ön işlemden geçirilmesi gerekmektedir.

2.5.6. Veri madenciliğinin yaygın olarak kullanıldığı sektörler

VM, çok geniş bir kullanım alanına sahiptir. Bu alanlar başlıklar halinde aşağıda ifade edilmiştir [1,59].

- *Finans Sektörü*

Finans sektöründe de diğer sektörlerde olduğu gibi mevcut olan müşteriyi elde tutma, müşteri memnuniyetini sağlama, devam ettirme ve buna göre müşteriye doğru zamanda alternatif kampanyalar sunma çabası vardır. Bu nedenle bu sektördeki müşterilerin geçmiş özellikleri incelenip, müşteri profilleri çıkarılarak ileriki zamanlarda nasıl bir yol izlenileceğine karar verilir.

Finans kuruluşları VM tekniklerini kullanarak henüz dolandırıcılık olayı daha meydana gelmeden yapılması muhtemel olan dolandırıcılık olaylarını önleyebilmektedirler. Bu işlem; geçmişte yapılan dolandırıcılık kayıtları esas alınarak bir model geliştirilmesi ve yeni gelen kayıtların bu modele uygulanması ile dolandırıcılıklara karşı önceden önlem alınması şeklinde işlemektedir.

- *Haberleşme Sektörü*

Telekomünikasyon sektöründe de en büyük sorun müşteri kaybıdır. Eğer kuruluşlar müşteri profillerini belirleyerek, hangi müşterilerini kaybedeceklerini önceden bildikleri takdirde, düşük maliyetli etkili ileriye dönük stratejiler geliştirip, kampanyalar düzenleyerek mevcut olan müşterilerini elde tutmayı başarabilirler. Örnek verecek olursak; ABD'nin en büyük kablosuz iletişim sağlayıcısı Verizon, müşteri kaybına neden olan faktörleri belirleme amaçlı bir VM çalışması yapmıştır. Müşteriler, müşterilerin yapmış oldukları aramalar (ev ve cep telefonu ile yapılan), attıkları mesajlar, hat bilgileri, kontör yükleme zamanları, fatura bilgileri ile ilgili bilgilerin hepsi veri tabanında tutulan kayıt altına alınmış bilgilerdir. Dolayısıyla bu konuya odaklı bir VA'nda, OLAP sorgulamaları ile birçok kural çıkartılabilir ve bu kurallara göre kampanyalar düzenlenebilir.

- *Sağlık Sektörü*

VM tekniklerinin kullanıldığı en önemli sektörlerden birisi, insan sağlığı odaklı olması nedeniyle sağlık sektörüdür. Hastane bünyelerinde de; fatura bilgileri, cerrahi işlem verileri, kişisel sağlık kayıtları, reçete kayıtları gibi birçok veri toplanmaktadır. Bu veri yığınlarından ileriye dönük tespitler yapılarak hastalıkların erken teşhisi, yapılan harcamalar ile ilgili anormal olan durumlar (Hastane–Doktor–İlaç yolsuzluğu), hastalıklara neden olan faktörler, ilaç reaksiyonları, insan sağlığı ile sosyal ve çevresel sorunların korelasyonları tespit edilebilir. Ayrıca insan vücudunda bulunan genlerin diziminden çeşitli hastalıkların (kanser gibi) ortaya çıktığı bilinmekte olup, bu genlerin dizim kuralları, VM teknikleri ile araştırılarak hangi genlerin hangi hastalıklara sebep olduğu da araştırılmaktadır. Yapılacak olan çalışmaya göre örnekleri arttırmak mümkündür. Örneğin; San Francisco Hearth Institute, hasta sonuçlarının iyileştirilmesi, hastaların hastanede kalma sürelerinin azaltılması vb. gibi amaçlarla bir çalışma başlatmış ve kurum bünyesindeki hastaların geçmiş verilerini VM teknikleri ile kullanarak bu verileri bilgiye dönüştürmüştür.

Diğer bir örnek; Down Sendromu gibi bazı hastalıkların % 100 kesin teşhisi konulamaz. Anne karnındaki bebeğin bu hastalığı taşıyıp taşımadığı ancak anne karnından alınacak olan sıvı ile tespit edilmekte fakat bu da anne karnındaki bebek için bir tehlike arz etmektedir. İşte VM bu aşamada devreye girer ve daha önceden bu işlem uygulanmış olan annelerin veri tabanlarındaki geçmiş verilerine VM teknikleri uygulanarak, doktorların daha sonra gelecek olan veriler ile hastalarına daha kolay teşhis koymasını mümkün kılar. Bu tür çalışmalar tıp alanında insan hayatı için çok önemli bir yer teşkil etmektedir.

- *Devlet Uygulamaları*

Karayollarında bölgelere ve zamana göre yapılan kazalar tespit edilerek bir model oluşturulup, kaza oranları asgariye indirilebilir.

Emniyet birimlerinde, suç işleyenlerin profilleri çıkarılarak buna uygun politikalar geliştirilebilir.

e-Devlet uygulamalarında ise, geçmişteki ziyaretçilerin sayfaları nasıl kullandığı incelenerek, buna göre kurumların web sayfalarının vatandaşın ihtiyacına göre yeniden düzenlenmesi VM teknikleri kullanılarak sağlanılabilir.

Ayrıca Hükümetler istatistik, nüfus sayımı ve vergilendirme durumları içinde VM tekniklerini kullanmaktadırlar.

- *Pazarlama Sektörü*

VM pazarlama sektöründe de birçok amaç için kullanılmaktadır [59];

Bu amaçlardan bir tanesi müşterilerin satın alma örüntülerini belirlemektir. Mesela, bir bilgisayar firmasından DVD ve yazıcı alan müşterinin sonraki haftalarda boş DVD ve DVD zarfı aldığı da görülmüştür. Doğal olarak DVD alan müşterinin boş DVD ile birlikte DVD zarfı de alması işletmeler için çok önemli bir bilgi keşfi

olmayıp normal bir süreç olarak değerlendirilir. Fakat DVD alan müşterinin elektronik ürün dışında başka bir ürün alması işletmeler için ilginç bir örüntü olarak değerlendirilebilir.

Müşterilerin demografik özellikleri (yaşı, cinsiyeti gibi.) ile satın aldıkları ürünler arasında bağıntı kurar.

Yapılacak olan kampanyaların daha etkili hale getirilmesi için, daha önce yapılan kampanyalara katılan müşterilerin profilleri belirlenerek, o profile uygun müşteriler için kampanyalar düzenlenmesi.

Mevcut olan müşterileri elde tutmak, kaybedilen müşterilerin profillerini çıkarmak ve ona göre kampanyalar düzenlemek.

Pazarlama sektöründe en çok kullanılan uygulamalardan biri de Pazar Sepet Analizi olup, genellikle marketlere veya büyük alışveriş merkezlerine giden müşterilerin aldıkları ürün ile beraber başka hangi ürünü de aldıkları, VM algoritmaları ile analiz edilir ve marketler bu şekilde en çok beraber satılan ürünleri aynı reyona koyarak satışlarını belli bir oranda arttırma imkânına sahip olurlar.

- *Eğlence Sektörü*

Bu sektörde de şirketler veri tabanlarındaki geçmiş fatura bilgilerinden yola çıkarak; hangi aktörün, hangi film türlerinin hangi bölgelerde daha çok izlendiğini tespit etmesi ile bölgelere göre yeni film projelerini başlatabilirler.

- *Bankacılık Sektörü*

Değişen ve zorlaşan yaşam koşulları insanların kredi kartı kullanarak geçimlerini idame ettirmeye zorladığından, bankacılık sektöründe de çok büyük bir veri topluluğu bulunmaktadır. Bankalar kredi kartı harcamalarına göre, hangi kredi kartını hangi müşterilerin kullandığını belirleyen bir çalışma ile geleceğe yönelik daha etkili

kampanyalar düzenleyebilirler. Ayrıca kredi kartı dolandırıcılık ihtimali taşıyan müşterilerini önceden yapılan bir VM çalışması ile belirleyerek, ona göre kredi kartı taleplerini değerlendirebilirler.

- *Astronomi Sektörü*

VM teknikleri görüntü, teleskop ve uydulardan gelen diğer veriler arasındaki nadir bulunan hatta önceden bilinmeyen astronomik ilişkilerin belirlenmesinde de kullanılır [6].

2003 yılında VM'nin sektörler bazında kullanımına ilişkin bir araştırmanın sonuçları Çizelge 2.2.'de gösterilmektedir [11, 14, 43]. Bu çizelgede 421 şirket VM tekniklerini kullanarak araştırma yapmıştır. Örneğin, Bankacılık sektöründe VM teknikleri ile 51 araştırma, Biyoteknoloji/Genomik sektöründe 11 araştırma yapılmıştır.

Çizelge 2.2. Veri madenciliğinin uygulandığı alanların dağılımı

Bankacılık (51)	12%
Biyoteknoloji / Genomik (11)	3%
Kredi Puanlama (35)	8%
CRM (52)	12%
Doğrudan Pazarlama / Fundraising (34)	8%
e-Ticaret (11)	3%
Eğlence / Müzik (4)	% 1
Sahtecilik Algılama (31)	7%
Kumar (2)	0%
Hükümet uygulamaları (12)	3%
Sigorta (24)	% 6
Yatırım / Hisse (5)	% 1
Önemsiz e-posta / Anti-spam (5)	% 1
Sağlık / HR (15)	% 4
İmalat (19)	5%
Medikal / İlaç (12)	3%
Perakende (25)	% 6
Bilim (17)	% 4
Güvenlik / Anti-terör (5)	% 1
Telekom (23)	5%
Seyahat / Hospitality (8)	% 2
Web (9)	% 2
Diğer (11)	3%

2.5.7. Veri madenciliği modelleri

VM'nde modeller, kullanılan algoritmalara dayalı olarak üç grupta toplanmıştır [42].

- *Tahmin edici modeller (Değer tahmini modeli)*

Bu tür modellerde, sonuçları bilinen verilerden hareket edilerek bir model geliştirilmesi ve kurulan bu modelden yararlanılarak sonuçları bilinmeyen veri kümeleri için sonuç değerlerin tahmin edilmesi amaçlanmaktadır.

Yani, kullanılan yazılım veri tabanını bir bütün olarak ele alır ve sonuçlar çıkarır. Örnek olarak, "bir sigortalıdan ne kadar prim toplayabilirim?", "yapılan işlemde veya hastaneye alınan sağlık malzemelerinde dolandırıcılık yapılıyor mu?" gibi sorulara cevap aranır.

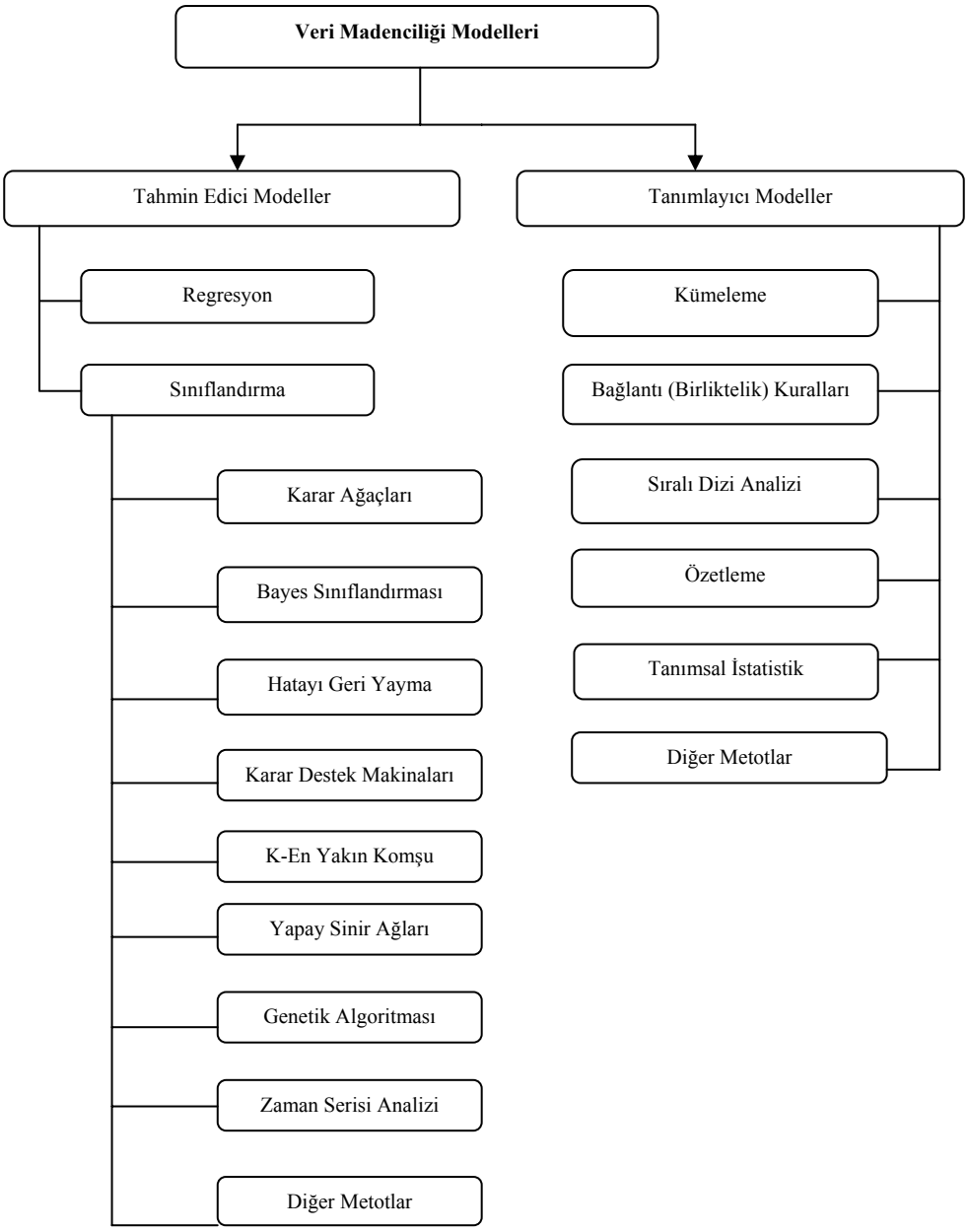
- *Tanımlayıcı modeller*

Bu modelde ise, karar vermeye rehberlik etmede kullanılabilecek, daha çok veriler arasındaki gizli kalmış bilgiler açığa çıkartılmaya çalışılır. Örnek olarak markette, çocuk bezi alan müşterilerin % 30'u bira da satın alır, gibi gizli bilgileri açığa çıkarmaya yönelik modellerdir. Birliktelik kuralı tanımlayıcı bir model olarak tanımlanır.

- *Her iki yaklaşımı da içeren modeller*

Bazen yukarıda belirtilen modeller tek başına yeterli gelmeyip, bu durumda her iki modeli de kullanan algoritmalara ihtiyaç duyulabilir.

Aşağıda VM'nde kullanılan modeller ve yöntemler (metotlar) Şekil 2.9. 'da ifade edilmiştir [42].



Şekil 2.9. Veri madenciliği modelleri ve uygulanan metotlar

VM modellerini gördükleri işleve göre; sınıflandırma ve regresyon modelleri, kümeleme modelleri ve birliktelik kuralları olmak üzere üç ana başlık altında incelemek mümkündür [53]. Sınıflandırma ve regresyon modelleri tahmin edici, birliktelik kuralları ise tanımlayıcı bir modeldir.

Birliktelik Kuralı

Birliktelik kuralı belirli türdeki veri ilişkilerini tanımlayan tanımlayıcı bir modeldir [59]. Veri tabanı içerisindeki kayıtların birbirleriyle olan ilişkilerini inceleyerek, hangi olayların eş zamanlı olarak birlikte gerçekleşebileceğini ortaya koymaya çalışır. Birliktelik Kuralında öğeler arasındaki bağıntı, destek ve güven kriterleri ile hesaplanır. Destek kriteri, veride öğeler arasındaki bağıntının ne kadar sık olduğunu, güven kriteri ise Y öğesinin hangi olasılıkla X öğesi ile beraber olacağını söyler [3]. Genellikle perakendecilik sektöründe kullanılır. Hangi ürünlerin birlikte alındığının bağıntısı bulunur. Örneğin; bir market geçmiş satışlarını inceleyerek traş bıçağı yanında pil de alındığı bilgisini çıkarsadıysa, bu iki reyonu yan yana getirerek satışlarını arttırabilir.

Sınıflandırma ve regresyon

Sınıflandırma ve regresyon tahmin edici bir modeldir. Sınıflandırma kategorik değerleri tahmin ederken, regresyon süreklilik gösteren değerlerin tahmin edilmesinde kullanılır [64].

Verilerin sınıflandırılması için belirli bir süreç izlenir. Öncelikle var olan veri tabanının bir kısmı eğitim amacıyla kullanılarak sınıflandırma kurallarının oluşturulması sağlanır. Daha sonra bu kurallar yardımıyla yeni bir durum ortaya çıktığında nasıl karar verileceği belirlenir. Temeli öğrenme algoritmasına dayanır. Veri tabanından rastgele seçilecek olan örnek veriler üzerinde test edilerek bir sınıflandırma modeli oluşturulur. Test verileri üzerinde sınıflandırma kuralları belirlenir ve örnek veri test edilir. Elde edilen modelin doğruluğu kabul edilirse, bu model diğer veriler üzerinde de uygulanır.

Genellikle resim, örüntü tanıma, hastalık tanıları, dolandırıcılık tespiti, kalite kontrol çalışmaları ve pazarlama konularında sınıflama teknikleri kullanılır.

Sınıflandırma tekniklerinde kullanılan algoritmalar; Bayesyen Sınıflandırma Algoritması, Karar Ağaçlarına Dayalı Algoritmalar, Yapay Sinir Ağları, K - En Yakın Komşu Algoritması gibi birçok algoritma kullanılmaktadır. Bankacılık sektöründeki kullanımı için bir örnek verecek olursak; veri tabanındaki bankanın müşterilerine vermiş olduğu kredi verileri ile karar ağaçları algoritması kullanılarak bir sınıflandırma yapılabilir. Ve bankaya yeni bir kredi talebi olduğunda bu karar ağacı algoritması ile birtakım kurallar oluşturularak, yeni müşterinin verileri karar ağacı algoritmasına girilerek müşteriye kredi verilip verilmeyeceği değerlendirilebilir.

Kümeleme

Çok değişkenli istatistiksel tekniklerden birisi olan kümeleme analizi, grup sayısı bilinmeyen ve gruplandırılmamış verilerin benzerliklerine göre sınıflandırılması amacıyla kullanılmaktadır [2]. Verilerin modellenmesinde kümeleme analizinin çok önemli bir yeri vardır. Verilerin kendi aralarındaki benzerlikleri göz önüne alınarak gruplandırma yapılır. Burada benzerlikten kasıt veriler arasındaki mesafe, uzaklık kastedilmektedir. Bilgisayar bilimlerinde; desen tanımlama, resim işleme, uzaysal harita verilerinin analizinde kullanılmaktadır. İstatistikte ise; çok değişkenli istatistiksel tahmin ve örüntü tanıma analizlerinde kullanılmaktadır. Ayrıca internet üzerinde web sayfalarının aranması, DNA analizi, coğrafi bilişim sistemi gibi alanlarda da kullanılmaktadır.

Kümeleme, denetimsiz bir öğrenmedir. Çünkü önceden belirlenmiş sınıflar yoktur, zaten sınıflar önceden belli olsaydı bu kümeleme değil, sınıflandırma olurdu.

2.5.8. Veri madenciliğinde kurulan modelin değerlendirilmesi

Verilerin önişlemden geçirilmesi aşamasından sonraki adım modelleme adımıdır. 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 [16].

VM'nde kullanılacak olan en uygun model belirlendikten sonra, bu modelin doğruluğunun test edilmesi gerekir. Bir modelin doğruluğunun test edilmesinde kullanılan en basit yöntem *basit geçerlilik testi (Simple Validation)* 'dir. Bu yöntemde verilerin % 5 ile % 33 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 hesaplanır (Doğruluk oranı=1–Hata Oranı).

Sınırlı miktarda veriye sahip olunması durumunda kullanılabilecek diğer bir yöntem, *çapraz geçerlilik testi (Cross Validation)'*dir. Bu yöntemde veri kümesi rastgele iki eşit parçaya ayrılır. İlk aşamada bir parça üzerinde model eğitimi ve diğer parça üzerinde test işlemi; ikinci aşamada ise ikinci parça üzerinde model eğitimi ve birinci parça üzerinde test işlemi yapılarak elde edilen hata oranlarının ortalaması kullanılır.

Bir kaç bin veya daha az satırdan meydana gelen küçük veri tabanlarında, verilerin *n* gruba ayrıldığı *n katlı çapraz geçerlilik testi* (*n Fold Cross Validation)* tercih edilebilir. Verilerin örneğin 10 gruba ayrıldığı bu yöntemde, ilk aşamada birinci grup test, diğer grupların öğrenim amaçlı kullanılması ile sürdürülür. Sonuçta elde edilen on hata oranının ortalaması, kurulan modelin tahmini hata oranı olacaktır [41].

2.5.9. Veri madenciliğinde kullanılan yöntemler

VM'nde kullanılacak olan modele göre uygulanan farklı yöntemler mevcuttur. [59].

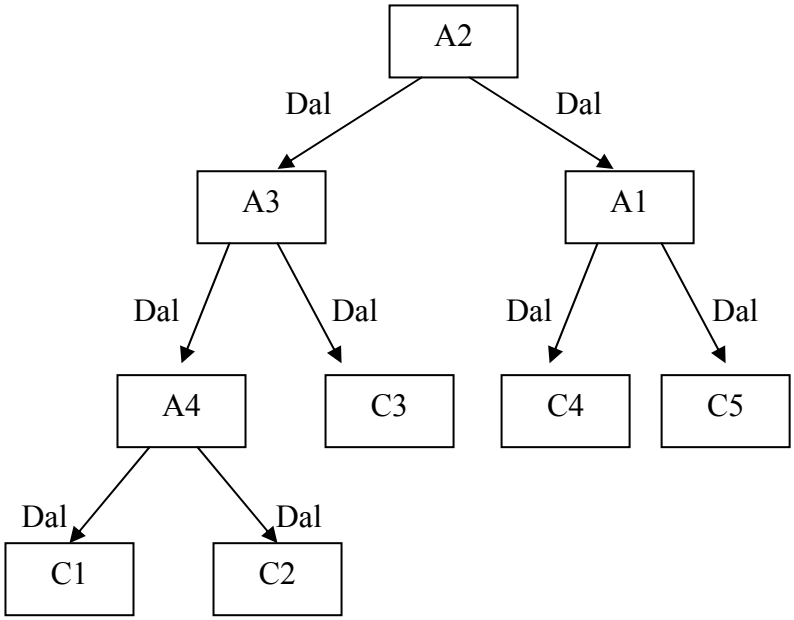
Sınıflandırma ve regresyon modelinde kullanılan yöntemler:

- *Karar Ağaçları (Decision Trees)*

Karar Ağaçları, kurgulanmasının, yorumlanmasının ve veri tabanları ile entegrasyonun kolaylığı nedeniyle sınıflandırma problemlerinde en yaygın kullanılan

ve adından da anlaşılacağı üzere ağaç görünümünde olan yöntemlerden bir tanesidir [76]. Bir karar ağacı algoritmasının prensipte görevi veriyi özyinelemeli olarak alt veri gruplarına dallanma yaparak bölmektir. Bu ayrım aşamasında oluşan her yeni dal bir kuralı ifade etmektedir [8]. Temel olarak iki adımdan oluşur. Birinci adım ağacın oluşturulması, diğer adım ise veri tabanındaki her bir kaydın bu ağaca uygulanarak verilerin sınıflandırılmasıdır.

Bir karar ağacının yapısı Şekil 2.10.'da gösterilmiştir.

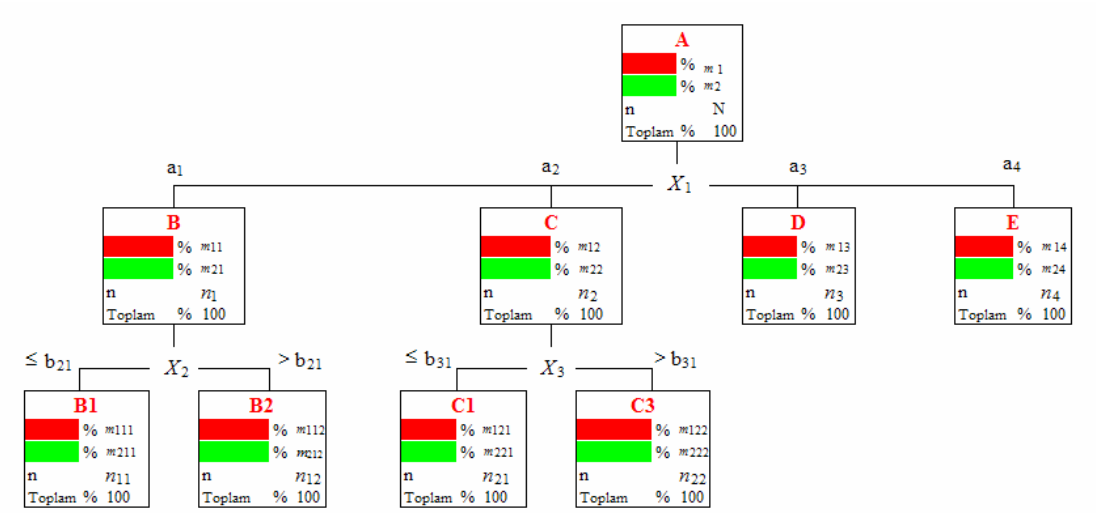


Şekil 2.10. Karar ağacı yapısı

Karar ağaçlarında her bir nitelik bir düğüm tarafından temsil edilir. Şekil 7.2'de düğümler A harfi ile ifade edilmiş ve her bir düğüm de kendinden sonra iki dala ayrılmıştır. Bu ayrılma sürecinde, A düğümü hakkında cevabı veri tabanında bulunacak bir soru sorulmakta ve verilen yanıta göre bir dal izlenmektedir. Dalın sonucunda eğer bir sınıflandırma elde edilebiliyorsa C ile gösterilen yapraklar elde edilmiş olur ve her biri bir sınıfı temsil etmektedir. Kısaca en üst yapı kök, en son yapı yaprak ve bunların arasında kalan yapılar ise dal olarak adlandırılır [59].

Karar Ağaçlarında uygulanacak algoritmaya göre ağacın şekli değişir, dolayısıyla da değişik ağaç yapıları değişik sınıflandırmalar meydana getirir. Kök düğümü olarak ifade edilen A'nın farklı olması, en uçtaki yaprağa ulaşırken izlenecek yolu ve dolayısıyla da sınıflandırmayı değiştirir. Bu nedenle kök düğümün hangi nitelikten dallara ayrılacağı çok önemlidir. Kök düğümün dallara ayrılması işleminde veri tabanında bulunan cevaplar evet/hayır biçiminde ise iki dala, evet/hayır/belki biçiminde ise üç dala ayrılması istenilir.

Karar ağaçlarında dallanmanın yapılacağı niteliği bulmak için entropi kavramından yararlanılır. Entropi, veri tabanındaki verinin sayısallaştırılması, verideki belirsizliğin ölçülmesidir ve 0–1 arasında değer alır. Her değişken için entropi kavramından yararlanarak değişkenlerdeki belirsizlik ölçülür. Sonraki aşamada ise her değişken için kazançlar hesaplanır ve hangi kazanç yüksek ise dallanma o değişkenden yapılır.



Şekil 2.11. Karar ağacı değişken ilişkisi

Karar ağacı değişken ilişkisi Şekil 2.11'de gösterilmektedir [39]. Hedef değişken Y olmak üzere;

- X_1, X_2, X_3 yani sadece üç değişken hedef Y değişkeni ile istatistik olarak önemli ilişkiye sahiptir.

- X_1 değişkeni, Y hedef değişkeni ile istatistik olarak en önemli ilişkiye sahiptir.

- X_2 değişkeni, X_1 değişkeni ile $X_1 = a_1$ olması koşuluyla istatistik olarak önemli ilişkiye sahiptir.

- X_3 değişkeni, X_1 değişkeni ile $X_1 = a_2$ olması koşuluyla istatistiksel açıdan önemli ilişkiye sahiptir.

Karar ağacı oluşturmak için geliştirilen algoritmalar arasında CHAID, Exhaustive CHAID, CART, ID3, C4.5, MARS, QUEST, C5.0, SLIQ, SPRINT yer almaktadır [76].

Bu çalışmada, söz konusu algoritmalardan CHAID algoritması kullanılmış olup, bir sonraki bölümde ayrıntılı olarak anlatılmıştır.

- *Bellek Tabanlı Sınıflandırma*

Bellek tabanlı veya örnek tabanlı bu yöntemler istatistikte 1950'li yıllarda önerilmiş olmasına rağmen o yıllarda gerektirdiği hesaplama ve bellek yüzünden kullanılamamış ama günümüzde bilgisayarların ucuzlaması ve kapasitelerinin artmasıyla, özellikle de çok işlemcili sistemlerin yaygınlaşmasıyla kullanılabilir olmuştur [1].

- *Yapay Sinir Ağları (Artifical Neural Networks)*

Yapay Sinir Ağları (YSA), temelde tamamen insan beyni örneklenerek geliştirilmiş bir teknolojidir. Bilindiği gibi; öğrenme, hatırlama, düşünme gibi tüm insan davranışlarının temelinde sinir hücreleri bulunmaktadır. İnsan beyninde tahminen 10^{11} adet sinir hücresi olduğu düşünülmektedir ve bu sinir hücreleri arasında sonsuz diyebileceğimiz sayıda sinirler arası bağ vardır. Bu sayıdaki bir birleşimi gerçekleştirebilecek bir bilgisayar sisteminin dünya büyüklüğünde olması gerektiği

söylenmektedir. Sinir ağları iki ya da üç katmandan oluşur. Bu katmanlar girdi, gizli ve çıktı katmanlarıdır [77].

- *Genetik Algoritmalar (Genetic Algorithms)*

Genetik Algoritmalar, benzer bir şekilde çalışan arama ve en iyileme yöntemidir. Problemlere tek bir çözüm üretmek yerine farklı çözümlerden oluşan bir çözüm kümesi üretir. Çözüm kümesindeki çözümler birbirinden tamamen bağımsızdır. Her biri çok boyutlu uzay üzerinde bir vektördür [77].

- *K–En Yakın Komşu (K–Nearest Neighbor)*

VM'nde sınıflama amacıyla kullanılan bir diğer teknik ise örnekleme yoluyla öğrenmeye dayanan k-en yakın komşu algoritmasıdır. Bu teknikte tüm örneklemler bir örüntü uzayında saklanır. Algoritma, bilinmeyen bir örneklemin hangi sınıfa dâhil olduğunu belirlemek için örüntü uzayını araştırarak bilinmeyen örnekleme en yakın olan *k* örneklemi bulur. Yakınlık öklid uzaklığı ile tanımlanır. Daha sonra, bilinmeyen örneklem, *k* en yakın komşu içinden en çok benzediği sınıfa atanır [77].

- *Lojistik Regresyon*

Lojistik regresyon analizinin kullanım amacı istatistikte kullanılan diğer model yapılandırma teknikleriyle aynıdır. En az değişkeni kullanarak en iyi uyuma sahip olacak şekilde sonuç değişkeni (bağımlı ya da cevap değişkeni) ile bağımsız değişkenler kümesi (açıklayıcı değişkenler) arasındaki ilişkiyi tanımlayabilen ve genel olarak kabul edilebilir modeli kurmaktır.

Son yıllarda yoğun bir şekilde kullanılan lojistik regresyon analizi, gözlemlerin gruplara atanmasında sık kullanılan üç yöntemden (diğerleri kümeleme analizi ve diskriminant analizi) birisidir. Kümeleme analizinde gözlemlerin atanacağı küme sayısı tam bilinmezken, diskriminant ve lojistik regresyon analizinde grup sayısı

bilinmekte, mevcut veriler kullanılarak bir ayrımsama modeli elde edilmekte ve kurulan bu model yardımıyla veri kümesine eklenen yeni gözlemlerin gruplara atanması mümkün olabilmektedir.

- *Naïve - Bayes*

Naive bayes algoritmasında her kriterin sonuca olan etkilerinin olasılık olarak hesaplanması temeline dayanmaktadır. VM işlemini en çok verilen örneklerden biri ile açıklayacak olursak elimizde tenis maçının oynanıp oynanmamasına dair bir bilgi olduğunu düşünelim. Ancak bu bilgiye göre tenis maçının oynanması veya oynanmaması durumu kaydedilirken o anki hava durumu, sıcaklık, nem ve rüzgâr durumu bilgileri de alınmış olsun. Biz bu bilgileri değerlendirdiğimizde varsayılan tahmin yöntemleri ile "hava bugün rüzgârlı, tenis maçı bugün oynanmaz" şeklinde kararları farkında olmasak'da veririz. Ancak, VM bu kararların tüm kriterlerin etkisi ile verildiği bir yaklaşımdır. Dolayısıyla biz ileride öğrettiğimiz sisteme "bugün hava güneşli, sıcak, nemli ve rüzgâr yok" şeklinde bir bilgiyi verdiğimizde sistem eğitildiği daha önce gerçekleşmiş istatistiklerden faydalanarak tenis maçının oynanma ve oynanmama ihtimalini hesaplar ve bize tahminini bildirir [77].

Kümelemede kullanılan yöntemler [42] :

- *Bölme Yöntemleri (Partitioning methods)*

  (Veriyi bölerek, her grubu belirlenmiş bir kritere göre değerlendirir.)

- *Hiyerarşik Yöntemler (Hierarchical methods)*

  (Veri kümelerini ya da nesneleri önceden belirlenmiş bir kritere göre hiyerarşik olarak ayırır.)

- *Yoğunluk Tabanlı Yöntemler (Denisty-base methods)*

  (Nesnelerin yoğunluğuna göre kümeleme oluşturur.)

- *Izgara Tabanlı Yöntemler (Grid-based methods)*

- *Model Tabanlı Yöntemler (Model-base methods)*

  (Her kümenin bir modele uyduğu varsayılır ve bu modellere uyan veriler gruplandırılır.)

CHAID algoritması

CHAID algoritması, 1980'de Kaas tarafından en iyi bölmeyi hesaplamak için istatistik olarak anlamlı bir farklılığın olmadığı, hedef değişkene uyan çiftlerde tahmin değişkeninin olası kategori çiftini birleştirmesiyle oluşturulmuştur. En uygun bölümleri seçmek için kullanılan *entropi* veya *gini metrikleri*[4] yerine *ki-kare (chi-square)* testi kullanılmaktadır [69]. En iyi bölmeyi hesaplamak için tahmin değişkenleri hedef değişkene uyan bir çiftin içinde istatistik olarak anlamlı bir fark kalmayıncaya kadar birleştirilmektedir. Bu analiz sonucunda bağımlı değişken üzerine etkisi istatistiksel olarak önemli bulunan bağımsız değişkenler içinde en yüksek *F* değerine sahip olan değişken, CHAID diyagramında ilk sırayı almaktadır [39].

CHAID algoritması, karar ağaçları algoritmaları içinde sürekli ve kategorik tüm değişken tipleri ile çalışabilmesi nedeni ile yaygın olarak kullanılmaktadır [20, 31, 32, 48]. Bu yöntemde, sürekli tahmin edici değişkenler bu uygulama tarafından otomatik olarak analizin amacına uygun olarak kategorize edilmektedir.

CHAID, *ki-kare* metriği vasıtasıyla ilişki düzeyine göre farklılık rastlanan grupları ayrı ayrı sınıflamakta ve ağacın yaprakları, ikili değil, verideki farklı yapı sayısı kadar dallanmaktadır.

---

[4] İkili bölünmelere dayalı bir sınıflandırma yöntemi

Karar ağacının oluşturulmasında kullanılan CHAID algoritmasının işleyişi aşağıda adımsal olarak ifade edilmiştir.

1.Adım: Her bir tahmin edici değişken *X* için, *X*'in, *Y* hedef değişkenini dikkate alan en az öneme sahip kategori çiftini bul (en büyük *p* değerine sahip).

Yöntem, *Y*'nin ölçüm düzeyine bağlı olarak *p* değerlerini hesaplayacaktır.

a. Eğer *Y* sürekli ise *F* testini[5] kullan.

b. Eğer *Y* isimsel ise, *X*'in kategorileri satırlarda ve *Y*'nin kategorileri sütunlarda olacak biçimde iki yönlü çapraz tablo düzenle. *Pearson ki-kare testini* veya *olabilirlik oranı testini* kullan.

c. Eğer *Y* sıralı ise, bir *Y* birliktelik modeli (Clogg ve Eliaisin, 1987; Goodman, 1979; Magidson,1992) uydur. Olabilirlik oranı testini kullan.

2.Adım: En büyük *p* değerine sahip *X*'in kategori çifti için, *p* değerini önceden belirlenmiş alfa düzeyi $\alpha_{birleş}$ ile kıyasla.

a. Eğer *p* değeri $\alpha_{birleş}$'den büyük ise bu çifti bir tek kategori altında birleştir. *X*'in yeni kategori kümesi için süreci Adım 1'den başlat.

b. Eğer *p* değeri $\alpha_{birleş}$'den küçük ise Adım 3'e git.

3.Adım: X'in ve *Y*'nin kategori kümesi için uygun Bonferroni çarpanını kullanarak, düzeltilmiş *p* değerini hesapla.

---

[5] İstatistik bilimi içinde bir sıra değişik problemlerde kullanılan parametrik çıkarımsal sınama yöntemidir

4.Adım: En küçük düzeltilmiş *p* değerine sahip *X* tahmin edici değişkenini seç (en önemli olan). Bunun *p* değerini önceden tanımlanmış alfa düzeyi $\alpha_{böl}$ ile kıyasla.

a. Eğer *p* değeri, $\alpha_{böl}$ değerinden küçük veya eşit ise düğümü *X*'in kategori kümesini temel alarak böl.

b. Eğer *p* değeri, $\alpha_{böl}$ değerinden büyük ise düğümü bölme. Bu düğüm uç düğümdür. Ağaç büyütme sürecini durma kuralları görülene kadar sürdür (SPSS,2001).

Karar ağaçlarının oluşturulmasında yaygın olarak kullanılan CHAID algoritmasının işleyişi, Çizelge 2.3.'de verilen "Hava problemi" örnek veri seti ile açıklanmaktadır [63].

Çizelge 2.3. "Hava problemi" örnek veri seti

HAVA	ISI	NEM	RÜZGÂR	OYUN
Güneşli	sıcak	yüksek	hafif	hayır
Güneşli	sıcak	yüksek	kuvvetli	hayır
Bulutlu	sıcak	yüksek	hafif	evet
Yağmurlu	ılık	yüksek	hafif	evet
Yağmurlu	soğuk	normal	hafif	evet
Yağmurlu	soğuk	normal	kuvvetli	hayır
Bulutlu	soğuk	normal	kuvvetli	evet
Güneşli	ılık	yüksek	hafif	hayır
Güneşli	soğuk	normal	hafif	evet
Yağmurlu	ılık	normal	hafif	evet
Güneşli	ılık	normal	kuvvetli	evet
Bulutlu	ılık	yüksek	kuvvetli	evet
Bulutlu	sıcak	normal	hafif	evet
Yağmurlu	ılık	yüksek	kuvvetli	hayır

Örnek veri setinde kullanılan hedef değişken (Y) "Oyun" (Evet/Hayır), açıklayıcı değişkenler (X_1, X_2, X_3, X_4); "Hava" (Güneşli/Yağmurlu/Bulutlu), "Isı" (Sıcak/Ilık /Soğuk), "Nem" (Normal/Yüksek), "Rüzgâr" (Hafif/Kuvvetli) değişkenlerinden oluşmaktadır. Bu örnek veri seti kullanılarak hedef değişken ve açıklayıcı

değişkenlere ilişkin oluşturulan frekans tablosu Çizelge 2.4. ve Çizelge 2.5.'de ifade edilmektedir.

Çizelge 2.4. "Hava problemi" örnek veri setinde kullanılan hedef değişkene ilişkin frekans tablosu

Hedef değişken (Y)	*Sayı*	*Yüzde (%)*
Y = Evet	9	64,29
Y = Hayır	5	35,71
Toplam	14	100,00

Çizelge 2.5. "Hava problemi" örnek veri setinde kullanılan açıklayıcı değişkenlere ilişkin frekans tablosu

	Hedef Değişken (Oyun)		
Açıklayıcı değişkenler	Y = Evet	Y = Hayır	*Toplam*
X_1 = Güneşli	2	3	5
X_1 = Bulutlu	4	0	4
X_1 = Yağmurlu	3	2	5
X_2 = Sıcak	2	2	4
X_2 = Ilık	4	2	6
X_2 = Soğuk	3	1	4
X_3 = Normal	6	1	7
X_3 = Yüksek	3	4	7
X_4 = Hafif	6	2	8
X_4 = Kuvvetli	3	3	6

Karar ağacı oluşumunda hedef değişkenden itibaren dallara ayrılma, en küçük *p* değerine sahip olan X_1 açıklayıcı değişkenden başlamaktadır. X_1 açıklayıcı değişkeni için $\left(\frac{3}{2}\right) =$ 3 alt grup meydana gelmektedir.

X_1 açıklayıcı değişkeninin tüm alt kategori çiftlerinin hedef değişken kategori çifti ile oluşturduğu çapraz tablolar sonucunda meydana gelen bu alt gruplar Çizelge 2.6 - Çizelge 2.7 ve Çizelge 2.8'da ifade edilmektedir.

Algoritmanın adımları gereğince her bir alt grup için elde edilmesi gerekli olan ki-kare (χ^2) değerleri ise $\chi^2 = \sum_{j=1}^{c} \sum_{i=1}^{r} \frac{(G_{ij} - B_{ij})^2}{B_{ij}}$ formülü ile hesaplanmaktadır. (2.1)

G_{ij} = *j*.sütun ve *i*.satırdaki gözlenen değer

B_{ij} = *j*.sütun ve *i*.satırdaki beklenen değer

c = sütun değişkeninin düzey sayısı

r = satır değişkeninin düzey sayısı

$$B_{ij} = \frac{(n_{.j})(n_{i.})}{n} \qquad s.d.^6 = (r\text{-}1)\,(c\text{-}1) \qquad (2.2)$$

1.alt grup;

Çizelge 2.6. "Oyun" hedef değişkeni (*Y*) ile "Hava" açıklayıcı değişken (X_1) 1.alt grup kategorilerinin oluşturduğu çapraz tablo

Hava (X_1)	*Y* = Evet	*Y* = Hayır	*Toplam*
X_1 = Güneşli	2	3	5
X_1 = Bulutlu	4	0	4
Toplam	6	3	9

[6] s.d. ifadesi serbestlik derecesi anlamına gelmektedir.

1.alt grup için gözlenen değerler tablodaki sayısal değerleri ifade etmekte olup, her bir sayısal değer için beklenen değerler aşağıda görüldüğü üzere hesaplanmaktadır.

$$B_{11} = \frac{6*5}{9} = \frac{30}{9} \qquad B_{12} = \frac{3*5}{9} = \frac{15}{9}$$

$$B_{21} = \frac{6*4}{9} = \frac{24}{9} \qquad B_{22} = \frac{3*4}{9} = \frac{12}{9}$$

$$\chi^2 = \sum_{j=1}^{c} \sum_{i=1}^{r} \frac{(G_{ij} - B_{ij})^2}{B_{ij}} = \frac{(2-\frac{30}{9})^2}{\frac{30}{9}} + \frac{(3-\frac{15}{9})^2}{\frac{15}{9}} + \frac{(4-\frac{24}{9})^2}{\frac{24}{9}} + \frac{(0-\frac{12}{9})^2}{\frac{12}{9}}$$

χ^2 =3.24 olarak bulunur. (*s.d.* =1 α[7]=0.05 *p*=0,05778)

2.alt grup;

Çizelge 2.7. "Oyun" hedef değişkeni (Y) ile "Hava" açıklayıcı değişken (X_1) 2.alt grup kategorilerinin oluşturduğu çapraz tablo

Hava (X_1)	Y = Evet	Y = Hayır	*Toplam*
X_1 = Güneşli	2	3	5
X_1 = Yağmurlu	3	2	5
Toplam	5	5	10

$$B_{11} = \frac{5*5}{10} = \frac{25}{9} \qquad B_{12} = \frac{5*5}{10} = \frac{25}{9}$$

$$B_{21} = \frac{5*5}{10} = \frac{25}{9} \qquad B_{22} = \frac{5*5}{10} = \frac{25}{9}$$

[7] α, anlamlılık düzeyini ifade etmektedir

$$\chi^2 = \sum_{j=1}^{c} \sum_{i=1}^{r} \frac{(G_{ij} - B_{ij})^2}{B_{ij}} = \frac{(2-\frac{25}{9})^2}{\frac{25}{9}} + \frac{(3-\frac{25}{9})^2}{\frac{25}{9}} + \frac{(3-\frac{25}{9})^2}{\frac{25}{9}} + \frac{(2-\frac{25}{9})^2}{\frac{25}{9}}$$

χ^2 =0.4 olarak bulunur. (*s.d.* =1 α=0.05 p=0,5271)

3.alt grup;

Çizelge 2.8. "Oyun" hedef değişkeni (Y) ile "Hava" açıklayıcı değişken (X_1) 3.alt grup kategorilerinin oluşturduğu çapraz tablo

Hava (X_1)	Y = Evet	Y = Hayır	*Toplam*
X_1 = Bulutlu	4	0	4
X_1 = Yağmurlu	3	2	5
Toplam	7	2	9

$$B_{11} = \frac{4*7}{9} = \frac{28}{9} \qquad B_{12} = \frac{2*4}{9} = \frac{8}{9}$$

$$B_{21} = \frac{5*7}{9} = \frac{35}{9} \qquad B_{22} = \frac{2*5}{9} = \frac{10}{9}$$

$$\chi^2 = \sum_{j=1}^{c} \sum_{i=1}^{r} \frac{(G_{ij} - B_{ij})^2}{B_{ij}} = \frac{(4-\frac{28}{9})^2}{\frac{28}{9}} + \frac{(0-\frac{8}{9})^2}{\frac{8}{9}} + \frac{(3-\frac{35}{9})^2}{\frac{35}{9}} + \frac{(2-\frac{10}{9})^2}{\frac{10}{9}}$$

χ^2 =2.057 olarak bulunur. (*s.d.* =1 α=0.05 p=0,0909)

X_1 açıklayıcı değişkeninin alt gruplarının hedef değişken ile oluşturdukları çapraz tablordan elde edilen değerler Çizelge 2.9'da ifade edilmektedir.

Çizelge 2.9. X_1 açıklayıcı değişkeninin alt gruplarının ki-kare değerleri, serbestlik dereceleri ve *p* değerleri

Alt gruplar (X_1 değişkeni için)	χ^2	*s.d.*	*p*
1.alt grup	3,6	1	0,05778
2.alt grup	0,4	1	0,5271
3.alt grup	2,057	1	0,0909

Çizelge 2.9'dan anlaşılacağı üzere, X_1 değişkeni için en büyük *p* değerine sahip olan grubun 2.alt grup olduğu görülmektedir. 2.alt grup kategorileri Çizelge 2.7.'den anlaşılacağı üzere birleştirilerek ($p > \alpha_{birleş}$ olduğundan), oluşan yeni çapraz tablo için χ^2 değeri ve bonferroni çarpanı ile düzeltilmiş *p* değeri hesaplanır.

Bonferroni çarpanı;

- Eğer açıklayıcı değişken kategorileri nominal ise;

$$B_{serbest} = \sum_{i=0}^{r-1} (-1)^i \frac{(r-i)^c}{r!(r-i)!} \tag{2.3}$$

- Eğer açıklayıcı değişkenin kategorileri sıra belirten bir ölçek düzeyinde ise;

$$B_{monoton} = \frac{(c-1)}{(r-1)} \tag{2.4}$$

formülleri ile hesaplanmaktadır. (*c* kategori sayısını, *r* ise $1 \leq r \leq c$ olmak üzere açıklayıcı değişkenin grup sayısını ifade etmektedir.)

X_1 açıklayıcı değişkeni bu örnekte nominal bir değişken olduğundan,

$$B_{serbest} = \sum_{i=0}^{2} (-1)^i \frac{(2-i)^3}{r!(r-i)!} = (-1)^0 \frac{(2-0)^3}{2!(2-0)!} + (-1)^1 \frac{(2-1)^3}{2!(2-1)!} + (-1)^2 \frac{(2-2)^3}{2!(2-2)!} = 1{,}5$$

$B_{serbest}$ = 1,5 ve düzeltilmiş *p* değeri 0,3786 (0,2524 * 1,5) olarak bulunmuştur.

Çizelge 2.10. X_1 değişkeni 2.alt grup kategorileri değerlerinin birleştirilmesi ile elde edilen çapraz tablo

Hava Durumu (X_1)	Y = Evet	Y = Hayır	*Toplam*
X_1 = Güneşli veya Yağmurlu	5	5	10
X_1 = Bulutlu	4	0	4
Toplam	9	5	14

$$B_{11} = \frac{9*10}{14} = \frac{90}{14} \qquad B_{12} = \frac{5*10}{14} = \frac{50}{10}$$

$$B_{21} = \frac{9*4}{14} = \frac{36}{14} \qquad B_{22} = \frac{5*4}{14} = \frac{20}{14}$$

$$\chi^2 = \sum_{j=1}^{c} \sum_{i=1}^{r} \frac{(G_{ij} - B_{ij})^2}{B_{ij}} = \frac{(5-\frac{90}{14})^2}{\frac{90}{14}} + \frac{(5-\frac{50}{14})^2}{\frac{50}{14}} + \frac{(4-\frac{36}{14})^2}{\frac{36}{14}} + \frac{(0-\frac{20}{14})^2}{\frac{20}{14}}$$

χ^2 = 3,11 olarak bulunur. (*s.d.* =1 α =0.05 *p*=0,0777)

Yukarıda yapılan işlemler her bir açıklayıcı değişken için uygulanır ve yaprağa ulaşıldığında ağaç sonlandırılır.

CHAID algoritması ile diğer yöntemler arasındaki farklar

- ID3, C4.5 ve CART ikili ağaçlar türetirken, CHAID ikili olmayan çoklu ağaçlar türetir.

- Sürekli ve kategorik tüm değişken tipleriyle çalışabilmektedir.

- Sürekli tahmin edici değişkenler otomatik olarak analizin amacına uygun olarak kategorize edilmektedir.

- Ki-kare metriği vasıtasıyla, ilişki düzeyine göre farklılık rastlanan grupları ayrı ayrı sınıflamaktadır.

2.5.10. Veri madenciliğinde kullanılan programlar

VM'nde kullanılan birçok program vardır. Bunlar arasında en çok tercih edilen programların başında SPSS Clementine ve SAS Enterprise Manager programları gelmektedir [29].

- *SPSS Clementine*

SPSS firmasının merkezi Chicago'da bulunmakta olup, 1967 yılından bu yana verilerdeki gizli bilgileri keşfetme ve stratejik karar desteği sağlama yönünde ileri analitik çözümler sunmaktadır. SPSS'in VM metodolojisi olarak kabul ettiği CRISP DM (cross industry standart processing for datamining) % 50'nin üzerinde bir kullanıma sahiptir. İnternet kayıtlarına ve elde edilen verilere, gelişmiş VM teknikleri uygulayarak, kullanıcılar ile birebir ilişki kurmayı sağlayacak öngörüler elde edilebilir.

- *SAS (Statistical Analysis Software) Enterprise Miner*

Şirketlerin çok büyük veri yığınlarından kritik bilgileri elde etmelerini sağlayan VM çözümlerinde dünyada önemli bir yere sahip olan SAS, veri üzerinde değil bilgi üzerinde düşünme ve strateji geliştirme avantajını bir adım öteye taşıyarak SAS Enterprise Miner programını geliştirmiştir. Regresyon, Sınıflama, İstatistiksel analiz gibi fonksiyonları içerir.

SAS programında komut yazmak gerektiğinden kullanımı SPSS programına göre biraz daha zor oldur. SAS, Kamu, Perakende, Sigorta, Bankacılık, Medya, Eğitim ve Telekomünikasyon ve benzeri sektörlerde kullanılmaktadır.

- *Statistica Data Miner*

VM için parametrik istatistik ve makine öğrenimi kombinasyonu zengin algoritmalar sunan bir programdır. Grafiksel programlama ara yüzüne sahiptir, tüm ortak VM görevleri için araçlar sunar, boyut azaltmada kullanılan güçlü araçlara sahiptir, daha geniş veri kümelerini diğer programlardan daha hızlı işleyebilir.

- *Darwin*

Darwin, Oracle firmasının geliştirdiği bir VM yazılımıdır. Yapay sinir ağları, bayesyen öğrenme, k-en yakın komşu, regresyon ağaçları, kümeleme ve sınıflama gibi farklı algoritmaları destekleyen bir yazılımdır.

- *DBMiner*

Kanada'da Simon Fraser Üniversitesi Veritabanı Araştırma Laboratuarında geliştirilmiş bir VM yazılımıdır. Araştırma ürünü olarak piyasaya çıkmış fakat ticari bir ürün haline gelmiştir. DBMiner'i diğer programlardan ayıran en önemli özelliği, OLAP yöntemleriyle VM algoritmalarını bütünleşik hale getirmiş olmasıdır.

3. UYGULAMA

3.1. Uygulama Konusu ve Amacı

"Sosyal güvenlik", insanlara bugün ve gelecekte, çalışma koşullarını yitirmesi hali de dâhil olmak üzere çeşitli risklere karşı, yaşamını sürdürebileceği sürekli bir gelir güvencesinin sağlanmasıdır. Sosyal güvenliğin gelişimi iş kazası, meslek hastalıkları ve analık sigortaları ile başlamış, daha sonra diğer hastalık, maluliyet, yaşlılık, ölüm ve işsizlik sigortası hakları kazanılmıştır. Ülkemizde de bu güvenliği sağlayan kurum Sosyal Güvenlik Kurumu (SGK) olup, 16 Mayıs 2006 tarihinde kabul edilen 5502 sayılı Sosyal Güvenlik Kurumu Kanunu ile Bağkur (BK), Sosyal Sigortalar Kurumu (SSK) ve Emekli Sandığı (ES) kurumlarının devredilmesi ile kurulmuş, Çalışma ve Sosyal Güvenlik Bakanlığı'na bağlı bir kurumdur.

Dolayısıyla yukarıda ifade edilen üç kurumun birleşmesiyle oluşan SGK, veri bakımından oldukça zengin bir kurum haline gelmiştir. SGK'ya kayıtlı bireylere ilişkin; doğum tarihi, cinsiyet, eğitim durumu, çalışma süresi, emeklilik, iş kazası ve meslek hastalığı, sağlık harcamaları, kimlik bilgileri, askerlik ve yurt dışı borçlanmaları, işyerlerine ilişkin ise; tescil (kanun kapsamına alınış), kanun kapsamından çıkış, mahiyet (özel–kamu), yapılandırma durumu, tahakkuk, tahsilât, icra ve borç bilgileri gibi oldukça detaylandırılmış bilgiler mevcuttur.

Bu mevcut bilgiler içerisinde uygulama konusu olarak, SSK'ya tabii tescili yapılmış olan işyerlerine tahakkuk eden prim borcu, idari para cezası (IPC) ve işsizlik borçlarının (tüm tahakkuk eden borçlarının 2005 yılına kadar ki tutarları) çıkarılan 5458 sayılı kanunla (04.03.2006 tarih ve 26098 sayılı Resmi Gazete) yeniden yapılandırılması, VM'nde en çok kullanılan sınıflandırma tekniklerinden biri olan karar ağaçları algoritmalarından CHAID kullanarak ele alınmıştır.

Yapılan çalışma ile, ülkemizde yapılandırma kanunundan yararlanan işyerlerinin profilleri belirlenmeye çalışılmış ve daha sonra çıkarılacak olan kanunların bu

yapılan çalışmada öngörülen bilgiler ışığında karar destek amaçlı olarak katkıda bulunması amaçlanmıştır.

3.2. Veri Ön İşlemi

SGK veri ambarında veri tabanı yönetim sistemi olarak Oracle kullanılmakta olup, uygulamada kullanılan veriler, veri tabanları birleştirilmiş olan üç kurumun ortak veri tabanından Oracle'ın kullanıcı ara yüzü Toad for Oracle 9,5 programı kullanılarak Yapısal Sorgulama Dili olan SQL ile çeşitli sorgular yazılarak, kullanılacak olan değişken ihtiyacına göre tablolardan çekilmiştir.

Başvuruda bulunan işyerleri arasında; yapılandırmadan yararlanan 322 543 işyeri, yapılandırmadan yararlanmayan ise 191 876 işyeri bulunmaktadır. Yapılandırmadan yararlanan işyerleri içerisinde halen yapılandırması devam eden işyerleri de bulunmaktadır.

Verilerin önişlemi;

- *Veri Temizleme*

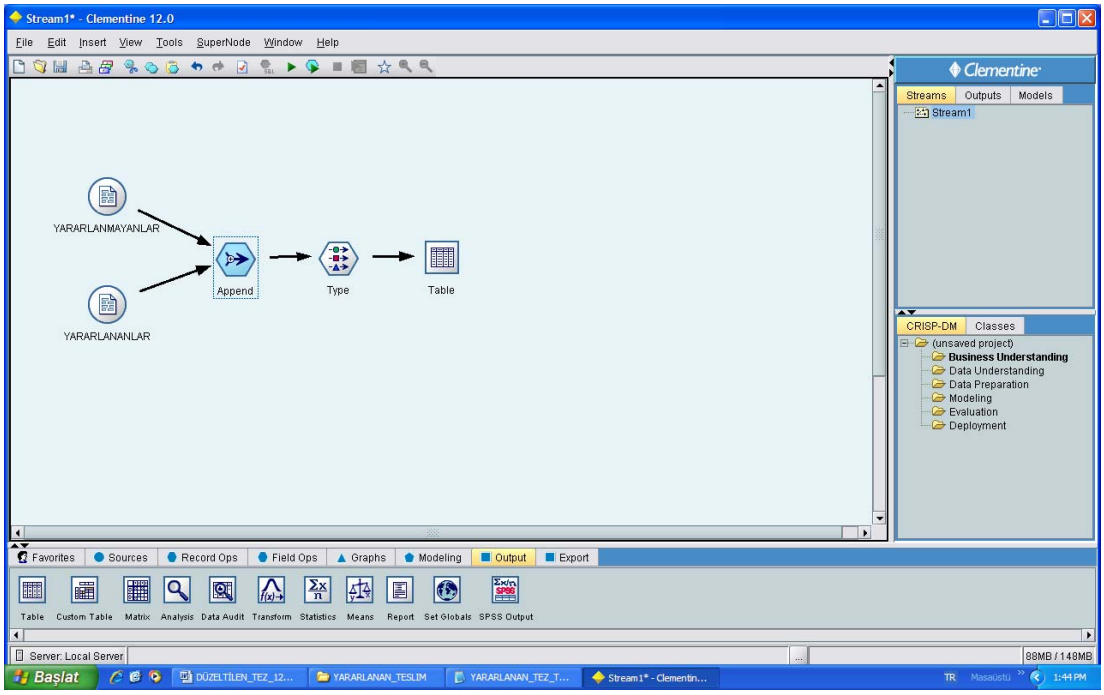
Veri temizliği, altıncı bölümde ifade edildiği üzere VM sürecinin en önemli aşamalarından birisidir. Yapılandırma durumu için oluşan yeni veri incelendiğinde, yapılandırmadan yararlanan bazı işyerleri için Kısım Id[8] bilgisinin boş olduğu tespit edilmiştir. Bu alanların başka tablolardan elde edilmesi mümkün olmadığından, boş olan 146 kayıt tablodan silinmiştir. Ayrıca sigortalı sayısı negatif olan 7 kayıt, işyeri yaşı negatif olan 7 kayıt, il kodu 86 olan 45 kayıt ile sigortalı sayısı 999 999 olan 1551 kayıt da deneme kayıtları olduğundan tablodan silinmiştir. Yapılandırmadan yararlanmayan işyerleri içerisinde ise; il kodu 86 olan 42 işyeri, sigortalı sayısı 999

[8] Ekonomik faaliyetlerin istatistikî sınıflandırılması sisteminin alfabetik bir kodla tanımlanan başlıklardan oluşan birinci seviyesi

999 olan 949 kayıt, işyeri yaşı negatif olan 6 kayıt ile nace kodu[9] bilinmeyen 78 kayıt da bu tablodan silinmiştir.

- *Veri Birleştirme*

Yapılandırmadan yararlanan ve yararlanmayan işyerlerine ilişkin veriler, Şekil 3.1'den görüleceği üzere Clementine 12.0 paket programı kullanılarak birleştirilmiştir [23, 24, 25,61].



Şekil 3.1. Yapılandırma durumu için veri kaynaklarının birleştirilmesi

- *Veri Dönüştürme*

Veri temizlemeden sonraki adım veri dönüştürmedir. Yapılandırma durumu için veri dönüştürme işleminde işyerlerinde çalışan sigortalıların sayısı ve işyeri yaşı, aşağıda ifade edilen SQL komutu ile aralıklara dönüştürülerek gruplandırılmıştır:

---

[9] Ekonomik faaliyetlerin istatistikî sınıflandırılması

Update *tablo_adı* set *kolon_adı* = 'A – B arası' where *kolon_adı* >A and *kolon_adı* <=B.

Ön işlemlerin tamamlanmasından sonra yapılandırmadan yararlanan 320 787, yararlanmayan 190 801 işyeri verisi üzerine CHAID algoritması uygulanmıştır.

3.3. Değişkenlerin Açıklanması

Veri tabanında işyerleri ile ilgili bilgiler farklı tablolarda tutulmaktadır. Nace kodu, borç türü, bölgeler ve işyeri türü boyut tabloları oluşturularak, bu boyut tabloları SQL fonksiyonları kullanılarak ana tablolarla birleştirilmiş ve tek bir tablo haline getirilmiştir.

Uygulamada kullanılan bağımlı/hedef değişken, yapılandırmadan yararlanmış (1) ve yapılandırmadan yararlanmamış (0) şeklinde iki seviyeli kategorik bir değişken olarak alınmıştır. Bağımsız/açıklayıcı değişkenlerimiz ise; Borç Türü Açıklaması (kategorik), Toplam Borç Miktarı (sürekli), Prim Nispet Oranı[10] (kategorik), İşyeri Türü (kategorik), Sigortalı Sayısı (kesikli), Kısım Id (kategorik), İşyeri Yaşı (sürekli) ve Bölge Kodu (kategorik) değişkenlerinden oluşmaktadır.

Uygulamada kullanılan hedef değişken ve açıklayıcı değişkenlere ilişkin tanımlayıcı istatistikler çizelgeler şeklinde aşağıda özetlenmektedir.

Çizelge 3.1. hedef değişken olarak tanımlanan "yapılandırma durumu"na göre frekans/sıklık dağılımını göstermektedir.

Çizelge 3.1. İşyerlerinin borçlarının yapılandırma durumu sıklık dağılımı

Yapılandırma Durumu		Sıklık	Yüzde	Geçerli Yüzde	Kümülatif Yüzde
0	YARARLANMAMIS	190 801	37,3	37,3	37,3
1	YARARLANMIS	320 787	62,7	62,7	100,0
	Toplam	511 588	100,0	100,0	

[10] İşyerinin tehlike, sınıf ve derecesine göre belirlenen oran

Çizelge 3.2. açıklayıcı değişken olarak tanımlanan “bölge kodu”na göre frekans/sıklık dağılımını göstermektedir.

Çizelge 3.2. İşyerlerinin bulunduğu bölgenin bölge koduna göre sıklık dağılımı

Bölge Kodu	Bölge Adı	Sıklık	Yüzde	Geçerli Yüzde	Kümülâtif Yüzde
1	Akdeniz Bölgesi	51 525	10,1	10,1	10,1
2	Doğu Anadolu Bölgesi	13 782	2,7	2,7	12,8
3	Ege Bölgesi	91 576	17,9	17,9	30,7
4	Güneydoğu Anadolu Böl.	14 495	2,8	2,8	33,5
5	İç Anadolu Bölgesi	78 435	15,3	15,3	48,8
6	Karadeniz Bölgesi	40 385	7,9	7,9	56,7
7	Marmara Bölgesi	221 390	43,3	43,3	100,0
	Toplam	511 588	100,0	100,0	

Çizelge 3.3. açıklayıcı değişken olarak tanımlanan “borç türü açıklaması”na göre frekans/sıklık dağılımını göstermektedir.

Çizelge 3.3. İşyerlerinin borçlarının borç türü açıklamasına göre sıklık dağılımı

Borç Türü Açıklaması	Sıklık	Yüzde	Geçerli Yüzde	Kümülâtif Yüzde
İdari Para Cezası Borcu (IPC)	33 226	6,5	6,5	6,5
İşsizlik Borcu	212 677	41,6	41,6	48,1
Prim Borcu	265 685	51,9	51,9	100,0
Toplam	511 588	100,0	100,0	

Çizelge 3.4. açıklayıcı değişken olarak tanımlanan “sigortalı sayısı”na göre frekans/sıklık dağılımını göstermektedir.

Çizelge 3.4. İşyerlerinde çalışan sigortalıların sayısının sıklık dağılımı

Sigortalı Sayısı	Sıklık	Yüzde	Geçerli Yüzde	Kümülâtif Yüzde
0–1 ARASI	295 450	57,8	57,8	57,8
2–3 ARASI	138 367	27,0	27,0	84,8
4 VE ÜZERI	77 771	15,2	15,2	100,0
Toplam	511 588	100,0	100,0	

Çizelge 3.5. açıklayıcı değişken olarak tanımlanan “prim nispet oranı”na göre frekans/sıklık dağılımını göstermektedir.

Çizelge 3.5. İşyerlerinde yapılan işin tehlike derecesine göre belirlenen prim nispet oranı sıklık dağılımı

Prim Nispet Oranı	Sıklık	Yüzde	Geçerli Yüzde	Kümülâtif Yüzde
1,0	247 786	48,4	48,4	48,4
1,5	39 336	7,7	7,7	56,1
2,0	38 168	7,5	7,5	63,6
2,5	38 439	7,5	7,5	71,1
3,0	140 723	27,5	27,5	98,6
3,5	4601	0,9	0,9	99,5
4,0	106	0,0	0,0	99,5
4,5	135	0,0	0,0	99,6
5,0	9	0,0	0,0	99,6
5,5	246	0,0	0,0	99,6
6,0	1533	0,3	0,3	99,9
6,5	506	0,1	0,1	100,0
Toplam	511 588	100,0	100,0	

Çizelge 3.6 açıklayıcı değişken olarak tanımlanan “işyeri yaşı”na göre frekans/sıklık dağılımını göstermektedir.

Çizelge 3.6. İşyeri yaşına göre sıklık dağılımı

İşyeri Yaşı	Sıklık	Yüzde	Geçerli Yüzde	Kümülâtif Yüzde
1–3 ARASI	127 377	24,9	24,9	24,9
10 VE UZERI	151 402	29,6	29,6	54,5
4–5 ARASI	89 812	17,6	17,6	72,0
6–9 ARASI	142 997	28,0	28,0	100,0
Toplam	511 588	100,0	100,0	

Çizelge 3.7. açıklayıcı değişken olarak tanımlanan "işyeri türü" frekans/sıklık dağılımını göstermektedir.

Çizelge 3.7. İşyeri türüne göre sıklık dağılımı

İşyeri kodu[11]	İşyeri türü	Sıklık	Yüzde	Geçerli Yüzde	Kümülâtif Yüzde
0	DIGER	383 999	75,1	75,1	78,0
1	GERÇEK KISI	46 417	9,1	9,1	87,1
2	ADI ORTAKLIK	1935	0,4	0,4	0,5
3	KOLL.	2198	0,4	0,4	87,6
4	ADI. KOM.	60	0,0	0,0	0,5
5	PAYLI KOM.	92	0,0	0,0	99,9
6	LMT.	61 052	11,9	11,9	99,9
7	ANONIM	9837	1,9	1,9	2,4
8	KOOP.	1933	0,4	0,4	88,0
9	KONSOLOSLUKLAR	8	0,0	0,0	87,6
10	KAMU TÜZEL KISILIKLERI	19	0,0	0,0	87,2
11	DERNEK	630	0,1	0,1	2,9
12	VAKIF	164	0,0	0,0	100,0
13	SENDIKA	74	0,0	0,0	100,0
14	APT. YÖN.	1949	0,4	0,4	2,8
15	S. PARTI.	16	0,0	0,0	99,9
16	BASIN	3	0,0	0,0	2,8
17	SPOR KL.	52	0,0	0,0	100,0
18	ÖZEL DIGER	469	0,1	0,1	0,1
21	GENEL BÜTÇE KAP.	317	0,1	0,1	78,0
22	ÖZEL BÜTÇE KAP.	51	0,0	0,0	0,0
23	DÜZENLEYICI-DENETLEYICI.	6	0,0	0,0	2,8
24	SGK	4	0,0	0,0	100,0
25	ÖZEL IDARELER	2	0,0	0,0	0,1
26	BELEDIYE	31	0,0	0,0	2,8
28	KAMU DIGER	270	0,1	0,1	87,2
Toplam		511 588	100,0	100,0	

[11] İşyeri kodu alanında operasyonel sistemde kullanılan kodlar esas alınmış olup, ayrıca bir kodlama yapılmamıştır

Çizelge 3.8. açıklayıcı değişken olarak tanımlanan Kısım Id frekans/sıklık dağılımını göstermektedir.

Kısım Id kodları ve bu kodlara karşılık gelen açıklamalar aşağıda verilmiştir:

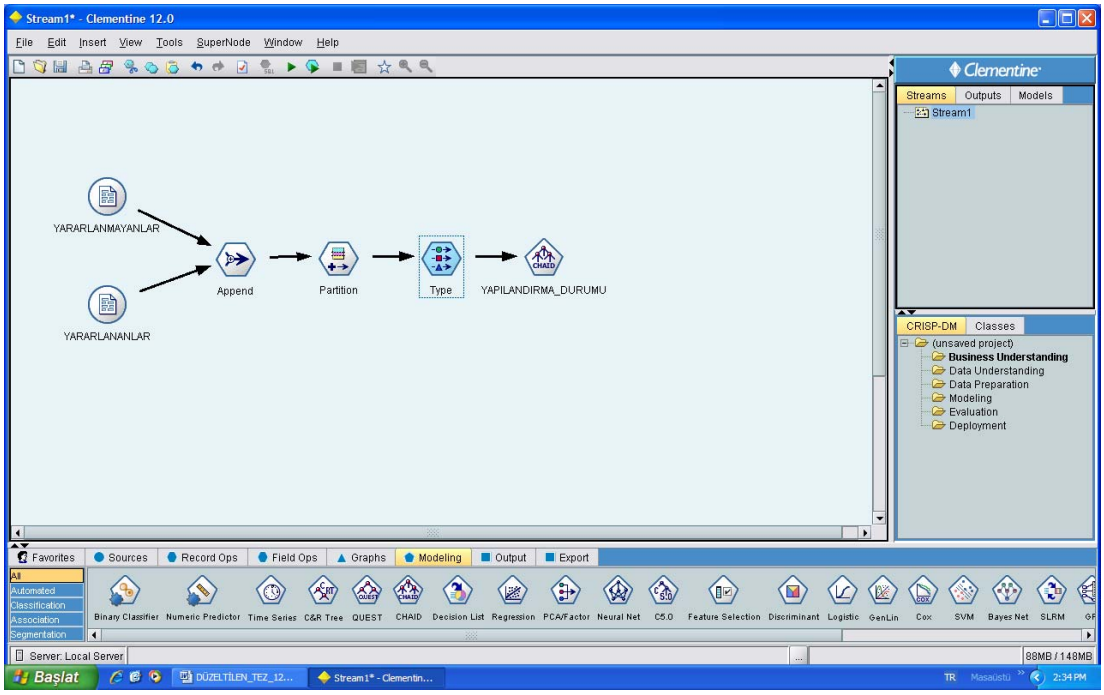
A = Tarım, Ormancılık ve Balıkçılık
B = Madencilik ve Taş Ocakçılığı
C = İmalat
D = Elektrik, Gaz, Buhar ve İklimlendirme Üretimi ve Dağıtımı
E = Su Temini, Kanalizasyon, Atık Yönetimi ve İyileştirme Faaliyetleri
F = İnşaat
G = Toptan ve Perakende Ticaret, Motorlu Kara Taşıtlarının ve Motosikletlerin Onarımı
H = Ulaştırma ve Depolama
I = Konaklama ve Yiyecek Hizmeti Faaliyetleri
J = Bilgi ve İletişim
K = Finans ve Sigorta Faaliyetleri
L = Gayrimenkul Faaliyetleri
M = Mesleki, Bilimsel ve Teknik Faaliyetler
N = İdari ve Destek Hizmet Faaliyetleri
O = Eğitim
P = İnsan Sağlığı ve Sosyal Hizmet Faaliyetleri
Q = Kültür, Sanat, Eğlence, Dinlence ve Spor
R = Diğer Hizmet Faaliyetleri
S = Hane Halklarının İşverenler olarak Faaliyetleri, Hane Halkları Tarafından Kendi Kullanımlarına Yönelik Olarak Ayrım Yapılmamış Mal ve Üretim Faaliyetleri
T = Uluslararası Örgütler ve Temsilciliklerinin Faaliyetleri

Çizelge 3.8. İşyerlerinin Kısım Id sıklık dağılımı

Kısım Id	Sıklık	Yüzde	Geçerli Yüzde	Kümülâtif Yüzde
A	3612	0,7	0,7	0,7
B	3022	0,6	0,6	1,3
C	131 343	25,7	25,7	27,0
D	4140	0,8	0,8	27,8
E	1268	0,2	0,2	28,0
F	101 701	19,9	19,9	47,9
G	96 537	18,9	18,9	66,8
H	31 294	6,1	6,1	72,9
I	2471	0,5	0,5	73,4
J	571	0,1	0,1	73,5
K	3455	0,7	0,7	74,2
L	30	0,0	0,0	74,2
M	45 093	8,8	8,8	83,0
N	20 449	4,0	4,0	87,0
O	5964	1,2	1,2	88,1
P	699	0,1	0,1	88,3
Q	36 099	7,1	7,1	95,3
R	22 064	4,3	4,3	99,7
S	1285	0,3	0,3	99,9
T	491	0,1	0,1	100,0
Toplam	511 588	100,0	100,0	

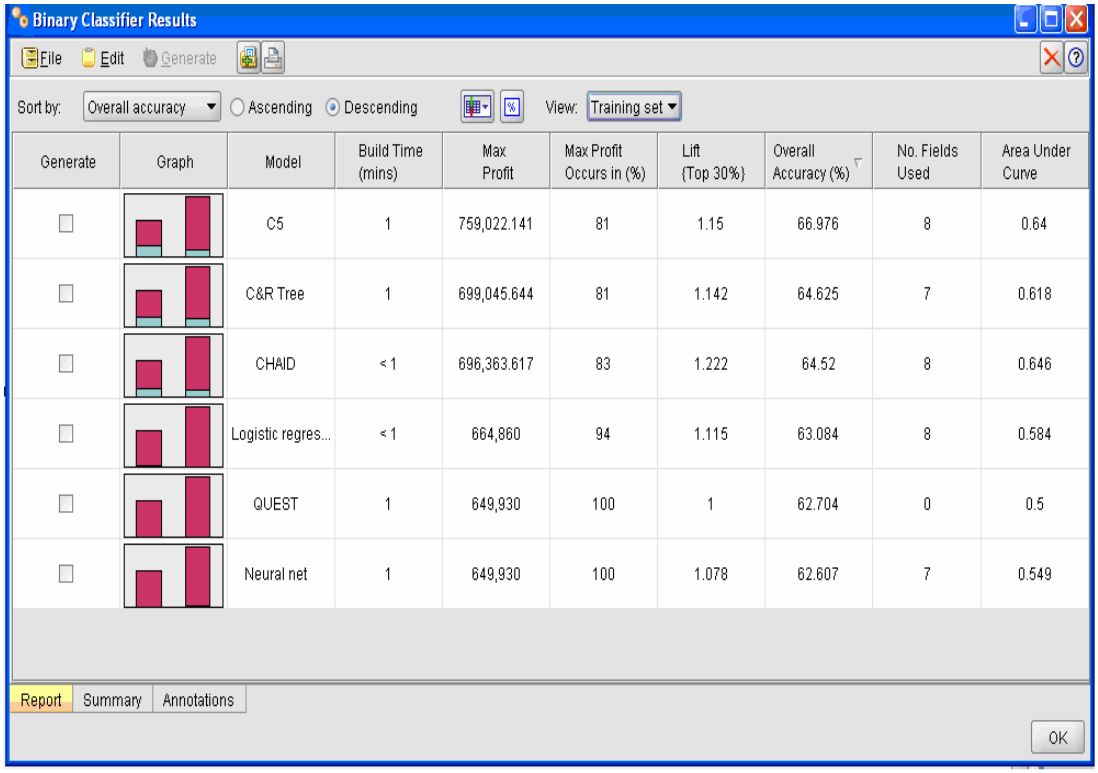
3.4. Modelin Kurulması ve Değerlendirilmesi

Şekil 3.2.'den anlaşılacağı üzere Clementine 12.0 paket programı ile 511 588 işyeri verisi, *eğitim* (*training)* (% 70) ve *test* (*testing)* (% 30) olmak üzere partition işlemcisi ile iki örnek gruba ayrılmıştır.



Şekil 3.2.Yapılandırma durumu veri setinin *eğitim (traning)* ve *test (testing)* olarak gruplandırılması

Verilerin % 70'ini oluşturan *eğitim* (*training)* veri setinde farklı sınıflandırma algoritmalarından C5.0, C&R Tree, CHAID, Lojistik Regresyon, QUEST, Neural Net algoritmaları denenerek model geliştirilmiş ve oluşturulan modellerin hepsi *test* (*testing)* veri seti ile test edilmiştir. *Eğitim* (*training)* veri seti üzerinde uygulanan sınıflandırma algoritmaları sonucunda kurulan modellerin doğruluk yüzdeleri Şekil 3.3.'de görülmektedir.

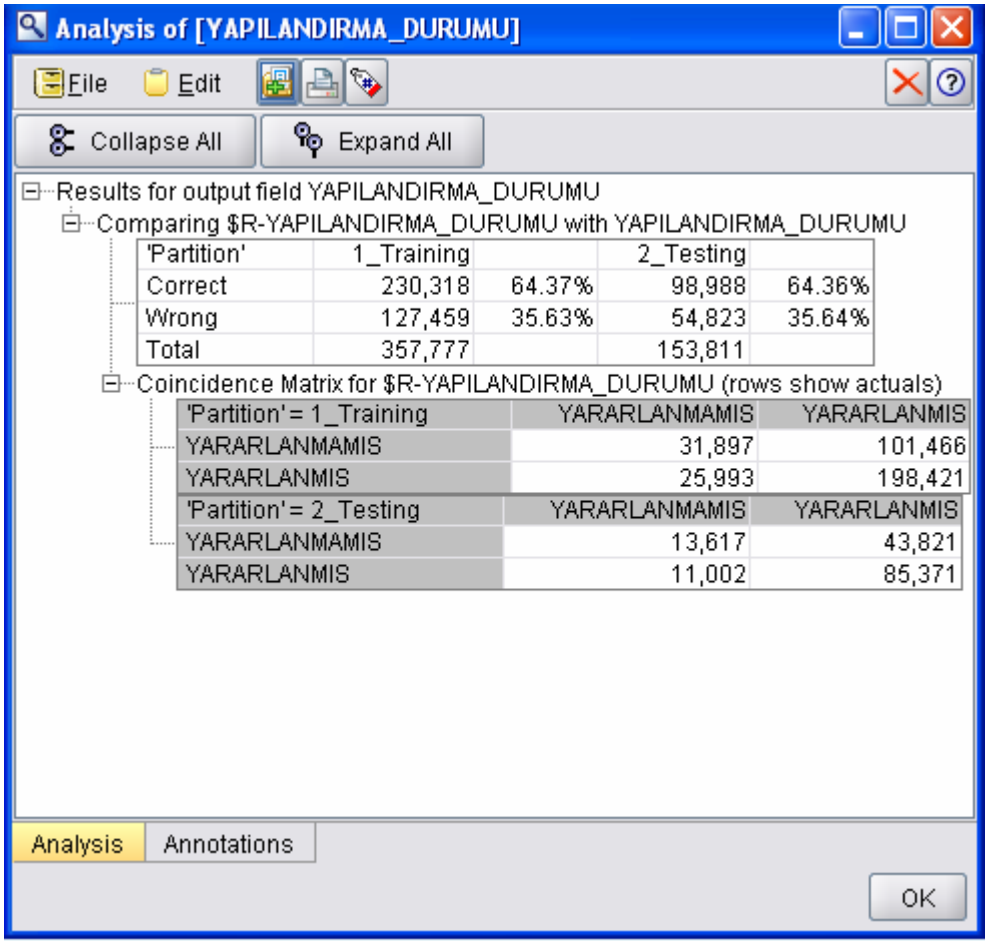


Şekil 3.3. Sınıflandırma algoritmalarının doğruluk yüzdeleri

Bu algoritmalar arasında CHAID algoritması çalışmanın amacına uygun olarak, değişkenler arasındaki ilişkinin ötesinde, değişkenlerin içerisindeki veri grupları arasındaki ilişkilerin en alt detayına kadar araştırılmasına olanak sağladığından yapılan çalışmanın amacına uygun bir algoritma olarak belirlenmiştir.

Test verileri model işlemcisinden geçirildiğinde doğruluk değeri bir miktar düşer, eğer doğruluk değeri geniş ölçüde azalıyorsa, *eğitim* verisinde modelin aşırı öğrenme (overfit) eğiliminde olduğunun bir göstergesi olabilir. Eğer doğruluk çok az bir miktarda düşüyorsa bu modelin gelecekte iyi çalışacağının kanıtıdır [61]. Yapılandırma durumu için uygulanan algoritmanın sonucu da bu modelin iyi çalıştığının bir göstergesi olmaktadır.

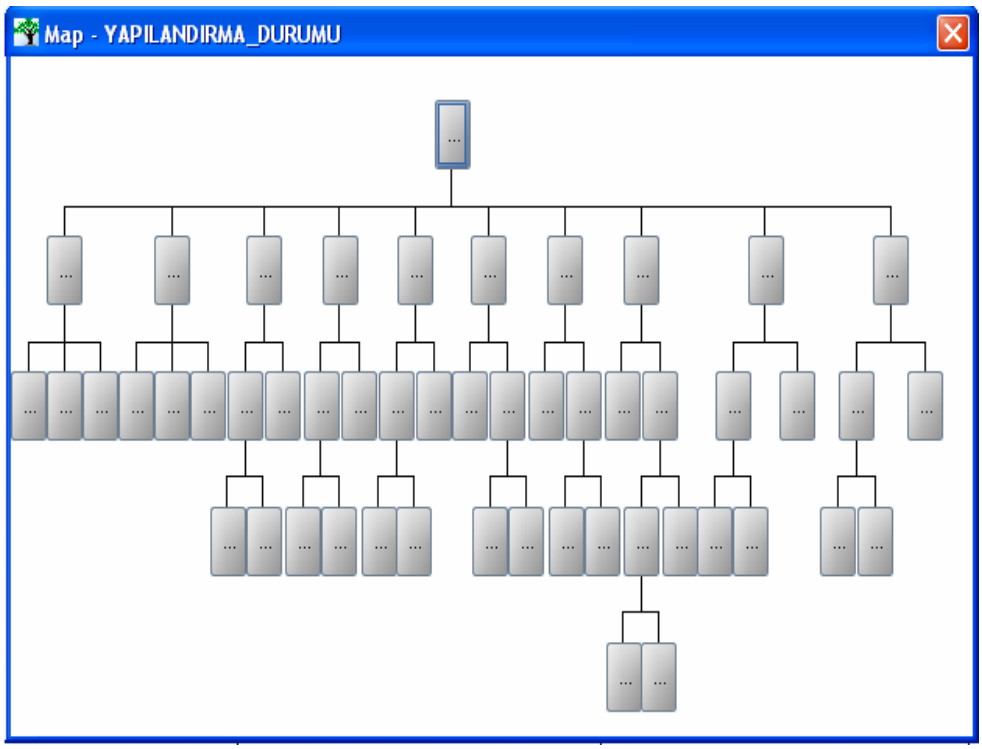
Şekil 3.4.'de görüldüğü üzere; kurulan modelin doğruluğunun değerlendirilmesinde basit geçerlilik testi kullanılmış olup, $\alpha_{\text{birlestir}} = \alpha_{böl} = 0.05$ ve ağaç derinliği 5 olarak alınmıştır.



Şekil 3.4. CHAID algoritması sonucu kurulan modelin doğruluk yüzdeleri

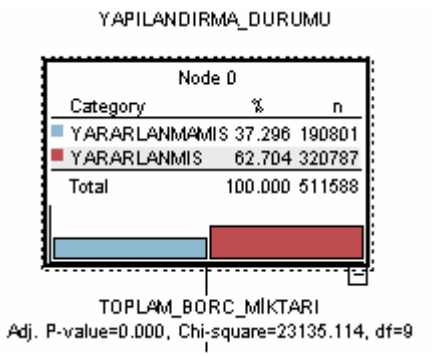
3.5. Algoritmanın Uygulanması

Yapılandırma durumu için uygulanan CHAID algoritması sonucunda elde edilen karar ağacı görünümü Şekil 3.5.'de gösterilmektedir.



Şekil 3.5. Yapılandırma durumu için elde edilen karar ağacının görünümü

CHAID algoritması sonucunda elde edilen karar ağacının başlangıç nodu (düğümü) aşağıda Şekil 3.6.'de gösterilmektedir.

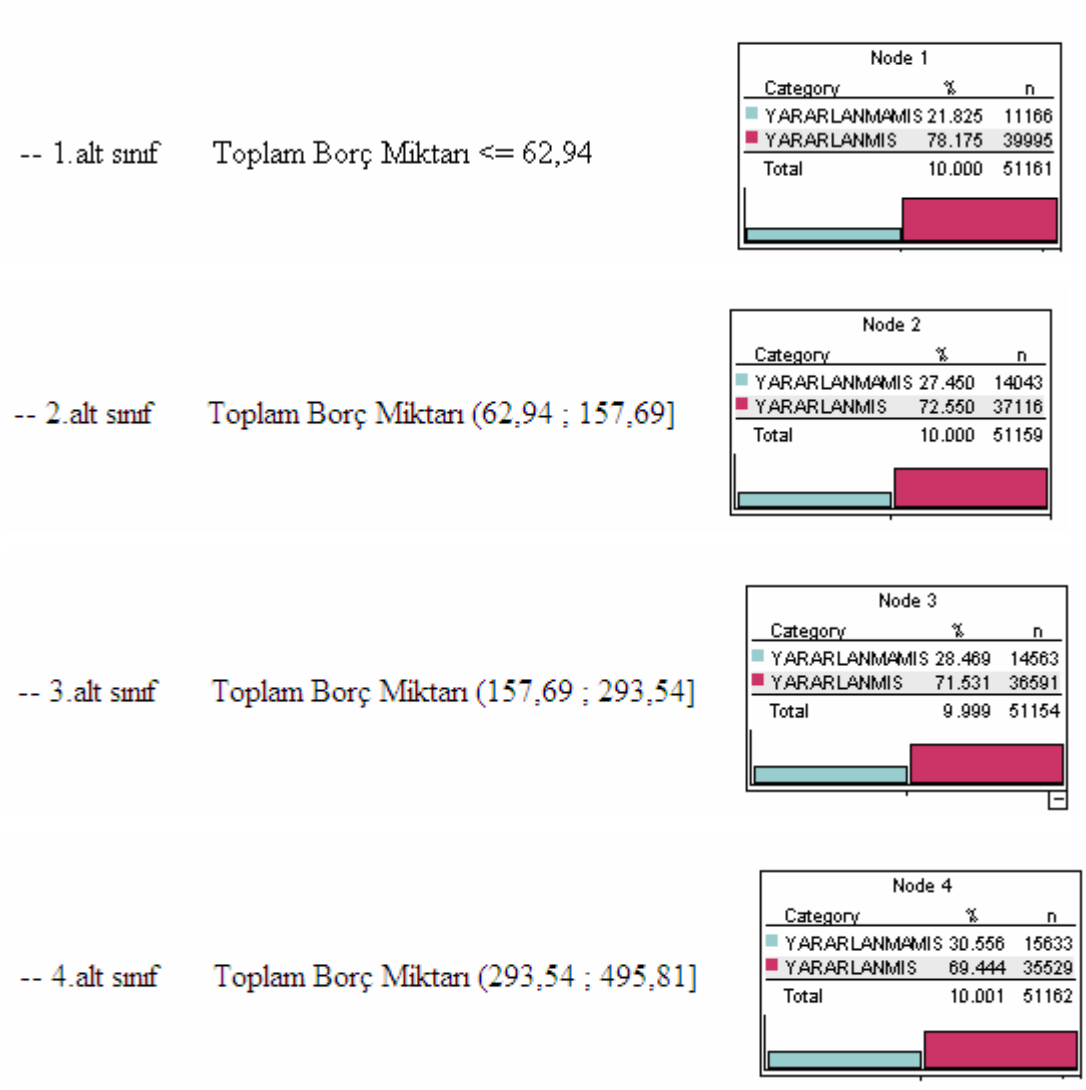


Şekil 3.6. CHAID algoritması sonucu elde edilen karar ağacının başlangıç nodu gösterimi

Dallanmanın (bölümlemenin) başladığı nodda toplam 511 588 işyeri arasından yapılandırmadan yararlanan 320 787 (% 62,7) işyeri, yapılandırmadan yararlanmayan 190 801 (% 37,3) işyeri bulunmaktadır.

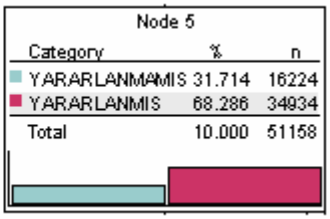
Yapılandırma durumu olarak tanımlanan hedef değişken için dallara ayırmada toplam borç miktarı açıklayıcı değişkeni önemli bulunmuştur. İlk dallanma, toplam borç miktarı değişkenine göre başlamakta ve borç aralıklarına göre kollara/dallara ayrılmaktadır.

Yapılandırma durumu için, önemli bulunan toplam borç miktarı açıklayıcı değişkenine göre Şekil 3.7.'den anlaşılacağı üzere 10 alt sınıf meydana gelmiştir.

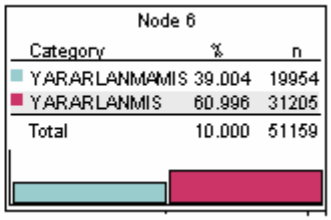


Şekil 3.7. Elde edilen 10 alt sınıfın gösterimi

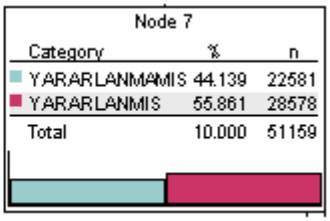
-- 5.alt sınıf Toplam Borç Miktarı (495,81 ; 808,23]



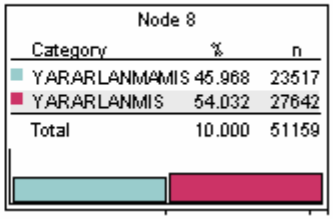
-- 6.alt sınıf Toplam Borç Miktarı (808,23 ; 1328,90]



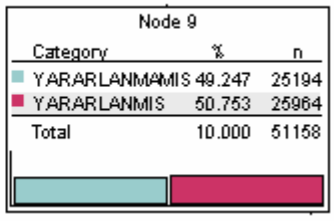
-- 7.alt sınıf Toplam Borç Miktarı (1328,90 ; 2276,01]



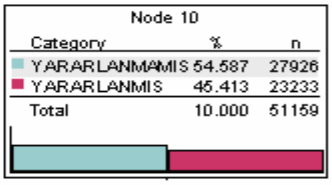
-- 8.alt sınıf Toplam Borç Miktarı (2276,01 ; 4206,92]



-- 9.alt sınıf Toplam Borç Miktarı (4206,92 ; 9282,81]



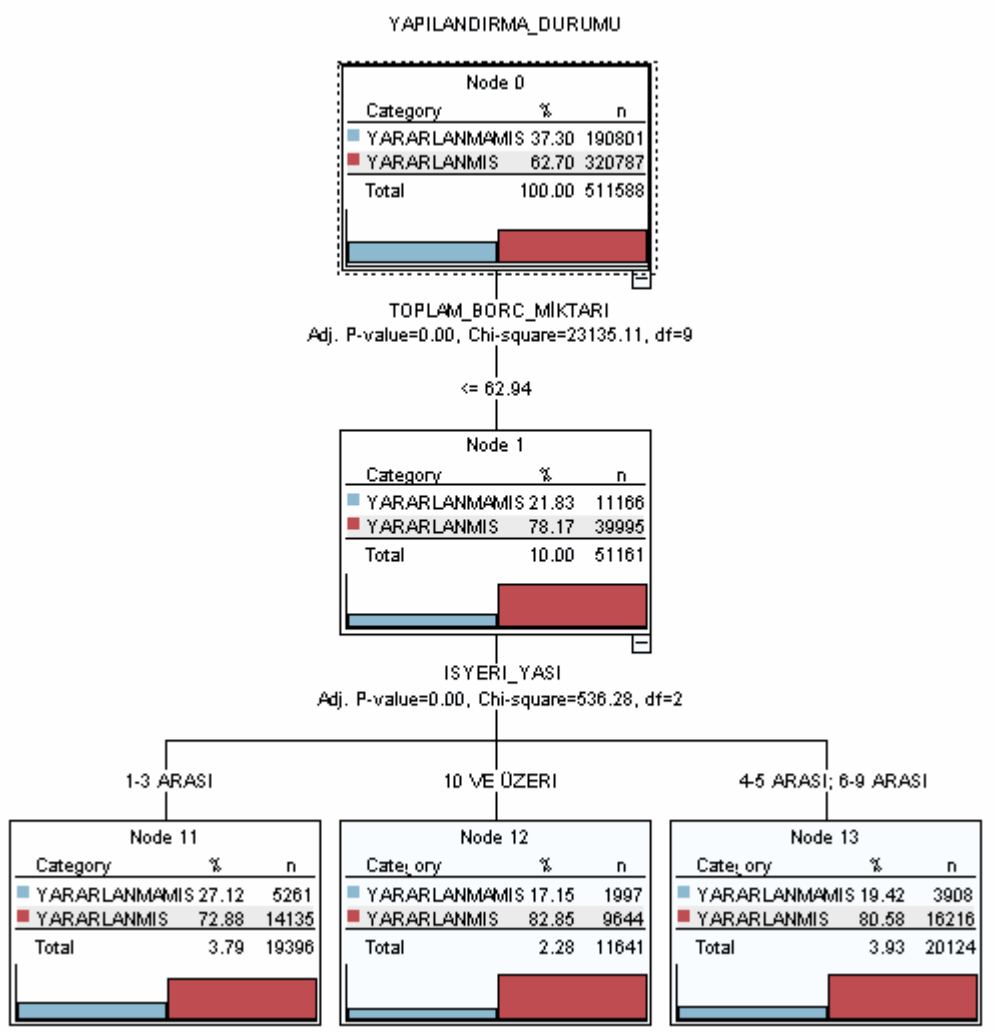
-- 10.alt sınıf Toplam Borç Miktarı > 9282,81



Şekil 3.7. (Devam) Elde edilen 10 alt sınıfın gösterimi

Başlangıç düğümü ile oluşan karar ağacı ana dalları aşağıdaki şekillerde başlıklar halinde sırasıyla ifade edilmektedir.

3.5.1. Karar ağacı birinci anadalı



Şekil 3.8. Karar Ağacı Birinci Anadalı

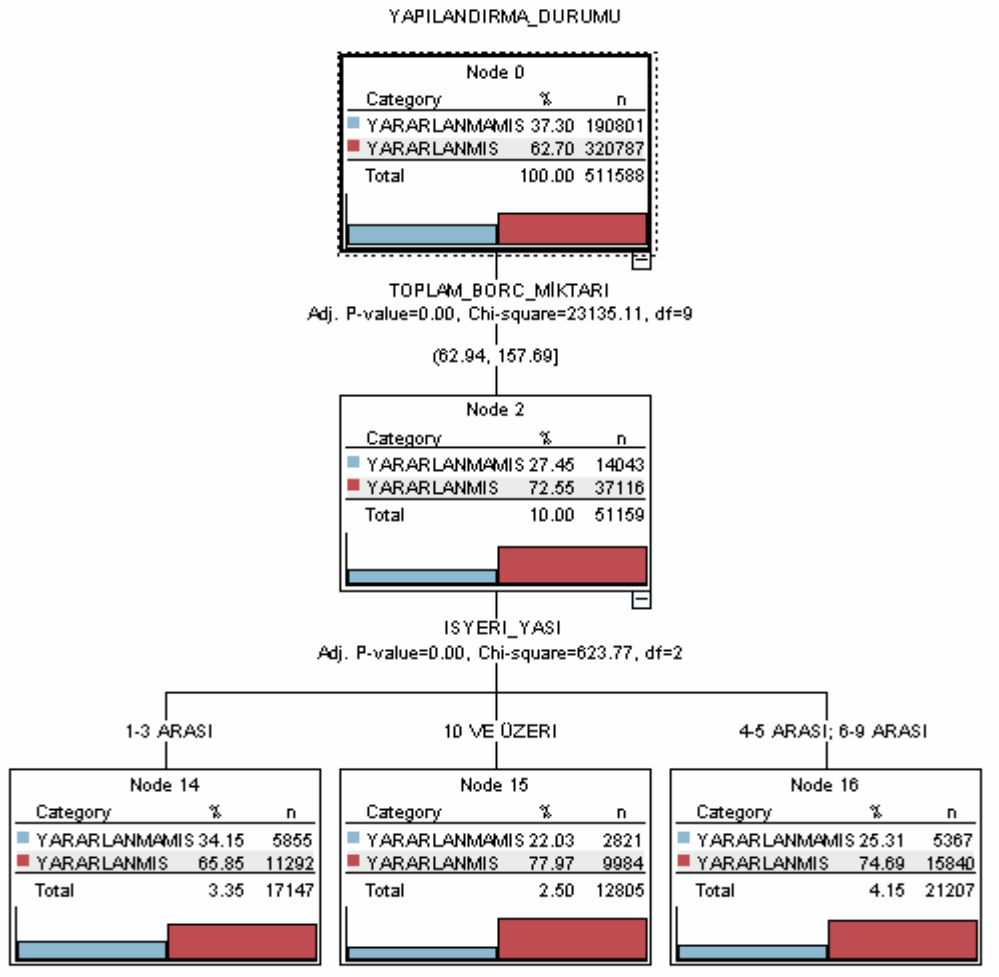
Şekil 3.8.'den görüldüğü gibi; toplam borç miktarı 62,94 liraya eşit veya daha küçük olan işyeri sayısı 51 161 olup, bu işyerlerinin 39 995'i (% 78) yapılandırmadan yararlanmış, 11 166'sı (% 22) ise yapılandırmadan yararlanmamıştır. Toplam Borç Miktarı 62,94 liraya eşit veya daha küçük olan işyerleri için dallara ayırmada önemli bulunan bir sonraki açıklayıcı değişken *işyeri yaşı* olup;

- İşyeri yaşı “1–3 ARASI” olan 19 396 işyerinin 14 135’i (% 73) yapılandırmadan yararlanmış, 5261’i (% 27) yapılandırmadan yararlanmamıştır,

- İşyeri yaşı “4–5 ARASI” ve “6–9 ARASI” olan 20 124 işyerinin 16 216’sı (% 81) yapılandırmadan yararlanmış, 3908’i (% 19) yapılandırmadan yararlanmamıştır,

- İşyeri yaşı “10 VE ÜZERİ” olan 11 641 işyerinin 9644’ü (% 83) yapılandırmadan yararlanmış, 1997’si (% 17) yapılandırmadan yararlanmamıştır.

3.5.2. Karar ağacı ikinci anadalı



Şekil 3.9. Karar Ağacı İkinci Anadalı

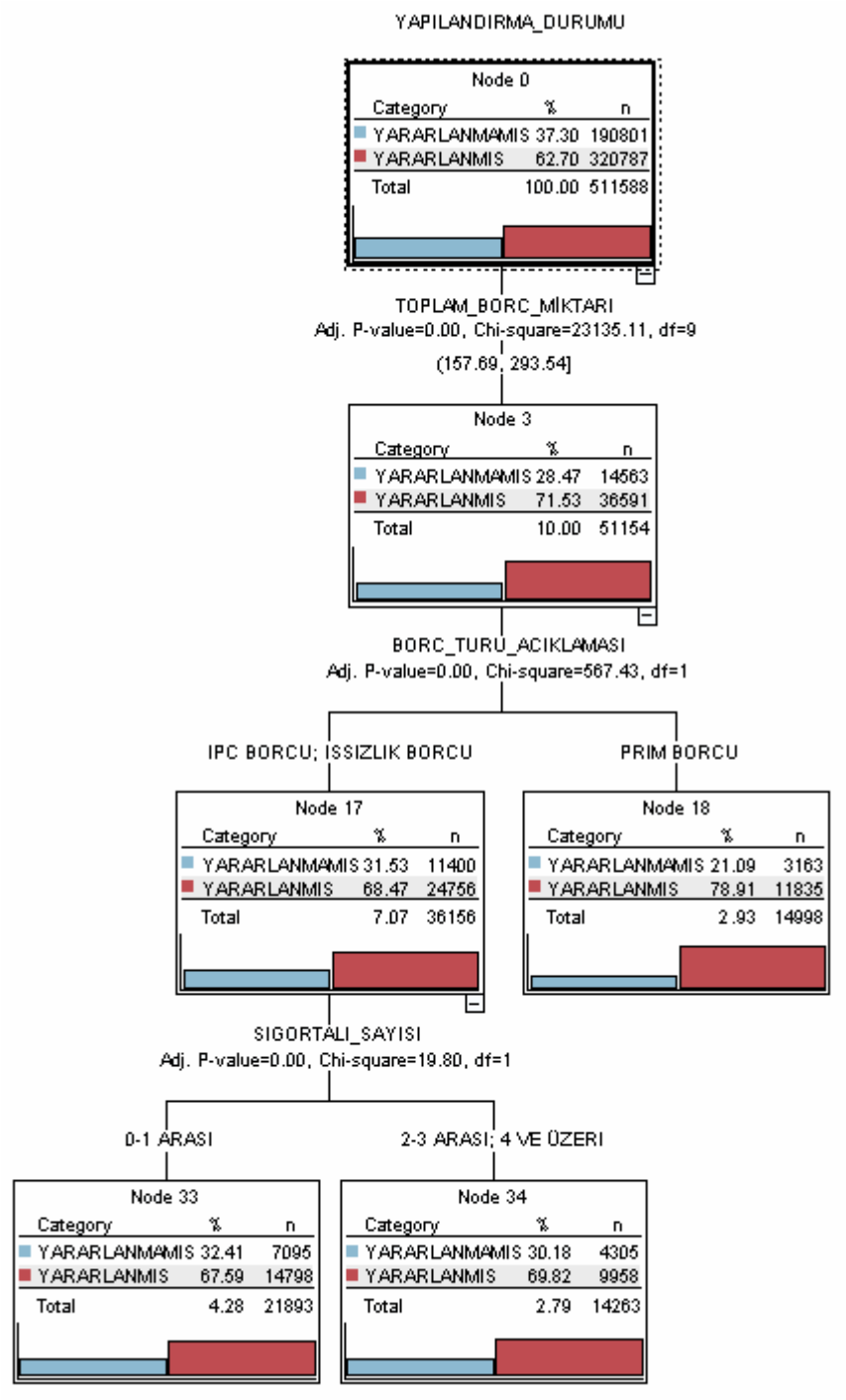
Şekil 3.9.'dan görüldüğü gibi; toplam borç miktarı 62,94 liradan büyük ve 157,69 liraya eşit veya daha küçük olan 51 159 işyeri olup; bu işyerlerinin 37 116'sı (% 73) yapılandırmadan yararlanmış, 14 043'ü (% 27) yapılandırmadan yararlanmamıştır. Toplam Borç Miktarı bu aralıkta olan işyerleri içerisinde, toplam borç miktarı açıklayıcı değişkeninden sonra dallara ayırmada önemli bulunan bir sonraki açıklayıcı değişken *işyeri yaşı* olup;

- İşyeri yaşı "1–3 ARASI" olan 17 147 işyerinin 11 292'si (% 66) yapılandırmadan yararlanmış, 5855'i (% 34) yapılandırmadan yararlanmamıştır,

- İşyeri yaşı "4–5 ARASI" ve "6–9 ARASI" olan 21 207 işyerinin 15 840'ı (% 75) yapılandırmadan yararlanmış, 5367'si (% 25) yapılandırmadan yararlanmamıştır,

- İşyeri yaşı "10 VE ÜZERİ" olan 12 805 işyerinin 9984'ü (% 78) yapılandırmadan yararlanmış, 2821'i (% 22) yapılandırmadan yararlanmamıştır.

3.5.3. Karar ağacı üçüncü anadalı



Şekil 3.10. Karar Ağacı Üçüncü Anadalı

Şekil 3.10.'dan görüldüğü gibi; toplam borç miktarı 157,69 liradan büyük ve 293,54 liraya eşit veya daha küçük olan 51 154 işyeri olup; bu işyerlerinin 36 591'i (% 72)

yapılandırmadan yararlanmış, 14 563'ü (% 28) yapılandırmadan yararlanmamıştır. Toplam Borç Miktarı bu aralıkta olan işyerleri içerisinde toplam borç miktarı açıklayıcı değişkeninden sonra dallara ayırmada önemli bulunan bir sonraki açıklayıcı değişken *borç türü açıklaması* olup;

- Borç türü açıklaması idari para cezası ve işsizlik borcu olan 36 156 işyerinin 24 756 'sı (% 68) yapılandırmadan yararlanmış, 11 400'ü (% 32) yapılandırmadan yararlanmamıştır,

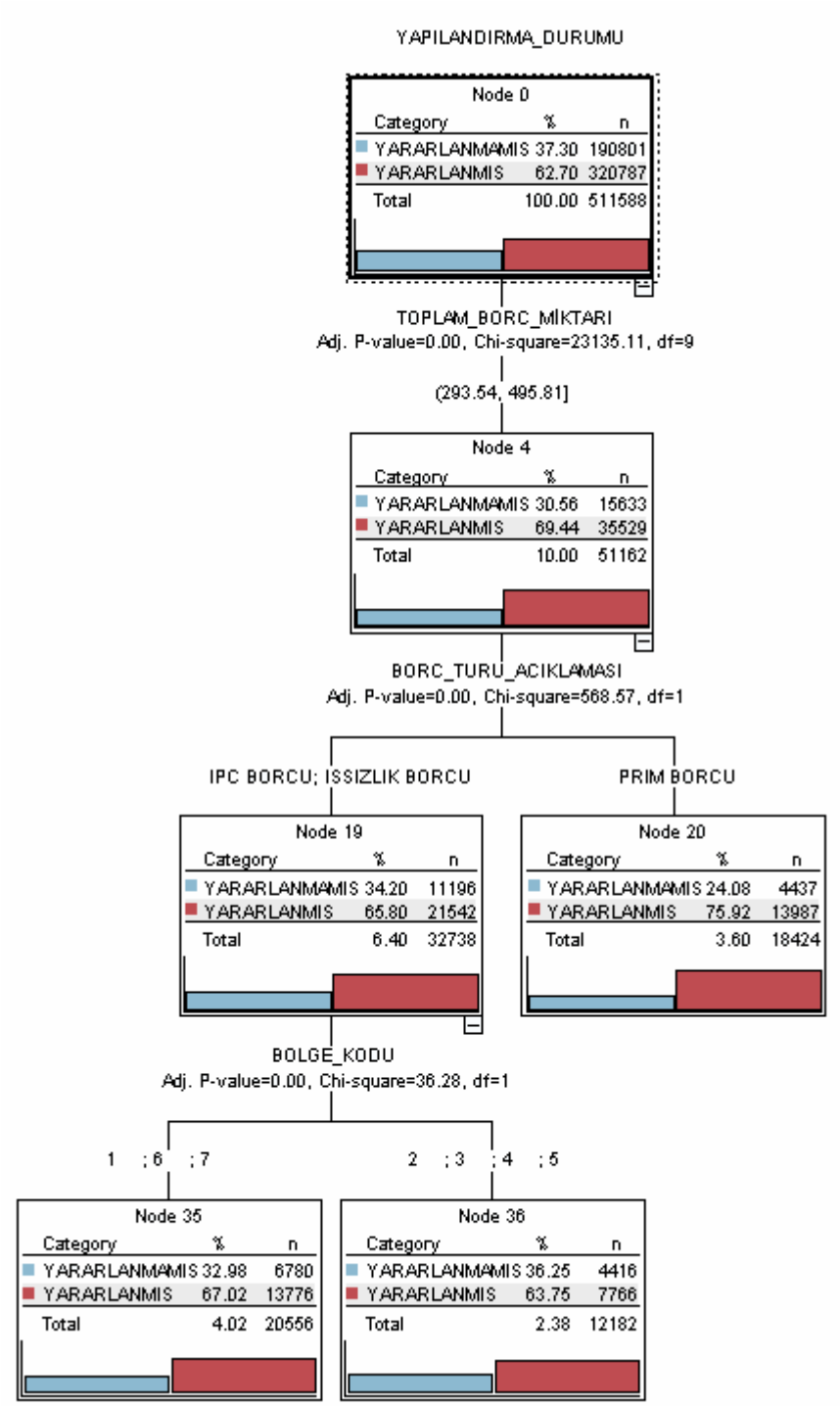
- Borç türü açıklaması prim borcu olan 14 998 işyerinin 11 835'i (% 79) yapılandırmadan yararlanmış, 3163'ü (% 21) yapılandırmadan yararlanmamıştır.

Borç türü açıklaması idari para cezası ve işsizlik borcu olan işyerlerinde dallara ayırmada önemli bulunan bir sonraki açıklayıcı değişken *sigortalı sayısı* olup;

- Sigortalı sayısı "0–1 ARASI" olan 21 893 işyerinin 14 798'i (% 68) yapılandırmadan yararlanmış, 7095'i (% 32) yapılandırmadan yararlanmamıştır,

- Sigortalı sayısı "2–3 ARASI" ve "4 VE ÜZERİ" olan 14 263 işyerinin 9958'i (% 70) yapılandırmadan yararlanmış, 4305'i (% 30) yapılandırmadan yararlanmamıştır.

3.5.4. Karar ağacı dördüncü anadalı



Şekil 3.11. Karar Ağacı Dördüncü Anadalı

Şekil 3.11.'den görüldüğü gibi; toplam borç miktarı 293,54 liradan büyük ve 495,81 liraya eşit veya daha küçük olan 51 162 işyeri olup, bu işyerlerinin 35 529'u (% 69) yapılandırmadan yararlanmış, 15 633'ü (% 31) yapılandırmadan yararlanmamıştır. Toplam Borç Miktarı bu aralıkta olan işyerleri içerisinde toplam borç miktarı açıklayıcı değişkeninden sonra dallara ayırmada önemli bulunan bir sonraki açıklayıcı değişken *borç türü açıklaması* olup,

- Borç türü açıklaması idari para cezası ve işsizlik borcu olan 32 738 işyerinin 21 542 'si (% 66) yapılandırmadan yararlanmış, 11 196'sı (% 34) yapılandırmadan yararlanmamıştır,

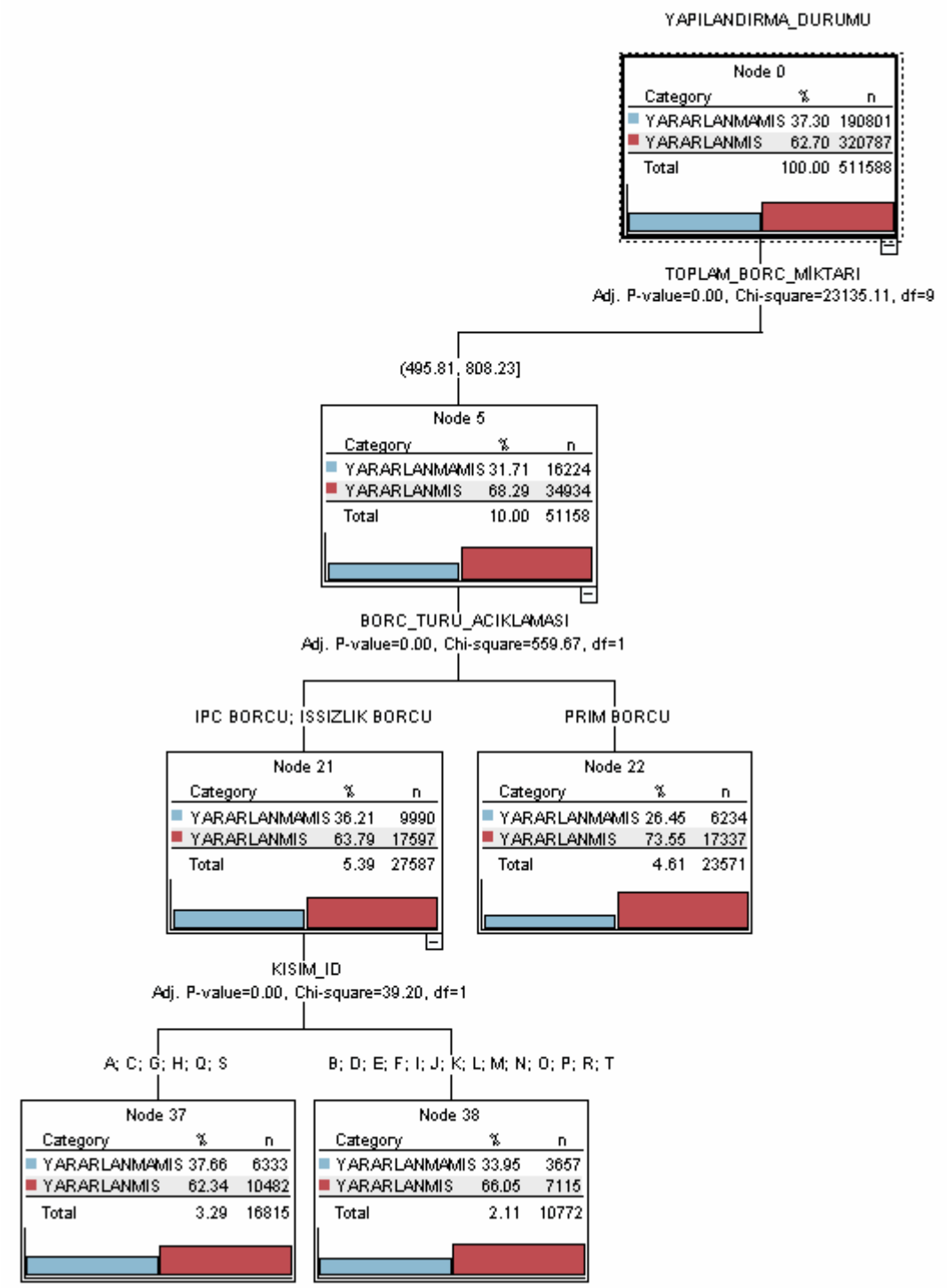
- Borç türü açıklaması prim borcu olan 18 424 işyerinin 13 987'si (% 76) yapılandırmadan yararlanmış, 4437'si (% 24) yapılandırmadan yararlanmamıştır.

Borç türü açıklaması idari para cezası ve işsizlik borcu olan işyerlerinde dallara ayırmada önemli bulunan bir sonraki açıklayıcı değişken *bölge kodu* olup;

- Bölge kodu "1; 6; 7" olan 20 556 işyerinin 13 776'sı (% 67) yapılandırmadan yararlanmış, 6780'i (% 33) yapılandırmadan yararlanmamıştır,

- Bölge kodu "2; 3; 4; 5" olan 12 182 işyerinin 7766'sı (% 64) yapılandırmadan yararlanmış, 4416'sı (% 36) yapılandırmadan yararlanmamıştır.

3.5.5. Karar ağacı beşinci anadalı



Şekil 3.12. Karar Ağacı Beşinci Anadalı

Şekil 3.12'den görüldüğü gibi; toplam borç miktarı 495,81 liradan büyük ve 808,23 liraya eşit veya daha küçük olan 51 158 işyeri olup, bu işyerlerinin 34 934'ü (% 68) yapılandırmadan yararlanmış, 16 224 'ü (% 32) yapılandırmadan yararlanmamıştır. Toplam Borç Miktarı bu aralıkta olan işyerleri içerisinde toplam borç miktarı açıklayıcı değişkeninden sonra dallara ayırmada önemli bulunan bir sonraki açıklayıcı değişken *borç türü açıklaması* olup;

- Borç türü açıklaması idari para cezası ve işsizlik borcu olan 27 587 işyerinin 17 597 'si (% 64) yapılandırmadan yararlanmış, 9990'ı (% 36) yapılandırmadan yararlanmamıştır,

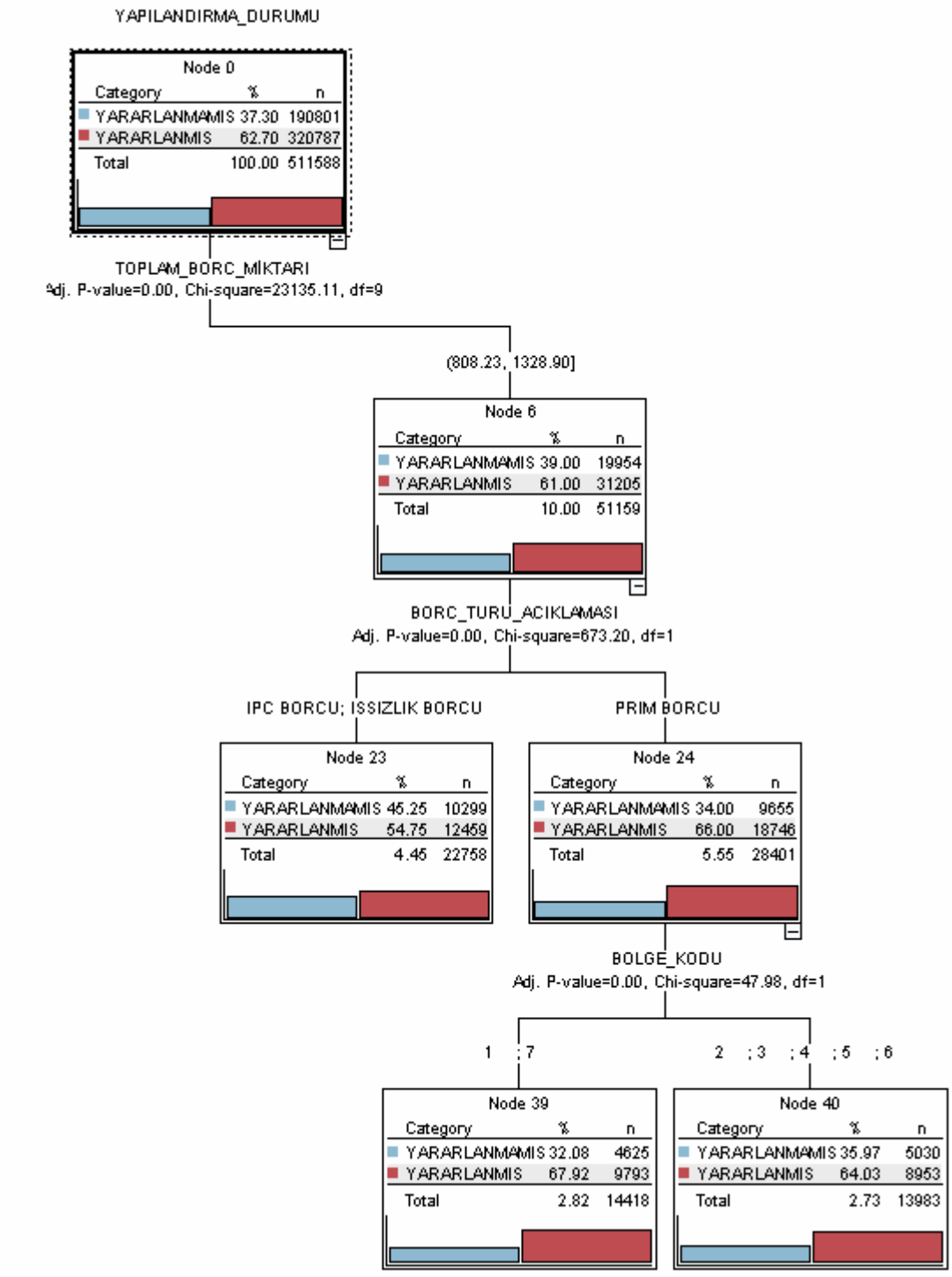
- Borç türü açıklaması prim borcu olan 23 571 işyerinin 17 337'si (% 74) yapılandırmadan yararlanmış, 6234'ü (% 26) yapılandırmadan yararlanmamıştır.

Borç türü açıklaması idari para cezası ve işsizlik borcu olan işyerlerinde dallara ayırmada önemli bulunan bir sonraki açıklayıcı değişken *Kısım Id* olup;

- Kısım Id "A; C; G; H; Q; S" olan 16 815 işyerinin 10 482'si (% 62) yapılandırmadan yararlanmış, 6333'ü (% 38) yapılandırmadan yararlanmamıştır,

- Kısım Id "B; D; E; F; I; J; K; L; M; N; O; P; R; T" olan 10 772 işyerinin 7115'i (% 66) yapılandırmadan yararlanmış, 3657'si (% 34) yapılandırmadan yararlanmamıştır.

3.5.6. Karar ağacı altıncı anadalı



Şekil 3.13. Karar Ağacı Altıncı Anadalı

Şekil 3.13.'den görüldüğü gibi; toplam borç miktarı 808,23 liradan büyük ve 1328,90 liraya eşit veya daha küçük olan 51 159 işyeri olup, bu işyerlerinin 31 205'i (% 61) yapılandırmadan yararlanmış,19 954'ü (% 39) ise yapılandırmadan yararlanmamıştır. Toplam Borç Miktarı bu aralıkta olan işyerleri içerisinde toplam borç miktarı açıklayıcı değişkeninden sonra dallara ayırmada önemli bulunan bir sonraki açıklayıcı değişken *borç türü açıklaması* olup;

- Borç türü açıklaması idari para cezası ve işsizlik borcu olan 22 758 işyerinin 12 459'u (% 55) yapılandırmadan yararlanmış, 10 299'u (% 45) yapılandırmadan yararlanmamıştır,

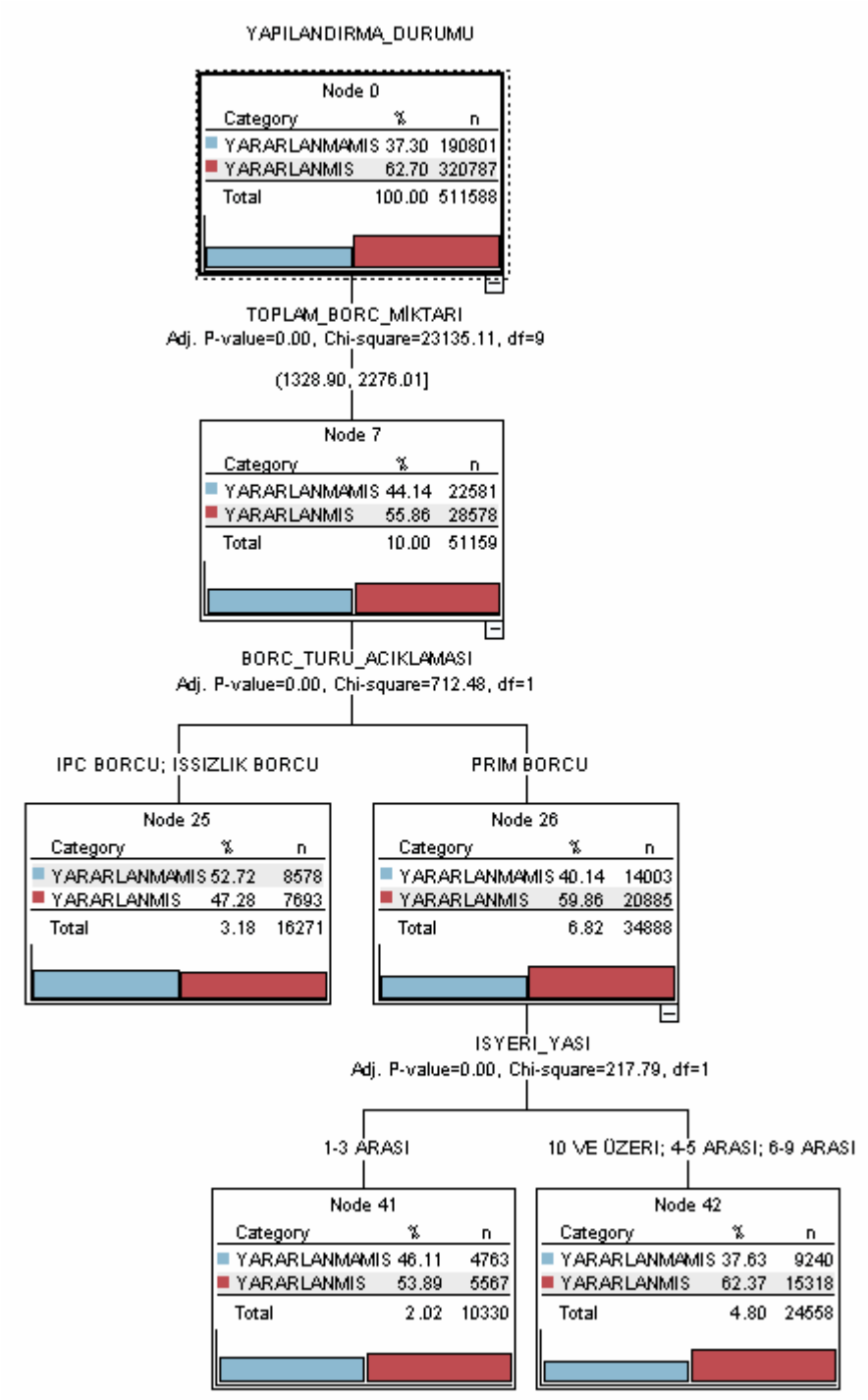
- Borç türü açıklaması prim borcu olan 28 401 işyerinin 18 746'sı (% 66) yapılandırmadan yararlanmış, 9655'i (% 34) yapılandırmadan yararlanmamıştır.

Borç türü açıklaması prim borcu olan işyerlerinde dallara ayırmada önemli bulunan bir sonraki açıklayıcı değişken *bölge kodu* olup;

- Bölge kodu "1; 7" olan 14 418 işyerinin 9793'ü (% 68) yapılandırmadan yararlanmış, 4625'i (% 32) yapılandırmadan yararlanmamıştır,

- Bölge kodu "2; 3; 4; 5; 6" olan 13 983 işyerinin 8953'ü (% 64) yapılandırmadan yararlanmış, 5030'u (% 36) yapılandırmadan yararlanmamıştır.

3.5.7. Karar ağacı yedinci anadalı



Şekil 3.14. Karar Ağacı Yedinci Anadalı

Şekil 3.14.'den görüldüğü gibi; toplam borç miktarı 1328,90 liradan büyük ve 2276,01 liraya eşit veya daha küçük olan 51 159 işyeri olup, bu işyerlerinin 28 578'i

(% 56) yapılandırmadan yararlanmış, 22 581'i (% 44) ise yapılandırmadan yararlanmamıştır. Toplam Borç Miktarı bu aralıkta olan işyerleri içerisinde toplam borç miktarı açıklayıcı değişkeninden sonra dallara ayırmada önemli bulunan bir sonraki açıklayıcı değişken *borç türü açıklaması* olup;

- Borç türü açıklaması idari para cezası ve işsizlik borcu olan 16 271 işyerinin 7693'ü (% 47) yapılandırmadan yararlanmış, 8578'i (% 53) yapılandırmadan yararlanmamıştır,

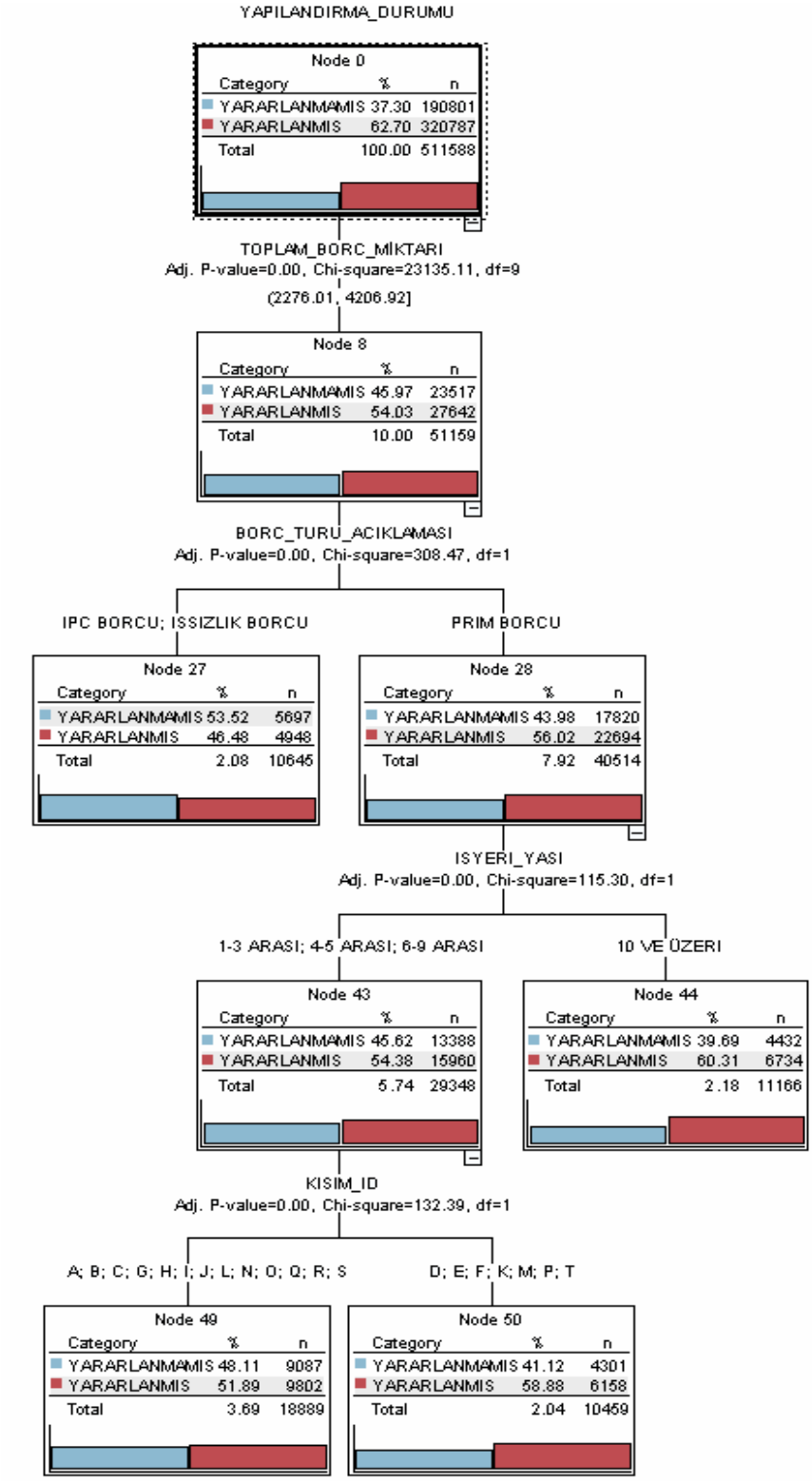
- Borç türü açıklaması prim borcu olan 34 388 işyerinin 20 885'i (% 60) yapılandırmadan yararlanmış, 14 003'i (% 40) yapılandırmadan yararlanmamıştır.

Borç türü açıklaması prim borcu olan işyerlerinde dallara ayırmada önemli bulunan bir sonraki açıklayıcı değişken *işyeri yaşı* olup;

- İşyeri yaşı "1–3 ARASI" olan 10 330 işyerinin 5567'si (% 54) yapılandırmadan yararlanmış, 4763'ü (% 46) yapılandırmadan yararlanmamıştır,

- İşyeri yaşı "4–5 ARASI", "6–9 ARASI" ve "10 VE ÜZERİ" olan 24 558 işyerinin 15 318'i (% 62) yapılandırmadan yararlanmış, 9240'ı (% 38) yapılandırmadan yararlanmamıştır.

3.5.8. Karar ağacı sekizinci anadalı



Şekil 3.15. Karar Ağacı Sekizinci Anadalı

Şekil 3.15.'den görüldüğü gibi; toplam borç miktarı 2276,01 liradan büyük ve 4206,92 liraya eşit veya daha küçük olan 51 159 işyeri olup, bu işyerlerinin 27 6142'si (% 54) yapılandırmadan yararlanmış, 23 517'si (% 46) yapılandırmadan yararlanmamıştır. Toplam Borç Miktarı bu aralıkta olan işyerleri içerisinde toplam borç miktarı açıklayıcı değişkeninden sonra dallara ayırmada önemli bulunan bir sonraki açıklayıcı değişken *borç türü açıklaması* olup;

- Borç türü açıklaması idari para cezası ve işsizlik borcu olan 10 645 işyerinin 4948'i (% 46) yapılandırmadan yararlanmış, 5697'si (% 54) yapılandırmadan yararlanmamıştır,

- Borç türü açıklaması prim borcu olan 40 514 işyerinin 22 694'ü (% 56) yapılandırmadan yararlanmış, 17 820'si (% 44) yapılandırmadan yararlanmamıştır.

Borç türü açıklaması prim borcu olan işyerlerinde dallara ayırmada önemli bulunan bir sonraki açıklayıcı değişken *işyeri yaşı* olup;

- İşyeri yaşı "1–3 ARASI", "4–5 ARASI"ve "6–9 ARASI" olan 29 348 işyerinin 15 960'ı (% 54) yapılandırmadan yararlanmış, 13 388'i (% 46) yapılandırmadan yararlanmamıştır,

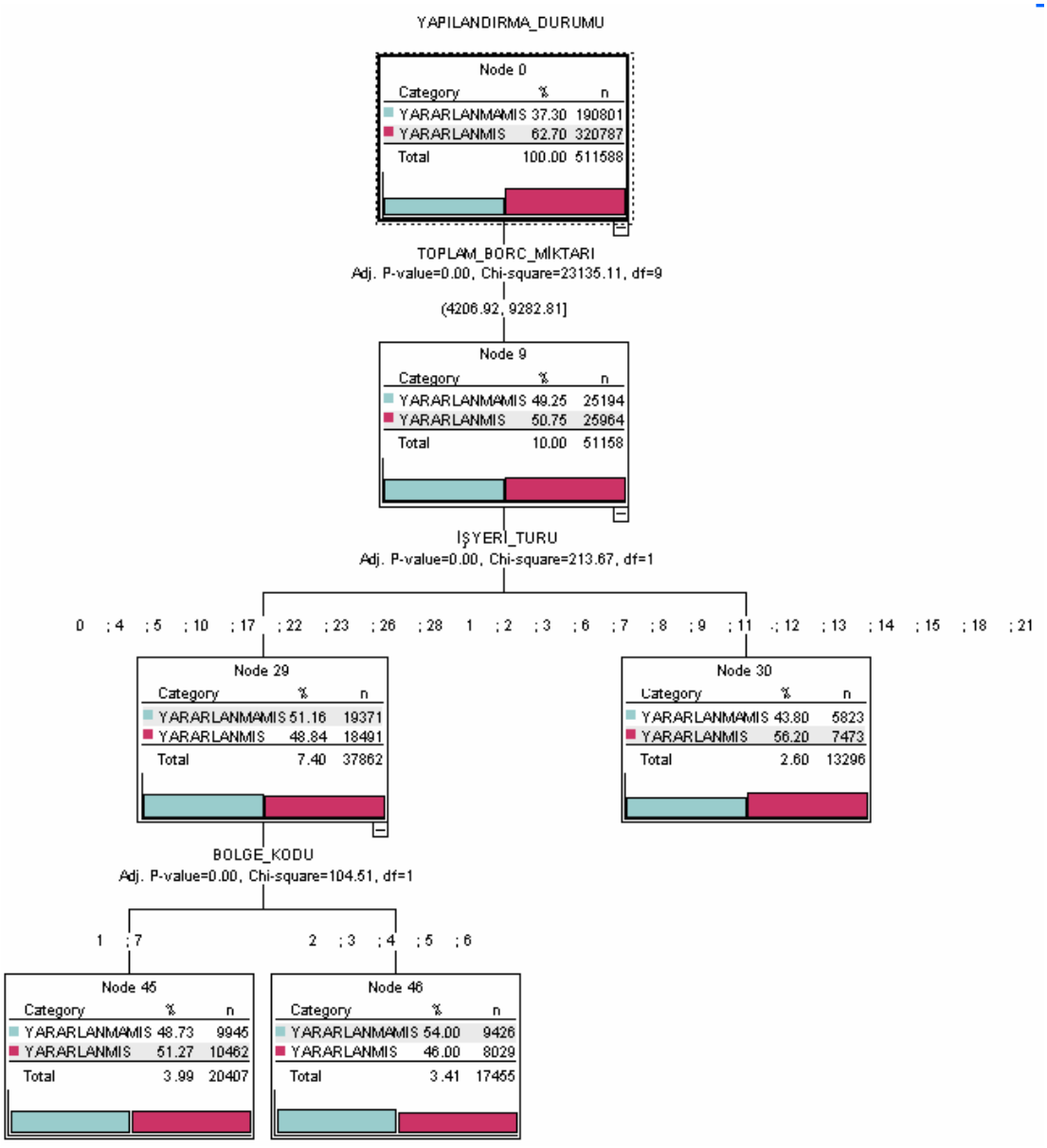
- İşyeri yaşı "10 VE ÜZERİ" olan 11 166 işyerinin 6734'ü (% 60) yapılandırmadan yararlanmış, 4432'si (% 40) yapılandırmadan yararlanmamıştır.

İşyeri yaşı "1 – 3 ARASI", "4 – 5 ARASI"ve "6 – 9 ARASI" olan işyerlerinde dallara ayırmada önemli bulunan bir sonraki açıklayıcı değişken *Kısım Id* olup;

- Kısım Id "A; B; C; G; H; I; J; L; N; O; Q; R; S" olan 18 889 işyerinin 9802'si (% 52) yapılandırmadan yararlanmış, 9087'si (% 48) yapılandırmadan yararlanmamıştır,

- Kısım Id “D; E; F; K; M; P; T” olan 10 459 işyerinin 6158’i (% 59) yapılandırmadan yararlanmış, 4301’i (% 41) yapılandırmadan yararlanmamıştır.

3.5.9. Karar ağacı dokuzuncu anadalı



Şekil 3.16. Karar Ağacı Dokuzuncu Anadalı

Şekil 3.16.’dan görüldüğü gibi; toplam borç miktarı 4206,92 liradan büyük ve 9282,81 liraya eşit veya daha küçük olan 51 158 işyeri olup, bu işyerlerinin 25 964’ü

(% 51) yapılandırmadan yararlanmış, 25 194'ü (% 49) yapılandırmadan yararlanmamıştır. Toplam Borç Miktarı bu aralıkta olan işyerleri içerisinde toplam borç miktarı açıklayıcı değişkeninden sonra dallara ayırmada önemli bulunan bir sonraki açıklayıcı değişken *işyeri türü* olup;

- İşyeri türü kodu "0; 4; 5; 10; 17; 22; 23; 26; 28" olan 37 862 işyerinin 18 491'i (% 49) yapılandırmadan yararlanmış, 19 371'i (% 51) yapılandırmadan yararlanmamıştır,

- İşyeri türü kodu "1; 2; 3; 6; 7; 8; 9; 11; 12; 13; 14; 15; 18; 21" olan 13 296 işyerinin 7473'ü (% 56) yapılandırmadan yararlanmış, 5823'ü (% 44) yapılandırmadan yararlanmamıştır.

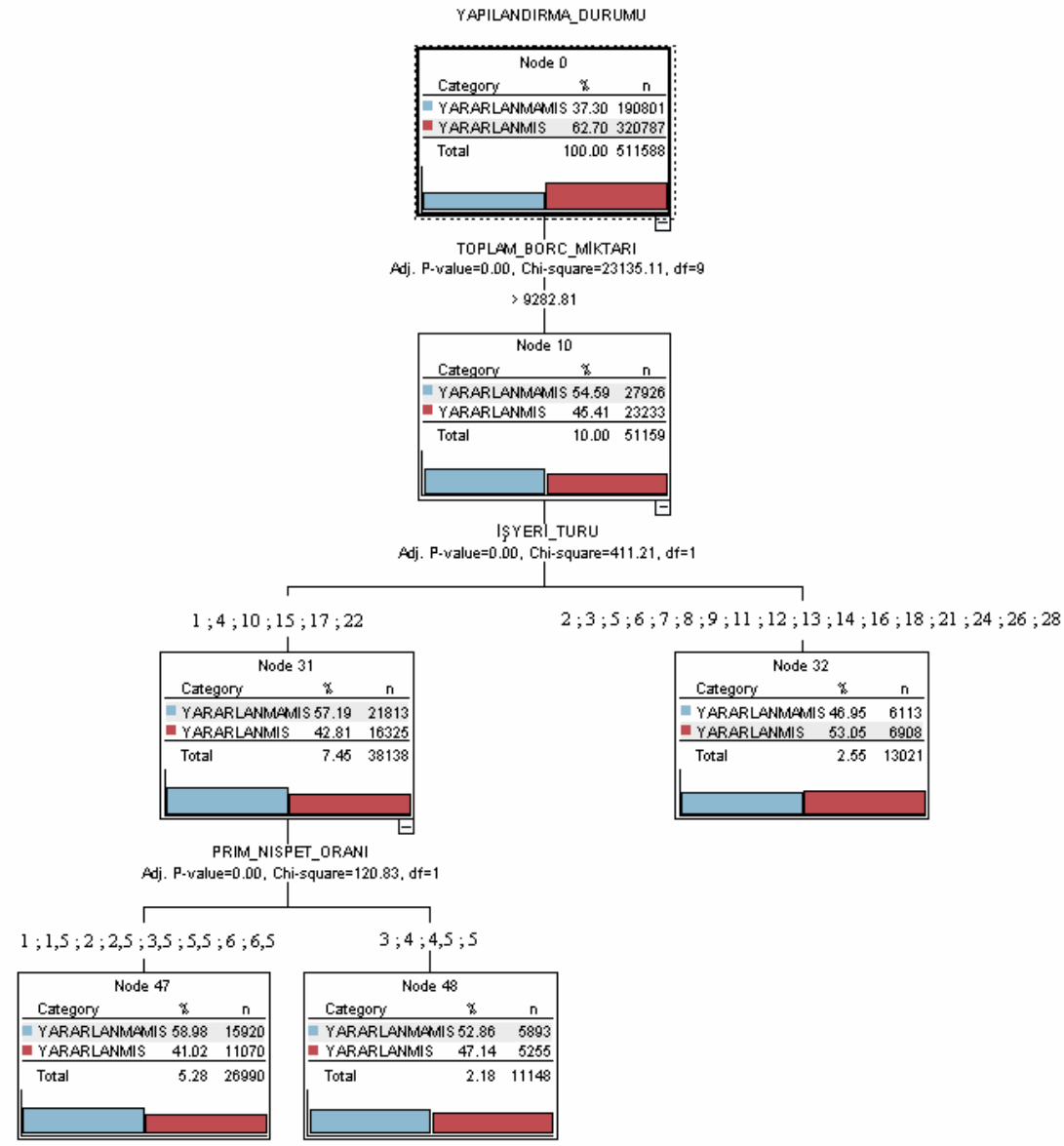
İşyeri türü kodu "0; 4; 5; 10; 17; 22; 23; 26; 28" olan işyerlerinde dallara ayırmada önemli bulunan bir sonraki açıklayıcı değişken *bölge kodu* olup;

- Bölge kodu "1; 7" olan 20 407 işyerinin 10 462'si (% 51) yapılandırmadan yararlanmış, 9945'i (% 49) yapılandırmadan yararlanmamıştır,

- Bölge kodu "2; 3; 4; 5; 6" olan 17 455 işyerinin 8029'u (% 46) yapılandırmadan yararlanmış, 9426'sı (% 54) yapılandırmadan yararlanmamıştır.

.

3.5.10. Karar ağacı onuncu anadalı



Şekil 3.17. Karar Ağacı Onuncu Anadalı

Şekil 3.17.'den görüldüğü gibi; toplam borç miktarı 9282,81 liradan büyük olan 51 159 işyeri olup, bu işyerlerinin 23 233'ü (% 45) yapılandırmadan yararlanmış, 27 926'sı (% 55) yapılandırmadan yararlanmamıştır. Toplam Borç Miktarı bu aralıkta olan işyerleri içerisinde toplam borç miktarı açıklayıcı değişkeninden sonra dallara ayırmada önemli bulunan bir sonraki açıklayıcı değişken *işyeri türü* olup;

- İşyeri türü kodu “1; 4; 10; 15; 17; 22” olan 38 138 işyerinin 16 325’i (% 43) yapılandırmadan yararlanmış, 21 813’ü (% 57) yapılandırmadan yararlanmamıştır,

- İşyeri türü kodu “2; 3; 5; 6; 7; 8; 9; 11; 12; 13; 14; 16; 18; 21; 24; 26; 28” olan 13 021 işyerinin 6908’i (% 53) yapılandırmadan yararlanmış, 6113’ü (% 47) yapılandırmadan yararlanmamıştır.

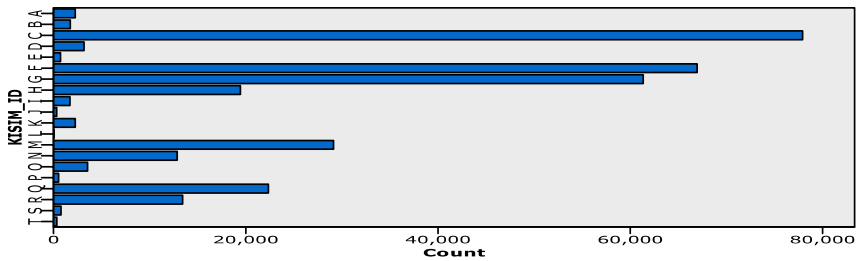
İşyeri türü kodu “1; 4; 10; 15; 17; 22” olan işyerlerinde dallara ayırmada önemli bulunan bir sonraki açıklayıcı değişken *prim nispet oranı* olup;

- Prim nispet oranı “1; 1,5; 2; 2,5; 3,5; 5,5; 6; 6,5” olan 26 990 işyerinin 11 070’i (% 41) yapılandırmadan yararlanmış, 15 920’si (% 59) yapılandırmadan yararlanmamıştır,

- Prim nispet oranı “3; 4; 4,5; 5” olan 11 148 işyerinin 5255’i (% 47) yapılandırmadan yararlanmış, 5893’ü (% 53) yapılandırmadan yararlanmamıştır.

Uygulamada kullanılan yapılandırmadan yararlanan işyeri verilerine ilişkin dağılım grafikleri ve dağılım yüzdeleri şekiller ve çizelgeler halinde aşağıda özetlenmektedir.

Yapılandırmadan yararlanan işyerlerinin sektörel dağılım grafiği ve dağılım yüzdeleri Şekil 3.18. ve Çizelge 3.9.’da ifade edilmektedir.



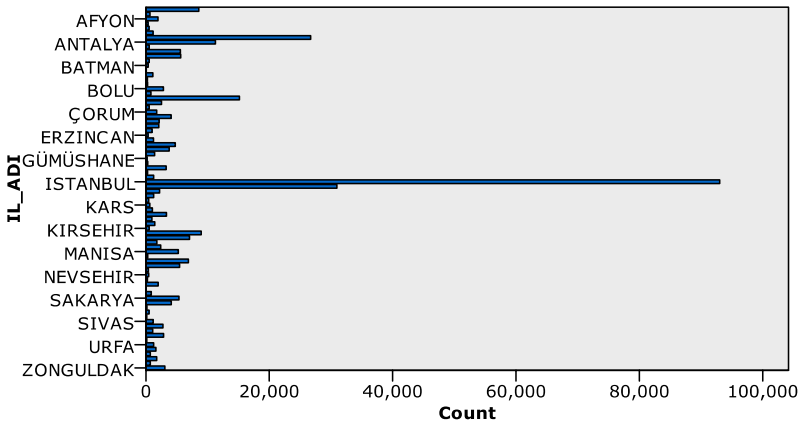
Şekil 3.18. Yapılandırmadan yararlanan işyerlerinin sektörel dağılım grafiği

Çizelge 3.9. Yapılandırmadan yararlanan işyerlerinin sektörel dağılım yüzdeleri

Kısım Id	Yüzde (%)	Sayı
A	0,70	2261
B	0,54	1743
C	24,29	77 924
D	0,99	3176
E	0,23	727
F	20,87	66 951
G	19,12	61 339
H	6,06	19 433
I	0,53	1709
J	0,10	323
K	0,70	2257
L	0,01	20
M	9,08	29 125
N	4,01	12 854
O	1,10	3538
P	0,16	526
Q	6,97	22 344
R	4,19	13 427
S	0,24	768
T	0,11	342
Toplam	100,0	320 787

Şekil 3.18. ve Çizelge 3.9.'dan anlaşılacağı üzere, yapılandırmadan en fazla yararlanan işyerlerinin Kısım Id alanı "C", "F" ve "G" olarak kodlanan, "İMALAT", "İNŞAAT" ve "TOPTAN VE PERAKENDE TICARET, MOTORLU KARA TASITLARININ VE MOTOSIKLETLERIN ONARIMI" alanında faaliyet gösteren işyerleri olduğu ve sırasıyla % 24,29 ; % 20,87 ve % 19,12 yüzdelik dağılım ile yapılandırmadan yararlandıkları görülmektedir.

Yapılandırmadan yararlanan işyerlerinin illere göre dağılım grafiği ve dağılım yüzdeleri Şekil 3.19. ve Çizelge 3.10.'da ifade edilmektedir.



Şekil 3.19. Yapılandırmadan yararlanan işyerlerinin illere göre dağılım grafiği

Çizelge 3.10. Yapılandırmadan yararlanan işyerlerinin illere göre dağılım yüzdeleri

İl adı	**Yüzde (%)**	**Sayı**
ADANA	2,67	8570
ADIYAMAN	0,21	673
AFYON	0,61	1949
AGRI	0,11	356
AKSARAY	0,15	496
AMASYA	0,36	1156
ANKARA	8,33	26 719
ANTALYA	3,51	11 263
ARTVIN	0,17	535
AYDIN	1,75	5598
BALIKESIR	1,76	5652
BARTIN	0,17	541
BATMAN	0,12	384
BAYBURT	0,02	59
BILECIK	0,35	1127
BINGÖL	0,08	251
BITLIS	0,08	263
BOLU	0,88	2837
BURDUR	0,25	811
BURSA	4,72	15 150
ÇANAKKALE	0,78	2517
ÇANKIRI	0,17	535
ÇORUM	0,55	1750
DENIZLI	1,27	4084
DIYARBAKIR	0,67	2152

Çizelge 3.10. (Devam) Yapılandırmadan yararlanan işyerlerinin illere göre dağılım yüzdeleri

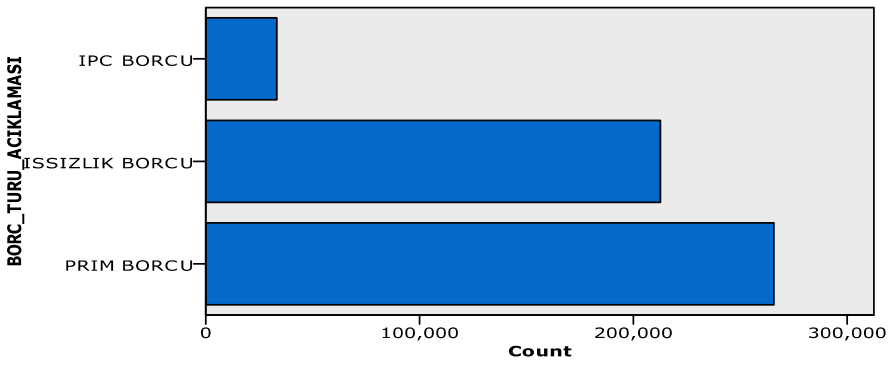
İl adı	Yüzde (%)	Sayı
EDIRNE	0,65	2092
ELAZIG	0,31	1000
ERZINCAN	0,11	358
ERZURUM	0,39	1241
ESKISEHIR	1,49	4764
GAZIANTEP	1,18	3783
GIRESUN	0,44	1408
GÜMÜSHANE	0,06	183
HAKKARI	0,09	286
HATAY	1,02	3288
IGDIR	0,08	266
ISPARTA	0,39	1261
ISTANBUL	29,00	93 029
IZMIR	9,65	30 960
KAHRAMANMARAS	0,69	2216
KARABüK	0,40	1268
KARAMAN	0,14	450
KARS	0,20	637
KASTAMONU	0,33	1051
KAYSERI	1,04	3331
KIRIKKALE	0,31	979
KIRKLARELI	0,45	1438
KIRSEHIR	0,17	542
KOCAELI	2,80	8966
KONYA	2,20	7063
KÜTAHYA	0,56	1784
MALATYA	0,74	2389
MANISA	1,64	5256
MARDIN	0,09	286
MERSIN	2,15	6889
MUGLA	1,70	5456
MUS	0,12	391
NEVSEHIR	0,14	458
NIGDE	0,08	252
ORDU	0,62	1991
OSMANIYE	0,00	2
RIZE	0,27	873
SAKARYA	1,67	5348
SAMSUN	1,28	4101
SIIRT	0,05	163
SINOP	0,16	513
SIRNAK	0,05	152

Çizelge 3.10. (Devam) Yapılandırmadan yararlanan işyerlerinin illere göre dağılım yüzdeleri

İl adı	Yüzde (%)	Sayı
SIVAS	0,37	1172
TEKIRDAG	0,86	2768
TOKAT	0,35	1111
TRABZON	0,89	2869
TUNCELI	0,04	137
URFA	0,39	1264
USAK	0,50	1615
VAN	0,22	718
YALOVA	0,54	1741
YOZGAT	0,22	720
ZONGULDAK	0,96	3080
Toplam	100,0	320 787

Şekil 3.19. ve Çizelge 3.10.'dan anlaşılacağı üzere, yapılandırmadan en fazla yararlanan illerin başında % 29 yüzdelik dağılım ile İstanbul ilinin, % 9,65 yüzdelik dağılım ile İzmir ilinin , % 8,33 yüzdelik dağılım ile Ankara ilinin, % 4,72 yüzdelik dağılım ile Bursa ve % 3,51 yüzdelik dağılım ile Antalya ilinin geldiği görülmektedir.

Yapılandırmadan yararlanan işyerlerinin borç türlerine göre dağılım grafiği ve dağılım yüzdeleri Şekil 3.20. ve Çizelge 3.11.'de ifade edilmektedir.



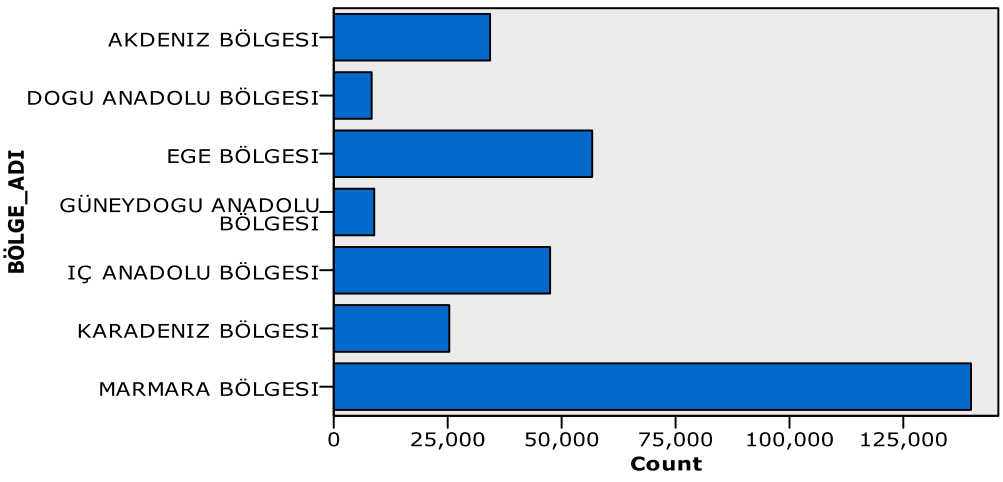
Şekil 3.20. Yapılandırmadan yararlanan işyerlerinin borç türü açıklamasına göre dağılım grafiği

Çizelge 3.11. Yapılandırmadan yararlanan işyerlerinin borç türü açıklamasına göre dağılım yüzdeleri

Borç türü açıklaması	Yüzde (%)	Sayı
IPC BORCU	6,10	19 574
ISSIZLIK BORCU	43,92	140 891
PRIM BORCU	49,98	160 322
Toplam	100,0	320 787

Şekil 3.20. ve Çizelge 3.11.'den anlaşılacağı üzere; işyerlerinin % 49,98'inin prim borcunu yapılandırdığı, % 43,92'sinin işsizlik borcunu yapılandırdığı ve % 6,10'unun ise idari para cezası borcunu yapılandırdığı görülmektedir.

Yapılandırmadan yararlanan işyerlerinin bulunduğu bölgenin bölge koduna göre dağılım grafiği ve dağılım yüzdeleri Şekil 3.21. ve Çizelge 3.12.'de ifade edilmektedir.



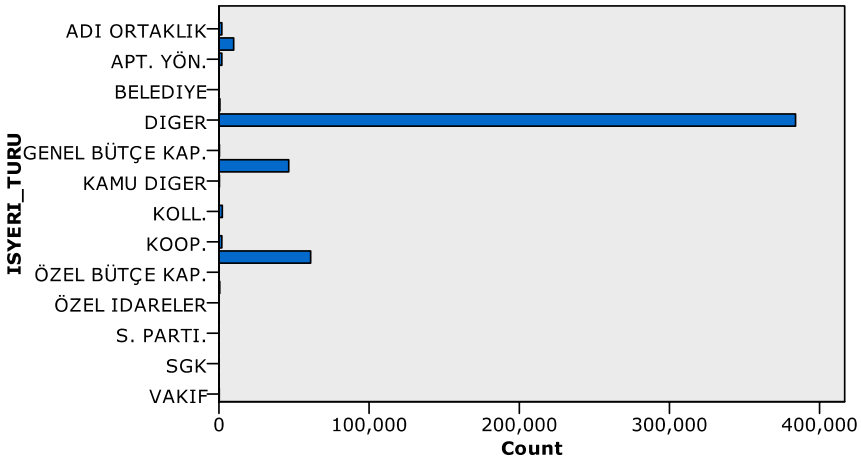
Şekil 3.21. Yapılandırmadan yararlanan işyerlerinin bulunduğu bölgelere göre dağılım grafiği

Çizelge 3.12. Yapılandırmadan yararlanan işyerlerinin bulunduğu bölgelere göre dağılım yüzdeleri

Bölge Adı	**Yüzde (%)**	**Sayı**
AKDENIZ BÖLGESI	10,69	34 300
DOGU ANADOLU BÖLGESI	2,59	8293
EGE BÖLGESI	17,68	56 702
GÜNEYDOGU ANADOLU BÖLGESI	2,76	8857
IÇ ANADOLU BÖLGESI	14,80	47 481
KARADENIZ BÖLGESI	7,89	25 326
MARMARA BÖLGESI	43,59	139 828
Toplam	100,0	320 787

Şekil 3.21. ve Çizelge 3.12.'den anlaşılacağı üzere; yapılandırmadan en fazla yararlanan işyerlerinin % 43,59'luk yüzdelik dağılım ile Marmara, % 14,80 yüzdelik dağılım ile İç Anadolu ve % 17,68 yüzdelik dağılım ile Ege Bölgesinde faaliyet gösteren işyerleri olduğu görülmektedir.

Yapılandırmadan yararlanan işyerlerinin işyeri türüne göre dağılım grafiği ve dağılım yüzdeleri Şekil 3.22. ve Çizelge 3.13.'de ifade edilmektedir.



Şekil 3.22. Yapılandırmadan yararlanan işyerlerinin işyeri türüne göre dağılım grafiği

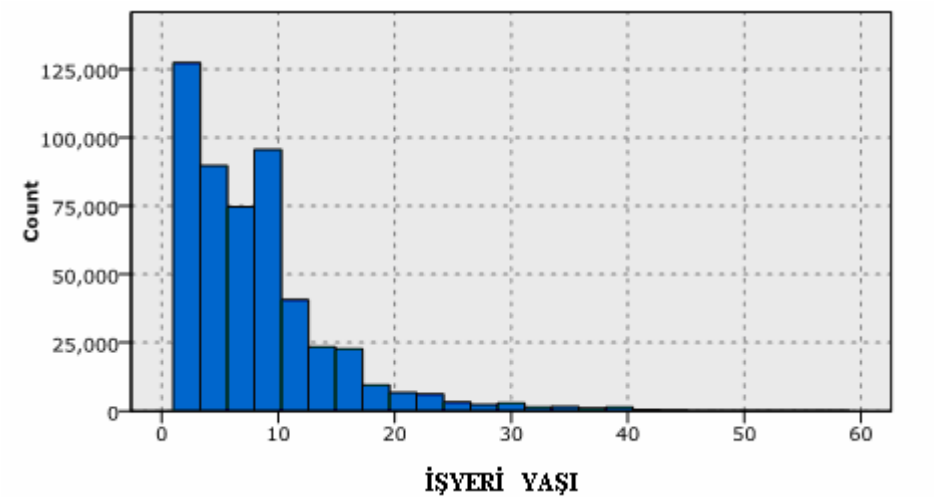
Çizelge 3.13. Yapılandırmadan yararlanan işyerlerinin işyeri türüne göre dağılım yüzdeleri

İşyeri türü	**Yüzde (%)**	**Sayı**
ADI. KOM.	0,01	38
ADI ORTAKLIK	0,43	1381
ANONIM	2,16	6944
APT. YÖN.	0,49	1560
BASIN	0,00	3
BELEDIYE	0,01	21
DERNEK	0,14	450
DIGER	73,81	236 766
DÜZENLEYICI-DENETLEYICI.	0,00	1
GENEL BÜTÇE KAP.	0,07	220
GERÇEK KISI	9,56	30 654
KAMU DIGER	0,05	168
KAMU TÜZEL KISILIKLERI	0,00	4
KOLL.	0,47	1507
KONSOLOSLUKLAR	0,00	8
KOOP.	0,42	1359
LMT.	12,17	39 053
ÖZEL BÜTÇE KAP.	0,01	26
ÖZEL DIGER	0,11	346
ÖZEL IDARELER	0,00	2
PAYLI KOM.	0,02	56
S. PARTI.	0,00	7
SENDIKA	0,02	65
SGK	0,00	4
SPOR KL.	0,01	17
VAKIF	0,04	127
Toplam	100,0	320 787

Şekil 3.22. ve Çizelge 3.13.'den anlaşılacağı üzere; yapılandırmadan yararlanan işyerlerinin başında % 73,81 yüzdelik dağılım ile mevcut bulunan işyerleri dışında kalan "DİĞER" olarak kodlanan işyerleri yararlanmış olup, bunu % 12,17 yüzdelik dağılım ile limited şirketlerin ve % 9,56 yüzdelik dağılım ile gerçek kişilerin takip ettiği görülmektedir.

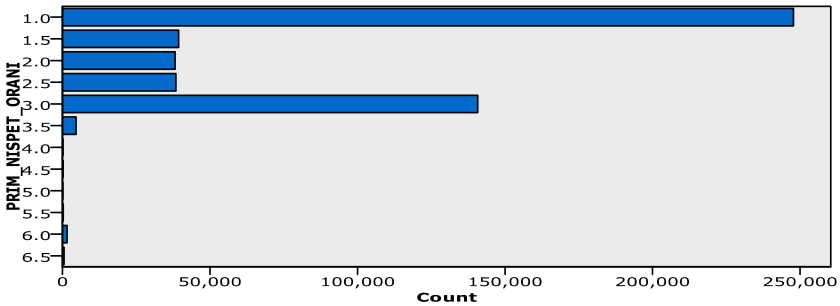
Yapılandırmadan yararlanan işyerlerinin işyeri yaşına göre dağılım grafiği Şekil 3.23.'de ifade edilmektedir.

Şekil 3.23. Yapılandırmadan yararlanan işyerlerinin işyeri yaşına göre dağılım grafiği



Şekil 3.23'den anlaşılacağı üzere, yapılandırmadan yararlanan işyerlerinin yaşları arttıkça yapılandırmadan daha az yararlandıkları görülmektedir. Yeni açılan işyerlerinin yapılandırmaya daha fazla başvurduğu anlaşılmakta olup, bu durum yeni açılan işyerlerinin ayakta kalma çabası olarak değerlendirilebilir.

Yapılandırmadan yararlanan işyerlerinin prim nispet oranlarına göre dağılım grafiği ve dağılım yüzdeleri Şekil 3.24. ve Çizelge 3.14.'de ifade edilmektedir.



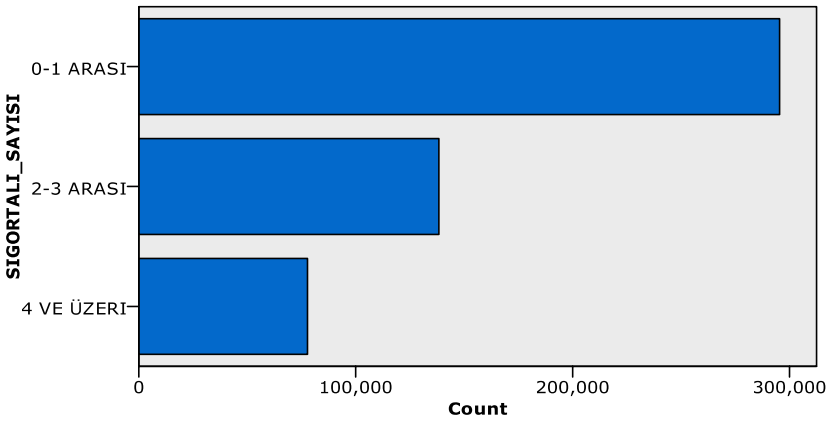
Şekil 3.24. Yapılandırmadan yararlanan işyerlerinin prim nispet oranına göre dağılım grafiği

Çizelge 3.14. Yapılandırmadan yararlanan işyerlerinin prim nispet oranına göre dağılım yüzdeleri

Prim nispet oranı	Yüzde (%)	Sayı
1	48,28	154 878
1,5	7,49	24 032
2	7,45	23 913
2,5	7,34	23 536
3	28,11	90 173
3,5	0,86	2743
4	0,02	77
4,5	0,02	80
5	0,00	7
5,5	0,05	161
6	0,29	922
6,5	0,08	265
Toplam	100,0	320 787

Şekil 3.24. ve Çizelge 3.14.'den anlaşılacağı üzere; % 48,28 yüzdelik dağılım ile prim nispet oranı 1 olan ve % 28,11 yüzdelik dağılım ile prim nispet oranı 3 olan işyerlerinin yapılandırmadan daha fazla yararlandıkları görülmektedir.

Yapılandırmadan yararlanan işyerlerinin çalıştırdıkları sigortalı sayısına göre dağılım grafiği ve dağılım yüzdeleri Şekil 3.25. ve Çizelge 3.15.'de ifade edilmektedir.



Şekil 3.25. Yapılandırmadan yararlanan işyerlerinin çalıştırdıkları sigortalıların sayısına göre dağılım grafiği

Çizelge 3.15. Yapılandırmadan yararlanan işyerlerinin çalıştırdıkları sigortalıların sayısına göre dağılım yüzdeleri

Sigortalı Sayısı	**Yüzde (%)**	**Sayı**
0–1 ARASI	58,78	188 552
2–3 ARASI	26,75	85 824
4 VE ÜZERI	14,47	46 411
Toplam	100,0	320 787

Şekil 3.25. ve Çizelge 3.15.'den anlaşılacağı üzere yapılandırmadan yararlanan işyerlerinin çalıştırdıkları sigortalıların sayısına göre dağılım grafiği incelendiğinde; bir işçi çalıştıran işyerlerinin % 58,78 yüzdelik dağılım ile yapılandırmadan daha fazla yararlandıkları görülmektedir.

4. SONUÇ VE ÖNERİLER

Bilişim teknolojisindeki hızlı değişim, sektörleri de etkilemiş ve aralarındaki rekabeti arttırmıştır. Bu nedenle kurumlar yeni teknolojilerden mümkün olduğunca faydalanarak, kurumların yararına olabilecek ileriye dönük kararlar almak istemektedirler.

Daha çok ticari sektörlerde yoğun olarak kullanılan VM'nin sosyal güvenliği ilgilendiren bir kamu kurumunda kullanılmasıyla farklı bir bakış açısı getirilmeye çalışılmıştır.

SGK'da gerçek veriler üzerinde yapılan VM uygulaması ile elde edilmiş olan sonuçlar, işverenlerin borç türlerinin çıkarılan kanunlar ile ne kadar yapılandırıldığını değerlendirmemize olanak sağlamaktadır.

Yapılandırma durumu hedef değişkeni için dallara ayırmada önemli bulunan açıklayıcı değişkenin "toplam borç miktarı" olduğu ve bu miktarın 10 alt sınıfa ayrıldığı görülmektedir. Bu alt sınıflardaki, 1. ve 2. düğümler için; "işyeri yaşı", 3. ve 8.düğümler arasındaki düğümler için "borç türü açıklaması", 9. ve 10. düğümler için ise "işyeri türü" açıklayıcı değişkenlerinin dallara ayırmada önemli bulunan açıklayıcı değişkenler olduğu görülmektedir.

511 588 işyeri arasında yapılandırmadan yararlanan 320 787 (% 63) işyeri olup, toplam borç miktarı açıklayıcı değişkeninden dallara ayrılan 10 alt sınıf yapılandırmadan yararlanan işyerlerinin durumu ele alınarak incelendiğinde;

- Bu işyerleri arasında toplam borç miktarı 62,94 liraya eşit veya daha küçük olan 39 995 (% 78) işyeri,

- Toplam borç miktarı 62,94 liradan büyük ve 157,69 liraya eşit veya daha küçük olan 37 116 (% 73) işyeri,

- Toplam borç miktarı 157,69 liradan büyük ve 293,54 liraya eşit veya daha küçük olan 36 591 (% 72) işyeri,

- Toplam borç miktarı 293,54 liradan büyük ve 495,81 liraya eşit veya daha küçük olan 35 529 (% 69) işyeri,

- Toplam borç miktarı 495,81 liradan büyük ve 808,23 liraya eşit veya daha küçük olan 34 934 (% 68) işyeri,

- Toplam borç miktarı 808,23 liradan büyük ve 1328,90 liraya eşit veya daha küçük olan 31 205 (% 61) işyeri,

- Toplam borç miktarı 1328,90 liradan büyük ve 2276,01 liraya eşit veya daha küçük olan 28 578 (% 56) işyeri,

- Toplam borç miktarı 2276,01 liradan büyük ve 4206,92 liraya eşit veya daha küçük olan 27 642 (% 54) işyeri,

- Toplam borç miktarı 4206,92 liradan büyük ve 9282,81 liraya eşit veya daha küçük olan 25 964 (% 51) işyeri,

- Toplam borç miktarı 9282,81 liradan daha büyük olan 23 233 (% 45) işyerinin olduğu ve bu işyerlerinin borçlarının % 78'inin 62,94 liraya eşit veya daha küçük borcu olan işyerlerinin oluşturduğu anlaşılmaktadır. Borç miktarı arttıkça yapılandırmadan yararlanma oranında bir düşüş olduğu gözlenmektedir. Bu durum toplam borç miktarı 62,94 liraya eşit veya daha küçük borcu olan işyerlerinin miktarı düşük olan borçlarını taksitlendirerek borçlarından kurtulmak istemeleri olarak değerlendirilebilir.

Karar ağacı sonucunda elde edilen 31 kural (sınıf) incelendiğinde;

- Toplam borç miktarı 62,94 liraya eşit veya daha küçük ve işyeri yaşı "10 VE ÜZERİ" olan işyerlerinin % 83'ünün yapılandırmadan yararlandığı görülmektedir.

- Toplam borç miktarı 62,94 liraya eşit veya daha küçük ve işyeri yaşı "4–5 ARASI", "6–9 ARASI" olan işyerlerinin % 81'inin yapılandırmadan yararlandığı görülmektedir.

- Toplam borç miktarı (157,69–293,54] aralığında olan işyerlerinin, % 79'unun prim borçlarını yapılandırdığı görülmektedir.

- Toplam borç miktarı (62,94–157,69] aralığında; işyeri yaşı "10 VE ÜZERİ" olan işyerlerinin % 78'inin yapılandırmadan yararlandığı görülmektedir.

- Toplam borç miktarı (293,54–495,81] aralığında; borç türü açıklaması prim borcu olan işyerlerinin % 76'sının yapılandırmadan yararlandığı görülmektedir.

- Toplam borç miktarı (62,94–157,69] aralığında; işyeri yaşı "4–5 ARASI", "6–9 ARASI" olan işyerlerinin % 75'inin yapılandırmadan yararlandığı görülmektedir.

- Toplam borç miktarı (495,81–808,23] aralığında, borç türü açıklaması prim borcu olan işyerlerinin % 74'ünün yapılandırmadan yararlandığı görülmektedir.

- Toplam borç miktarı 62,94 liraya eşit veya daha küçük ve işyeri yaşı "1–3 ARASI" olan işyerlerinin % 73'ünün yapılandırmadan yararlandığı görülmektedir.

- Toplam borç miktarı (157,69–293,54] aralığında, borç türü açıklaması idari para cezası ve işsizlik borcu olup, “2–3 ARASI” ve “4 VE ÜZERİ” sigortalı çalıştıran işyerlerinin % 70’inin yapılandırmadan yararlandığı görülmektedir.

- Toplam borç miktarı (808,23–1328,90] aralığında, borç türü açıklaması prim borcu olup, bölge kodu “1; 7” olan işyerlerinin % 68’inin yapılandırmadan yararlandığı görülmektedir.

- Toplam borç miktarı (157,69–293,54] aralığında, borç türü açıklaması idari para cezası ve işsizlik borcu olup, “0–1 ARASI” sigortalı çalıştıran işyerlerinin % 68’inin yapılandırmadan yararlandığı görülmektedir.

- Toplam borç miktarı (293,54–495,81] aralığında; borç türü açıklaması idari para cezası ve işsizlik borcu olup, bölge kodu “1; 6; 7” olan işyerlerinin % 67’sinin yapılandırmadan yararlandığı görülmektedir.

- Toplam borç miktarı (495,81–808,23] aralığında, borç türü açıklaması idari para cezası ve işsizlik borcu olup, Kısım Id “B; D; E; F; I; J; K; L; M; N; O; P; R; T” olan işyerlerinin % 66’sının yapılandırmadan yararlandığı görülmektedir.

- Toplam borç miktarı (62,94–157,69] aralığında; işyeri yaşı “1–3 ARASI” olan işyerlerinin % 66’sının yapılandırmadan yararlandığı görülmektedir.

- Toplam borç miktarı (808,23–1328,90] aralığında, borç türü açıklaması prim borcu olup, bölge kodu “2; 3; 4; 5; 6” olan işyerlerinin % 64’ünün yapılandırmadan yararlandığı görülmektedir.

- Toplam borç miktarı (293,54–495,81] aralığında; borç türü açıklaması idari para cezası ve işsizlik borcu olup, bölge kodu “2; 3; 4; 5” olan işyerlerinin % 64’ünün yapılandırmadan yararlandığı görülmektedir.

- Toplam borç miktarı (1328,90–2276,01] aralığında, borç türü açıklaması prim borcu olup, işyeri yaşı “4–5 ARASI”, “6–9 ARASI” ve “10 VE ÜZERİ” olan işyerlerinin % 62’sinin yapılandırmadan yararlandığı görülmektedir.

- Toplam borç miktarı (495,81–808,23] aralığında, borç türü açıklaması idari para cezası ve işsizlik borcu olup, Kısım Id “A; C; G; H; Q; S” olan işyerlerinin % 62’sinin yapılandırmadan yararlandığı görülmektedir.

- Toplam borç miktarı (2276,01–4206,92] aralığında prim borcu olup, işyeri yaşı “10 VE ÜZERİ” olan işyerlerinin % 60’ının yapılandırmadan yararlandığı görülmektedir.

- Toplam borç miktarı (2276,01–4206,92] aralığında prim borcu olup, işyeri yaşı “1–3 ARASI”, “4–5 ARASI” ve “6–9 ARASI” olan ve Kısım Id “D; E; F; K; M; P; T” olan işyerlerinin % 59’unun yapılandırmadan yararlandığı görülmektedir.

- Toplam borç miktarı (4206,92–9282,81] aralığında olup, işyerinin türü “1, 2, 3, 6, 7, 8, 9, 11, 12, 13, 14, 15, 18, 21” olarak kodlanan işyerlerinin % 56’sının yapılandırmadan yararlandığı görülmektedir.

- Toplam borç miktarı (808,23–1328,90] aralığında, borç türü açıklaması idari para cezası ve işsizlik borcu olan işyerlerinin ise % 55’inin yapılandırmadan yararlandığı görülmektedir.

- Toplam borç miktarı (1328,90–2276,01] aralığında, borç türü açıklaması prim borcu olup, işyeri yaşı “1–3 ARASI” olan işyerlerinin % 54’ünün yapılandırmadan yararlandığı görülmektedir.

- Toplam borç miktarı 9282,81 liradan büyük ve işyerinin türü “2, 3, 5, 6, 7, 8, 9, 11, 12, 13, 14, 16, 18, 21, 24, 26, 28” olarak kodlanan işyerlerinin % 53’ünün yapılandırmadan yararlandığı görülmektedir.

- Toplam borç miktarı (2276,01–4206,92] aralığında prim borcu olup, işyeri yaşı “1–3 ARASI”, “4–5 ARASI” ve “6–9 ARASI” olan ve Kısım Id “A; B; C; G; H; I; J; L; N; O; Q; R; S” olan işyerlerinin % 52’sinin yapılandırmadan yararlandığı görülmektedir.

- Toplam borç miktarı (4206,92–9282,81] aralığında olup, işyerinin türü “0; 4; 5; 10; 17; 22; 23; 26; 28” olarak kodlanan ve bölge kodu “1; 7” olan işyerlerinin % 51’inin yapılandırmadan yararlandığı görülmektedir.

- Toplam borç miktarı (1328,90–2276,01] aralığında, borç türü açıklaması idari para cezası ve işsizlik borcu olan işyerlerinin % 47’sinin yapılandırmadan yararlandığı görülmektedir.

- Toplam borç miktarı 9282,81 liradan büyük ve işyerinin türü “1; 4; 10; 15; 17; 22” olarak kodlanan ve prim nispet oranı “3; 4; 4,5; 5” olan işyerlerinin % 47’sinin yapılandırmadan yararlandığı görülmektedir.

- Toplam borç miktarı (2276,01–4206,92] aralığında borç türü açıklaması idari para cezası ve işsizlik borcu olan işyerlerinin % 46’sının yapılandırmadan yararlandığı görülmektedir.

- Toplam borç miktarı (4206,92–9282,81] aralığında olup, işyerinin türü “0; 4; 5; 10; 17; 22; 23; 26; 28” olarak kodlanan ve bölge kodu “2; 3; 4; 5; 6” olan işyerlerinin % 46’sının yapılandırmadan yararlandığı görülmektedir.

- Toplam borç miktarı 9282,81 liradan büyük ve işyerinin türü “1; 4; 10; 15; 17; 22” olarak kodlanan ve prim nispet oranı “1; 1,5; 2; 2,5; 3,5; 5,5; 6; 6,5” olan işyerlerinin % 41’inin yapılandırmadan yararlandığı görülmektedir.

Yapılandırmadan yararlanan işyerleri için genel olarak sonuçlar aşağıdaki gibidir:

- Yapılandırmadan yararlanma durumu “işyeri yaşı” bazında incelendiğinde; daha çok yeni açılan işyerlerinin yapılandırmadan yararlandıkları görülmekte olup, bu durum daha çok yeni açılan işyerlerinin ayakta kalma çabası olarak değerlendirilebilir.

- Yapılandırmadan yararlanma durumu “işyeri türü” bazında incelendiğinde; daha çok mevcut bulunan işyerleri arasında “diğer” sınıfında bulunan işyerlerinin (% 73.81) yapılandırmadan yararlandığı, bu işyerlerini gerçek kişi ve limited işyerlerinin takip ettiği görülmektedir.

- Yapılandırmadan yararlanma durumu “bölge” bazında incelendiğinde; yapılandırmadan en fazla % 43.59’luk yüzdelik dilim ile Marmara Bölgesinde faaliyet gösteren işyerlerinin yararlandığı ve bunu sırasıyla Ege ve İç Anadolu Bölgelerinde faaliyet gösteren işyerlerinin izlediği görülmektedir.

- Yapılandırmadan yararlanma durumu işyerlerinin “borç türü” bazında incelendiğinde; işyerlerinin % 49.98’inin prim borcunu yapılandırdığı, % 43.92’sinin işsizlik borcunu yapılandırdığı ve % 6.10’unun ise idari para cezası borcunu yapılandırdığı görülmektedir.

- Yapılandırmadan yararlanma durumu “il” bazında incelendiğinde; % 29’luk dilim ile İstanbul ilinin ilk sırada yer aldığı, % 9.65 ile İzmir ilinin ikinci sırada yer aldığı, % 8.33 ile Ankara ilinin üçüncü sırada yer aldığı, % 4.72 ve % 3.51 yüzdelik dilimler ile de Bursa ve Antalya illerinin bu illeri takip ettiği görülmektedir.

- Yapılandırmadan yararlanma durumu "prim nispet oranı" bazında incelendiğinde; % 48.28'lik yüzdelik dilim ile prim nispet oranı 1 olan işyerlerinin ilk sırada, % 28.11 yüzdelik dilim ile de prim nispet oranı 3 olan işyerlerinin yapılandırmaya daha çok başvurdukları görülmektedir.

- Yapılandırmadan yararlanma durumu "sigortalı sayısı" bazında incelendiğinde; % 58.78 yüzdelik dilim ile bir işçi çalıştıran işyerlerinin yapılandırmadan daha fazla yararlandıkları görülmektedir.

- Yapılandırmadan yararlanma durumu "sektör" bazında incelendiğinde, "İMALAT" sektörünün yapılandırmadan yararlanan işyerleri arasında % 24.29 yüzdelik dilim ile ilk sırada yer aldığı, "İNŞAAT" sektörünün % 20.87 ile ikinci sırada yer aldığı ve "TOPTAN VE PERAKENDE TİCARET, MOTORLU KARA TAŞITLARININ VE MOTOSİKLETLERİN ONARIMI" sektörünün de % 19.12 ile üçüncü sırada yer aldığı gözlenmektedir. İnşaat sektöründe faaliyet gösteren işyerlerinin, yeni ihalelere girerek iş yapabilmeleri için kurumdan "Borcu yoktur" yazısı almaları gerektiğinden yapılandırmaya daha çok başvurmaları dikkat çekici bir durumdur.

Sonuç olarak, SGK veri tabanında tutulan bu bilgiler ışığında gerçek veriler ile yapılan bu çalışmada birçok karar ağacı algoritması mevcut veriler üzerinde uygulanmış, kurulan modellerin başarı yüzdeleri birbirine çok yakın sonuçlar elde edilmiştir. Modelin başarısının düşük olması elde edilen karar kurallarının anlamsız olması manasına gelmemekle birlikte, kurulan modelin başarısını arttırmaya yönelik olarak da Türkiye genelinde veya pilot bölgeler bazında bir anket çalışması yapılarak, işyerlerine ilişkin daha fazla bilgi toplanarak elde edilen yeni değişkenlere ilişkin veriler ile karar ağacı algoritmaların tekrar bir araya getirilen veriler üzerinde uygulanması, model başarısını arttırmaya yönelik bir çalışma alanı olarak değerlendirilebilir.

Bu çalışma daha sonra çıkarılması tasarlanan yapılandırma kanunlarına bir fikir vermesi açısından değerlendirildiğinde ise; 5458 sayılı yapılandırma kanunundan yararlanan işyerleri içerisinde en fazla küçük işletmelerin yer alması bu işletmelerin ve küçük esnafın desteklenmesi gerekliliğini ortaya koymakta olup, bu çarpıcı durum göz ardı edilmemelidir.

KAYNAKLAR

1. İnternet: Hacettepe Üniversitesi,"Veri Madenciliğine Giriş", http://yunus.hacettepe.edu.tr/~hcingi/ist376a/6Bolum.doc

2. İnternet: Dumlupınar Üniversitesi, "Kümeleme Analizi Teknikleri İle İllerin Kültürel Yapılarına Göre Sınıflandırılması ve Değişimlerinin İncelenmesi", http://sbe.dpu.edu.tr/12/15-36.pdf

3. İnternet: Fırat Üniversitesi, "Apriori Algoritması İle Öğrenci Başarısı Analizi", http://www.emo.org.tr/ekler/24f4c5eef7ec01c_ek.pdf

4. İnternet: İstanbul Teknik Üniversitesi, "Veri Madenciliği, Genel Bilgiler", htpp://www.cs.itu.edu.tr/~gunduz/courses/verımaden/

5. İnternet: Gebze Yüksek Teknoloji Enstitüsü, "Veri Madenciliği Teknik Ve Uygulamaları", http://www.bilmuh.gyte.edu.tr/BIL454/birinci_ders.pdf, (2008).

6. İnternet: Department of Computer Science, "Data Mining", http://datamining.anu.edu.au/talks/2005/datamining-comp2340-2005.pdf, (2005).

7. İnternet: SPSS, "SPSS Türkiye", http://www.spss_brosur_mart_2007.pdf/, (2007).

8. İnternet: Karabük Üniversitesi, "Öğrenci Seçme Sınavında (ÖSS) Öğrenci Başarımını Etkileyen Faktörlerin Veri Madenciliği Yöntemleriyle Tespiti", http://iats09.karabuk.edu.tr/press/bildiriler_pdf/IATS09_01–01_603.pdf (2009).

9. İnternet: Erciyes Üniversitesi, "Mekânsal Uygulamalar İçin Veri Madenciliği Yaklaşımı", http://uzalcbs2008.erciyes.edu.tr/pdf/77.pdf (2008).

10. İnternet : "Data Mining", http://datamining.anu.edu.au/talks/2005/datamining-comp2340–2005.pdf (2005).

11. İnternet: Current data mining applications / percentage in different industries, http://www.kdnuggets.com/polls/2005/successful_data_mining_applications.htm

12. Alpaydın, E., "Zeki Veri Madenciliği", ***Bilişim 2000 Eğitim Semineri***, İstanbul, 1-10 (2000).

13. Akpınar, H., "Veritabanlarında Bilgi Keşfi ve Veri Madenciliği", ***İ.Ü. İşletme Fakültesi Dergisi***, 29: 1-22 (2000).

14. Akbulut, S., "Veri Madenciliği Teknikleri İle Bir Kozmetik Markanın Ayrılan Müşteri Analizi Ve Müşteri Segmentasyonu", Yüksek Lisans Tezi, ***Gazi Üniversitesi Fen Bilimleri Enstitüsü***, Ankara, 3-62 (2006).

15. Aköz, E., "Otomotiv Sektöründe Veri Ambarı Ve Bir Uygulama", Yüksek Lisans Tezi, ***Beykent Üniversitesi Fen Bilimleri Enstitüsü*****,** İstanbul, 2-16 (2007).

16. Argüden, Y., Erşahin, B., "Veri Madenciliği", ***Ar-Ge Danışmanlık / 10, Alkim Yayınevi***, İstanbul, 15-67 (2008).

17. Arslan, H., "Sakarya Üniversitesi Web Sitesi Erişim Kayıtlarının Web Madenciliği İle Analizi", Yüksek Lisans Tezi, ***Sakarya Üniversitesi Fen Bilimleri Enstitüsü***, Sakarya, 3-34 (2008).

18. Aydoğan, F., "E-Ticarette Veri Madenciliği Yaklaşımlarıyla Müşteriye Hizmet Sunan Akıllı Modüllerin Tasarımı ve Gerçekleştirmi", Yüksek Lisans Tezi, ***Hacettepe Üniversitesi Fen Bilimleri Enstitüsü*****,** 7-16, 34-37, 66-86 (2003).

19. Bardakcı, G., "Çevrimiçi Analitik Veri İşleme (OLAP)' nin İstatistikteki Yeri ve Bir Uygulama", Yüksek Lisans Tezi, ***Hacettepe Üniversitesi Fen Bilimleri Enstitüsü***, 2-14 (2008).

20. Boysan, M., Kayri, M., "Using Chaid Analysis in Researches and an Application Pertaining to Coping Strategies", ***Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi***, 40 (2): 133-149 (2007).

21. Bayam, E., Liebowitz, J., William, A., "Older drivers and accidents: A meta analysis and data mining application on traffıc accident data", ***Expert Systems with Applications***, 29: 598-629 (2005).

22. Bijmolt, T., H., A., Claassen, W., Brus, B., "Childrens Understanding of TVAdvertising: Effects of Age, Gender and Parental Influence", ***Journal of Consumer Policy***, 21: 171-194 (1998).

23. "Clementine İle İleri Modelleme", ***SPSS Türkiye***, Ankara, 1-76 (2009).

24. "Clementine Temel Eğitimi", ***SPSS Türkiye***, Ankara, 1-112 (2009).

25. "Clementine Veri Manipulasyon Yöntemleri", ***SPSS Türkiye***, Ankara, 1-67 (2009).

26. Chan, F., Cheing, G., Chan, J., Y., C., Rosenthal, D., A., Chronister, J., "Predicting employment outcomes of rehabilitation clients with orthopedic disabilities: A CHAID analysis", ***Disability and Rehabilitation,*** 28 (5): 257-270 (2006).

27. Çınar, M., S., "E-Ticarette Veri Madenciliği Teknikleri İçeren Sistemin Tasarımı ve Gerçekleştirimi", Bitirme Projesi, ***Selçuk Üniversitesi Mühendislik Mimarlık Fakültesi***, Konya, 1-71 (2007).

28. Diepen, M., V., Franses, P., H., "Evaluating chi-squared automatic interaction detection", ***Information Systems,*** 31: 814-831 (2006).

29. Dolgun, Ö.M., "Büyük Alışveriş Merkezleri İçin Veri Madenciliği Uygulamaları", Yüksek Lisans Tezi, ***Hacettepe Üniversitesi Fen Bilimleri Enstitüsü***, Ankara, 1-73 (2006).

30. Döşlü A., "Veri Madenciliğinde Market Sepet Analizi ve Birliktelik Kurallarının Belirlenmesi", Yüksek Lisans Tezi, ***Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü***, İstanbul, 6-24 (2008).

31. Doğan, N., Özdamar, K., "Chaid Analizi ve Aile Planlaması İle İlgili Bir Uygulama", ***Türkiye Klinikleri Tıp Bilimleri Dergisi***, 23 (5): 392-398 (2003).

32. Doğan, İ., "Holştayn ırkı ineklerde süt verimine etki eden faktörlerin CHAID analizi ile incelenmesi", ***Ankara Üniversitesi Veteriner Fakültesi Dergisi,*** 50: 65-70 (2003).

33. Erdoğan, Ş.Z., "Veri Madenciliği ve Veri Madenciliğinde Kullanılan K-Means Algoritmasının Öğrenci Veritabanında Uygulanması", Yüksek Lisans Tezi, ***İstanbul Üniversitesi Sosyal Bilimler Enstitüsü***, İstanbul, 3-25 (2004).

34. Gebhardt, M. , Jarke, M. , Jeusfeld, M.A. , Quix, C. , Sklorz, S., "Tools for Data Warehouse Quality", ***10.Bilimsel ve İstatistiksel Veritabanı Yönetimi Uluslararası Konferansı IEEE***, Almanya, 1-4 (1998).

35. Gürüler, H., Karahasan, M., İstanbullu, A., "Üniversite Öğrencilerinin Profilini Belirleme: Muğla Üniversitesi Veri Tabanları Üzerinde Bir Durum Çalışması", ***Muğla Üniversitesi Sosyal Bilimler Enstitüsü Dergisi***, 18: 27-37 (2007).

36. Gencer, C., Kızılkaya Aydoğan, E., Akbulut, S., "Veri Madenciliği Teknikleri İle Bir Kozmetik Markanın Ayrılan Müşteri Analizi ve Müşteri Bölümlenmesi", ***Sigma***, 26 (1): 43-57 (2008).

37. Ho, S., H., Jee, S., H., Lee, J., E., Park, J., S., "Analysis on risk factors for cervical cancer using induction technique", ***Expert Systems with Applications***, 27: 97-105 (2004).

38. Horner, S., B., Fireman, G., D., Wang, E., W., "The relation of student behavior, peer status, race and gender to decisions about school discipline using CHAID decision trees and regression modeling", ***Journal of School Psychology.***, 48: 135-161 (2010).

39. Koyuncugil, A.S., "Borsa Şirketlerinin Sektörel Risk Profillerinin Veri Madenciliğiyle Belirlenmesi", ***Sermaye Piyasası Kurulu Araştırma Raporu***, ***Ankara***, 1-17 (2007).

40. Köktürk, F., Ankaralı, H., Sümbüloğlu, V., "Veri Madenciliği Yöntemlerine Genel Bakış", ***Biyoistatistik***, 1: 20-25 (2009).

41. Kılıç, N. , Gümüş, E., Kusayeva, N., "Veri Madenciliği", Bitirme Proje Tezi, ***Selçuk Üniversitesi Bilgisayar Mühendisliği Bölümü***, Konya, 3-43 (2007).

42. Kaya, H. , Köymen, K., "Veri Madenciliği Kavramı ve Uygulama Alanları", ***Doğu Anadolu Bölgesi Araştırmaları***, 6 (2): 159-164 (2008).

43. Kayaalp, K., "Asenkron Motorlarda Veri Madenciliği İle Hata Tespiti", Yüksek Lisans Tezi, ***Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü***, Isparta, 12-30 (2007).

44. Koyuncugil, A., S., Ozgulbas, N., "Detecting Road Maps for Capacity Utilization Decisions by Clustering Analysis and CHAID Decision Tress, ***Journal of Medical Systems***, DOI 10.1007/s10916-009-9258-9 (2009).

45. Lin, Q., "Mobile Customer Clustering Analysis Based on Call Detail Records", ***Communications of the IIMA***, 7 (4): 95-100 (2007).

46. Lopez, H., Zitto, T., Bare, P., Vidal, G., Vukasovic, J., Gomez, R., "Prevalence of Anti-Hepatitis A Antibodies in an Urban Middle Class Area of Argentina: Some Associated Factors", ***International Journal of Infectious Diseases***, 4: 34-37 (1999).

47. McCarty, J., Hastak, M., "Segmentation approaches in data-mining: A comparison of RFM, CHAID and logistic regression", ***Journal of Business Research.***, 60: 656-662 (2007).

48. McMahon, B., T., Hurley, J., E., Chan, F., Rumrill Jr., P., D., Roessler, R., "Drivers of Hiring Discrimination for Individuals with Disabilities", ***Journal of Occupational Rehabilitation***, 18: 133-139 (2008).

49. Oğuzlar, A., "Veri Önişleme", ***Erciyes Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi***, 21: 67-76 (2003).

50. Özdamar, E.Ö., "Veri Madenciliğinde Kullanılan Teknikler ve Bir Uygulama", Yüksek Lisans Tezi, ***Mimar Sinan Üniversitesi Fen Bilimleri Enstitüsü*****,** İstanbul, 1-8 (2002).

51. Özkan, Y. "Veri Madenciliği Yöntemleri", Dr.Rifat Çölkesen, Dr.Cengiz Uğurkaya, ***Papatya Yayıncılık***, İstanbul, 1-115 (2008).

52. Özkan, Y., "Veri Tabanı Sistemleri, 3^{th}", Ayşe D. Tüzel, ***Alfa Yayınevi***, İstanbul, 2-37 (2003).

53. Özekes, S., "Veri Madenciliği Modelleri Ve Uygulama Alanları", ***İstanbul Ticaret Üniversitesi Dergisi***, 3: 65-82 (2003).

54. Özçakır, F., C., Çamurcu, A.Y., "Birliktelik Kuralı Yöntemi İçin Bir Veri Madenciliği Yazılımı Tasarımı Ve Uygulaması", ***İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi***, 6 (12): 21-37 (2007).

55. Pehlivan, G., "Chaid Analizi Ve Bir Uygulama", Yüksek Lisans Tezi, ***Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü***, İstanbul, 13-45 (2006).

56. Rygielski, C., Wang, J., C., Yen, D., C., "Data mining techniques for customer relationship management", ***Technology in Society***, 24: 483-502 (2002).

57. Rakowski, W., Clark, M., A., "Do Groups of Women Aged 50 to 75 Match the National Average Mammography Rate?", ***American Journal of Preventive Medicine***, 15: 187-197 (1998).

58. Sarıkan B. , "Veri Madenciliği ve İstatistik", Yüksek Lisans Tezi, ***Hacettepe Üniversitesi Fen Bilimleri Enstitüsü***, Ankara, 64-127 (2003).

59. Silahtaroğlu, G., "Kavram ve Algoritmalarıyla Veri Madenciliği", Dr.Rifat Çölkesen, Dr.Cengiz Uğurkaya, ***Papatya Yayıncılık***, İstanbul, 9-97 (2008).

60. Sullivan, D., J., Zyl, M., A., V., "The well-being of children in foster care: Exploring physical and mental health needs", ***Children and Youth Services Review***, 30: 774-786 (2008).

61. "SPSS Modeler Temel Eğitimi", ***SPSS Türkiye***, Ankara, 1-134 (2010).

62. Seibt, R., Spitzer, S., Blank, M., Scheuch, K., "Predictors of work abilitiy in occupations with psychological stress", ***Journal of Public Health***, 17: 9-18 (2009).

63. Soman, K., P., Diwakar, S., Ajay, V., "Insight into Data Mining Theory and Practice", ***PHI Learning Pvt. Ltd.***, New Delhi, 71-76 (2006).

64. Şen, F. , "Veri Madenciliği ile Birliktelik Kurallarının Bulunması", Yüksek Lisans Tezi, ***Sakarya Üniversitesi Fen Bilimleri Enstitüsü*****,** Sakarya, 3-22 (2008).

65. Şişaneci, İ., Karabağ, R., Kavza, U., "Veri Ön İşleme ve Tekil Değer Ayrışımı", ***Gebze İleri Teknoloji Enstitüsü Bilgisayar Mühendisliği Bölümü, Gebze/Kocaeli***, 3-9 (2007).

66. Şentürk, A., "Veri Madenciliği Kavram ve Teknikler", ***Ekin Yayınevi***, Bursa, 1-114 (2006).

67. Şen, O., N., "Oracle (9i) 4^{th} ", ***Beta Yayınevi***, İstanbul, 1-480 (2004).

68. Şimşek, U.T., Timor, M., "Veri Madenciliğinde Sepet Analizi İle Tüketici Davranışı Modellemesi", ***İstanbul Üniversitesi İşletme Fakültesi İşletme İktisadi Enstitüsü Yönetim Dergisi***, 19 (59): 3-10 (2008).

69. Thomas, E., H., Galambos, N., "WHAT SATISFIES STUDENTS? Mining Student – Opinion Data With Regression and Decision Tree Analysis", ***Research in Higher Education***, 45 (3): 251-269 (2004).

70. Ture, M., Tokatli, F., Kurt, I., "Using Kaplan-Meier analysis together with decision tree methods (C&RT, CHAID, QUEST, C4.5 and ID3) in determining recurrence-free survival of breast cancer patients", ***Expert Systems with Applications***., 36: 2017-2026 (2009).

71. Türe, M., Aktürk, Z., Kurt, İ., Dağdeviren, N., "The Effect of Health Status, Nutrition and Some Other Factors on Low School Performance Using Induction Technique", ***Trakya Üniversitesi Tıp Fakültesi Dergisi***, 23 (1) : 28-38 (2006).

72. Türe, M., Kurt, İ., Kürüm, T., "Analysis of intervariable relationships between major risk factors in the development of coronary artery disease: a classification tree approach", ***Anadolu Kardiyoloji Dergisi,*** 7: 140-145 (2007).

73. Ulaş, M.A., "Market Basket Analysis for Data Mining", Yüksek Lisans Tezi, ***Boğaziçi Üniversitesi Bilgisayar Mühendisliği***, İstanbul, 1-61 (2001).

74. Yalçıntaş, G., "Veri Madenciliği", Yüksek Lisans Tezi, ***Gazi Üniversitesi Fen Bilimleri Enstitüsü***, Ankara, 1-87 (2003).

75. Yağız, Z., " Chaid Analizi", Yüksek Lisans Tezi, ***Gazi Üniversitesi Fen Bilimleri Enstitüsü,*** Ankara, 1-72 (2003).

76. Yılmaz, Koltan, Ş., Albayrak, S.A., "Veri Madenciliği: Karar Ağacı Algoritmaları Ve İMKB Verileri Üzerine Bir Uygulama", ***Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*****,** 14 (1): 31-52 (2009).

77. Yavuz, U., Özdemir, A., Ayık, Z., Y., "Lise Türü Ve Lise Mezuniyet Başarısının, Kazanılan Fakülte İle İlişkisinin Veri Madenciliği Tekniği İle Analizi", ***Atatürk Üniversitesi Sosyal Bilimler Enstitüsü Dergisi***, 10 (2): 441-454 (2007).

ÖZGEÇMİŞ

Kişisel Bilgiler

Soyadı, adı:	DEMİREL, Betül
Uyruğu:	T.C.
Doğum tarihi ve yeri:	28.06.1973 Eskişehir
Medeni hali:	Evli
Telefon:	0 (312) 207 85 86
Faks:	0 (312) 207 87 86
E-mail:	bdemirel@sgk.gov.tr

Eğitim

Derece	Eğitim Birimi	Mezuniyet tarihi
Lisans	Hacettepe Üniversitesi / İstatistik Bölümü	1998
Lise	Deneme Lisesi	1991

İş Deneyimi

Yıl	Yer	Görev
2005–	Sosyal Güvenlik Kurumu Başkanlığı	İstatistikçi
2005–2000	Yargıtay Başkanlığı	Memur
2000–1997	Mamak Belediye Başkanlığı	Memur

Yabancı Dil

İngilizce