# VERİ MADENCİLİĞİ SÜRECİ VE GERÇEK BİR VERİ SETİ ÜZERİNDE UYGULANMASI

# DATA MINING PROCESS AND AN APPLICATION OF IT ON A SAMPLE DATA SET

**FATMA MELTEM KOCABAŞ**

Hacettepe Üniversitesi

Lisansüstü Eğitim – Öğretim ve Sınav Yönetmeliğinin

İSTATİSTİK Anabilim Dalı İçin Öngördüğü

YÜKSEK LİSANS TEZİ

olarak hazırlanmıştır.

2010

Bu çalışma jürimiz tarafından **İSTATİSTİK ANABİLİM DALI 'nda YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Başkan :….....

Doç. Dr. Turhan MENTEŞ

Üye (Danışman) :….....

Yrd. Doç. Dr. Canan HAMURKAROĞLU

Üye :….....

Doç. Dr. Durdu KARASOY

Üye :….....

Yrd. Doç. Dr. Serpil CULA

Üye :….....

Yrd. Doç. Dr. İbrahim ZOR

ONAY

Bu tez ...../...../..... tarihinde Enstitü Yönetim Kurulunca kabul edilmiştir.

Prof.Dr. Adil DENİZLİ

Fen Bilimleri Enstitüsü Müdürü

**VERİ MADENCİLİĞİ SÜRECİ VE GERÇEK BİR VERİ SETİ ÜZERİNDE UYGULANMASI**

**Fatma Meltem Kocabaş**

## ÖZ

Veri tabanı büyüklüklerinin terabaytlarla ifade edildiği günümüzde verinin anlamlı bir şekilde ilişkilendirilebilmesi ve "değerli veri"nin bu veritabanından çıkarılabilmesi için yeni yöntemlere ihtiyaç duyulmuş ve veri madenciliği kavramı ve süreci bu ihtiyaç sonunda ortaya çıkmıştır.

Bu tez çalışmasında veri madenciliği sürecinin tarihçesi, OLAP ve istatistik gibi disiplinlerle ilişkisi açıklanmış olup süreç örneklerle detaylı olarak ele alınmış ve veri madenciliği yöntemleri ayrıntılı olarak incelenmiştir.

Veri madenciliği yöntemleri sınıflandırma, kümeleme ve birliktelik kuralları ana başlıkları altında ele alınmış, karar ağacına ait ID3, C4.5, CART, k-en yakın komşuluk algoritmaları ve market sepet analizi örneklerle incelenmiştir.

Uygulama bölümünde gerçek bir veri seti üzerinde Clementine programı kullanılarak veri madenciliği uygulaması yapılmış ve sonuçları yorumlanmıştır.

**Anahtar Kelimeler:** Veri madenciliği, veri madenciliği süreci, veri madenciliği yöntemleri, sınıflandırma, karar ağacı, ID3, C4.5, CART, kümeleme, k-NN, birliktelik kuralları.

Danışman: Yrd. Doç. Dr. Canan Hamurkaroğlu Hacettepe Üniversitesi, Fen Fakültesi, İstatistik Bölümü

**DATA MINING PROCESS AND AN APPLICATION OF IT ON A SAMPLE DATA SET**

**Fatma Meltem Kocabaş**

## ABSTRACT

Database size are expressed with terabytes today. In order to draw meaningful conclusions about the data which is “valuable”, new techniques and methods are required and data mining process appeared for this necessity.

In this thesis, history of data mining process, its relations with the disciplines such as OLAP and statistics are dealed with and data mining techniques are detailed with samples.

Classification, desicion trees, clustering, association rules, market basket analyses are studied in detail with the algorithms ID3, C4.5, CART, k-NN (k nearest neighbour).

In the last section, data mining application has been made on a real sample data set by using Clementine software.

**Keywords**: Data mining, data mining process, data mining techniques, classification, decision trees, ID3, C4.5, CART, k-NN (k-nearest neighbour), association rules.

Advisor: Assist. Prof. Dr. Canan Hamurkaroğlu, Hacettepe University, Faculty of Science, Department of Statistics

**TEŞEKKÜR**

Bu tezi hazırlamamda görüş ve katkılarıyla bana yol gösteren ve sürekli cesaretlendiren danışmanım Yrd. Doç. Dr. Canan Hamurkaroğlu'na, tez konusunda ve uygulamasında benden yardımını esirgemeyen Kadir Korhan Babadağ'a ve çalışmam boyunca manevi desteklerini esirgemeyen aileme içtenlikle teşekkür ederim.

## İÇİNDEKİLER DİZİNİ

Sayfa

## SİMGELER VE KISALTMALAR DİZİNİ

OLAP : Çevrimiçi Analitik İşleme

SQL : İlişkisel Veritabanlarını Sorgulama Dili

## ŞEKİLLER DİZİNİ

Sayfa

## ÇİZELGELER DİZİNİ

Sayfa

# 1 GİRİŞ

Günümüzde bilgisayarların çok yaygın kullanılmasıyla birlikte her türlü veri sayısal ortamda kayıt altına alınmaya başlanmıştır. Süpermarkette yapılan alışverişten, bankacılık işlemlerine, telekomünikasyon hizmetlerinden sağlık sektörüne kadar her bir işlem veritabanlarında birer kayıt olarak karşılık bulmaktadır. Sonuç olarak terabaytlar düzeyinde veritabanları oluşmakta ve bu verinin miktarı günden güne artmaktadır. Dolayısıyla büyük hacimli veriler arasında stratejik öneme sahip bilgi nasıl elde edilebilir. Bu sorunun yanıtı Veri Madenciliğidir [11;37;45].

Veri madenciliği bilgi teknolojilerinin doğal evriminin sonuçlarından biri olarak değerlendirilebilir. Başlangıçta basit dosya yapıları üzerine kurulu veri toplama işlemi, veritabanı sistemlerinin ortaya çıkışıyla bir aşama kaydetmiş ve ileri veritabanı sistemleriyle birlikte veri ambarının oluşumuna neden olmuştur. Sonuç olarak veri madenciliği çok miktardaki veriden anlamlı bilgi çıkarma işlemi olarak bu evrimin son aşamasıdır [13].

Bu çalışmada amaç; son yıllarda hızla önem kazanan veri madenciliğine ilişkin süreçleri ve süreçlerin alt bileşenlerini dikkate alarak veri madenciliği tekniklerini gerçek bir veri seti üzerinde uygulamaktır.

Çalışmanın ikinci bölümünde veri madenciliği kavramı ve süreci, bu kavrama temel oluşturan veri tabanı ve veri ambarı kavramlarının üzerinde durulmuştur.

Üçüncü bölümde veri madenciliği teknikleri ve kullanılan algoritmalar incelenmiştir.

Son bölümde de bir Internet Servis Sağlayıcısı firmasından alınan gerçek bir veri seti üzerinde özellikle tahmin edici modelleme yapısı ele alınarak uygun modeller kullanılıp veri karakteristiği incelenmiştir.

# 2 VERİ MADENCİLİĞİ

Veri madenciliği, bilgisayarlarda depolanan yüksek miktardaki verinin analizi ile ilgilidir. Örneğin, yapılan alışverişler sonucu süpermarketlerde çok miktarda veri oluşur. Müşteriler barkod teknolojisinin de katkısıyla perakende alışverişlerini ve ödemelerini kolayca yaparlar. Fiyatlara bilgisayarlar aracılığı ile hızlı bir şekilde ulaşılır ve alışveriş işlemleri pratik bir şekilde sona erdirilir. Her bir barkod okuma işlemi veritabanına bir kayıt olarak girmekte ve müşterinin alışveriş alışkanlıklarına ait bir gözlem olarak yerini almaktadır. Aynı bilgisayarlar ile marketler, stoklarda satılan maldan ne kadar kaldığı ile bilgiye yani envanter bilgisine ulaşabilir. Bunun yanında tedarikçi firma ile temasa geçip stok durumuna göre yeni sipariş geçebilir. Buna ek olarak maliyetler ve elde edilen kar ile ilgili muhasebe sisteminden faydalanır. Tüm bu bilgilere erişmek temelde her bir ürünün sahip olduğu barkod sayesinde mümkün olabilmektedir. Veri madenciliği analizinde de diğer veri kaynaklarının yanında barkod ile elde edilen bu veriler kullanılabilir [22;28;30].

Veri madenciliğinin uygulanması yukarıda bahsedilen market alışverişi ve ticaret ile sınırlı değildir. Uygulama alanlarına örnek olarak pazarlama, bankacılık, sigortacılık, elektronik ticaret ve sağlık gibi temel sektörlerin yanında iklim değişikliği, basketbolda oyun stratejileri geliştirilmesi, TV izleyicileri profili çıkarımı gibi özel konular da verilebilir. İşletmelerde yoğun olarak kullanılan veri madenciliği uygulamalarının başlıcaları aşağıda özetlenmiştir [2;3;12;21;29;39]:

Pazarlama:

- Müşterilerin satın alma örüntülerinin belirlenmesi,
- Müşterilerin demografik özelliklerinin bulunması,
- Müşteri İlişkileri Yönetimi, müşterinin elde tutulması,
- Satış tahmini.

Bankacılık:

- Farklı finansal göstergeler arasındaki gizli ilişkinin bulunması,
- Kredi kartı dolandırıcılıklarının tespiti,
- Kredi kartı harcalamarına göre müşteri segmentlerinin belirlenmesi,

- Kredi taleplerinin değerlendirilmesi.

Sigortacılık:

- Yeni poliçe alacak müşterilerin tahmin edilmesi,
- Sigorta dolandırıcılıklarının tespiti,
- Riskli müşteri gruplarının belirlenmesi.

Gartner Grubun tanımına göre veri madenciliği, istatistiksel ve matematiksel yöntemlerin yanında örüntü tanıma teknolojilerininde kullanılarak depolanan yüksek miktardaki verilerin süzülerek aralarındaki anlamlı ilişkileri, örüntüleri ve eğilimleri ortaya çıkarma işlemidir[29;44].

Veri madenciliği için yaygın olarak kullanılan tanımlar;

- Veri madenciliği, veri için geçerli tahminlerde bulunmak amacıyla üzerinde çeşitli veri analiz araçlarını kullanarak örüntüleri ve ilişkileri ortaya çıkarma işlemidir [45].

- Veri madenciliği, veri içinde tahmin edilemeyen ilişkilerin bulunması amacıyla yüksek miktarda gözlemsel veri setlerinin analizi ve veri sahibine faydalı olabilecek şekilde yeni yöntemlerle özetlenmesidir [14].

- Veri madenciliği otomatik öğrenme, örüntü tanıma, istatistik, veritabanı ve görselleştirme tekniklerini bir araya getirerek büyük veritabanlarından bilgi çıkarmaya yarayan bir ara disiplin alanıdır [8]

biçimindedir.

Tüm bu tanımların ortak noktaları olarak; veri madenciliği uygulanacak veri setinin çok büyük olması, kullanılan analiz yöntemlerinin farklı uzmanlık alanlarından geliyor olması ve örüntü/ilişkilerin çıkarılmasının asıl amaç olması gösterilebilir. Ancak veri madenciliğini diğer analiz yöntemlerinden ayıran en önemli özellik elde edilecek bilginin önceden bilinmiyor, tahmin edilemiyor olmasıdır. Tahmin edilebilen, beklenen bilgi için veri madenciliğini kullanmak verimli olmayacaktır [25].

Daha önce tahmin edilemeyen, ya da tahmin edilemeyenle ilgili en ünlü örnek bira-çocuk bezi örneğidir: Amerika'da önemli market zincirlerinden birinde yapılan bir veri madenciliği sonuçlarına göre bira ile çocuk bezi satışları arasında güçlü bir ilişki vardır. Özellikle cuma günleri çocuk bezi satın alan kişilerin büyük bir çoğunluğu aynı zamanda bira da satın almaktadırlar [8;38].

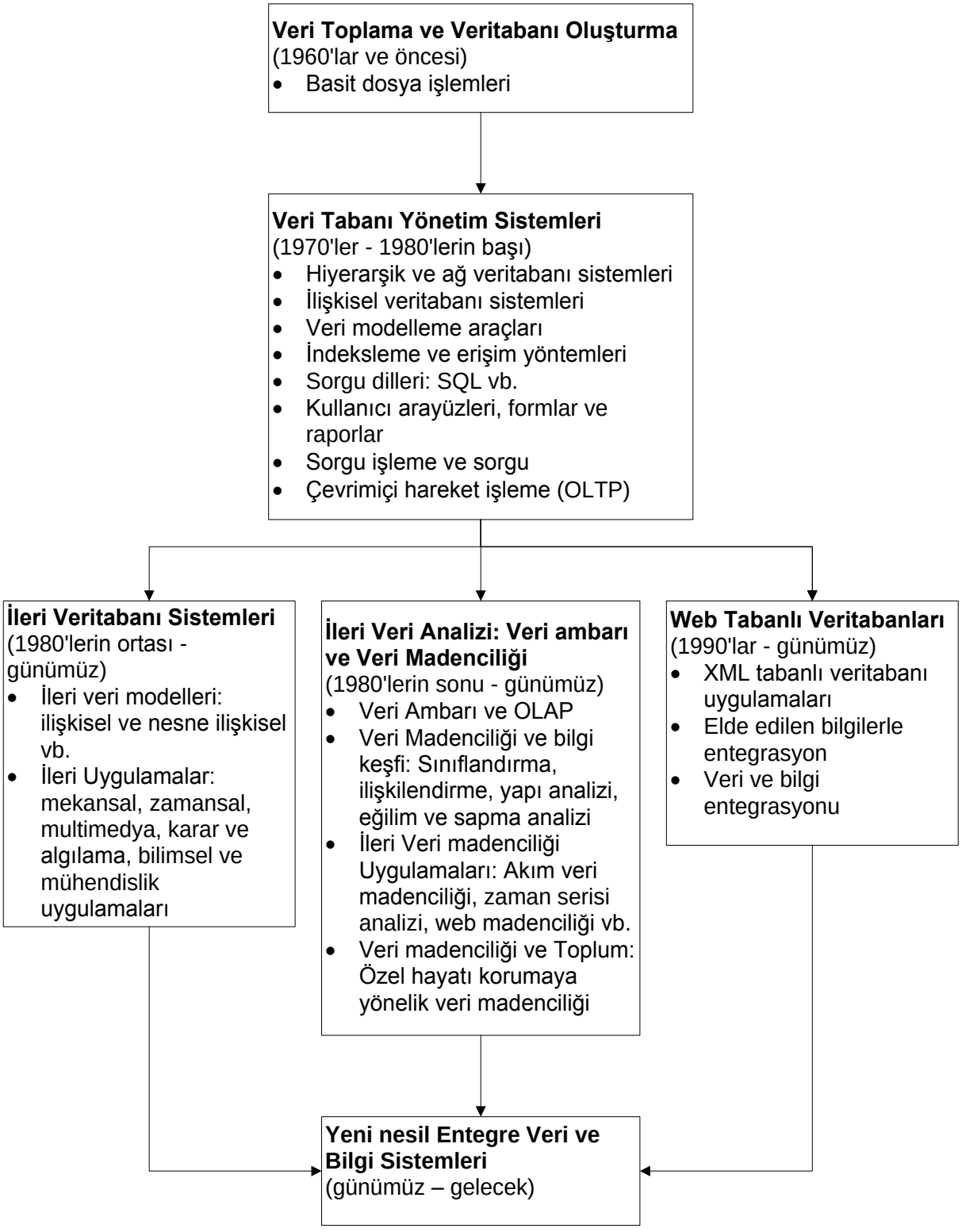
## 2.1 Veri Madenciliği Tarihi

Veri madenciliği bilgi teknolojilerinin doğal evriminin bir sonucu olarak da nitelendirilebilir. Veritabanı sistemleri Şekil 2.1'de görülen evrimsel yolu izleyerek veri toplama, veritabanı oluşturma, veri yönetimi (veri saklama ve geri erişim dahil) ve yüksek veri analizi (veri ambarı ve veri madenciliğini içeren) aşamalarından geçerek günümüze gelmiştir [13;23].

1960'lı yıllardan itibaren veritabanı ve bilgi teknolojileri basit dosya işlemlerinden gelişmiş ve güçlü veritabanı yapılarına doğru gelişim göstermiştir. 1970' li yıllarda başlayan veritabanı sistemlerindeki araştırma ve geliştirme çalışmaları hiyerarşik ve ağ veritabanı yapılarından ilişkisel veritabanı, veri modelleme araçlar ve indeksleme yapısına geçişi sağlamıştır. Bununla birlikte kullanıcılar sorgulama dilleri ve arayüzler sayesinde esnek veri erişim imkanına sahip olmuşlardır [13].

1980'li yılların ortasından itibaren ilişkisel teknolojilerle birlikte yeni ve güçlü veritabanı sistemleri üzerinde durulmuştur. İleri veri modellerindeki araştırmalarla nesneye yönelik, nesne-ilişkisel ve tümdengelim yöntemlerinde gelişmeler sağlanmıştır. Bununla beraber heterojen veritabanı sistemleri ve internet tabanlı sistemler bilgi teknolojileri endüstrisinde çok önemli bir rol oynamıştır. Son otuz yılda donanım teknolojisindeki şaşırtıcı gelişmeler güçlü bilgisayarların, veri toplama donanımlarının ve bellek medyalarının varlığına yol açmıştır [6].

Çizelge 2.1'de görüleceği üzere 1960'lı yıllardan günümüze veri madenciliği kavramının oluşmasına kadar gelişim adımları, cevap aranan sorular, kullanılan teknolojiler ve özellikler tablo halinde verilmiştir. Başlangıçta statik veriyi elde etmek için gösterilen çabalar günümüzde yerini ileriye dönük tahminlerde bulunmak amacıyla ileri düzey analize bırakmıştır [36].




Şekil 2.1. Veritabanı sistemleri teknolojisinin evrimi

Günümüzde veri çok çeşitli veritabanları üzerinde depolanabilmektedir. Depolama mimarisi veri ambarı kavramının, dolayısı ile veri tutarlılaştırma, veri entegrasyonu ve OLAP işlemlerinin de oluşmasını sağlamıştır. OLAP sayesinde veri analizi, özetleme, birleştirme ve entegrasyon bileşenleri ile çok boyutlu bir şekilde yapılabilir. Tüm bunlara rağmen sınıflandırma, gruplama, veri niteliğinin zamanla değişimini gözlemleme gibi ayrıntılı analiz yapmak için ek olarak veri analiz araçları gerekmektedir. Çünkü donanım ve depolama teknolojilerindeki göz kamaştırıcı gelişmeler veri zengini ancak bilgi yoksunu bir durumun oluşmasına sebep olmuştur. Çok miktarda verinin ve arşivin olduğu böyle bir ortamda önemli kararlar bu verilere göre değil karar vericilerin sezgilerine göre alınmaktadır. Zira karar alıcıların çok yüksek miktardaki bu verinin içinde gömülü değerli bilgiyi çıkarmak için araçlara ihtiyaçları vardır. İşte bu veri ve bilgi arasındaki açığı kapatacak olan yaklaşım veri madenciliği ve veri madenciliği araçlarıdır [13].

Çizelge 2.1. Veri madenciliği kavramının tarihsel gelişimi

| **Gelişim Adımları** | **Cevap Aranan Sorular** | **Kullanılan Teknolojiler** | **Özellikler** |
|---|---|---|---|
| **Veri Toplama (1960'lar)** | "Şirketin son 3 yıldaki satışları toplamı ne kadar?" | Bilgisayarlar, Teypler, Diskler | Geçmiş dönük, statik veri dağıtımı |
| **Veri Erişimi (1980'lar)** | "Ankara'da geçen Ocak ayında birim satışları ne kadardı?" | İlişkisel veritabanları, SQL, ODBC | Kayıt bazında geçmişe dönük, dinamik veri dağıtımı |
| **Veri Ambarları ve Karar Destek Sistemleri (1990'lar)** | "Ankara'da geçen Ocak ayında birim satışlarının bir önceki yıl Ocak ayındaki birim satış oranı kaç?" | OLAP, Çok boyutlu Veritabanı Sistemleri, Veri ambarları | Çoklu bazda, geçmişe dönük dinamik veri dağıtımı |
| **Veri Madenciliği (Bugün)** | "Gelecek ay Ankara'da birim satışlar ne kadar olabilir, neden?" | İleri düzey algoritmalar, çok işlemcili bilgisayarlar, gelişmiş veritabanları | Geleceğe dönük, öngörülü bilgi dağıtımı |

## 2.2 Veri Madenciliği Ne Değildir?

Yazılım firmaları analiz ürünlerini tak-ve-çalıştır şeklinde herhangi bir insan müdahelesi ve etkileşimi olmadan kullanılmak üzere pazarlamaktadır. Bu yaklaşım da kullanıcılara ürünlerin herhangi bir insan katkısı olmadan otomatik olarak çalışacağı önyargısına sebep olmaktadır. Ancak veri madenciliği bir üründen daha çok, yönetilmesi gereken bir disiplindir ve sihirli bir değnek değildir. Veritabanının içine kurgulanıp veri sahibine ilginç bir örüntü ile karşılaştığında haber verecek bir sistem değildir. Verinin iyi anlaşılması, analitik yöntemlerin anlaşılması ve işin bilinmesi gibi önemli ihtiyaçları ortadan kaldırmaz. İş analistine verideki örüntü ve ilişkileri bulmasında yardımcı olur ve analistin veri madenciliğinin tüm fazlarında aktif olarak yer almasını zorunlu kılar [22;45].

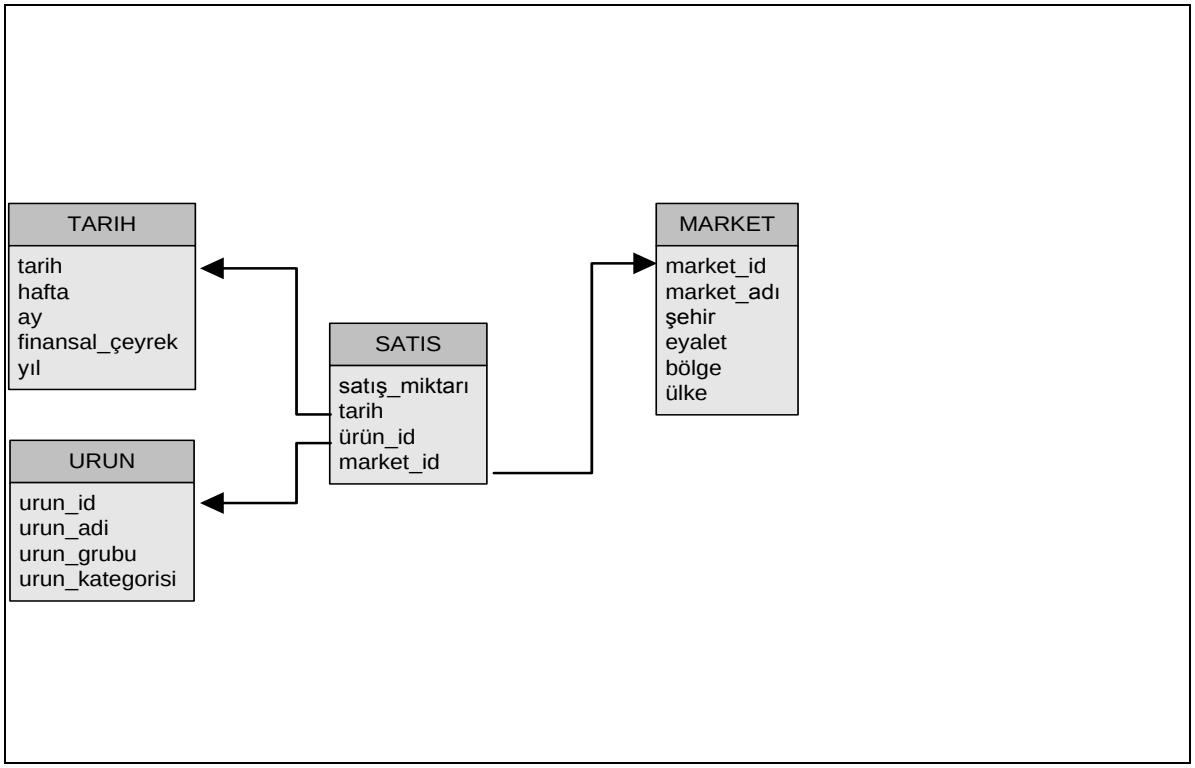
Veri madenciliği sonucu bulunan tahmini ilişki ve örüntüler bir davranışın ve aksiyonun sebebi olarak algılanmamalıdır. Örneğin veri madenciliği ile belirli gazetelere abone, gelir düzeyleri aylık 3.000 TL ile 5.000 TL arası erkeklerin belirli bir ürünü almak için potansiyel olduğu tahmin edilebilir. Ancak bu kriterler bahsi geçen özelliklere sahip kişilerin piyasadaki o belirli ürünü alması için belirleyici faktör olarak algılanmamalıdır.

## 2.3 Veri Madenciliği ve OLAP

Veri Madenciliği kavramı ile yeni tanışmış iş birimleri çok geniş veritabanları üzerinde yaptıkları karmaşık sorgu cümlecikleri ile veri madenciliği alanında çalışma yaptıklarını zannederler. Ancak sorgulama ve raporlama tümdengelim bir analizdir. Örneğin “Hangi marketler son çeyrekte taşınabilir DVD oynatıcısı sattı?” ve “Herbiri ne kadar sattı?” sorguları oldukça sıradan ve sıkça rastlanan türden sorgulardır. Ve bu soruları yanıtlamak için yapılan sorgulamalar ve raporlamalar Veri Madenciliği sayılmazlar. Bu spesifik soruların cevabına Şekil 2.2 veritabanı şemasına sahip bir yapıda aşağıda Çizelge 2.2’de görünen SQL sorgusu/cümleciği ile ulaşılabilir [20].

Çizelge 2.2. Sorgu cümlesi

```
SELECT market.market_adı, sum (satış_miktarı) AS toplam_satış
FROM market, satış, tarih, ürün
WHERE ürün.ürün_adı = 'taşınabilir DVD oynatıcısı'  AND
  ürün.ürün_id = satış.ürün_id   AND
  satış.market_id = market.market_id  AND
  satış.tarih = tarih.tarih  AND
  tarih.yıl = '2008'  AND
  tarih.çeyrek = 4
GROUP BY market.market_id
ORDER BY toplam_satış
```




Şekil 2.2. Satış, ürün, market ve tarih veritabanı şeması

Şekil 2.2'de ki veritabanı şemasında dört tablo yer almaktadır: Satış, Ürün, Market ve Tarih. Satış tablosu yapılan satış miktarını ve diğer üç tablo ile ilişkili referans alanlarını içerir: Satılan ürün, satıldığı market ve satıldığı tarih. Ürün tablosu ürünün adı ile birlikte gruplama ve sınıflandırma amacıyla ürünün kategori ve grup bilgisini de içerir. Örneğin 330 ml. lik bir kola, "içecek" grubunda ve "gazlı içecek" kategorisindedir [20].

Çizelge 2.3. Tablo içerik örnekleri

a. Ürün için,

| **ürün_id** | **ürün_adi** | **ürün_grubu** | **ürün_kategorisi** |
|---|---|---|---|
| 1 | 330 ml. kola | İçecek | Gazlı içecek |
| 2 | Taşınabilir DVD oynatıcısı | Elektronik | DVD oynatıcısı |
| 3 | Bayan Kot Ceket | Giyim | Bayan giyim |
| ... | ... | ... | ... |

b. Market için,

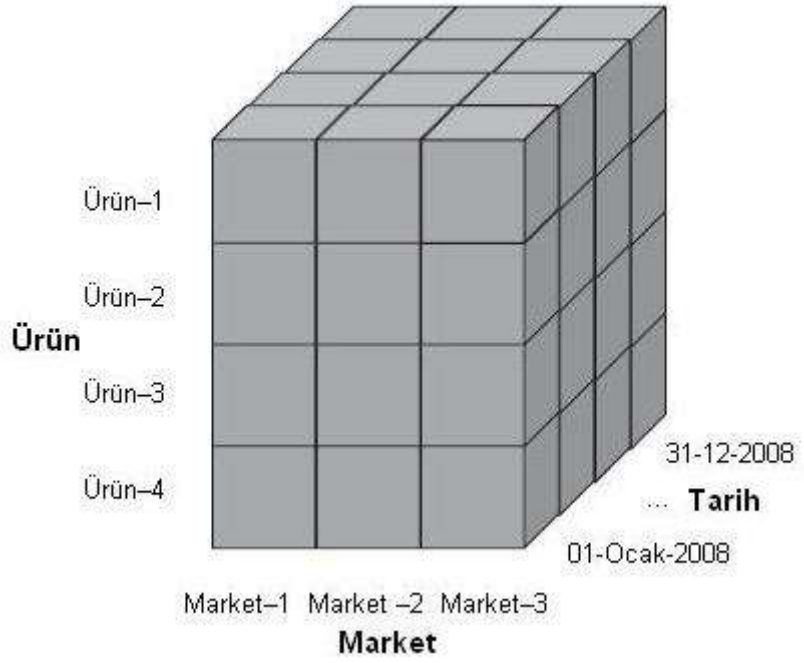
| **market_id** | **market_adı** | **şehir** | **eyalet** | **bölge** | **ülke** |
|---|---|---|---|---|---|
| 101 | Mall abc | Ohio | | North | USA |
| 102 | Mall klm | Missuri | | South | USA |
| 103 | Mall xyz | LA | | East | USA |
| ... | ... | ... | ... | ... | ... |

c. Satış için,

| **satış_miktarı** | **tarih** | **ürün_id** | **market_id** |
|---|---|---|---|
| 1309 | 20.08.2008 | 1 | 101 |
| 2303 | 20.08.2008 | 1 | 102 |
| 4 | 20.08.2008 | 2 | 101 |
| 14 | 21.08.2008 | 3 | 102 |
| 23 | 22.08.2008 | 3 | 103 |
| ... | ... | ... | ... |

d. Tarih için,

| tarih | hafta | ay | finansal_çeyrek | yıl |
|---|---|---|---|---|
| 20.08.2008 | 34 | 8 | 4 | 2008 |
| 21.08.2008 | 34 | 8 | 4 | 2008 |
| 22.08.2008 | 34 | 8 | 4 | 2008 |
| ... | ... | ... | ... | ... |




Şekil 2.3. OLAP küpü örneği

Çizelge 2.3'te verilen veri ve tablo yapıları üzerinde OLAP kullanılarak çok boyutlu veri analizi yapılabilir. Çok boyutlu veri analizinde veri farklı boyutlardan, ürün, satış ve tarih açısından incelenir ve Şekil 2.3'teki gibi bir küp oluşturulur. Küp oluşturulduktan sonra iş analisti dilimle/çevir (slice and dice) işlemi yaparak veriyi farklı açılardan inceleyebilir. Bir ürün kategorisine ait ürünün bölgesel bazda ve aylık dağılımlarla satış miktarını sorgulamak, A Kategorisindeki Ürün-1'in Market-2'de 2008 yılının son çeyreğindeki satışlarının miktarı bu işleme bir örnek olarak verilebilir.

Genel bir tanımlama ile OLAP özellikle ilişkisel veritabanlarındaki verilere çok hızlı bir şekilde erişebilme ve çok boyutlu analiz ihtiyaçlarının teminine yönelik geliştirilen bir teknolojidir. Dolayısı ile kavramsal olarak Veri Madenciliğine benzetilmekte, arasındaki farklar uzmanlar için zaman zaman soru işareti olmaktadır [43].

Veri madenciliği ise OLAP'ın aksine tahmin edilemeyen ve görülemeyen örüntü ve ilişki bilgisinin keşfine yöneliktir. Veri analizine tümevarımsal bir bakış açısı katar. Mevcut veriden daha çok örneğin gelecek yıl ilk finansal çeyrekte ne kadar taşınabilir DVD çalıcısı satılacak ve hangi müşteriler bu ürünlerden satın alacak, neden satın alacak sorularına yanıt bulmaya yöneliktir [20].

OLAP veritabanları üzerinde çeşitli stratejik kararlar almaya yardımcı olacak analiz ve sorgu işlemleridir. Geleneksel sorgu ve raporlama araçları, veritabanında "Ne?" sorusuna yanıt almaya çalışırken OLAP bir kademe daha ilerisine yönelir ve "Niçin?" sorusunu ispatlamak için kullanılır. Örneğin bir analist kredi borcu ödeme güçlüğüne sebep olan risk faktörlerini belirlemek istiyor olsun. Öncelikle düşük gelirli kişilerin kredi riskinin yüksek olacağı şeklinde bir hipotez ileri sürebilir ve veritabanını bunun doğruluğunu göstermek için analiz edebilir. Eğer doğruluğunu ispat edemezse hipotezini değiştirir. Yüksek borç sahibi olmanın risk faktörü olduğunu düşünerek bunu doğrulamaya çalışır. Eğer bunu da doğrulayamazsa her iki faktörün birlikte kredi riskinde etkili olduğu tezini araştırabilir. Yani analist örüntü ve ilişkilerle ilgili bir seri hipotez üretir ve bunların doğruluğunu veya yanlışlığını ispat etmeye çalışır. Bu yüzden OLAP tümdengelimsel bir işlemdir. Ancak incelenmesi gerekli değişken ve parametre sayısı düzinelerce yüzlerce olduğu zaman etkili hipotezler ileri sürmek ve bunları OLAP ile doğrulamak çok daha zorlaşır [45;37].

Veri madenciliği bu açıdan OLAP'dan farklıdır. Çünkü hipotez ileri sürerek bunu doğrulamaya çalışmak yerine doğrudan veriyi bu tip örüntüleri ve ilişkileri açığa çıkarmak için kullanır. Esas olarak Veri Madenciliği tümevarımsal bir yöntemdir. Örneğin, analistin kredi ödeme borcu ödeme güçlüğüne sebep olan risk faktörlerini belirlemek için Veri Madenciliği programı kullandığını varsayalım. Veri Madenciliği programı yüksek borçlu ve düşük gelirli insanların kredi riskinin yüksek olduğunu bulabilir. Ancak daha da fazlasını, analistin hiç hesaba katmadığı bir faktörü, örneğin yaş faktörünün belirleyici bir faktör olduğunu ortaya çıkarabilir. İşte bu noktada Veri

Madenciliği ve OLAP birbirlerini tamamlarlar. Ayrıca OLAP bilgi keşif sürecinin ilk safhalarında tamamlayıcı bir rol oynar. Çünkü verinin araştırılmasına, önemli değişkenlere odaklanarak keşfedilmesine, etkileşimleri bulmaya yardımcı olur [24;45].

Çizelge 2.4'de görüldüğü üzere OLAP geçmişe ait bilgilendirici veriler verirken Veri Madenciliği geleceğe dönük tahminler üzerinde yoğunlaşmaktadır [5;41].

Çizelge 2.4. OLAP ile veri madenciliği kavramlarının karşılaştırılması

| **OLAP** | **Veri Madenciliği** |
|---|---|
| Postalarımıza geri dönüş oranı nedir? | Gelecekteki postalarımıza yanıt verme potansiyeline sahip müşteri profili nedir? |
| Yeni ürünümüzden mevcut müşterilerimize ne kadar sattık? | Yeni ürünümüzü hangi müşterilerimiz alma eğilimine sahiptir? |
| Geçen ay hangi müşterilerimiz poliçelerini yenilemedi? | Önümüzdeki 6 ayda hangi müşterilerimiz rakip firmalara gidebilir? |
| Geçen yılki en iyi 10 müşterim kimlerdi? | Hangi 10 müşteri en büyük kar profili potansiyeline sahiptir? |
| Hangi müşteriler geçen yıl borçlarını ödemedi? | Bu müşteri ödeme riskine sahip bir müşteri midir? |
| Son çeyrekte bölgedeki satış cirosu ne kadardı? | Gelecek yıl bölgedeki satış cirosu tahmini nedir? |
| Dün üretilen parçaların yüzde kaçı hatalı idi? | Arızalı parçaları azaltmak ve iş çıkarma yeteneğini artırmak için ne yapabilirim? |

## 2.4 Veri Madenciliği ve İstatistik

İstatistik her zaman veriyi analiz etmek amacıyla yöntemler oluşturmak için kullanılmıştır. İstatistiksel yöntemlerin veri madenciliğinde geliştirilen yöntemlerden en önemli farkı veri setinin hacmidir. Klasik bir istatistikçi için "geniş" sayılabilecek veri seti yüzlerce veya birkaç bin adet veriden oluşmaktadır. Veri madenciliğinde ise bu sayı milyonları, milyarları ve gigabayt/terabaytlarlarca yer tutan veriye ulaşmaktadır. Diğer önemli bir fark ise istatistiksel yöntemlerin çoğunlukla analiz edilen veri üzerinde ve kavramsal referans örneklemler üzerinde oluşturulmasıdır. Bu özellik, istatistiksel yöntemlerin tutarlı ve kesin sonuçlar vermesinde etkili olsa da, bilgi

teknolojileri alanındaki hızlı gelişmelerle yeni metodolojilerin oluşturulma gereksinimlerine çabuk uyum sağlamasını engellemiştir [14;15;26].

Veri madenciliği ile istatistiksel analizi birbirinden ayıran özellikler:

a) İstatistiksel analiz belirli bir hipotezin doğruluğunu ispatlamak, önceden belirlenmiş soruları cevaplamak amacıyla temel veri üzerinde yoğunlaşırken, veri madenciliği başka amaçlar için toplanmış ikincil derecedeki veri ile de ilgilenir.

b) Veri madenciliği hacim olarak çok büyük verinin analizini içerir ve bu özellik istatistiksel analiz için yeni yaklaşımları zorunlu kılar. Bilgisayar dünyasındaki verimlilik ve yeterlilik kısıtlarından dolayı bir çok uygulamada tüm veritabanındaki veriyi analiz etmek mümkün değildir. Bu sebeple analiz, gerçek veri içinden seçilecek bir örnek veri üzerinde yapılmak zorundadır. Örneklem üzerinde yapılacak analiz ise geleneksel istatistiksel analiz yöntemleri ile değil veri madenciliği analiz yöntemleri ile yapılabilir.

c) Başta internet'in kaynak olarak kullanıldığı veriler olmak üzere bir çok veritabanı istatistiksel veri organizasyonu yapısına uygun değildir. Bu da istatistikte olmayan yeni yöntemlerin oluşturulmasını gerekli kılar.

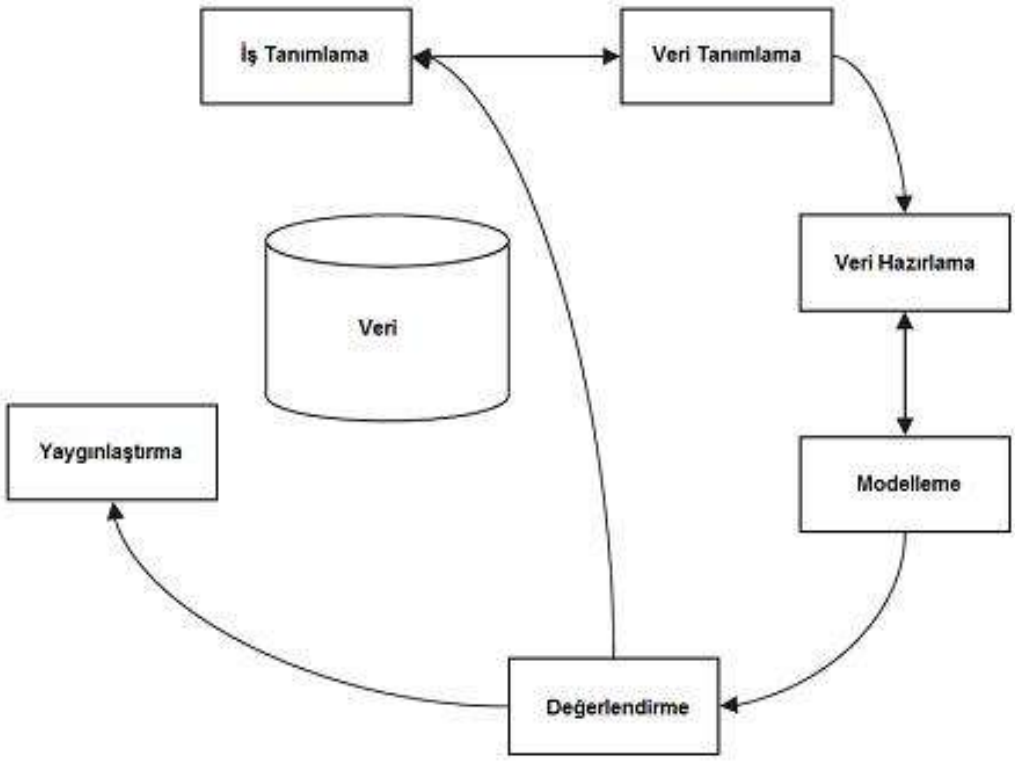
Tüm bu farklara rağmen istatistik veri madenciliğinde önemli bir rol oynar ve veri madenciliğinin ayrılmaz bir fonksiyonudur [14;16;26;46].

## 2.5 Veri Madenciliği Süreci

Veri madenciliği işlemi veri üzerinde anlık uygulanan ve doğrusal bir işlem değildir. Bu yüzden tekrarlıdır ve bir süreç olarak ele alınmalıdır. Başarılı bir veri madenciliği yapabilmek için sistematik bir yaklaşım sergilemek gerekmektedir. Bununla birlikte verinin kalitesi, uygulama ve bilgi teknoloji uzmanlarının ortak çalışması ve sabır sürecin başarılı olmasındaki faktörlerdir. Sistematik bir süreç geliştirmek amacıyla bazı danışmanlık şirketleri temelde benzer adımları içeren kendilerine özgü yöntemler geliştirmişlerdir. Bu yöntemlerden biri de 1996 yılında DaimlerChrysler, SPSS ve NCR firmalarının ortak çalışmasıyla oluşturulan endüstri, araç ve uygulama

bağımsız bir standart olan CRISP-DM Endüstri Bağımsız Veri Madenciliği Süreci Standartı (Cross-Industry Standard Process for Data Mining)'dır. "Endüstri Bağımsız" ifadesi ile iş dünyasından, ilaç sektörüne, yüksek eğitimden, sağlık dünyasına kadar her sektörde bir kullanım alanı bulabilmesidir [4;10;18;20;22;29].

Şekil 2.4'te görüldüğü üzere CRISP-DM Süreci 6 adımdan oluşmaktadır: İş Tanımlama, Veri Tanımlama, Veri Hazırlama, Modelleme, Değerlendirme, Yaygınlaştırma.




Şekil 2.4. CRISP-DM süreci

### 2.5.1 İş tanımlama fazı

CRISP-DM'deki en önemli faz verinin ve iş hedeflerininin tanımlama fazı olan iş tanımlama fazıdır. Bu fazda proje planlama ve işyükü tahmini gibi çalışmaları içeren çözülecek problemin tanımlanması da bulunur. CRISP-DM iş hedeflerini ve veri madenciliği hedeflerini birbirinden ayırır. İş hedefleri iş dünyası terminolojisi ile ifade edilir, örneğin "sigorta poliçelerindeki dolandırıcılık maliyetlerini düşürmek" gibi. Veri

madenciliği hedefleri ise teknik terimlerle ifade edilir; örneğin “Poliçedeki hangi değerler ve faktörler dolandırıcılık üzerinde etkilidir ve hangi poliçelerde bu tip durumların çıkması muhtemeldir?”. Dolayısı ile başlangıçtaki kısıt ve gereksinimlerin teknik bir dile ve tanıma dönüştürülmesi işi de bu fazda tamamlanmalıdır.

Birinci faz veri madenciliği sürecinin en önemli ve en zor fazıdır. Hangi problemin çözülmek istendiği anlaşılmaz ve sonuçların nasıl kullanılacağı bilinmez ise yürütülen tahminler gerçekçi olmaz. Bu yüzden iş birimleri ve teknik birimler bu tanımı yapmak için birlikte çalışırlar. Kampanya eğilim modeli, tanımlanması kolay bir örnektir. Ancak örneğin telekomünikasyon sektöründe “hizmet iptali” problemi çok ayrıntılı bir tanımlamayı ve çalışmayı gerektirir.

İş tanımlama/anlama fazının diğer önemli bir parçası insan, donanım, yazılım ve veri kaynaklarını tanımlamaktır. Hangi kaynakların bu problem için tahsis edileceğini bilmek ilk aşamadır. Diğer yandan bilişim kaynaklarına, donanım ve özellikle kullanılacak yazılıma karar vermek veri madenciliği projesinin başarısıyla doğrudan ilgili olduğu için ilk fazda tanımlanması ayrıca önemlidir [20;22;45].

### 2.5.2 Veri tanımlama fazı

Problemi tanımladıktan ve anladıktan sonraki adım, veriyi toplamak, veri kalitesinin yeterli olup olmadığını ve iş hedeflerine uygun olup olmadığını belirlemektir. Veri yapısı daha iyi anlaşıldıkça problem tanımı değiştirilebilir veya yeniden yapılabilir. Problem için incelenecek önemli bir kısım eksik veya bozuk olabilir, bu tip veri “kirli veri” olarak isimlendirilir. Bu ise verinin temizlenmesini veya yeni verinin elde edilmesini zorunlu kılar.

Çoğu kez farklı kaynaklardan gelen verilerin analiz edilmeden önce entegre edilmesi gerekir. Bu aşamada tablolardaki birincil anahtar bilgileri düzgün bir şekilde girilmemişse veri tutarsızlıkları olabilir. Örneğin farklı iki kaynaktaki müşteri tabloları birleştirilecekse ve müşteri bilgisi olarak “müşteri isim” alanı kullanılıyorsa, “Ahmet Çelik” bilgisi sık kullanılan bir isim olduğu için bu veri kaliteli bir veri olmayabilir. Birincil anahtar olarak farklı bir alanın kullanılması süreci kolaylaştıracaktır [20;22].

### 2.5.3 Veri hazırlama fazı

Problem ve hedefler tanımlandıktan sonra sıra veri hazırlama fazındadır. Aynı alana girilmiş farklı bilgilerin tümleştirilmesi, örneğin "Medeni Hal" alanında bulunan "Evli", "evli", "Evlenmiş" veya "EVLİ" gibi bilgilerin standart bir şekilde "Evli" ye dönüştürülmesi bu fazda yapılır. Bu tip veri temizleme ve tutarlılaştırma işlemleri hatalı tahminde bulunmayı engellemek ve gerçekçi tahminde bulunabilmek için bu aşamada yapılması gereken işlemlerdendir [20;22].

Bu fazda yapılan işlemler 4 ana adımdan oluşmaktadır:

a) Değişken seçimi,

b) Kayıt/satırların seçimi,

c) Yeni değişkenlerin oluşturulması,

d) Değişkenlerin dönüştürülmesi.

#### 2.5.3.1 Değişken seçimi

Bu aşamadaki ideal yöntem eldeki tüm değişkenlerin veri madenciliği yazılımına girilmesi ve en iyi öngörü/tahmin değişkenlerinin bulunmasıdır. Ancak uygulamada çalışması kolay değildir. Çünkü değişken sayısı arttıkça model oluşturma süresi de artacaktır. Diğer bir neden ise hiç bir kısıt gözetmeksizin model içine konulan çok sayıdaki değişkenin yanlış modelin oluşmasına sebep olabileceğidir. En sık görülen hatalardan birisi, ancak sonuç değişkenin değeri bilindiğinde öngörü/tahmin değişken değerin kullanılmasıdır. Örneğin, "Doğum tarihi" bilgisinin "yaş" bilgisini "tahmin etmek" için kullanılması ve gerçek değerin farkedilmemesi buna örnek olarak verilebilir.

Normal şartlar altında bazı veri madenciliği algoritmaları otomatik olarak ilgisiz değişkenleri saptayacak ve sadece ilgili değişkenleri hesaba katacaktır. Ancak pratikte sadece veri madenciliği aracının yöntemine güvenmekten kaçınılmalıdır. Çoğunlukla soruna ilişkin bilgiler kullanıcıya seçimleri doğru yapma şansı tanıyacaktır [45].

#### 2.5.3.2 Kayıt/Satırların seçimi

Değişken seçiminde olduğu gibi model oluşturma sırasında tüm kayıtların kullanılması istenebilir. Ancak veri miktarı çok büyükse model oluşturmak çok zaman alacaktır veya daha güçlü işlemcilere ihtiyaç duyulacaktır.

Sonuç olarak veritabanının büyük olduğu durumlarda örneklem kullanmak daha etkili bir yöntemdir. Bu durumda bilgi kaybının önlenmesi açısından örneklemin rastgele alınması önemlidir.

Ayrıca sapan veya aykırı değerleri atmak gerekebilir. Bazı durumlarda bu değerler modelin oluşturulmasına ilişkin önemli bilgiler içerebilir ama problemin anlaşılma düzeyine göre bunlar gözardı edilebilir. Örneğin, yanlış girilmiş bir veri veya sadece olağanüstü bir durumda oluşmuş veri gibi [45].

#### 2.5.3.3 Yeni değişkenlerin oluşturulması

Genellikle ham veriden yeni değişkenler oluşturmak gerekir. Örneğin kredi riskinin tahmin edilmesinde borç ve gelir değişkenlerini ayrı ayrı değerlendirmek yerine borç/gelir oranını temsil eden yeni bir değişkeni değerlendirmek daha etkin ve anlaşılır sonuçlar verecektir. Yanlız başlarına daha az etkileri olan değişkenler toplama, oran gibi aritmetik ve cebirsel yöntemlerle diğer değişkenlerle birleştirilebilir. Bununla birlikte bazı değişkenlerin zamana göre değişimi o değişkenlerin anlık değerlerinden daha önemli olabilir. Örneğin yanlız başına “gelir” değişkenini kullanmak yerine “gelirin zamana göre değişimi” gibi bir değişkeni kullanmak gerekebilir [45].

#### 2.5.3.4 Değişkenlerin dönüştürülmesi

Veri madenciliği için seçilen araç verinin temsil edilmesi noktasında ve kategorilendirmede bazı yönlendirmeler yapılabilir. Değişkenler 0 ile 1 şeklinde ölçeklendirilebilir. Örneğin gelirin sınıflandırmasında “yüksek”, “orta”, “düşük” kategorileri kullanılabilir. Bu seçilen sınıflandırma modelin sonuçlarını etkileyecektir [45].

### 2.5.4 Modelleme fazı

Modelleme fazı temel olarak uygun modelleme tekniklerinin belirlenmesi ve uygulanması, eniyileme için model değişkenlerinin düzenlenmesinden oluşur. Gerektiği durumlarda veri hazırlama fazına dönülebilir ve aynı veri madenciliği problemi için birden fazla teknik kullanılabilir [22].

Modelleme ile ilgili en önemli nokta bu sürecin tekrarlanan bir süreç olduğudur. Alternatif algoritmalar ve teknikler kullanılabileceği için kullanıcılar en iyi sonuca hangi yöntemle ulaşacaklarına deneme yanılma yöntemi ile ulaşabilirler. Tahmin için örneğin sınıflandırma veya regresyon analizi seçildikten sonra modelleme için de bir yöntem seçilmelidir. Seçilecek olan yöntem ne tip bir verinin hazırlanacağı ve nasıl ilerleneceği konusuyla doğrudan ilgilidir. Veya kullanılacak olan veri madenciliği aracı hazırlanacak olan verinin spesifik bir formatta olmasını zorunlu kılar [20;45].

### 2.5.5 Değerlendirme fazı

Modelleme yapıldıktan sonra bu modelin başlangıçta belirlenen iş hedeflerinin ne kadarını karşıladığı ölçülmeli ve ortaya çıkan sonucun yaygınlaştırma fazından önce kalite ve etkisi değerlendirilmelidir. Bununla birlikte problemde ele alınacak noktaların yeterli derecede dikkate alınıp alınmadığı kontrol edilmeli ve sonuçların kullanılıp kullanılmayacağı ile ilgili net karar verilmelidir [20;22;45].

Sınıflandırma problemleri için hata matrisi, sonuçları değerlendirmek için oldukça etkili bir araçtır.

Çizelge 2.5. Hata matrisi

| | *Güncel* | | |
|---|---|---|---|
| *Tahmin* | **Sınıf A** | **Sınıf B** | **Sınıf C** |
| **Sınıf A** | 45 | 2 | 3 |
| **Sınıf B** | 10 | 38 | 2 |
| **Sınıf C** | 4 | 6 | 40 |

Çizelge 2.5'teki matris örnek bir hata matrisidir. Sütunlar güncel sınıfları, satırlar da tahmini sınıfları göstermektedir. Köşegen değerleri de doğru tahminleri içermektedir. Bu matris modelin ne kadar iyi tahmin edildiğini göstermesinin yanında parametrelerin nerede yanıldığı konusunda da fikir verir. Çizelge 2.5 incelendiğinde modelin 46 Sınıf B değerinden 38'ini doğru tahmin ettiğini ve 2 si Sınıf A, 6 sı Sınıf C olmak üzere toplam 8 tanesini ise yanlış sınıflandırdığı görülebilir. Bu ayrıntıdaki bir bilgi ise toplam doğruluk oranının %82 (150 durumda 123 doğru sınıflandırma) olduğunu belirtmekten daha bilgi verici ve açıklayıcı bir bilgidir.

Genellikle farklı hata tipleri için farklı maliyetler oluştuğu durumlarda doğruluk oranı düşük bir model, doğruluk oranı yüksek ve daha maliyetli bir modele tercih edilebilir. Örneğin, yukarıdaki hata matrisinde her doğru yanıt 10$ değerinde olsun, her bir yanlış yanıt Sınıf A için 5$, Sınıf B için 10$, Sınıf C için 20$ maliyet getiriyor olsun. Matrisin net değeri:

$$(123 \times 10\$) - (5 \times 5\$) - (12 \times 10\$) - (10 \times 20\$) = 885\$$$

Çizelge 2.6 daki hata matrisinde doğruluk oranı %79'a (150 durumda 118 doğru sınıflandırma) düşürülmüştür. Diğer yandan bir önceki matrise uygulanan değer ve maliyetler uygulandığında net değer:

$$(118 \times 10\$) - (22 \times 5\$) - (7 \times 10\$) - (3 \times 20\$) = 940\$$$

olarak elde edilir.

Çizelge 2.6. Hata matrisi

| | *Güncel* | | |
|---|---|---|---|
| *Tahmin* | **Sınıf A** | **Sınıf B** | **Sınıf C** |
| **Sınıf A** | 40 | 12 | 10 |
| **Sınıf B** | 6 | 38 | 1 |
| **Sınıf C** | 2 | 1 | 40 |

Böylece modelin değeri ile ilgili doğruluk oranı ve maliyet arasında kullanıcı bir seçim yapabilir [22].

### 2.5.6 Yaygınlaştırma fazı

Yaygınlaştırma fazının başarı oranı oluşturulan modelden yararlanılması ile doğru orantılıdır. Ayrıca bu aşamada veri madenciliği çalışmasının sonuçlarının varsa proje sponsoruna raporlanması gerekmektedir. Veri madenciliği çalışması değerlendirilmesi gereken yeni bir bilgiyi ortaya çıkarır ve bu bilginin proje hedefleri ile birleştirilmesi gerekir.

Ayrıca araştırma sonucunda elde edilen bilginin zamanla değişebileceği de göz önünde bulundurulmalıdır. Müşteri davranışları zamanla değişebilir ve buna bağlı olarak oluşacak verinin yapısı da değişebilir. Bu yüzden yaygınlaştırma fazında genel parametrelerde radikal bir değişiklik olup olmadığı izlenmelidir [28;45].

## 2.6 Örnek Çalışma: K-grup Kümeleme Analizi Kullanarak Turizm Pazarı Profil Oluşturma Çalışması

### 2.6.1 İş tanımlama fazı

Alberta'daki eyalet içi turist tercihlerini araştıran Kanada Alberta Calgary üniversitesi araştırmacılarının amaçları, turistlerin karar ve tercihlerine dayanarak yerli Alberta turistlerinin profillerini oluşturmaktır. Travel Alberta firması tarafından sponsorluğu yapılan bu çalışmanın nihai amacı gelişen turizm pazarına dayanak teşkil edecek eyalet içi nicel bir temel oluşturmaktır. Hangi faktörlerin gidilecek yerlerin seçiminde ve tatil yapma tercihleri üzerinde etkili olduğunu belirlemek istenmektedir.

### 2.6.2 Veri tanımlama fazı

Çalışmada kullanılan veri 13.445 kişi ile yapılan tele-anket sonucunda 1999 yılında toplanmıştır. Katılımcıların 18 yaşın üzerinde olup olmadığı ve son bir yıl içinde boş vakitlerinde en az bir günlüğüne ve en az 80 km. mesafe uzaklığa gidip gitmedikleri sorulmuştur. 13.445 kişiden sadece 3.071 kişi tele-anketi tamamlamış ve bu veriler çalışmaya girdi olarak kullanılmıştır.

### 2.6.3 Veri hazırlama fazı

Katılımcılara tatil fikirleri üzerinde 13 faktörlük liste arasında en çok hangi faktörün etkili olduğu sorulmuştur. Kalınacak otelin ve yerin kalitesi, okul tatil günleri ve hava koşulları gibi bu faktörler kümeleme analizinde değişkenler olarak değerlendirilmiştir.

### 2.6.4 Modelleme fazı

Kümeleme, segment profillerini oluşturmanın doğal bir yöntemidir. Araştırmacılar k-grup kümeleme yöntemini tercih etmişlerdir. Çünkü bu algoritma çalışma sonucu oluşacak küme sayısının tahmin edilebildiği durumlarda hem hızlı hem de etkili bir yöntemdir. Çeşitli küme sayıları üzerinde çalıştıktan sonra 5'li küme yapısının gerçeği en çok yansıtan model olduğuna karar vermişlerdir. Kümelerin özet profili:

Küme 1: Genç, şehir dışı ağırlıklı turizm pazarı.

Küme 2: Şehir içi, boş zaman turizm pazarı.

Küme 3: Önce çocuk tercihli turizm pazarı,

Küme 4: Güzel hava dostları turizm pazarı.

Küme 5: Yaşlı, maliyet bilinçli turizm pazarı.

### 2.6.5 Değerlendirme fazı

Kümeleme kategorizasyonunu doğrulamak için ayrıştırma analizi (discriminant analyses) kullanılmış ve %93 oranda doğru bir kümeleme yapıldığı görülmüştür. Değerlendirme analizi ayrıca kümeler arasındaki farkların istatistiksel olarak önemli olduğunu da göstermiştir.

### 2.6.6 Yaygınlaştırma fazı

Tüm bu çalışmalar veri madenciliğinde küme yapısına dayanarak yeni bir pazarlama kampanyası olarak; "Alberta, ısmarlayın" biçiminde sonuçlanmıştır. Hükümet ve iş dünyasının iş birliği ile 80 den fazla proje başlatılmıştır. Televizyon reklamları 55 yaşından küçük yetişkinlerin %92'si tarafından 20 kez izlenmiş ve Travel Alberta

firması daha sonra Alberta'yı turizm merkezi olarak düşünen kişi sayısında %20 artış olduğunu farketmiştir [22].

## 2.7 Örnek Çalışma: Karar Ağacı Kullanarak Şirket İflaslarını Tahmin Etme

### 2.7.1 İş tanımlama fazı

Doğu Asya'daki ekonomik kriz bölgede ve dünyada çok sayıda şirketin iflasına neden olmuştur. Kyonggi Üniversitesi'nden Tae Kyung Sung, Seul Üniversitesi'nden Namsik Chang ve Sogang Üniversitesi'nden Gunhee Lee şirket iflaslarını önceden tahmin edebilen sonuçlar için yorumlanabilir bir model geliştirmişlerdir.

Bu bağlamda araştırmacılar analiz yöntemi olarak saydamlığı ve yorumlanabilirliği sebebiyle Karar Ağacı'nı kullanmayı tercih ettiklerini ifade etmişlerdir [22].

### 2.7.2 Veri tanımlama fazı

Veri iki grupta incelenmiştir: Görece olarak daha kararlı bir dönem olan 1991-1995 arası iflas eden Koreli şirketler ve ekonomik kriz şartları geçerli olan 1997-1998 döneminde iflas eden Koreli şirketler. Çeşitli tarama prosedürlerinden sonra çoğunluğu üretim sektöründen 29 şirket seçilmiştir. Finansal veriler menkul kıymetler borsasından toplanmış ve Kore Bankası ile Kore Endüstri Bankası tarafından doğrulanmıştır.

### 2.7.3 Veri hazırlama fazı

İflas tahmini üzerinde taranan literatür sonucunda 56 finansal rasyo belirlenmiş, kayıtların 16 tanesi mükerrer olduğu için geriye kalan 40 finansal rasyo üzerinde durulmuştur. İncelenen kriterler büyüme, karlılık, sermaye ve verimlilik olmuştur.

### 2.7.4 Modelleme fazı

Karar ağacı modelleri "normal koşullar" verisine ve "kriz koşulları" verisine ayrı ayrı uygulanmıştır. Karar ağacı modelleri ile kurulan kural kümelerinden bazıları aşağıda verilmiştir. Normal koşullar için;

- Sermayenin verimi 19.65'ten büyük olduğunda iflas etmeme ihtimali %86'dır.

- Nakit akış/toplam aktif oranı -5.65'ten büyük olduğunda iflas etmeme ihtimali %95 tir.

- Sermayenin verimi 19.65'ten küçük veya eşitse ve nakit akış/toplam aktif oranı -5.65'ten küçükse iflas etme ihtimali %84'tür.

Kriz koşulları için;

- Sermayenin verimi 20.61'ten büyük olduğunda iflas etmeme ihtimali %91'dir.

- Nakit akış/pasif oranı 2.64'ten büyük olduğunda iflas etmeme ihtimali %85'tir.

- Duran varlıklar/öz kaynaklar oranı 87.23'ten büyükse iflas etmeme ihtimali ise %86'dır.

- Sermayenin verimi 20.61'den küçük, nakit akış/pasif oranı 2.64'ten küçük; duran varlıklar/öz kaynaklar oranı 87.23'ten küçükse iflas etme ihtimali %84'tür.

Nakit akışı ve sermayenin veriminin ekonomik şartlardan bağımsız olarak önemli parametreler olduğu sonucu elde edilmiştir.

### 2.7.5 Değerlendirme fazı

Uzmanların katıldığı bir panelde sermaye veriminin firmaların iflası üzerinde yoğunlaşılmıştır. Dolayısıyla bir karar ağacı uygulaması sonucunda tahmin edilemeyen bir kriterin etkili olduğu görülebilir.

Modelin Kore'deki tüm üretim firmaları üzerinde uygulanıp uygulanamayacağını anlamak için iflas etmemiş örnek firmalardan bir grup seçilmiş ve parametreler veri setindeki firmaların parametreleri ile karşılaştırılmıştır. Örnek seçilen firmaların varlıklarının ve çalışan sayılarının veri setindeki değerlerin %20 sınırları içerisinde olduğu görülmüştür.

Sonuç olarak araştırmacılar performans göstergesi olarak çoklu diskriminant analizi uygulamışlar ve 40 finansal rasyodan çoğunun, iflası tahmin etmek için önemli kriterler olduğunu elde etmişlerdir.

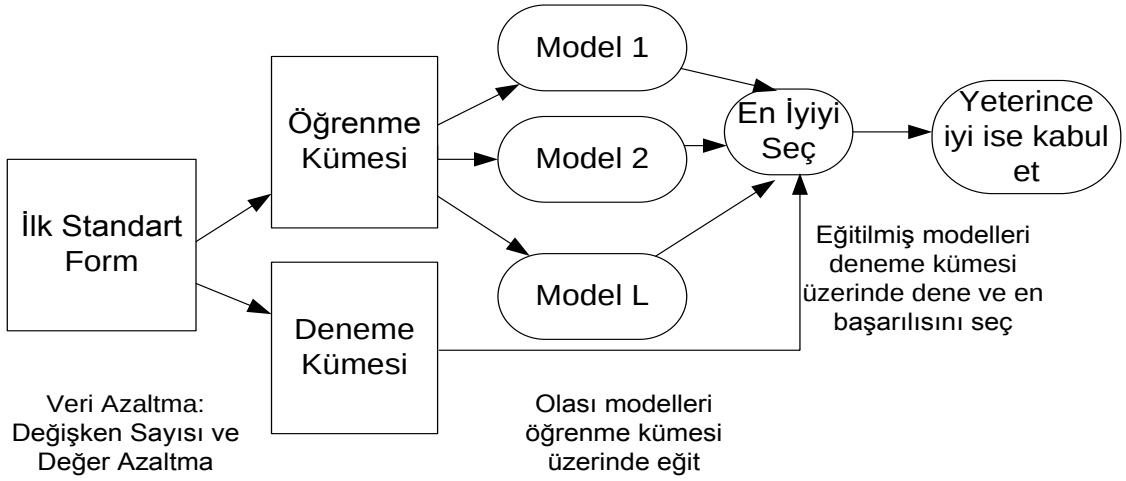
### 2.7.6 Yaygınlaştırma fazı

Bu çalışma için kullanıcının ve uygulayıcının insiyatifinde olarak bir yaygınlaştırma fazı gerçekleşmemiştir. Ancak bu araştırma sebebiyle Kore'deki finansal kurumlar ekonomik kriz şartlarında firmaların iflas etme ihtimallerini daha iyi tahmin edebilmişler ve bu sorun karşısında daha hazırlıklı olabilmişlerdir [22].

## 2.8 Metodoloji

Bir veri madenciliği çalışmasında kullanılan metodoloji Şekil 2.5'de verilmiştir. Standart form içinde verilen veri, öğrenme ve deneme olmak üzere ikiye ayrılır. Her uygulamada kullanılabilecek birden çok teknik vardır ve önceden hangisinin en başarılı olacağını kestirmek olası değildir. Bu yüzden öğrenme kümesi üzerinde tümü için birbirinden farklı teknikler kullanılarak L tane model oluşturulur. Daha sonra bu L model deneme kümesi üzerinde denenerek en başarılı olanı, yani deneme kümesi üzerindeki tahmin başarısı en yüksek olanı seçilir.

Eğer bu en iyi model yeterince başarılıysa kullanılır, aksi takdirde başa dönerek çalışma tekrarlanır. Tekrar sırasında başarısız olan örnekler incelenerek bunlar üzerindeki başarının nasıl arttırılabileceği araştırılır. Örneğin standart forma yeni alanlar ekleyerek programa verilen bilgi arttırılabilir; veya olan bilgi değişik bir şekilde kodlanabilir; veya amaç daha değişik bir şekilde tanımlanabilir [4].




Şekil 2.5. Veri madenciliği çalışmasında kullanılan metodoloji

# 3 VERİ MADENCİLİĞİ TEKNİKLERİ VE YÖNTEMLERİ

Veri madenciliği konusunda bir çok teknik ve yöntem bulunmaktadır ve hangi yöntemin kullanılması gerektiği problemin tanımına ve verinin yapısına göre değişir. Bu yüzden “en iyi yöntem” veya “en iyi algoritma” gibi bir kavramdan sözedilemez.

Temel olarak veri madenciliği teknikleri üç ana başlık altında gruplanabilir:

a) Sınıflandırma (Classification) ve Regresyon (Regression)

b) Kümeleme (Clustering)

c) Birliktelik Kuralları (Association Rules)

Bu ana başlıklar içerisinde standart istatistik tabanlı yöntemlerin yanında yapay zeka ve mühendisliğin de kullanıldığı yeni teknoloji yöntemler de bulunmaktadır. Yeni teknoloji veri madenciliği yöntemlerinin ortak özelliği, ilişki bulma mekanizmalarının kullanıcı odaklı yerine veri odaklı olmalarıdır. Bu sayede kişiye bağımlılık ortadan kalkmış olur ve veri madenciliği yazılımları sadece veri üzerine yoğunlaşarak örüntüleri ve ilişkileri ortaya çıkarır [29;37;45].

## 3.1 Sınıflandırma ve Regresyon

Sınıflandırma yöntemi kampanya geri dönüşleri, müşteri segmentasyonu, müşteri kaybı ve kredi analizi gibi bir çok alanda sıkça kullanılan bir yöntemdir. Sınıflandırma, her gözlemin, karakteristiği önceden tanımlanmış gruplardan hangisine ait olduğunu bulmayı amaçlar. Bu yöntemle hem varolan veri anlaşılabileceği gibi hem de bir değerin nasıl davranacağı tahmin edilir.

Sınıflandırma iki aşamalı bir süreçtir. Birinci aşama sabit bir kategori listesine göre bir veri setini veya durumu sınıflara ayırma işlemidir. Kategori belirli bir değerler kümesidir. Örneğin “Müşteri kampanyaya katılacak mı?” sorusunun iki kategori değeri vardır, “Evet” ve “Hayır”. Örnek bir veri seti üzerinde (Çizelge 3.1) sınıflandırma modeli incelendiğinde öngörme değerleri olarak yaş, gelir, çocuk sayısı gibi demografik değerler de alınabilir [34].

Çizelge 3.1. Sınıflandırma için kullanılan bir veri seti tanımlaması

| **Müşteri ID** | **Gelir** | **Yaş** | **Çocuk** | **Ev geçindiriyor** | **...** | **Kampanya (1=Evet, 0=Hayır)** |
|---|---|---|---|---|---|---|
| 1001 | 3.000 | 30 | 2 | E | | 1 |
| 1002 | 5.500 | 67 | 3 | H | | 1 |
| 1003 | 2.500 | 23 | 0 | H | | 0 |
| 1004 | 5.000 | 44 | 1 | E | | 0 |
| | | | | | | |
| | | | | | | |
| *$X_1$* | *$X_2$* | *...* | *...* | *...* | *$X_m$* | *Y* |
| *Durum belirteci* | *Öngörme Değerleri/Özellikleri* | | | | | *Hedef Değer* |

İkinci aşamada ise veri setinin üzerinde sınıflandırma kuralları belirlenir ve aşağıdaki fonksiyon elde edilir:

$$Y = f(X_2, ..., X_m) \quad (3.1)$$

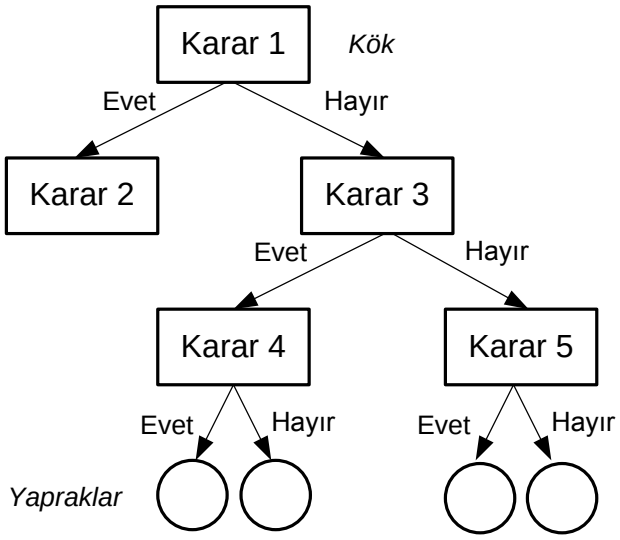
Söz konusu kurallar bu veri seti üzerinde uygulanır. Örneğin yeni bir müşterinin kampanyaya katılım cevabının olumlu olup olmadığı sorgulanıyor olsun. Bu müşterinin kampanyaya nasıl davranacağını belirlemek için örnek verilerden elde edilen karar kuralı doğrudan uygulanır.

Sınıflandırma modelinin Eşitlik 3.1'e bağlı olarak kalitesi güncel ve hedef değerlerinin doğru bir şekilde karşılaştırılmasına bağlıdır [13;20;45].

### 3.1.1 Karar ağaçları

Tahmin edici ve tanımlayıcı özelliklere sahip olan karar ağaçları akış şemalarına benzeyen yapılardır. Herbir nitelik bir düğüm tarafından temsil edilir. En üst yapı

“kök”, en son yapı “yaprak” ve bu yapılar arasındakiler de “dal” olarak isimlendirilir. (Şekil 3.1)




Şekil 3.1. Karar ağacı

Karar ağaçları temel olarak iki aşamadan oluşmaktadır; ağacın kurulması ve verilerin teker teker ağaca uygulanarak sınıflandırması biçimindedir. İstatistiksel yöntemlerde veya yapay sinir ağlarında verinin uyduğu fonksiyon belirlendikten sonra bu fonksiyonun insanlar tarafından anlaşılabilecek bir kural olarak yorumlanması zordur. Karar ağaçları ise veriden oluşturulduktan sonra yukarıdaki şekilde olduğu gibi ağaç kökten yaprağa doğru inilerek kurallar (*IF-THEN kuralları*) yazılabilir [4;7;37].

Örnek 1. Belirli ürünlerin temini için sözleşmelerle ilgili teklif verip vermeme konusunda karar aşamasında olan bir firmanın üç alternatifi bulunduğu varsayılsın:

1. Sadece MS1 sözleşmesi için teklif vermek (maliyet £50.000) veya

2. Sadece MS2 sözleşmesi için teklif vermek (maliyet £14.000) veya

3. MS1 ve MS2 sözleşmeleri için teklif vermek (maliyet £55.000)

Eğer teklif kabul edilirse ürünlerin temin maliyetleri sırasıyla MS1 sözleşmesi için £18.000, MS2 sözleşmesi için £12.00, her ikisi için ise £24.000‘dir.

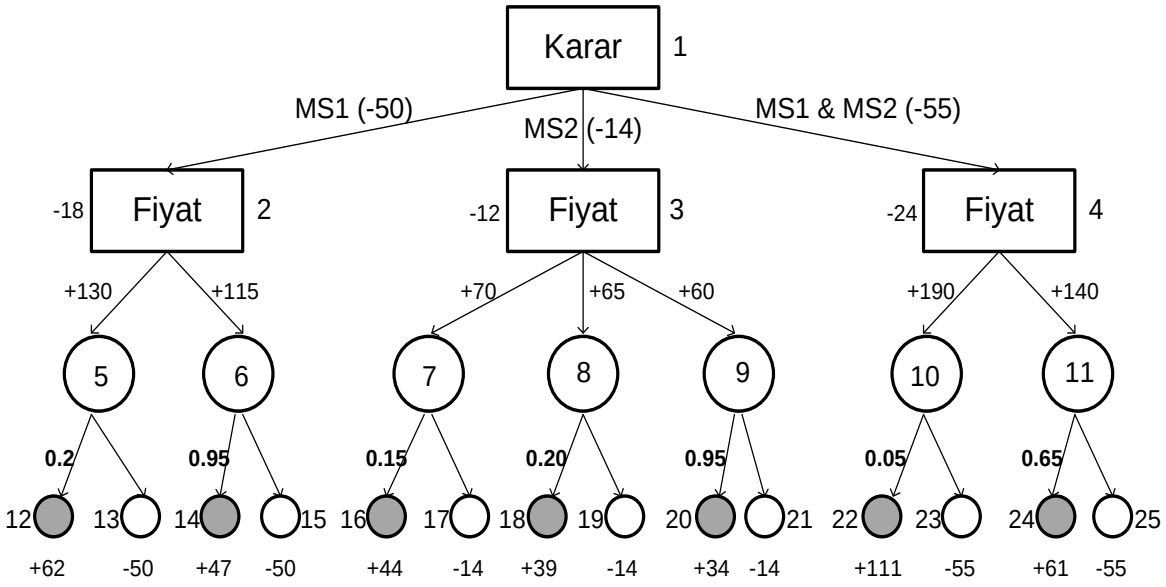
Ek olarak belirli fiyatlar için tekliflerin kabul edilip edilmeme olası tablosu ise

Çizelge 3.2’de verilmiştir.

Çizelge 3.2. Sözleşme teklif ve ihtimal tablosu

| Opsiyon | Muhtemel Teklif Fiyatları (£) | Sözleşme Yapma İhtimali |
|---|---|---|
| MS1 | 130.000<br>115.000 | 0.20<br>0.85 |
| MS2 | 70.000<br>65.000<br>60.000 | 0.15<br>0.80<br>0.95 |
| MS1 ve MS2 | 190.000<br>140.000 | 0.05<br>0.65 |

Bu örnek için bir karar ağacı Şekil 3.2’de verilmiştir.




Şekil 3.2. Sözleşmeler için teklif verme karar ağacı.

Kazanan nodlara ulaşan yollar 12, 16, 2 ve 3 nodlar için aşağıdaki gibi açıklanabilir:

Nod 12 için: Sadece MS1 için fiyat teklifi 130 verilir (maliyet 50), sözleşme imzalanır ve 18 birim tedarik masrafı oluşur. Toplam kar 130-50-18=62 biçiminde elde edilir.

Nod 16 için: Sadece MS2 için fiyat teklifi 70 verilir (maliyet 14), sözleşme imzalanır ve 12 birim tedarik masrafı oluşur. Toplam kar 70-14-12=44 biçiminde elde edilir.

Nod 22 için: MS1 ve MS2 için fiyat teklifi 190 verilir (maliyet 55), sözleşme imzalanır ve 24 birim tedarik masrafı oluşur. Toplam kar 190-55-24=111 biçiminde elde edilir.

Nod 2 için karar verme aşamasında olasılıklar kullanılırsa:

Nod 5: 0.2 x (62) + 0.8 x (-50) = -27.6

Nod:6: 0.85 x (47) + 0.15 x (-50) = 32.45

Bu yüzden Nod 2 de iken en iyi karar fiyat teklifini £115 olarak vermektir.

Aynı şekilde Nod 3 te karar aşamasında iken en iyi karar fiyat teklifini 60, Nod 4 te iken 140 olarak vermek olduğu görülür.

Karar ağaçları kuruluşlarının ve yorumlanmalarının kolay olması sebebiyle yaygın bir kullanıma sahiptir. Dezavantajı ise sadece bir dağımlı değişken kullanımına olanak verdiği için her bir değişken için ayrı ayrı karar ağacı kurgulamayı gerektirir [7;2].

Karar ağacının oluşturulmasında CHAID (Chi-squared Automatic Interaction Detection), CART (Classification ve Regression), ID3, QUEST (Quick, Unbiased, Efficient Statistical Tree), C4.5, C5.0, gibi algoritmalar kullanılır. Bu algoritmaların bazıları aynı zamanda regresyon için de uyarlanabilir. Çeşitli algoritmaların ortaya çıkış sebebi, karar ağacı oluşturulurken herhangi bir kökten itibaren ayrışmanın ve dallanmanın hangi kritere göre yapılacağı sorununa farklı yaklaşımlarda bulunmasından kaynaklanmaktadır.

#### 3.1.1.1 ID3

ID3 algoritması temelde sınıflamayı yapmak için kullanılacak değişkenin belirlenmesinde etkilidir. Bu belirlemeyi yaparken matematiksel bazı formüller ve entropi kavramı baz alınmaktadır. Bir sistemdeki belirsizliğin ölçüsüne entropi adı verilir, matematiksel olarak aşağıdaki biçimde ifade edilir:

$p_1, p_2, \ldots, p_n$ olasılıkları ile D öğrenme veritabanını olduğu varsayılsın.

$\sum_{i=1}^{i=n} p_i = 1$ olmak üzere;

Entropi: $H(p_1, p_2, \ldots, p_n) = -\sum_{i=1}^{i=n} p_i \log_2(p_i)$ (3.2)

ve

Kazanım: $H(D) - \sum_{i=1}^{i=n} P(D_i)\, H(D_i)$ (3.3)

biçimindedir.

Eşitlik 3.3'teki kazanım ifadesinin; verilerin ham halinin (başlangıçtaki) entropiyle herbir alt bölümün entropilerinin ağırlıklı toplamı arasındaki farka bağlı olduğu görülür.

Bu iterasyonlar sırasında her bir değişken için bu hesaplamalar yapılır ve kazanımın maksimum olduğu değerlere götüren değişkenler kök olarak seçilir.

Çizelge 3.3. ID3 algoritması uygulanacak eğitim veri seti

| Hava | Isı | Nem | Rüzgar | Aktivite |
|---|---|---|---|---|
| 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 |

Hava durumu ve aktivite ile ilgili Çizelge 3.3'teki veriler üzerinden örnek bir karar ağacı yapılandırması ve kök bulma iterasyonları aşağıdaki gibi gösterilebilir [29;31;37].

Bu tabloda Aktivite özelliği hedef sınıf değerlerini içermektedir. Bu yüzden hedef değerler kümesi olarak:

Aktivite = {*hayır, hayır, evet, evet, evet, hayır, evet, hayır, evet, evet, evet, evet, evet, hayır*}

Beş adet "hayır" değeri için $p_1$ = 5/14, dokuz adet evet değeri için $p_2$= 9/14 olmak üzere, olasılık ve entropi değerleri:

$P_{Aktivite} = \left(\frac{5}{14}, \frac{9}{14}\right)$

$H(Aktivite) = \left(\frac{5}{14} \log_2 \frac{5}{14} + \frac{9}{14} \log_2 \frac{9}{14}\right) = 0.940$

olarak hesaplanır.

Hava, Isı, Nem ve Rüzgar değişkenleri için entropi ve kazanım değerleri ayrı ayrı bulunur, en yüksek kazanım değerine sahip değişken kök olarak belirlenir.

Hava niteliği için kazanım değeri;

$|Hava_{güneşli}| = 5$

$|Hava_{yağmurlu}| = 5$

$|Hava_{bulutlu}| = 4$

tekrarlanma sayılarıdır. Hava niteliği için entropi değeri:

$H(Hava, Aktivite) = \frac{5}{14} H(Hava_{güneşli}) + \frac{4}{14} H(Hava_{bulutlu}) + \frac{5}{14} H(Hava_{yağmurlu})$

Bu ifadedeki entropi değerleri de:

$H(Hava_{güneşli}) = -\left(\frac{3}{5} \log_2 \frac{3}{5} + \frac{2}{5} \log_2 \frac{2}{5}\right) = 0.971$

$H(Hava_{yağmurlu}) = -\left(\frac{2}{5} \log_2 \frac{2}{5} + \frac{3}{5} \log_2 \frac{3}{5}\right) = 0.971$

$H(Hava_{bulutlu}) = -\left(\frac{4}{4} \log_2 \frac{4}{4}\right) = 0$

biçiminde elde edilir. Formülde değerler yerlerine konulduğunda:

$H(Hava, Aktivite) = \frac{5}{14}(0.971) + \frac{4}{14}(0) + \frac{5}{14}(0.971)$

$= 0.694$

$Kazanım\ (Hava, Aktivite) = H(Aktivite) – H(Hava, Aktivite)$

= 0.040 – 0.693

= 0.247

biçiminde elde edilir.

Isı, Nem ve Rüzgar için bu hesaplamalar tekrarlandığında ise:

Kazanım(Isı, Aktivite) = 0.029

Kazanım(Nem, Aktivite) = 0.151

Kazanım(Rüzgar, Aktivite) = 0.048 olarak bulunur.

Bu değerler karşılaştırıldığında en büyük kazanımın hava niteliği seçildiğinde elde edildiği görülür [29].

Benzer hesaplama yöntemi hava niteliğinin güneşli değeri için ve diğer tüm düğümler için kullanılır. Dolayısı ile her bir düğümdeki değişkenler ve karar ağacının genel yapısı oluşturulmuş olur.

#### 3.1.1.2 C4.5

C4.5 algoritması ID3 algoritmasının sayısal niteliklere ilişkin uyarlanabilir bir biçimidir ve yine Quinlan tarafından geliştirilmiştir. Böylece sayısal değerler içeren veritabanları üzerinde de karar ağaçlarının oluşturulma olanağı sağlanmıştır.

Sayısal nitelikleri belirli aralıklara bölme konusunda bazı zorluklar görülebilir. Ancak en uygun t eşik değerini hesaplamak için çeşitli yöntemler bulunmaktadır. Nitelik değerleri sıralanır ve $\{v_1, v_2, \ldots, v_n\}$ şeklini alır. Nitelik değerler kümesi iki parçaya ayrılır ve Eşik değeri olarak $[v_i, v_{i+1}]$ aralığının orta noktası alınabilir:

$$t_i = (v_i + v_{i+1}) / 2 \qquad (3.4)$$

Örneğin bir eğitim kümesinde yer alan:

(70, 90, 85, 95, 70, 90, 78, 65, 75, 80, 70, 80, 70, 96)

değerleri için nitelik kümesi {65, 70, 75, 80, 85, 90, 95, 96} değerlerine sahiptir. Kümenin orta noktaları olan (80, 85) aralığının orta noktası olan:

$t_i = (v_i + v_{i+1}) / 2 = (80 + 85) / 2 \approx 83$ noktası eşik değer olarak alınabilir.

Bu durumda bu değerleri içeren değişken için (nitelik $\leq$ 83) veya (nitelik > 83) testi uygulanarak hesaplamalara devam edilir ve entropi, kazanım değerleri bulunur [29].

#### 3.1.1.3 C5.0

C5.0 algoritması temelde C4.5 algoritması ile aynı ancak bazı özellikleri C4.5 algoritmasına göre daha gelişmiş ticari bir versiyondur. Artı özellikleri aşağıdaki şekilde sıralanabilir:

- C5.0 algoritması C4.5'e göre farkedilebilir derecede daha hızlıdır.
- Daha küçük karar ağaçları oluşturulabilir.
- Verinin ayrıştırılması ve kirli verinin iyileştirilmesinde daha etkilidir.

Bellek kullanımında C4.5'e göre daha az bellek kullanır [35].

#### 3.1.1.4 CART

Sınıflandırma ve Regresyon Ağaçları (Classification and Regression Trees) 1984 yılında Breiman tarafından ortaya çıkarılmıştır ve karar ağacının her bir düğümde iki dala ayrılma ilkesine dayandırılmıştır. Φ(s|t), t düğümündeki s aday bölünmelerinin uygunluk ölçüsü olsun:

$$\Phi(s|t) = 2P_{sol}P_{sağ}\sum_{j=1}^{j=n} \left|P(j \mid t_{sol}) - P(j \mid t_{sağ})\right| \qquad (3.5)$$

$t$ = Dallanmanın yapılacağı düğüm

$t_{sol}$ = $t$ düğümünün sol alt dalı

$t_{sağ}$ = $t$ düğümünün sağ alt dalı

$P_{sol}$ = ($t_{sol}$ daki kayıtların sayısı) / (Eğitim kümesindeki kayıtların sayısı)

$P_{sağ}$ = ($t_{sağ}$ daki kayıtların sayısı) / (Eğitim kümesindeki kayıtların sayısı)

$P(j \mid t_{sol})$ = ($t_{sol}$ daki kayıtların j sınıfları sayısı) / t deki kayıt sayısı

$P(j \mid t_{sağ})$ = ($t_{sağ}$ daki kayıtların j sınıfları sayısı) / t deki kayıt sayısı

Tüm değişkenler ve ayrıştırma kriterleri Eşitlik 3.5'e göre hesaplanır ve en büyük olanı seçilir. Bu değerin ilgili olduğu aday bölünme satırı dallanmanın yapılacağı satırı verecektir.

Çizelge 3.4. Kredi risk durumu sınıflandırması için eğitim veri seti

| Müşteri | Tasarruf | Varlık | Gelir | Kredi Risk Durumu |
|---|---|---|---|---|
| 1 | orta | yüksek | 75.000 | iyi |
| 2 | düşük | düşük | 50.000 | kötü |
| 3 | yüksek | orta | 25.000 | kötü |
| 4 | orta | orta | 50.000 | iyi |
| 5 | düşük | orta | 100.000 | iyi |
| 6 | yüksek | yüksek | 25.000 | iyi |
| 7 | düşük | düşük | 25.000 | kötü |
| 8 | orta | orta | 75.000 | iyi |

Kredi risk durumunun belirtildiği Çizelge 3.4'teki eğitim seti üzerinde CART algoritması kullanılarak karar ağacı oluşturulmak isteniyor olsun.

CART algoritması ikili sistem üzerinde ayrıştırma yöntemi üzerine olduğu için kök düğüm ayrıştırma aday değerleri Çizelge 3.5'teki gibi olacaktır.

Çizelge 3.5. *t* = Kök düğüm için ayrıştırma adayları

| Aday | Sol alt düğüm, $t_{sol}$ | Sağ alt düğüm, $t_{sağ}$ |
|---|---|---|
| 1 | tasarruf = düşük | tasarruf Є {orta, yüksek} |
| 2 | tasarruf = orta | tasarruf Є {düşük, yüksek} |
| 3 | tasarruf = yüksek | tasarruf Є {düşük, orta} |
| 4 | varlık = düşük | varlık Є {orta, yüksek} |
| 5 | varlık = orta | varlık Є {düşük, yüksek} |
| 6 | varlık = yüksek | varlık Є {düşük, orta} |
| 7 | gelir ≤ 25.000 | gelir > 25.000 |
| 8 | gelir ≤ 50.000 | gelir > 50.000 |
| 9 | gelir ≤ 75.000 | gelir > 75.000 |

Çizelge 3.5'teki tüm adaylar için Φ(s|t) değerleri bulunur ve Çizelge 3.6'daki değerlere ulaşılır.

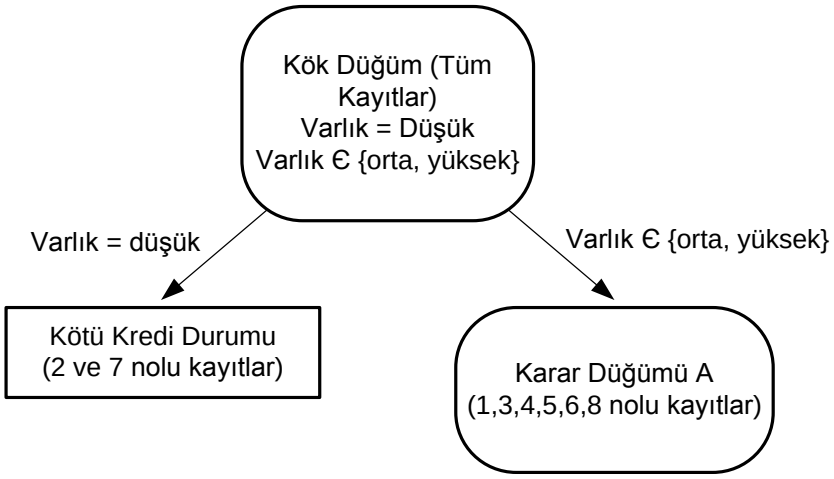
Çizelge 3.6. Kök düğüm için Φ(s|t) değerleri

| Aday | $P_{sol}$ | $P_{sağ}$ | $P(j \mid t_{sol})$ | $P(j \mid t_{sağ})$ | $2P_{sol}P_{sağ}$ | Q(s\|t) | Φ(s\|t) |
|---|---|---|---|---|---|---|---|
| 1 | 0.375 | 0.625 | İ: 0.333 K: 0.667 | İ: 0.8 K:0.2 | 0.46875 | 0.934 | 0.4378 |
| 2 | 0.375 | 0.625 | İ: 1.0 K:0 | İ:0.4 K:0.6 | 0.46875 | 1.2 | 0.5625 |
| 3 | 0.25 | 0.75 | İ:0.5 K:0.5 | İ:0.667 K:0.333 | 0.375 | 0.334 | 0.1253 |
| 4 | 0.25 | 0.75 | İ:0 K:1.0 | İ:0.833 K:0.167 | 0.375 | 1.667 | 0.6248 |
| 5 | 0.5 | 0.5 | İ:0.75 K:0.25 | İ:0.5 K:0.5 | 0.5 | 0.5 | 0.25 |
| 6 | 0.25 | 0.75 | İ: 1.0 K:0 | İ:0.5 K:0.5 | 0.375 | 1 | 0.375 |
| 7 | 0.375 | 0.625 | İ:0.333 K:0.667 | İ:0.8 K:0.2 | 0.46875 | 0.934 | 0.4378 |
| 8 | 0.625 | 0.375 | İ:0.4 K:0.6 | İ: 1.0 K:0 | 0.46875 | 1.2 | 0.5625 |
| 9 | 0.875 | 0.125 | İ:0.571 K:0.429 | İ: 1.0 K:0 | 0.21875 | 0.858 | 0.1877 |

$$Q(s|t) = \sum_{j=1}^{j=n} |P(j \mid t_{sol}) - P(j \mid t_{sağ})| \text{ olsun. Q(s|t) değeri } P(j \mid t_{sol}) \text{ ve } P(j \mid t_{sağ}) \quad (3.6)$$

arasındaki mesafenin maksimum olduğu yerde maksimum değerini alır. $2P_{sol}P_{sağ}$ ise maksimum değerine $P_{sol}$ ve $P_{sağ}$ değerlerinin birbirine yakın olduğu zaman yani birbirlerine eşit olduğu zaman ulaşacaktır. Böylece Φ(s|t) birbiriyle dengeli eşit sayıda eleman sayısına sahip iki dala ayrışmış bir düğüm olacaktır. $2P_{sol}P_{sağ}$ in teorik olarak maksimum değeri 2x(0.5)x(0.5) = 0.5 tir.

Çizelge 3.6'daki örnekte Φ(s|t) maksimum değerini 4 nolu satırda 0.6248 olarak alır. Bu yüzden CART başlangıç sınıflandırma için 4 nolu adayı yani varlık = düşük'e karşılık varlık Є {orta, yüksek} ayrımını seçer ve Şekil 3.3'deki gibi bir karar ağacı oluşur.



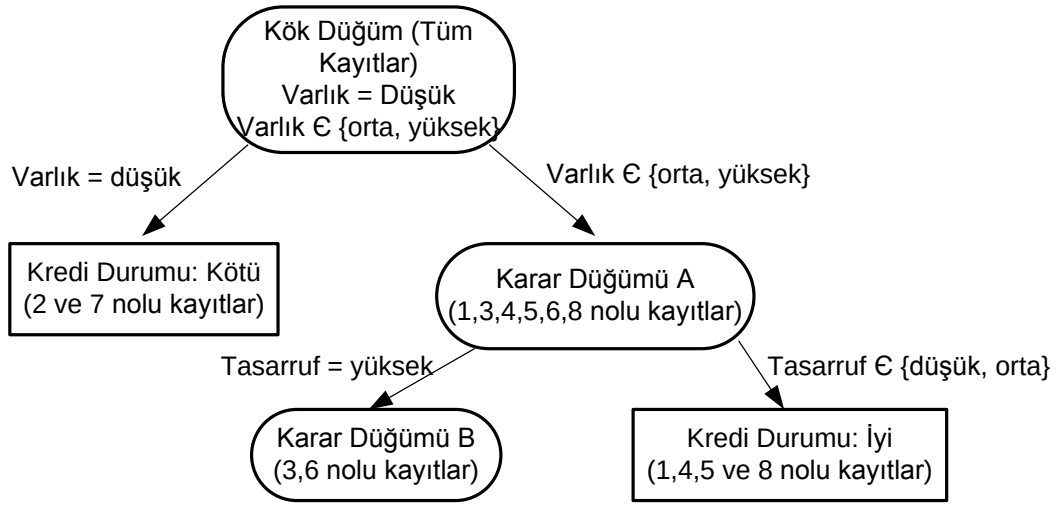

Şekil 3.3. İlk sınıflandırma sonrası CART karar ağacı

Bu karar ağacında sol taraf Kötü Kredi Durumu şeklinde sonlanır, sağ taraf ise daha fazla sınıflandırma için uygun bir yapıdadır. Çizelge 3.6'daki gibi aday düğümler üzerinde (4 nolu aday hariç) tekrar bir hesaplama tablosu çıkarılır ve Çizelge 3.7'deki değerlere ulaşılır.

Çizelge 3.7. Karar düğümü A için Φ(s|t) değerleri

| Aday | $P_{sol}$ | $P_{sağ}$ | $P(j \mid t_{sol})$ | $P(j \mid t_{sağ})$ | $2P_{sol}P_{sağ}$ | Q(s\|t) | Φ(s\|t) |
|---|---|---|---|---|---|---|---|
| 1 | 0.167 | 0.833 | İ: 1.0 K:0 | İ: 0.8 K:0.2 | 0.2782 | 0.4 | 0.1112 |
| 2 | 0.5 | 0.5 | İ: 1.0 K:0 | İ:0.667 K:0.333 | 0.5 | 0.6666 | 0.3333 |
| 3 | 0.333 | 0.667 | İ:0.5 K:0.5 | İ: 1.0 K:0 | 0.4444 | 1 | 0.4444 |
| 5 | 0.667 | 0.333 | İ:0.75 K:0.25 | İ: 1.0 K:0 | 0.4444 | 0.5 | 0.2222 |
| 6 | 0.333 | 0.667 | İ: 1.0 K:0 | İ:0.75 K:0.25 | 0.4444 | 0.5 | 0.2222 |
| 7 | 0.333 | 0.667 | İ:0.5 K:0.5 | İ: 1.0 K:0 | 0.4444 | 1 | 0.4444 |
| 8 | 0.5 | 0.5 | İ:0.667 K:0.333 | İ: 1.0 K:0 | 0.5 | 0.6666 | 0.3333 |
| 9 | 0.167 | 0.833 | İ: 0.8 K:0.2 | İ: 1.0 K:0 | 0.2782 | 0.4 | 0.1112 |

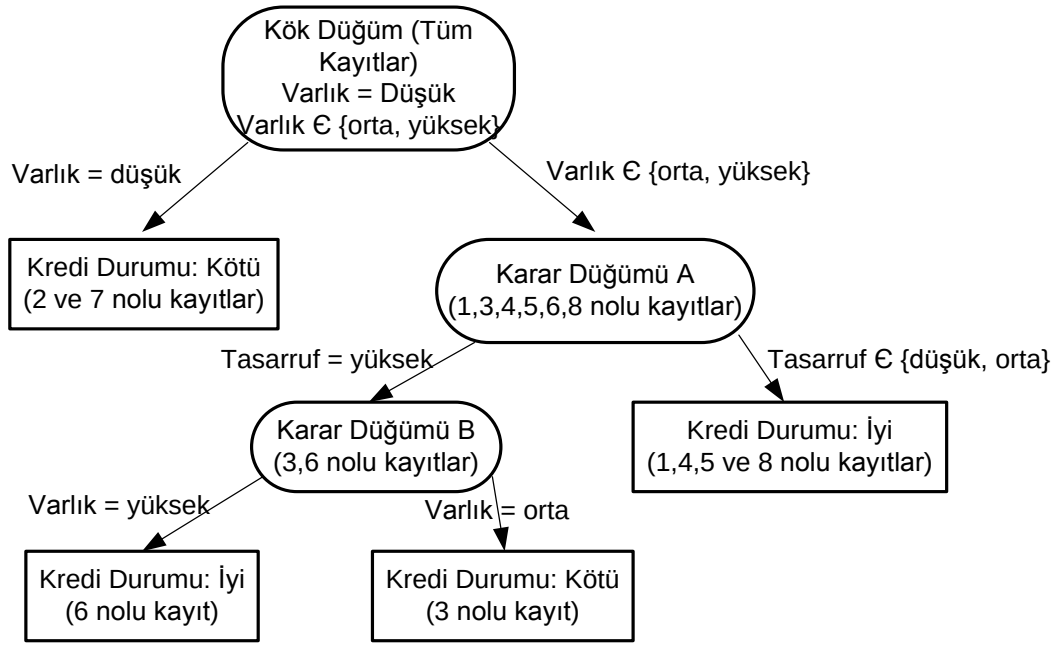
Çizelge 3.7'de ise 3 ve 7 nolu adaylar maksimum Φ(s|t) değerini 0.4444 paylaşırlar. Bu durumda 3 nolu aday seçilebilir, tasarruf = yüksek ve tasarruf {düşük, orta} ayrımı ile Şekil 3.4'deki karar ağacına ulaşılır [22;29;37].




Şekil 3.4. İkinci sınıflandırma sonrası CART karar ağacı

Karar Düğümü B'de tasarruf ve gelir değerleri aynı olup tek değişken varlıktır.

Varlık = orta ve varlık = yüksek değerlerine göre kırılım eklendiğinde karar ağacının son hali Şekil 3.5'deki gibi olur ve her bir dal bir değerde sonlanmış olur.




Şekil 3.5. CART karar ağacı son hali

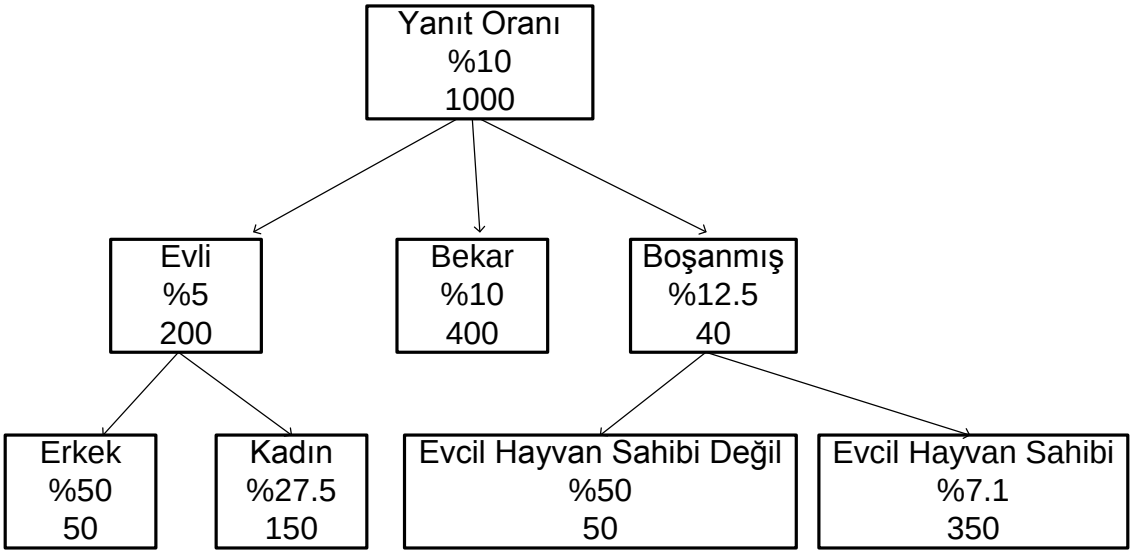
#### 3.1.1.5 CHAID

CHAID (Otomatik Ki-kare Etkileşim Belirleyicisi) algoritması 1980 yılında Gordon V.Kass tarafından bulunmuştur. CHAID analizi büyük hacimli verileri daha anlaşılabilir bir şekilde yorumlayabilmek için değişkenler arasındaki ilişkiyi bulma ve bağımlı değişken varyasyonlarını kullanarak büyük veri kümesini alt kümelere bölme işlemidir [9;33].

CHAID analizi, kategorik değişkenleri içeren veri kümesini homojen alt gruplara böler. Bu bölme işlemiyle bağımlı değişkeni ayrıntılı bir şekilde açıklar. Bölme işlemi sonucu oluşan alt kümeler, küçük tahmin edici gruplardan oluşur.

Başlangıç değişkenleri bağımsız olarak yeniden kategorilendirilerek en iyi tahmin sonucu elde edilmeye çalışılır. Bu işlem için Ki-kare analizi kullanılır. Benzer kategorileri birleştirme işlemi, değişkenler arasında daha fazla birleştirme sağlanamayacağı sonucuna, istatistiksel olarak varılıncaya kadar devam eder [32].

1000 kişi üzerinde yapılan bir araştırma sonucu yanıt oranının %10 olduğu görülmüştür. CHAID yöntemi sonucu oluşan karar ağacı Şekil 3.6'da verilmiştir. Bu sonuca göre yanıt oranı üzerinde birinci derecede medeni hal, ikinci ve üçüncü derecede cinsiyet ve evcil hayvan sahibi olup olmama değişkenlerinin etkili olduğu görülmüştür [42].




Şekil 3.6. CHAID algoritması için karar ağacı örneği.

CHAID yönteminin diğer yöntemlere göre üstünlükleri:

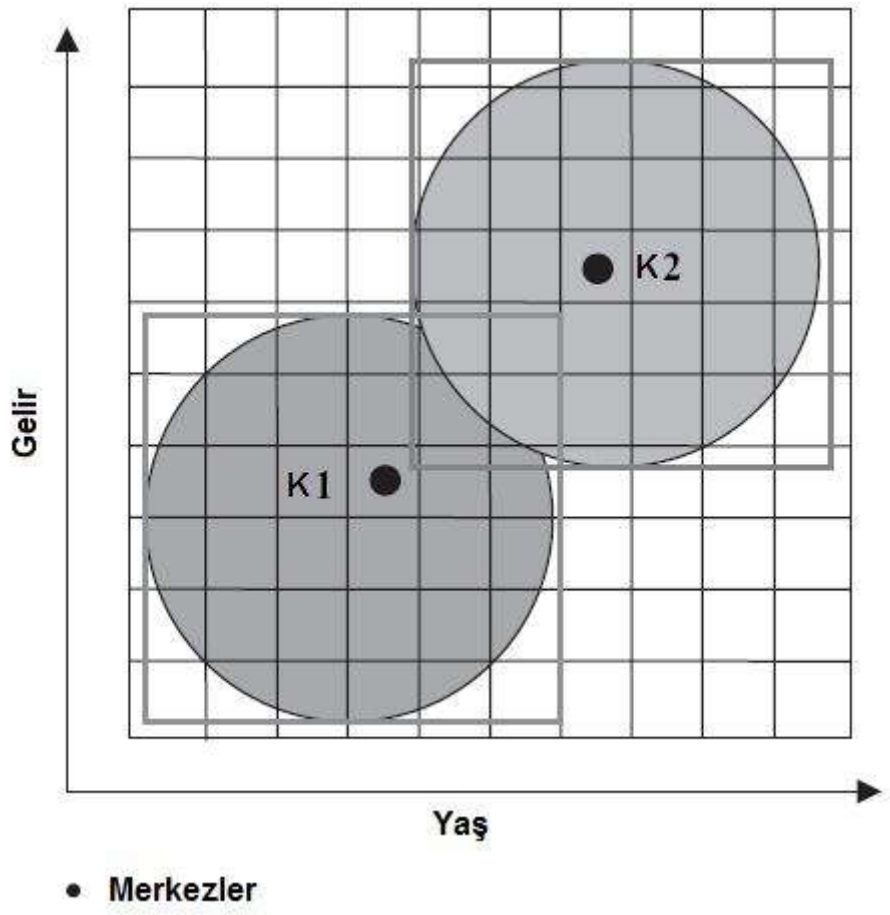
- Değişken tiplerinin nominal, sıralı ve aralıklı olması,
- Bağımsız değişkenlerin tümünün aynı düzeyde ölçülmesine gerek olmaması,
- Bağımsız değişkenlerdeki kayıp değerlerin, değişken kategori (floating category) olarak kullanılabilmesi,
- Uygun istatistiksel kriterler kullanıldığında şansa bağlı kalmadan çok etkili sonuçlara ulaşılabilmesidir.

## 3.2 Kümeleme

Kümeleme (Clustering), müşteri segmentasyonu, gen ve protein analizi, ürün gruplama gibi birbirinden çok farklı alanlarda kullanılabilen bir yöntemdir. Temelde verilerin birbirlerine olan benzerliklerine ve farklılıklarına göre gruplara ayrılma

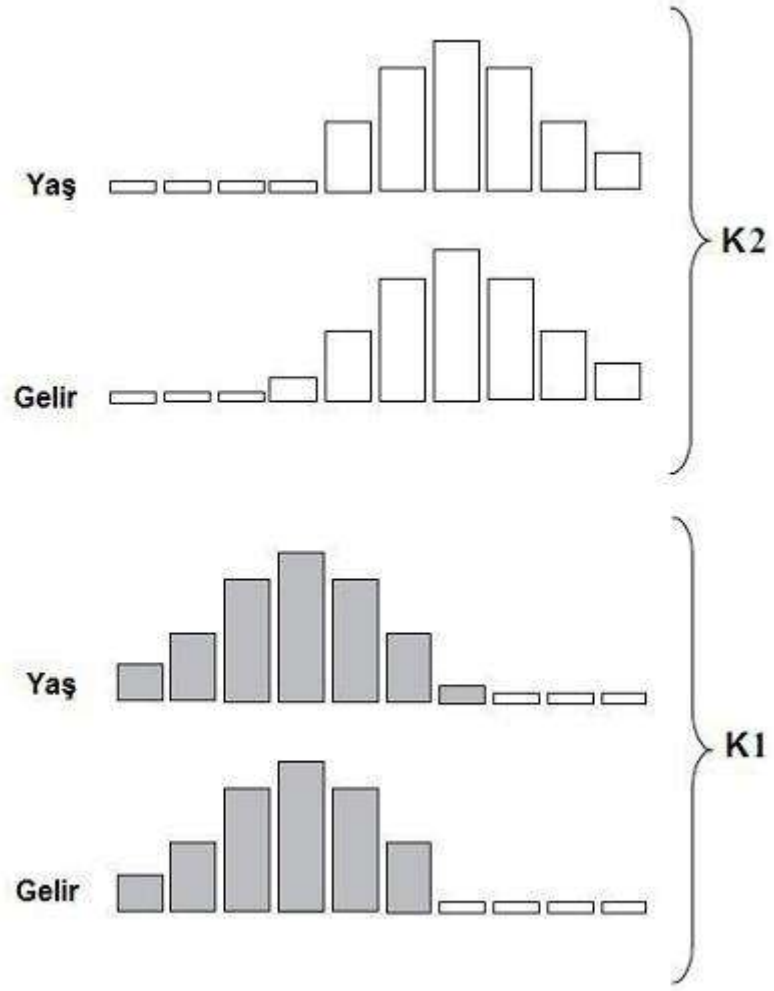
işlemidir. Kümelemenin sınıflamadan farkı, başlangıçta oluşacak küme hakkında bir fikrin olmaması ve verilerin hangi özelliğe göre kümeleneceğinin bilinmemesidir.

Aslında kümeleme analizi veri setindeki kümeleri ve grupları belirler. Şekil 3.7'de kolaylıkla görülebileceği üzere kümeler iki boyutludur. Bu örnekte iki adet nümerik nitelik bulunmaktadır: Gelir ve Yaş. K1, düşük gelire sahip ve yaşça küçük insanların yoğunlukta olduğu merkezi, K2 ise yüksek gelire sahip yaşça büyük insanların yoğunlukta olduğu merkezi temsil eder.




Şekil 3.7. Küme ağırlık merkezleri

Bu özelliklerin Şekil 3.8'deki histogram grafikleri incelendiğinde bu küme merkezlerine yakın değerlerin sayıca çok olduğu görülür.




Şekil 3.8. Küme histogramları

Bu basit örnekte kümeleri ve grupları belirlemek için veri madenciliği algoritması kullanılmasına gerek yoktur. Görsel olarak da bu iki grubun birbirinden net bir şekilde ayrıştığı görülmektedir. Üç boyutlu bir grafikte de benzer başarılı gruplamanın ve ayrışmanın yapılması mümkündür. Ancak nitelik sayısı onlarla, yüzlerle hatta binlerle ifade edildiği zaman görsel olarak bu grupları belirlemek mümkün olmayacaktır. İşte bu noktada kümeleme algoritmaları devreye girmekte ve grupları otomatik olarak belirlemektedir. Şekil 3.8'dan basit şekilde bir kural şu şekilde çıkarılabilir:

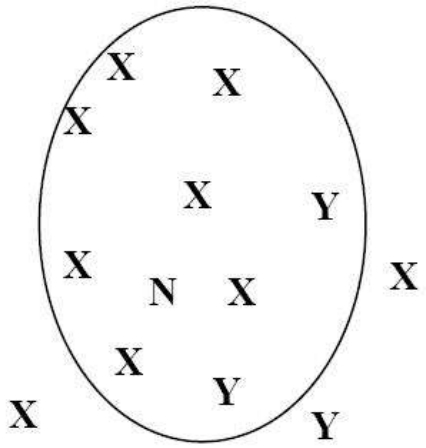
K1: 0 < gelir < 50.000 ve 0 < yaş < 35
K2: 40.000 < gelir < 100.000 ve 31 < yaş < 57

Bu iki küme dikkatlice incelendiğinde bazı değerler için kesişmeler olduğu görülmektedir. Bu durumda o değerin hangi küme merkezine daha yakın olduğu, yani mesafe ölçümü gruplamada etkili olacaktır. Yaş, gelir gibi nümerik değerlerin hangi

değerlere yakın olduğunu ölçmek kolaydır. Ancak medeni hal, renk tercihi gibi özelliklerin yakınlık ve uzaklığını bulmak nümerik değerlere göre zordur [7;20;37;45].

### 3.2.1 K- en yakın komşuluk algoritması

Mevcut eğitim verisine yeni bir gözlem veya değer eklenmek istenildiğinde eklenmek istenen gözlemin diğer gözlemlere olan yakınlığını esas alan bir algoritmadır. Bu yakınlığı ise k-NN (k nearest neighbour) algoritması diğer gözlemlere olan mesafeyi ölçerek bulmaktadır. Şekil 3.9'da N gözleminin diğer gözlemlere olan mesafesi dikkate alındığında N komşuluğunda Y den daha çok X olduğu görülür ve N, X grubuna atanır [1;27].




Şekil 3.9. N gözleminin diğer gözlemlere uzaklığı

Burada görsel olarak ifade edilen ve gözlemlenen uzaklık kavramı için bir çok hesaplama yöntemi bulunmaktadır. K-NN algoritması için çoğunlukla Öklid uzaklık bağıntısı kullanılır. Öklid uzaklık bağıntısından başka uzaklık bağıntısı olarak Manhattan uzaklığı, Minkowski uzaklığı bulunmaktadır [7;37;45].

Bu uzaklık ölçütlerinden herhangi birisi kullanılarak sınıflara atamalar yapılır. Bu yöntemin önemli dezavantajlarından birisi çok sayıda veriyi içeren durumlarda, her yeni bir gözlem için bu hesaplamayı tekrarlaması ve böylece bir hesaplama yükü getirmesidir [7;37;45].

### 3.3 Birliktelik Kuralları

Birliktelik kuralları (Association Rules) büyük veri kümeleri içerisindeki örüntü ve ilişkileri ortaya çıkarmak amacıyla yapılan veri analizine dayanır. Örneğin hangi ürün ve servisleri müşteriler birlikte satın almaya eğilimlidir, bu ürünlerin birlikte satın alınmasında bir ilişki var mıdır? Bu şekildeki ilişkileri ortaya çıkarma işlemine market-sepet analizi denir. Bu tip birlikteliklerin ve ilişkilerin ortaya çıkarılması sonucu müşterilerin alışveriş alışkanlıkları, eğilimleri ortaya çıkarılmış olur ve bu doğrultuda pazarlama stratejileri geliştirilir [17;19].

Örneğin, bir kişinin turizm ile ilgili bir kitap aldığı zaman %20 ihtimalle cep sözlüğü de alması biçiminde verilebilir.

Her bir kuralın destek ve güven şeklinde iki ölçütü vardır. Bu ölçütlerin hesaplanmasında destek sayısı adı verilen değer kullanılır. Birliktelik kuralı $(A \rightarrow B)$ biçimde gösterilir.

1.Destek: Örüntünün sıklığını öğelerin ne sıklıkta bir arada bulunduğunu ifade eder. Turizm kitabı-Cep Sözlüğü örneğinde destek %20 dir.

Destek ölçütü;

$$destek(A \rightarrow B) = \frac{sayı(A,B)}{N} \tag{3.7}$$

biçiminde ifade edilir. Eşitlik 3.7 $sayı(A,B)$ destek sayısı A ve B ürün gruplarını birlikte içeren alışveriş sayısını göstermektedir. N ise tüm alışverişlerin sayısını göstermektedir.

2.Güven: Birlikteliğin gücünü ve etki derecesini gösterir. Belirli bir öğenin diğerine ne derece bağımlı olduğunu ifade eder. Bir ürün grubunu alan müşterilerin diğer ürün grubunu da alma olasılığını ortaya koyar.

Güven ölçütü ;

$$güven(A \rightarrow B) = \frac{sayı(A,B)}{sayı(A)} \tag{3.8}$$

biçiminde ifade edilir.

Örneğin, bir mağazada 10 müşterinin bir defada yaptığı alışveriş bilgilerinden yararlanarak birliktelik kuralı şu şekilde hesaplanır:

A = { Ekmek,Peynir } , B ={ Süt }

Müşterilerin A ve B gruplarındaki ürünleri beraber satın alma sayısı 3, sadece A grubundaki ürünleri satın alma sayısı 4 olarak verilmiştir. Bu durumda Eşitlik 3.7'ye göre destek ölçütü:

$$destek(\text{Ekmek,Peynir} \rightarrow Süt) = \frac{sayı(\text{Ekmek, Peynir}, Süt)}{Müsterisayısı} = \frac{3}{10} = 0.3 = \% \ 30$$ olarak hesaplanır.

Güven ölçütü ise:

$$güven(\text{Ekmek,Peynir} \rightarrow Süt) = \frac{sayı(\text{Ekmek, Peynir}, Süt)}{sayı(\text{Ekmek, Peynir})} = \frac{3}{4} = 0.75 = \% \ 75$$ olur [29].

Birliktelik kurallarının üretilmesi için kullanılan yöntemlerden biri de IBM tarafından *APrioriAll* olarak da isimlendirilen algoritmadır. Bu algoritma temel olarak aşağıdaki adımları içerir:

Adım 1. Minimum destek değerinin üzerinde sık görülen / tekrarlayan öğeler bulunur.

Adım 2. Tekrarlayan öğeler kümesinden birliktelik kuralı uygulamak için yeteri derecede güven ölçütüne sahip değerler seçilir.

Adım 3. Tüm değerler taranır, yüzde olarak belirlenen destek düzeyinin üzerindeki değerler bulunur. Bu küme $L_1$ biçiminde ifade edilir.

Adım 4. $L_1$ kümesinden öğe çiftleri oluşturulur. Bu küme $C_2$ aday kümesidir.

Adım 5. C2 kümesi içindeki tüm değerler taranır tekrarlayan çiftler bulunur. Bu küme de L2 kümesi olarak isimlendirilir

Genel olarak 4 ve 5. adımlar için,

1. $L_{k-1}$ kümesinden k öğe oluşturulan küme $C_k$ 'dır

2. $C_k$ kümesinde tüm değerler taranarak tekrarlanan öğenin oluşturduğu küme $L_k$ kümesi biçiminde ifade edilir [7;29;37].

Algoritma taslak kod olarak Çizelge 3.8'de verilmiştir [17;37].

Çizelge 3.8. Apriori algoritması

```
L1 = {Geniş 1-nesne kümeleri}
for (k = 2; Lk-1 ≠ 0; k++) do
begin
        Ck = apriori-gen(Lk-1);      // Yeni adaylar
        for all işlemler t Є D do
           begin
                Ct = altküme(Ck, t); //t içindeki adaylar
            for all adaylar c Є Ct do
                        c.count++
            end
  end
answer = Uk Lk;
```

### 3.3.1 Apriori Algoritma Uygulaması

Bir marketten Çizelge 3.9'daki alışverişler yapılmış olsun. Bu veriler için %30 destek, %60 güven ölçütlerine göre birliktelik kurallarının arandığı varsayılsın.

Çizelge 3.9. Müşterilerin alışveriş sepetleri

| Müşteri ID | Ürün |
|---|---|
| 1 | A, B, C, D |
| 2 | A, B |
| 3 | C, E, F |
| 4 | G, A, H, J |
| 5 | G, A |
| 6 | C, D, K |
| 7 | G, A, B |

Çizelge 3.9'da tekrar eden ürünlere bakıldığında tüm alışverişlerin %30 unda yer alan ürünlerin yalnızca Çizelge 3.10'daki ürünler olduğu görülmektedir (Bkz. Bölüm 3.3, Adım 3).

Çizelge 3.10. %30'un üzerinde tekrar eden ürünler

| Ürün | Sıklık |
|---|---|
| G | 3 |
| A | 5 |
| B | 3 |

Çizelge 3.10'daki üç ürün, üç farklı kombinasyonda çiftler oluşturur:

C2 : { {G, A}, {A, B} ve {G, B} } kümesinde öğelerin sıklıkları ise ilk iki öğenin %30 destek düzeyinden fazla destek düzeyine sahip olduğudur (Bkz. Bölüm 3.3, Adım 4).

Çizelge 3.11. Çift-sıklık değerleri

| Çift | Sıklık |
|---|---|
| {G, A} | 3 |
| {A, B} | 3 |
| {G, B} | 1 |

Güven düzeyleri ise;

$güven(G \rightarrow A)$ = %100 (G'nin olduğu durumda, A'nın da olma yüzdesi)

$güven(A \rightarrow G)$ = %60

$güven(A \rightarrow B)$ = %60

$güven(B \rightarrow A)$ = %100

olarak bulunur. Bu durumda her bir çiftin başlangıçta belirlenen güven düzeyinin üzerinde olduğu için kabul edilebilir düzeydedir (Bkz. Bölüm 3.3, Adım 5).

$L_2$ kümesinde tekrarlayan çiftler Çizelge 3.12'de verilmiştir.

Çizelge 3.12. Çift-sıklık değerleri

| Çift | Sıklık |
|---|---|
| {G, A} | 3 |
| {A, B} | 3 |

Bu çiftler, her üç ürünü içeren farklı kümeler üretmek için kullanılır, örneğin $C_3$. Uygulamada ise sadece bir kümenin bu özelliğe sahip olduğu görülür: {G, A, B}. Bu kümenin sıklığı ise 1 dir ve %30 destek düzeyinin altında olduğu için geçerli değildir [7;29].

### 3.4 Hangi Veri Madenciliği Yöntemi ?

Veri Madenciliği algoritması olarak birçok farklı algoritma bulunmaktadır (Çizelge 3.13). Hangi algoritmanın kullanılacağı mevcut verinin karakteristiğine ve özelliğine bağlıdır. Tüm bu algoritmaların üstünlükleri, kısıtları ve özellikleri Çizelge3.13'te verilmiştir[7].

**Çizelge 3.13** Veri madenciliği yöntemlerinin karşılaştırılması

| Teknoloji | Üstünlükler | Kısıtlar | Ne zaman kullanılacağı |
|---|---|---|---|
| Kural Tabanlı Analiz (Birliktelik kuralları, market sepet analizi) | Anlaşılması kolay, If ... then kuralları konabilen, sürekli ve kategorik veri için etkili | Veri içinde kural tabanlı kuvvetli ilişkiler olmayabilir. Tüm kombinasyonlar gözönüne alındığında sayıca incelenmesi gerekli bir çok ilişki ortaya çıkabilir. | Bir arada gruplanacak iyi tanımlanabilir kümeler üzerinde etkili (perakende sektörü ve zaman serileri gibi) |
| Karar Ağaçları | Anlaşılması en kolay olanlardan, kompleks yapıdaki değişkenler üzerine geliştirilmiş belirli hedef değerlerde ön plana çıkıyor. | Sürekli değişken üzerindeki değerleri tahmin etme aşamasında ve zaman serileri için çok uygun değil. | Kayıtların sınıflandırılması ve çıktıların tahmin edilmesi durumlarında etkili, |
| Sinir Ağları | Çok yönlü, lineer olmayan komplike verilerde ve büyük hacimli verilerde dahi etkili, kirli ve eksik veri setlerinde kullanılabilir. | Bulunan ilişkilerin açıklamasında yetersiz, nümerik veri üzerinde çalışılabiliyor aksi takdirde dönüştürme gerekiyor. Girdi sayısı çok olduğu durumlarda örüntü bulunamıyor. | Tahmin ve sınıflandırma durumlarında etkili, çok sayıda girdi olduğu durumlarda diğer yöntemler yardımıyla (birliktelik) sadece belirli özellikler ve değişkenler üzerinde çalışılabilir. |
| Kümeleme (K-Ortalama, ...) | Uygulaması kolay, veri madenciliği sürecine iyi bir başlangıç, veri tabanı üzerinde bir önbilgi gerektirmiyor, kategorik, nümerik ve metin değerleri üzerinde etkili | Bellek ihtiyacı çok fazla, mesafe ve ağırlık değerlerini bulmak zor, K-Ortalama yöntemi birbirlerine yakın değerler için uygun değil. | Çok değişkenli ve kompleks veri setlerinde kullanılabilir. |

# 4 UYGULAMA

## 4.1 Uygulamada Kullanılan İstatistik Programı: SPSS Clementine 10.1

Uygulamada veri madenciliğinde temel programlardan biri olması, veriye erişim aşamasında açık bir çözüm olması, veri büyüklüklerinde bir kısıtlamaya gerek duyulmaması, modelleme konusunda zengin bir içeriğe sahip olması, verinin kalitesini anlama ve verinin görsel olarak incelenmesinin kolay olması nedeniyle yaygın istatistik programlarından SPSS Clementine kullanılmıştır. Veri madenciliği uygulamalarında kullanılan diğer programlar: SAS Enterprise Miner, Weka.

Veri madenciliği araçları son derece esnek oldukları için, veri madenciliği çalışmalarında kılavuz olabilecek bir veri madenciliği metodolojisi geliştirilmiştir. Bu metodolojiye “Cross-Industry Standart Process for Data Mining- CRISP-DM” adı verilmektedir. Clementine bu metodolojiye göre geliştirilmiştir [40].

## 4.2 Uygulamada Kullanılan Veri

Çalışmada bir internet servis sağlayıcı firmanın 2009 yılında ilk ve ikinci 6 aylık dönemlerde ADSL hizmetini iptal ettiren abone bilgileri kullanılmıştır.

Uygulamada, abonelerin iptal nedenleri modellenmiştir. Bu veri madenciliği sürecinde iş tanımlama fazına karşılık gelmektedir.

Elde edilen verinin veri madenciliği çalışması yapılmasına uygun hale getirilmesi veri tanımlama ve veri hazırlama fazıdır. Bunun için veri üzerinde şu çalışmalar yapılmıştır:

- Veriler bir yılın 6 aylık iki dönemini içerdiği için ‘Yıllık Dilim’ değişkeni eklenmiş, değişkene ilk 6 aylık dönem için ‘1’, ikinci 6 aylık dönem için ‘2’ değeri atanmıştır.
- İptal eden abone bilgilerinde il değişkeni yer almaktadır. Hem yorumlanması hem de anlaşılabilir olması için bu illerin ait olduğu ‘TÜİK Bölgeler’ belirlenmiştir.
  Bu bilgi yeni bir değişken olarak eklenmiştir.

- İptal nedeni yılın ilk ve ikinci diliminde farklı kategorilere ayrılmıştı. Bunlar özelliklerine göre 6 adet iptal nedeninde birleştirilmiştir.
- İptal eden abone bilgilerinde meslek kodları Sarı Sayfalar iş kollarında kullanılan tanımlamalar doğrultusunda 15 ana grup altında toplanmıştır.
- İptal eden abone paket bilgileri ticari ve bireysel ana başlıklarında limitli olup olmamalarına göre 4 grupta birleştirilmiştir.

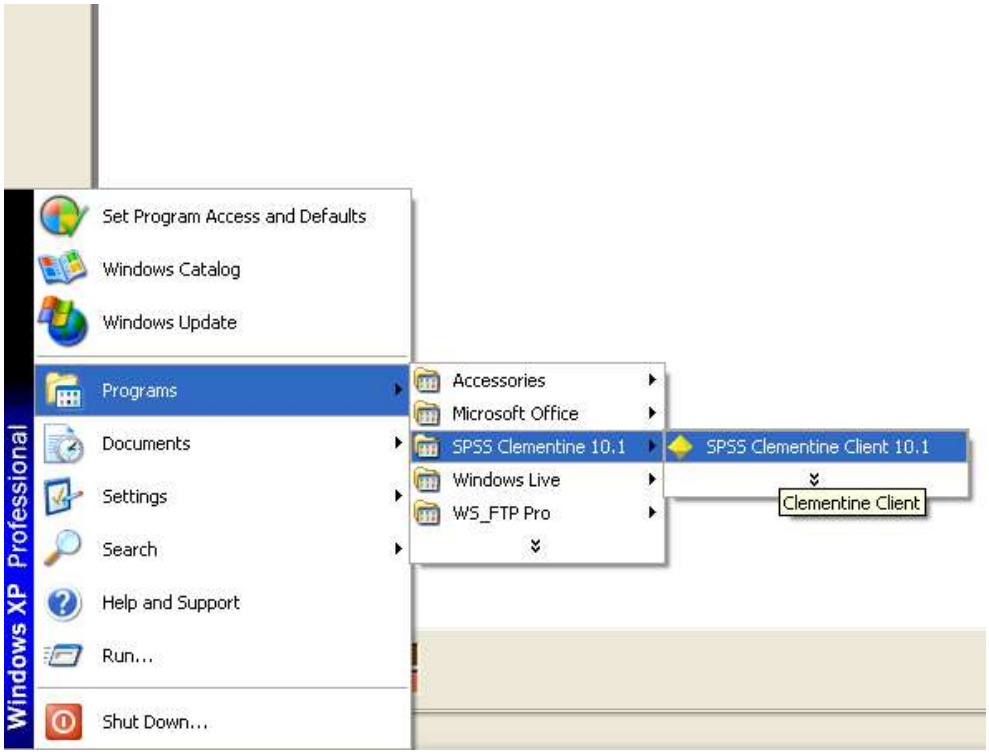
Uygulamada kullanılan değişken sayısı 7 olarak belirlenmiştir:

1. TÜİK Bölgeler : İstatistik Enstitüsü tarafından belirlenen 26 bölge içermektedir. (EK 1),
2. Cinsiyet (EK 2),
3. Yıllık dilim : İki düzey içermektedir. (EK 3),
4. Paket kodu : Abonelerin iptal ettirdikleri paket kodları 4 grupta toplanmıştır. (EK 4)
5. Meslek kodu : Sarı sayfalar tarafından belirlenen 15 meslek grubunu içermektedir. (EK 5),
6. Eğitim Durumu : Eğitim düzeylerine göre 10 adet eğitim kodu vardır. (EK 6),
7. İptal Nedeni : Abonelerin belirttiği iptal nedenleri 6 grupta toplanmıştır. (EK 6).

### 4.3 Uygulamanın Yapılışı

Uygulama için öncelikle veri madenciliği programı olan Clementine'nın çalıştırılması gerekmektedir.

Clementine default olarak local modda açılır. Başlat menüsünden Programs..SPSS Clementine 10.1... SPSS Clementine Client 10.1 tıklanır (Şekil 4.1).




Şekil 4.1. Clementine programının çalıştırılması

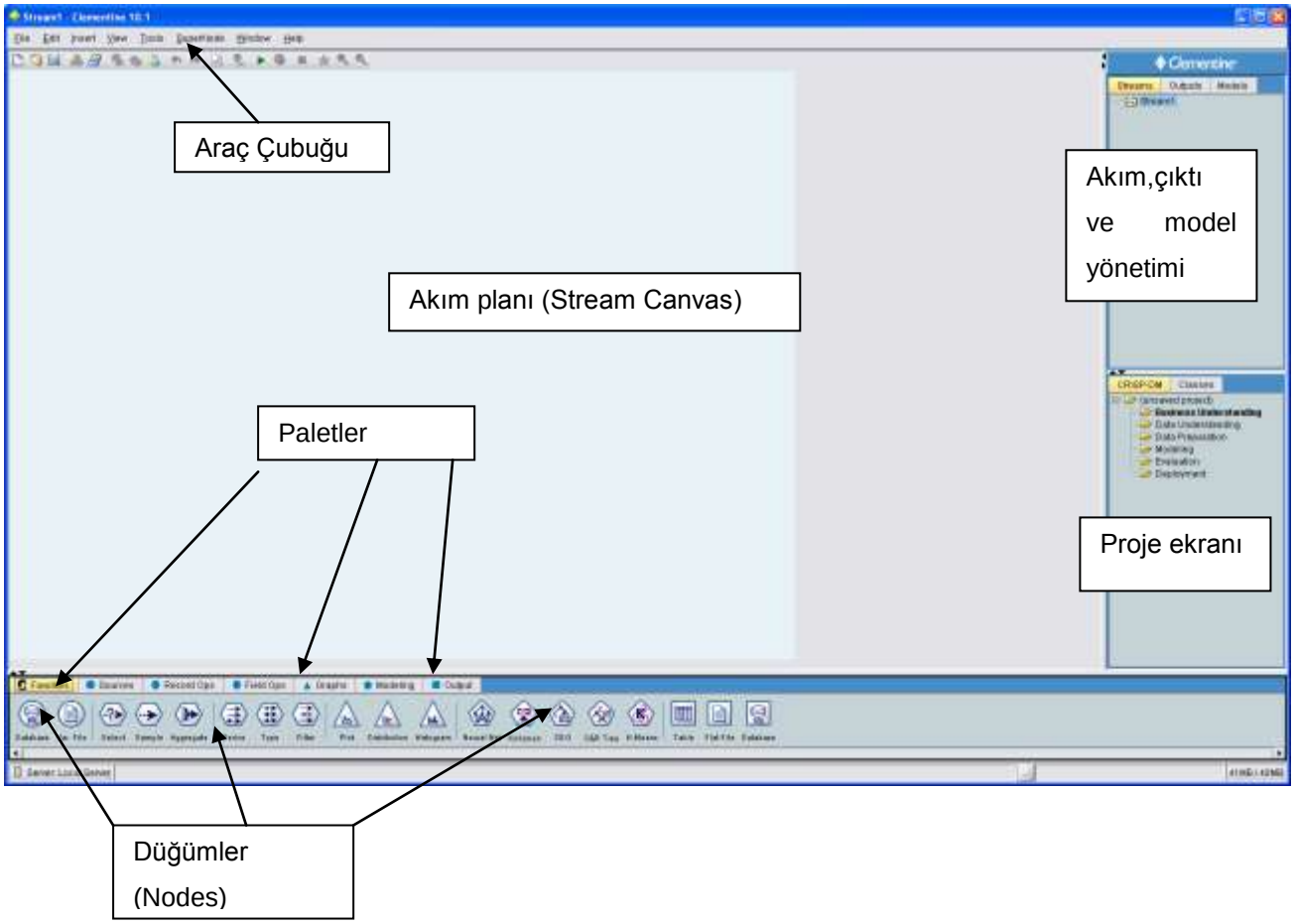
SPSS Clementine ilk açıldığında esas çalışma alanı Akım Planı (Stream Canvas) açılır. Bu alan görsel programlama tekniklerini kullanarak veri madenciliği uygulanmasını sağlar.

Nodlar veri üzerinde yapılacak işlemleri niteler. Her bir paletin içinde kendisi ile ilgili nodlar yer alır. Nodlar Akım Planına yerleştirildikten sonra birbiriyle bağlanarak akımlar oluşturulur. Akımlar nodlardan veri akışını simgeler ve her akım bir çıktı yada modelle son bulur.

Clementine penceresinin sağ üst köşesinde üç tip yönetim aracı vardır. Akımlar, çıktılar ve modeller. Akımları açmak, saklamak, adlarını değiştirmek için Akımlar Tab'ı kullanılır. Çıktıları (grafik ve tablolar) Çıktı Tabı'nda saklanır. Model Tab'ı oluşturulan modelleri saklamak için kullanılır.

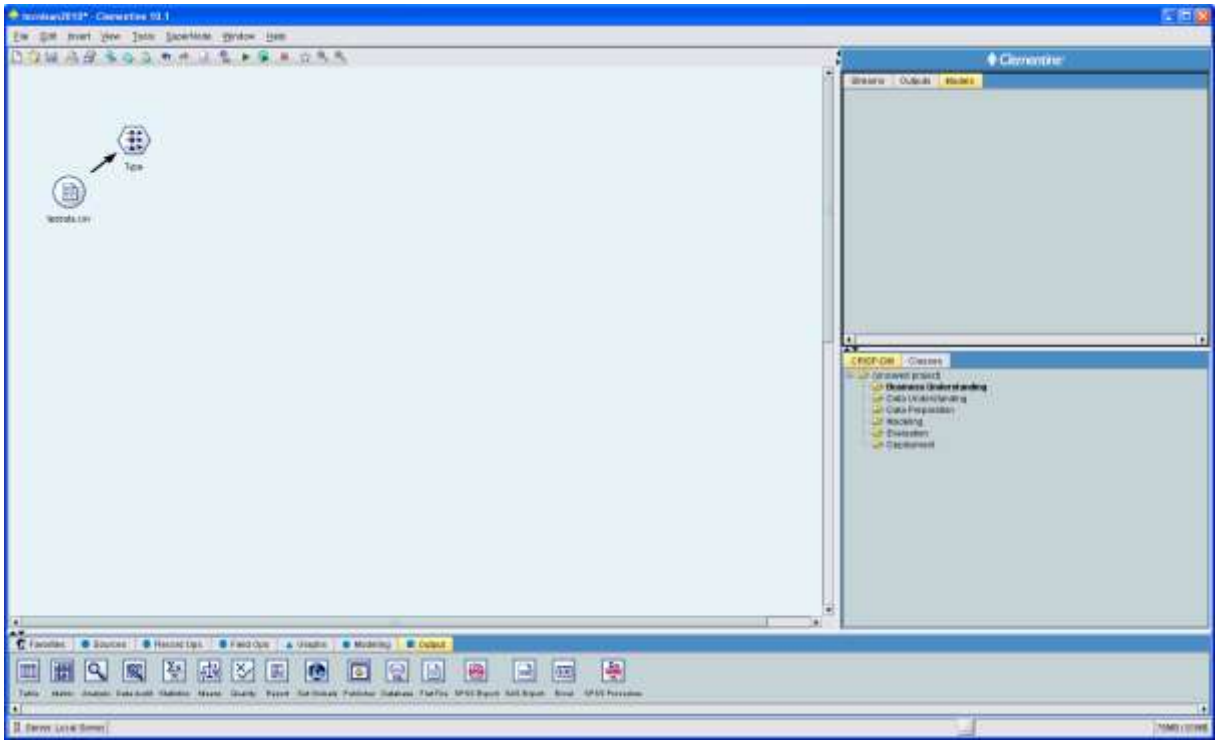
Clementine'de sağ alt köşede veri madenciliği çalışmalarının organize edilebileceği proje penceresi bulunur. Burada yer alan CRISP-DM Tab'ı akımların ve çıktıların fazlarına uygun olarak düzenlenmesini, Sınıflar (classes) Tab'ı ise nesnelerin kategorilere uygun olarak düzenlenmesini sağlar.

Sol üst kısımda yer alan araç çubuğu ekleme, silme, düzenleme ve yardım menülerini içerir [40] (Şekil 4.2).



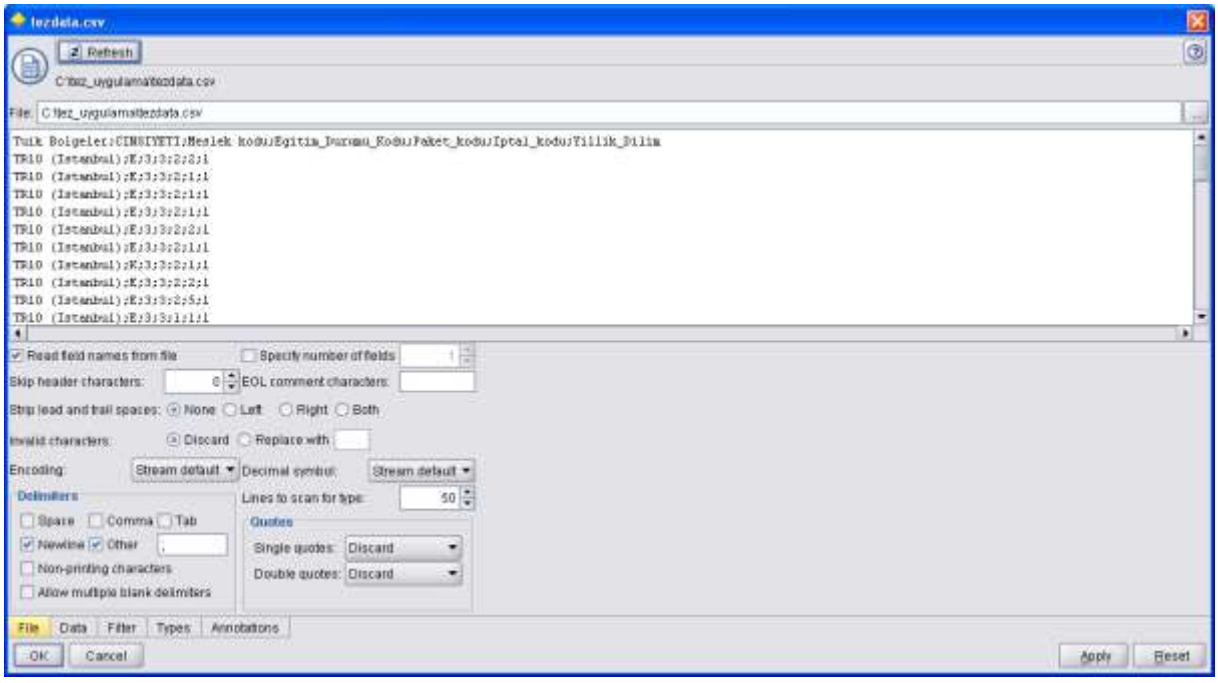

Şekil 4.2. Clementine arayüzü

Clementine üzerine veriyi aktarabilmek için excel formatında olan verinin CSV (Comma delimited) uzantılı olarak kaydedilmesi gerekir. Kaynaklar (sources) paletinden var.doc nodu seçilir ve Akım Planına eklenir. Değişken tiplerinin ayrıntılı olarak görülmesini ve değişken isim değişikliklerinin yapılabilmesini sağlayan “type” nodu eklenerek iki nod bağlanır (Şekil 4.3).



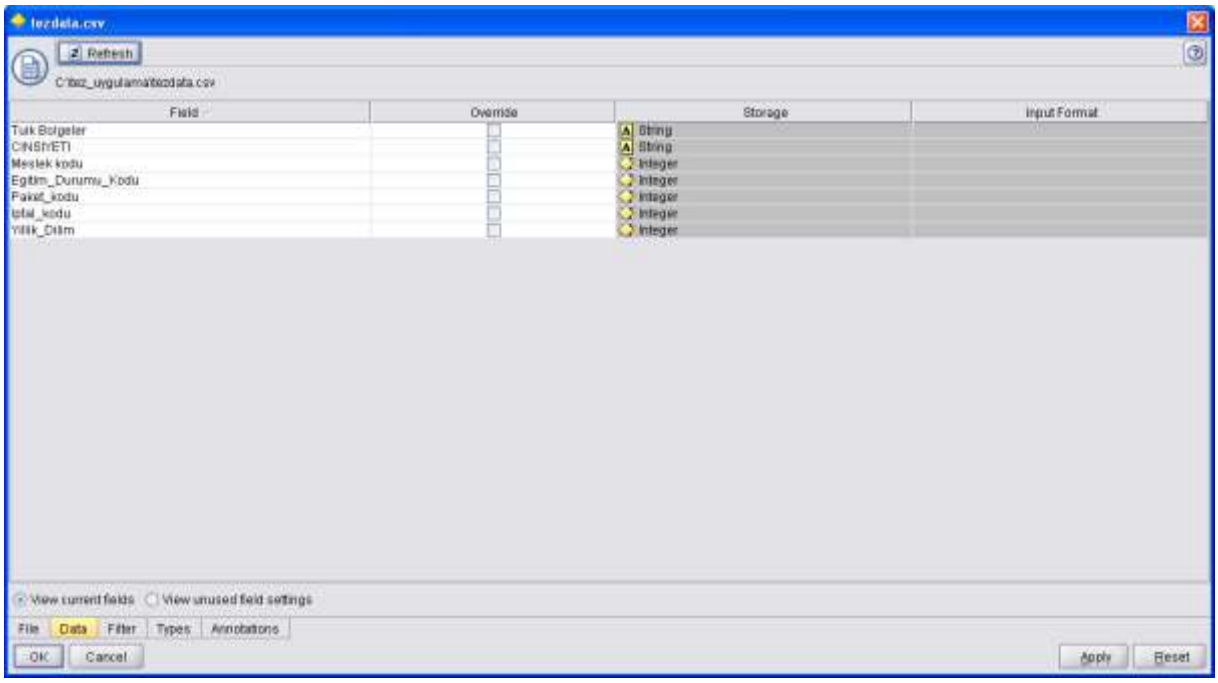
Şekil 4.3. Veri aktarım nodunun eklenmesi

Veri içeri okutma işlemi var.doc nodu içerisinde file alanından verinin yer aldığı klasör seçilerek gerçekleştirilir. (Şekil 4.4)



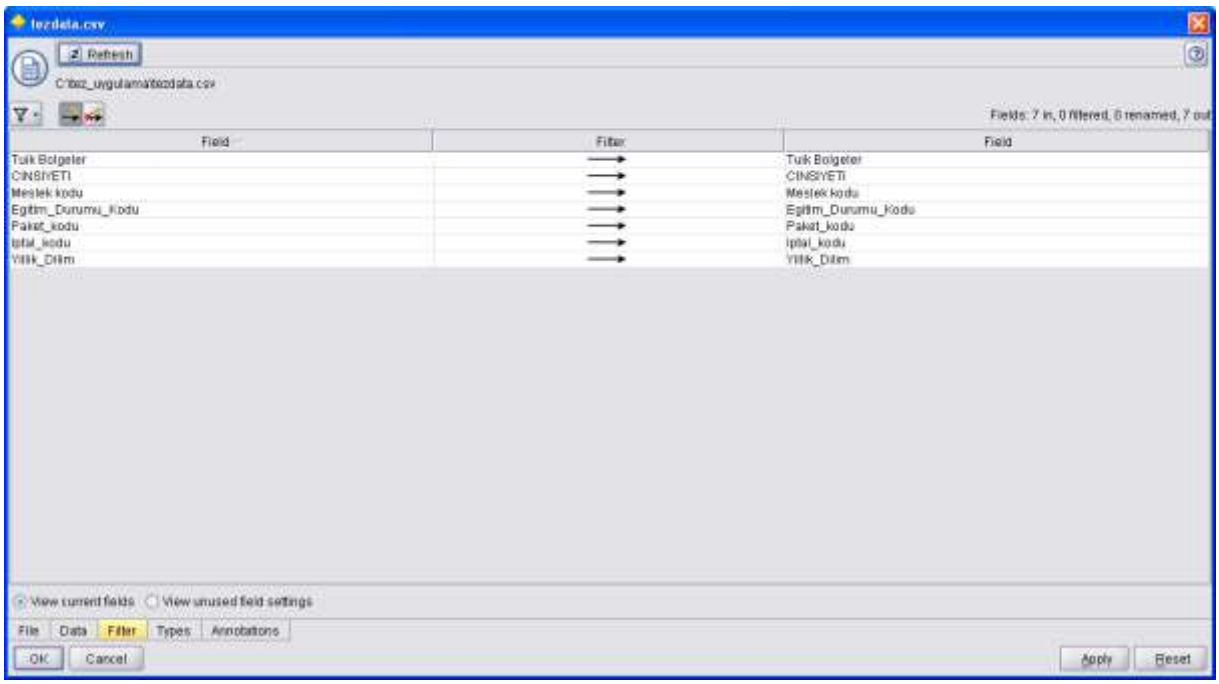

Şekil 4.4. Veri okuma işlemi

Veride yer alan değişkenlerin özellikleri “data” bölümünde görülebilir. (Şekil 4.5)



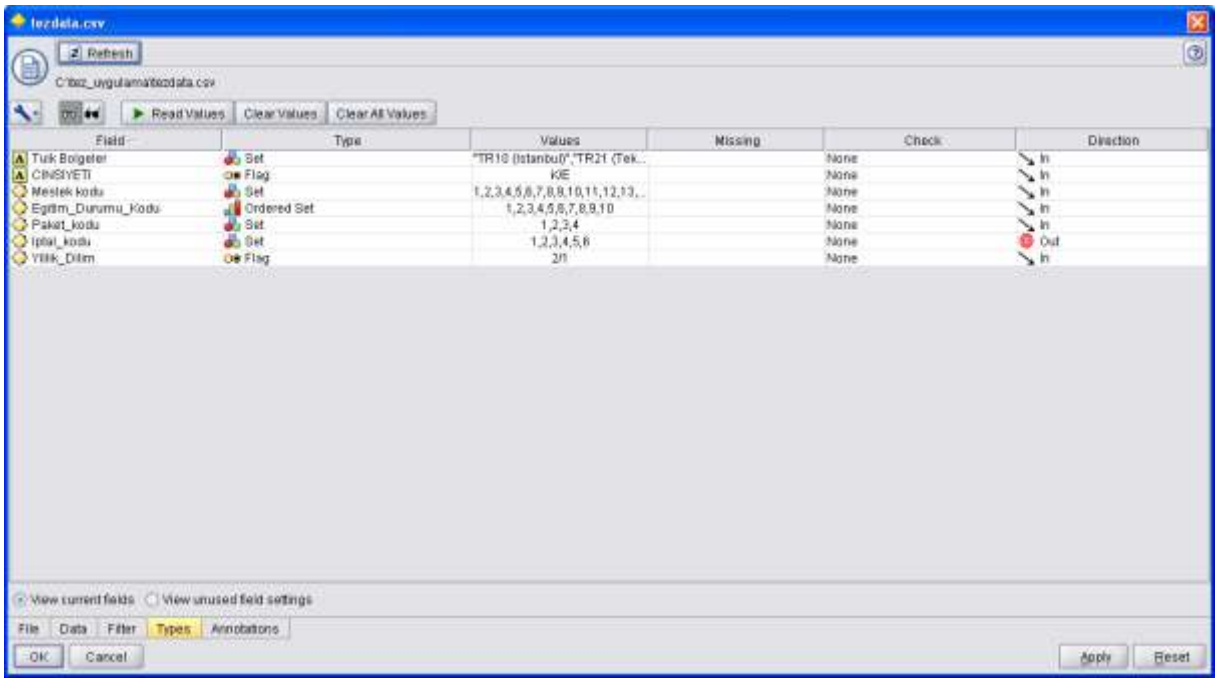

Şekil 4.5. Değişkenlerin özellikleri

Eğer bazı alan adları değiştirilmek yada analize dahil edilmemek isteniyorsa “filter” tabı’nda bu işlem gerçekleştirilebilir. Çalışmada tüm değişkenler kullanılacağı için bu işleme gerek duyulmamıştır. (Şekil 4.6)



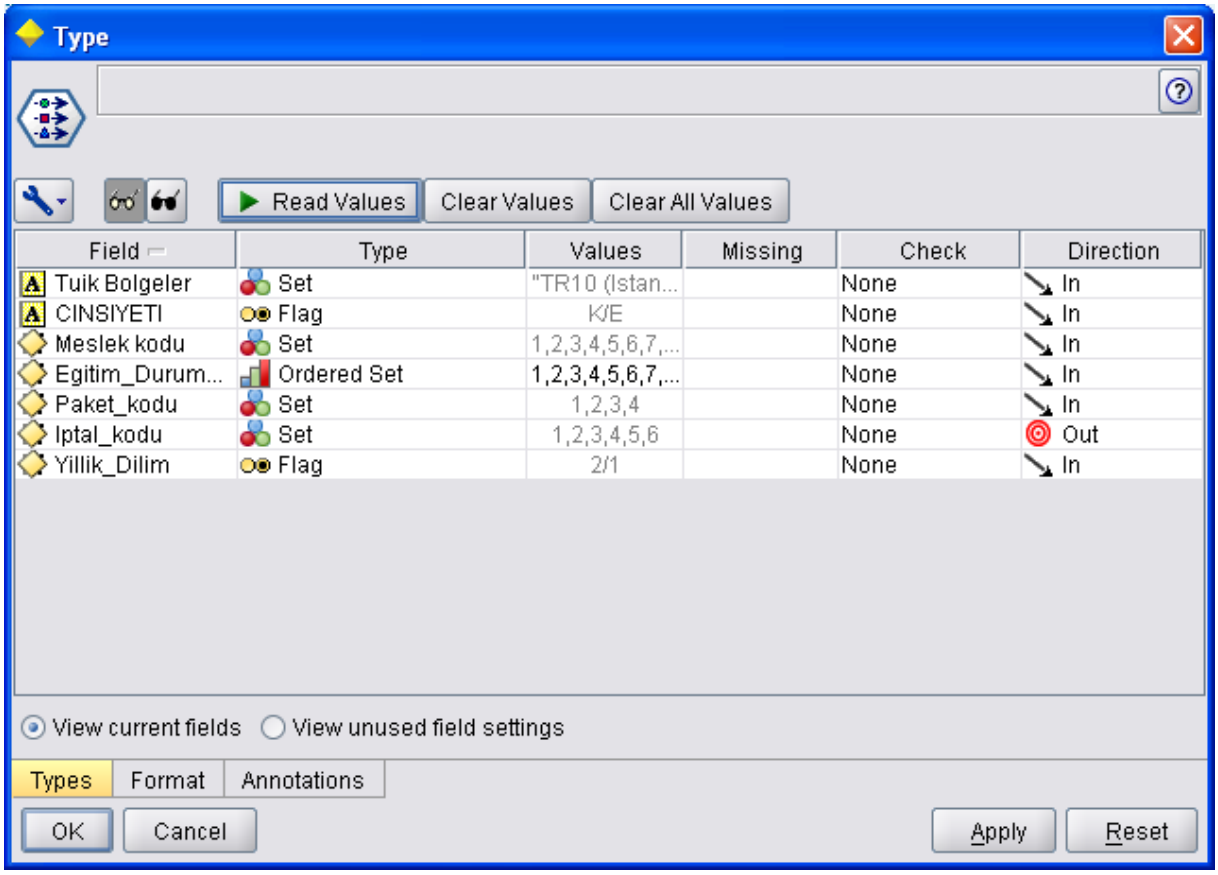

Şekil 4.6. Analize dahil edilecek veya edilmeyecek değişkenlerin seçimi

Verinin doğru bir şekilde analiz edilebilmesi için “types” tabı’nda alanların özellikleri belirtilir. Evet/hayır gibi iki değer alabilen değişkenlerde “flag” tipini kullanılır. Cinsiyet ve yıllık dilim değişkenleri bu şekilde iki değer aldığı için “flag” olarak tanımlanmıştır. Bir doğal düzenleme içinde birden çok değer alabilen maaş skalası, öğrenim durumu gibi verileri tanımlamak için “order (sıralı) set” tipi kullanılır. Eğitim durumu bu yapıya uygun olduğu için tipi “order set” seçilmiştir. Birden çok farklı değerde veri içeren değişken tanımlamalarında “set” tipi kullanılır. Diğer değişkenlerimiz için “set” tipi seçilmiştir. Açıklanmak istenen abonelerin iptal nedenleri olduğu için iptal nedenleri çıktı, diğer değişkenler girdi olarak seçilmiştir. (Şekil 4.7)



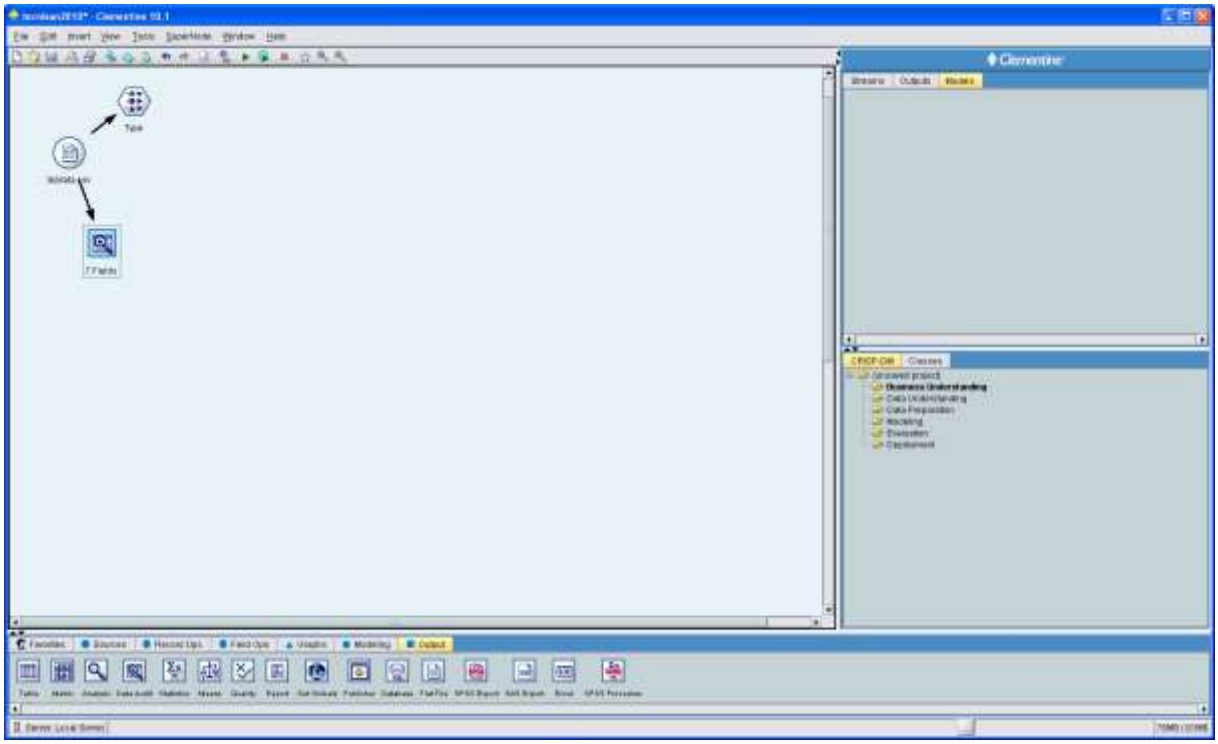

Şekil 4.7. Değişken tiplerinin belirlenmesi

Öte yandan, değişken tiplerinin seçilmesi, format yapılarında değişiklik yapılması type nodu ile de gerçekleştirilebilir. (Şekil 4.8)



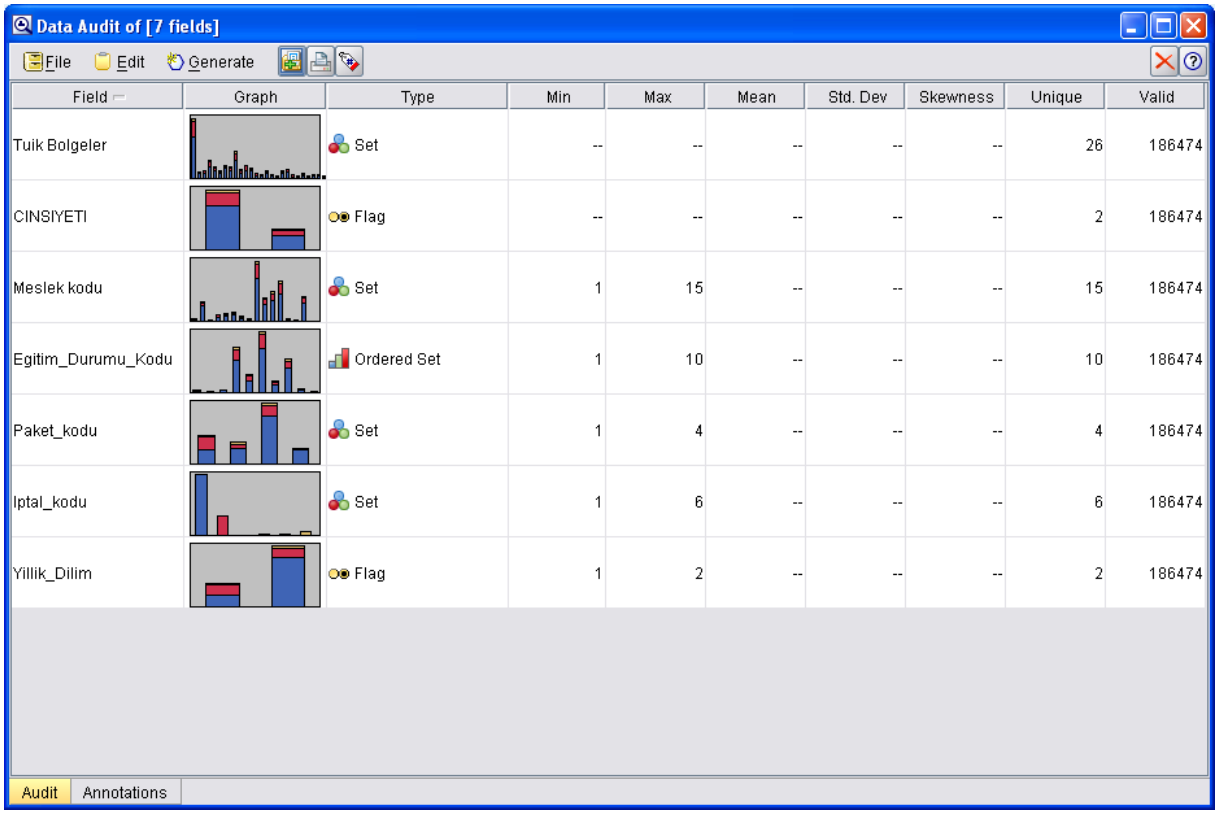

Şekil 4.8. Değişken tiplerinin type nodunda belirlenmesi

Verinin genel yapısı hakkında bilgi sahibi olabilmek için var.doc noduna “data audit “ nodunu eklenir. Data audit nodunda her satır bir alana karşılık gelmektedir. Her sütun da grafik tip bilgisi ve istatistiksel özellikler içermektedir. (Şekil  4.9)



Şekil 4.9. Veri noduna data audit nodunun eklenmesi

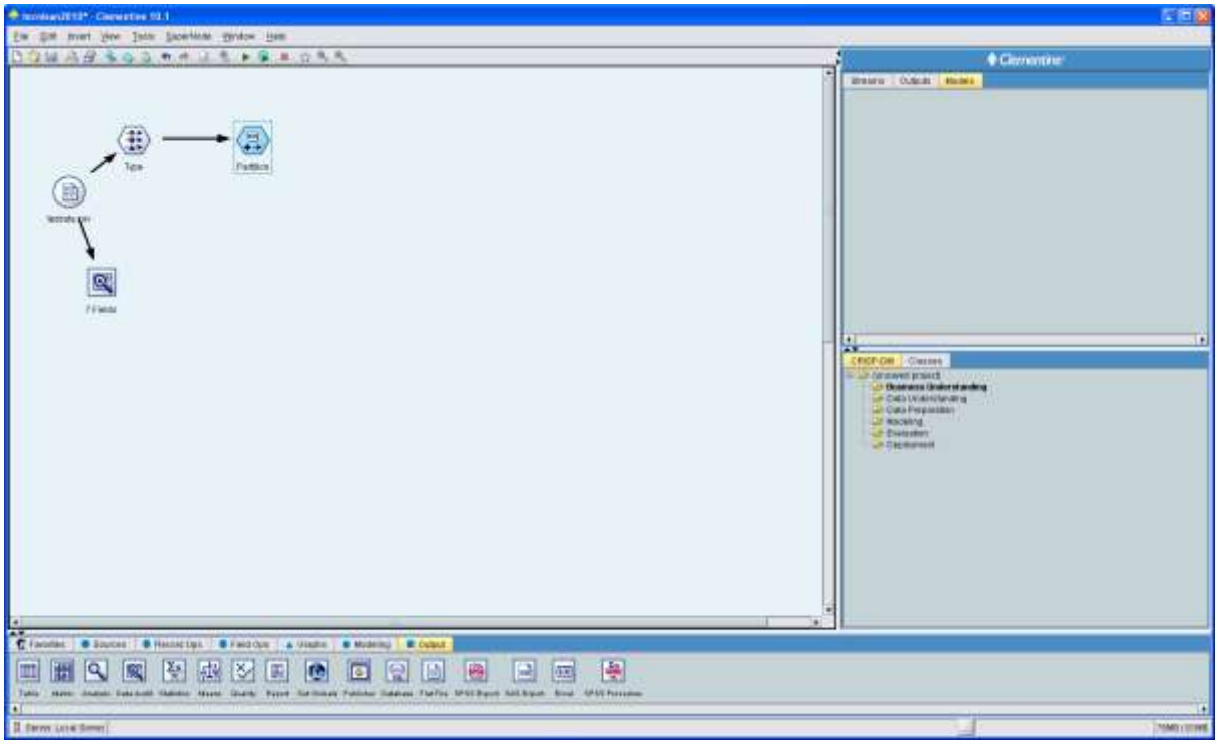
Verinin genel yapısı 7 değişkenden oluşmaktadır. TÜİK Bölgeler 26 adet bölge bilgisi içermektedir. Yıllık dilim ilk ve ikinci altı aylık döneme ait olmasına göre 2 çeşittir. Meslek kodu 15 meslek grubunu içermektedir. Paket bilgisi 4 çeşit paket ürünü kapsamaktadır. Eğitim durumu 10 düzeylidir. İptal kodu ise abonelerin hizmeti iptal ederken belirttikleri 6 nedeni içermektedir. (Şekil 4.10)



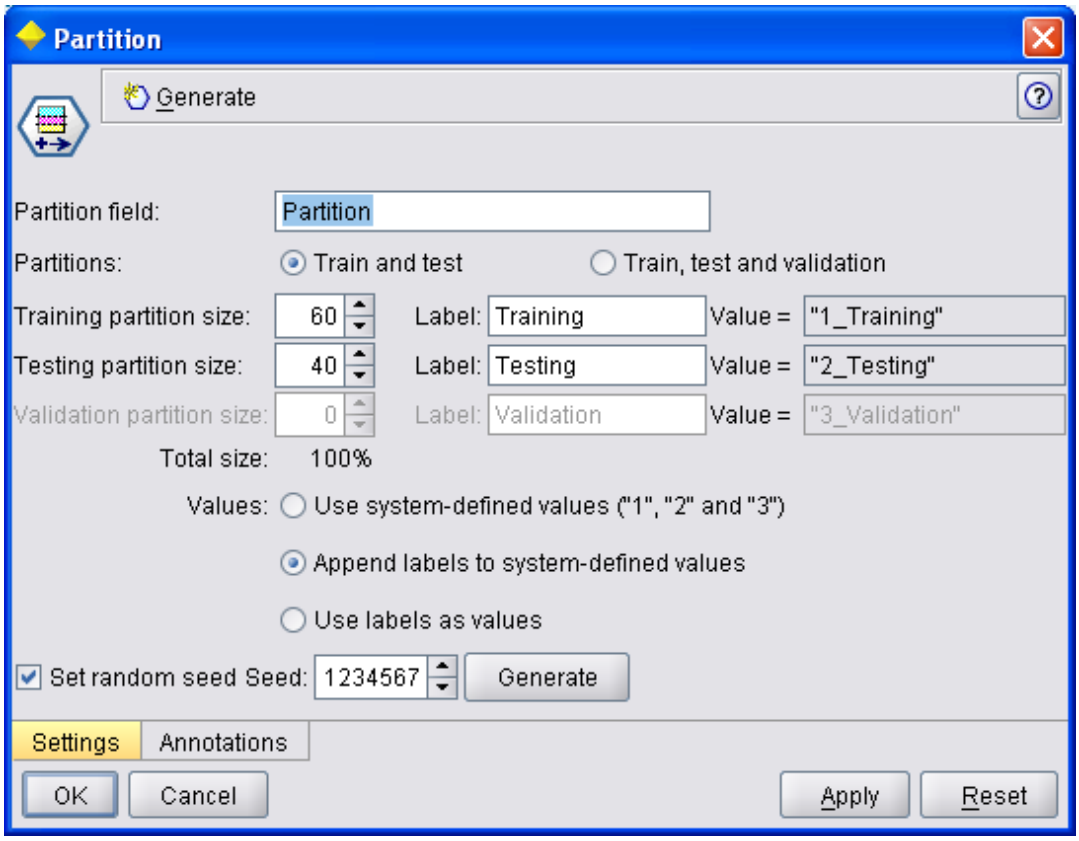

Şekil 4.10. Verinin genel yapısı

Model oluşturmadan önce verinin bölme işlemi “partition nodu” ile gerçekleştirilir. (Şekil 4.11)

Veri bölme işleminin yapılması modelin veriyi öğrenmemesini sağlar. Veri 2 yada 3 parçaya bölünebilir. Çalışmada veri “training” ve “testing “olarak iki parçaya bölünmüştür. İlk bölümde model çalıştırılırken diğer bölümde kontrol edilmesi sağlanır. Bu yeni bir veri seti üzerinde modelin başarılı olmasını sağlar. Böylece veri ezberlenmemiş olur. (Şekil 4.12)



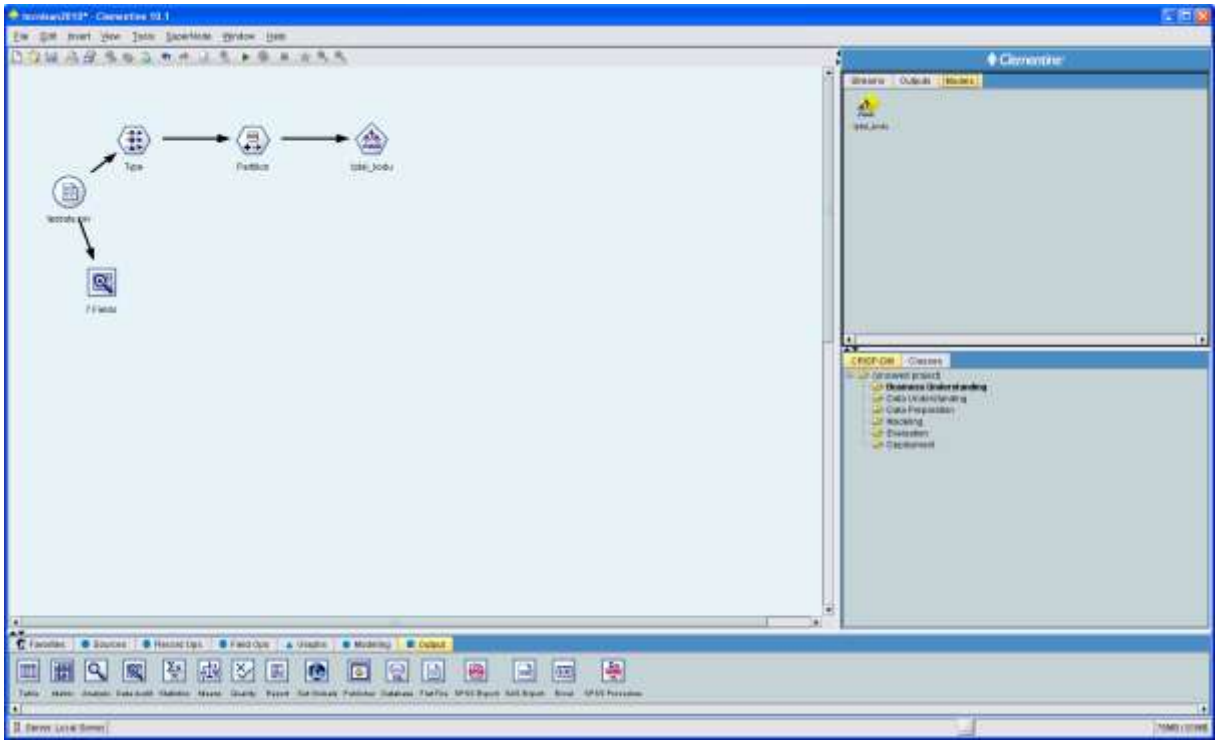
Şekil 4.11. Data bölme işlemi (partition) eklenmesi




Şekil 4.12. Data bölme işlemi (partition)

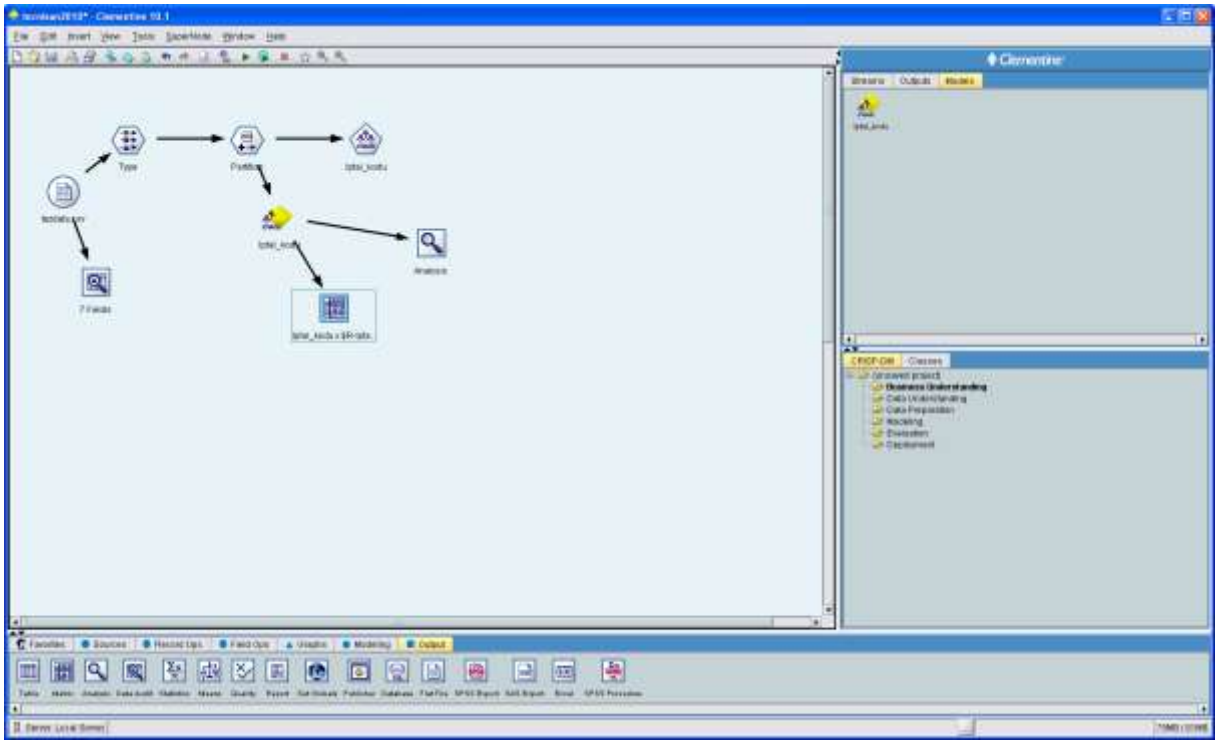
### 4.3.1 CHAID Algoritması

Chaid algoritmasını çalıştırmak için “partition nodu”’na chaid modeli bağlanır ve çalıştırılır. (Şekil 4.13)



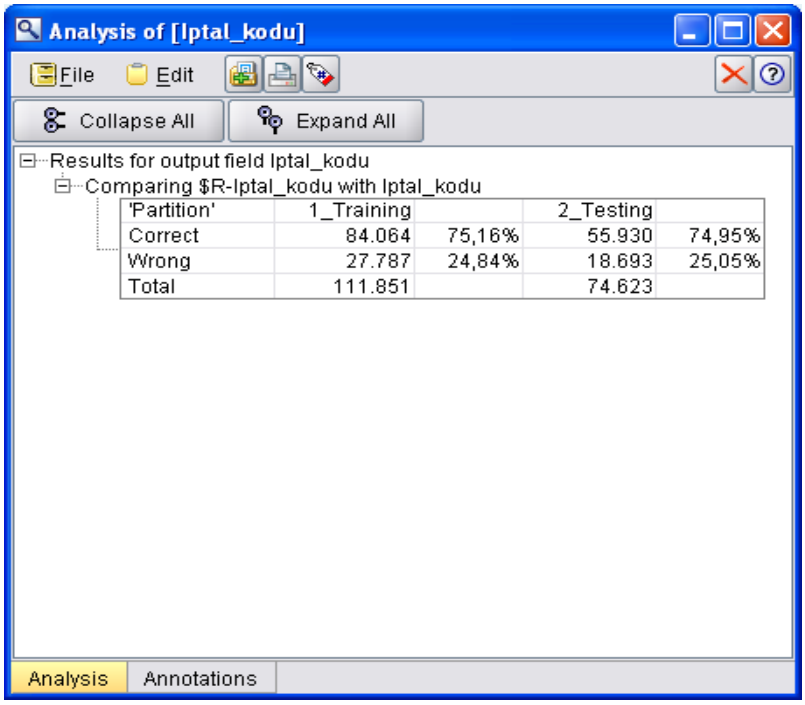
Şekil 4.13. CHAID nodunun eklenmesi

Chaid modeline ait sonuçlar analiz ve matris nodları eklenerek incelenir. (Şekil 4.14)



Şekil 4.14. CHAID modeline analiz ve matris nodlarının eklenmesi

Analiz sonucuna göre modelin test ve training sonuçlarının birbirine yakın çıktığı ve %75 doğru skorlama yaptığı görülmektedir. (Şekil 4.15)




Şekil 4.15. CHAID analiz sonucu

Model skorlama sonuçları ile gerçek değerler karşılaştırıldığında iptal nedeni 1 (Bireysel Tercih) olan verilerin model sonucunda iptal nedeni 1 tahmin edilme oranı %66 (tüm veri dikkate alındığında), iptal nedeni 2 (Ekonomik Nedenler) olan verilerin model sonucunda iptal nedeni 2 tahmin edilme oranı %9'dur. Model sonucu tahminler sadece iptal nedeni 1 ve iptal nedeni 2'yi içermektedir. Bunun nedeni verinin bu iki neden üzerine dominant olmasından kaynaklanmaktadır. (Şekil 4.16)

Matrix of Iptal_kodu by $R-Iptal_kodu #3

File Edit Generate

$R-Iptal_kodu

| Iptal_kodu | | 1 | 2 |
|---|---|---|---|
| 1 | Count | 123257 | 11801 |
| | Row % | 91.262 | 8.738 |
| | Column % | 78.409 | 40.309 |
| | Total % | 66.099 | 6.328 |
| 2 | Count | 27822 | 16737 |
| | Row % | 62.439 | 37.561 |
| | Column % | 17.699 | 57.170 |
| | Total % | 14.920 | 8.976 |
| 3 | Count | 30 | 73 |
| | Row % | 29.126 | 70.874 |
| | Column % | 0.019 | 0.249 |
| | Total % | 0.016 | 0.039 |
| 4 | Count | 97 | 141 |
| | Row % | 40.756 | 59.244 |
| | Column % | 0.062 | 0.482 |
| | Total % | 0.052 | 0.076 |
| 5 | Count | 1029 | 170 |
| | Row % | 85.822 | 14.178 |
| | Column % | 0.655 | 0.581 |
| | Total % | 0.552 | 0.091 |
| 6 | Count | 4963 | 354 |
| | Row % | 93.342 | 6.658 |
| | Column % | 3.157 | 1.209 |
| | Total % | 2.661 | 0.190 |

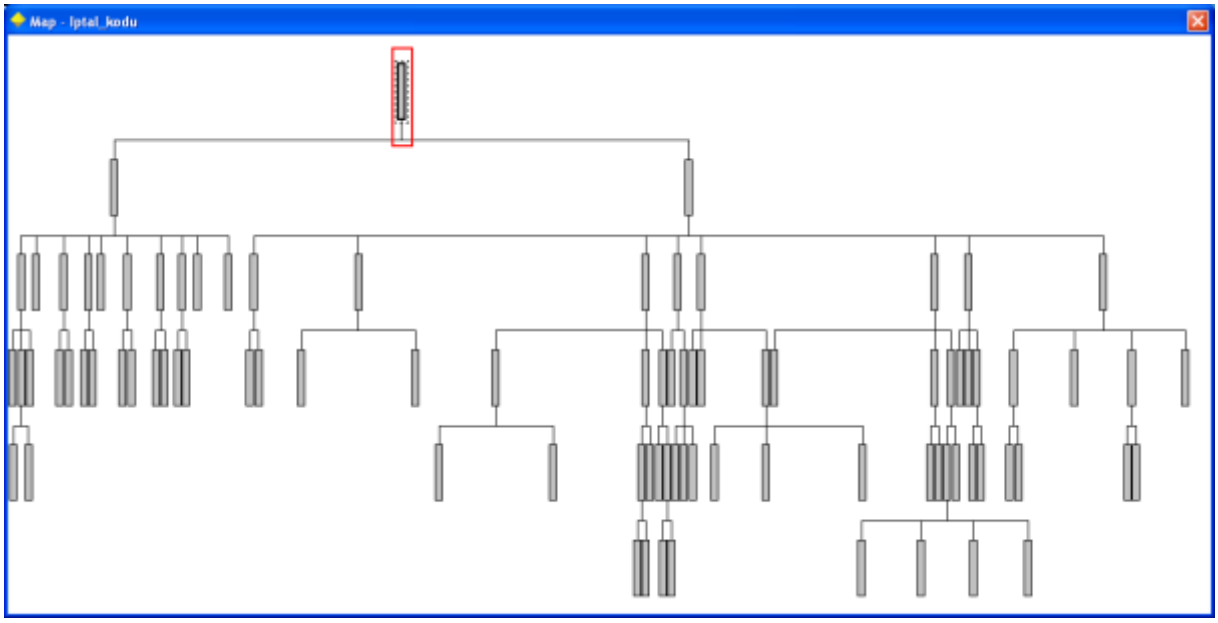
Cells contain: cross-tabulation of fields (including missing values)

Chi-square = 21.945,492, df = 5, probability = 0

Matrix Appearance Annotations

Şekil 4.16. CHAID karşılaştırma matrisi

Chaid modeli sonucunda oluşan karar ağacının genel yapısı Şekil 4.17'de görülmektedir.

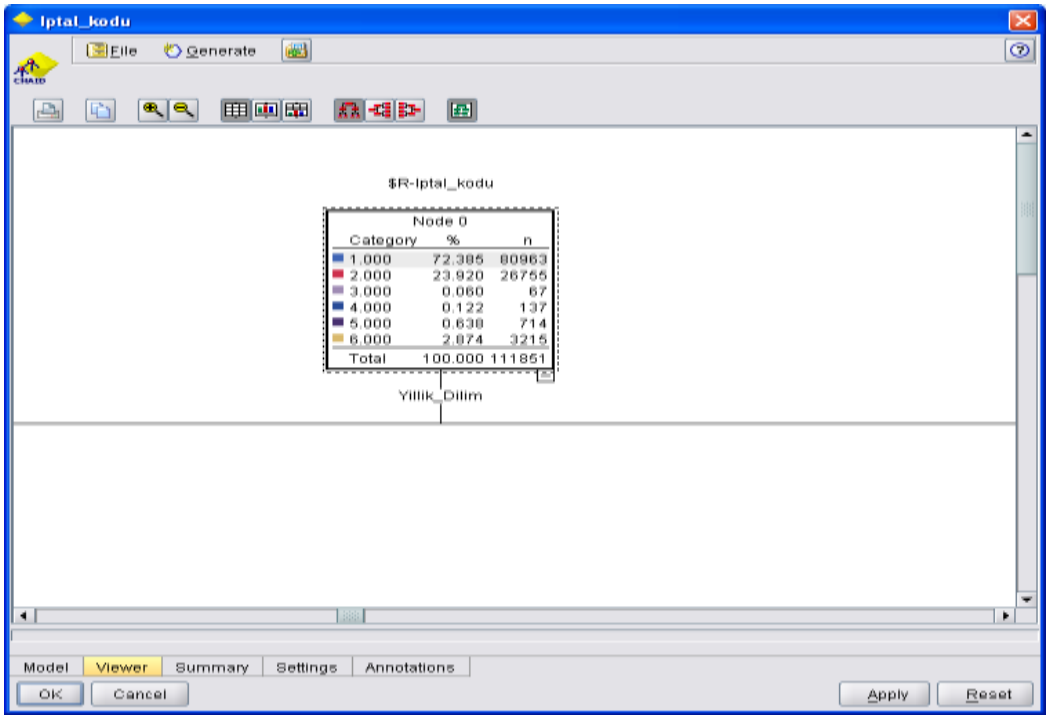



Şekil 4.17. CHAID karar ağacı genel yapısı

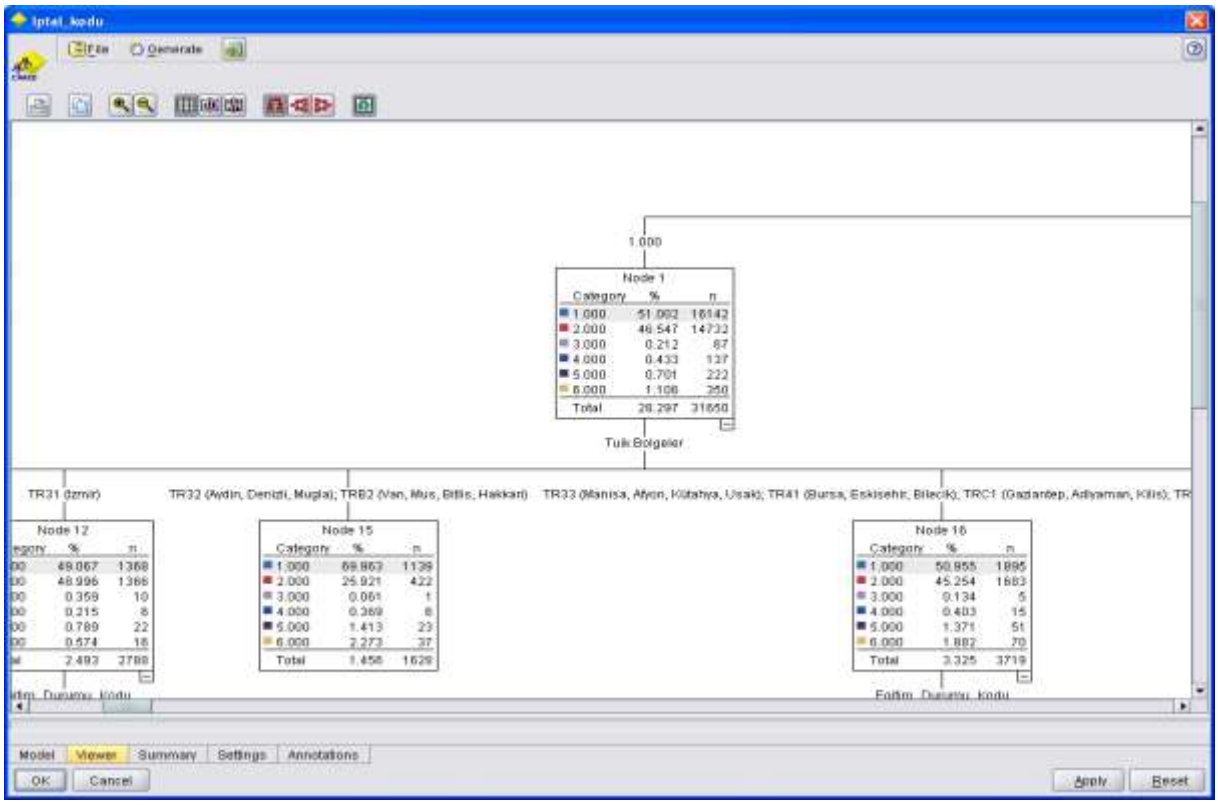
Karar ağacı incelendiğinde başlangıç noktası olan nod 0'da tüm veri dikkate alındığında iptal nedeni 1 olanların oranının %72, iptal nedeni 2 olanların oranının %23,92, iptal nedeni 6 (Telefon İptali Sebebiyle) olanların oranının % 2,84 olduğu görülmektedir. Diğer nedenlerden dolayı iptal edenlerin oranı %0,8'dir. (Şekil 4.18)

Karar ağacında ilk ayırım yıllık dilime göredir. Yılın ilk yarısını içeren nod 1'de iptal nedeni 1 oranının %51, iptal nedeni 2 oranının %46,54 olarak değiştiği görülmektedir. İptal nedeni 6 oranı ise %1 'e düşmüştür. Yani veri bütün olarak düşünüldüğünde aboneler yoğun olarak bireysel tercih nedeniyle aboneliklerini iptal ettirirken, yılın ilk 6 aylık diliminde bireysel tercih ve ekonomik nedenlerden iptal birbirine yakın oranlardadır. (Şekil 4.19)

Karar ağacında yılın ilk yarısını içeren bölümde ağaç yapısı TÜİK bölgeler – meslek kodu- eğitim durumu olarak dallara ayrıldığı görülmektedir.




Şekil 4.18. Nod 0 (Başlangıç düğümü)

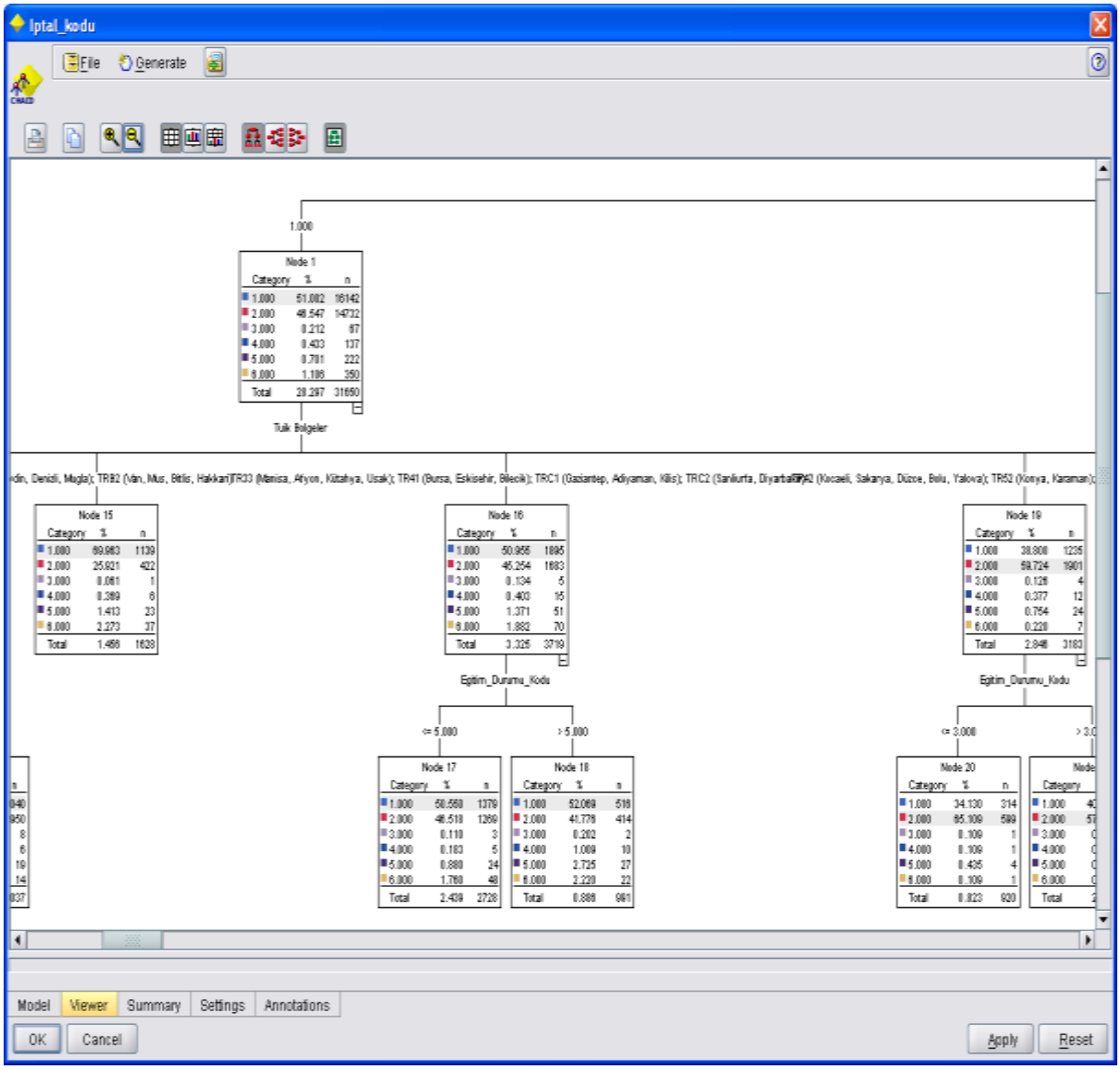



Şekil 4.19. Nod 1

İlk 6 aylık dilimi içeren nod 1'de bireysel nedenler iptal oranı %51, ekonomik nedenler iptal oranı %46 olduğu halde bu oranın alt düzeylerde değiştiği gözlenmiştir.

Nod 15 için: Bireysel nedenlerden iptal oranının %69 'a yükseldiği gözlenmiştir. Bu nod Van, Muş, Bitlis, Uşak, Bilecik gibi illeri içermektedir. (Şekil 4.20)

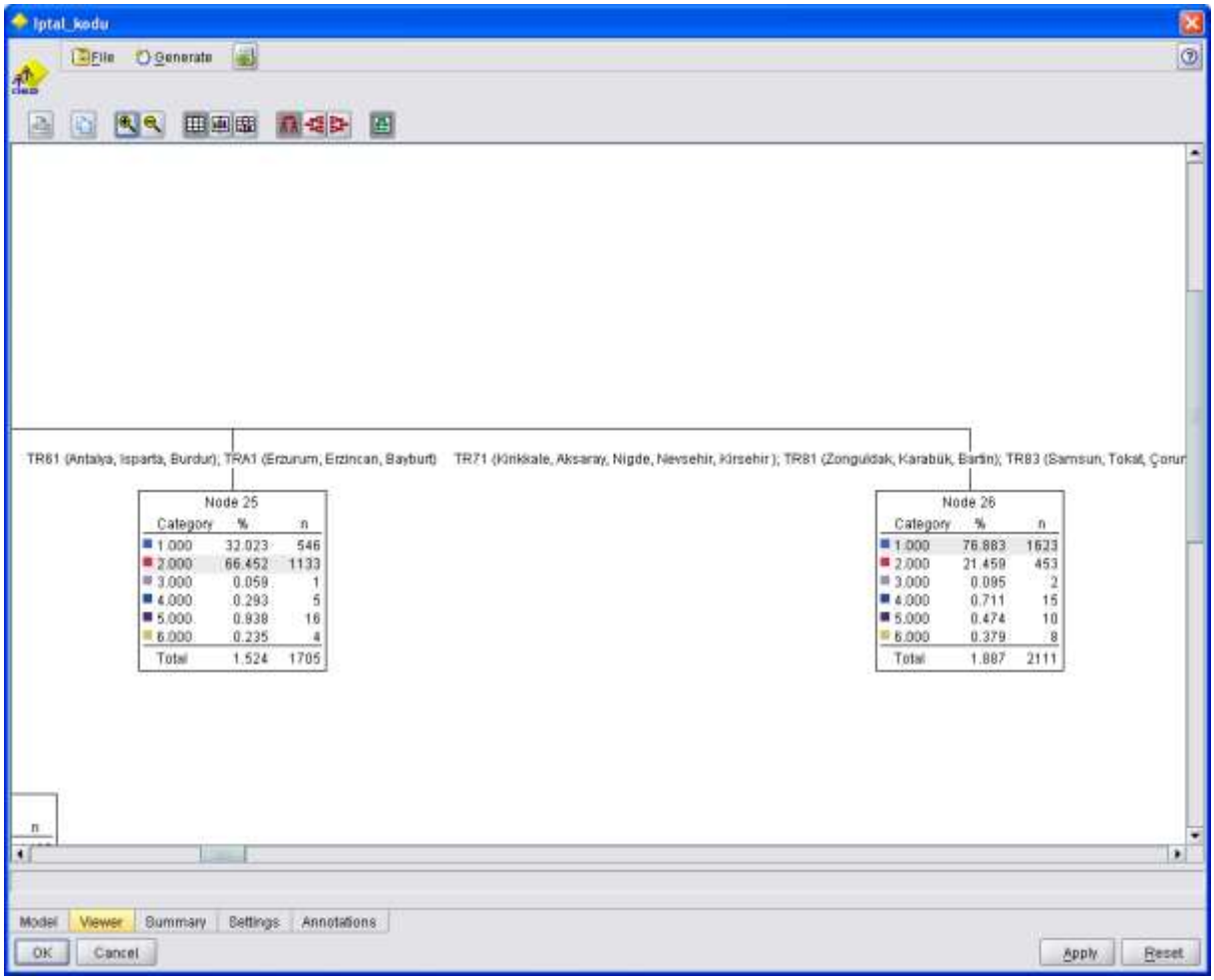
Nod 19 için: Ekonomik nedenlerden iptal oranının %59'a hatta bir alt düzeyde yer alan nod 20'de eğitim durumu ilkokul ve daha alt düzey olanlarda bu oranın % 65 'e yükseldiği görülmektedir. Eğitim seviyesi azaldıkça ekonomik nedenlerle iptal oranı artmaktadır. (Şekil 4.20)




Şekil 4.20. Nod 1 altında yer alan nod 15 ve nod 19

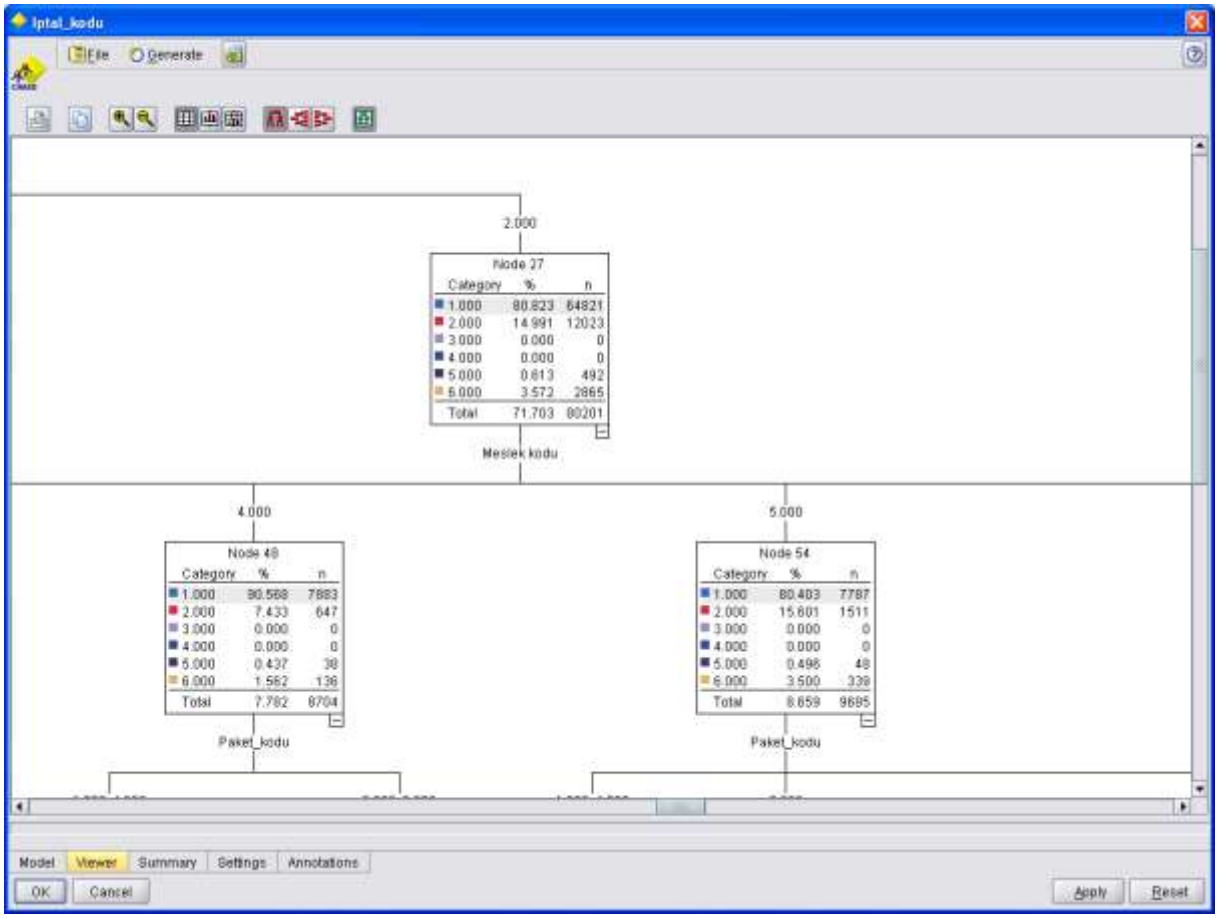
Nod 25 için: Ekonomik nedenler iptal oranı %66 ya yükselmiştir. Bu nod Doğu Anadolu ve Akdeniz Bölgesi'ndeki illeri içermektedir. (Şekil 4.21)

Nod 26 için: Karadeniz Bölgesi'ndeki illeri içermektedir ve bu nodda bireysel nedenlerden iptal oranının %76 ya yükseldiği görülmektedir. (Şekil 4.21)




Şekil 4.21. Nod 1 altında yer alan nod 25 ve nod 26

Karar ağacında yılın ikinci yarısını içeren nod 27'de iptal nedeni 1 oranının %80,82, iptal nedeni 2 oranının %14,99 olarak değiştiği görülmektedir. İptal nedeni 6 oranı ise %3,57 'e yükselmiştir. Veri bütün olarak düşüldüğünde aboneler yoğun olarak bireysel tercih nedeniyle aboneliklerini iptal ettirirken, yılın ikinci 6 aylık diliminde bireysel tercih nedeniyle iptal oranı daha da artmış ekonomik nedenlerden iptal oranı azalmıştır. (Şekil 4.22)



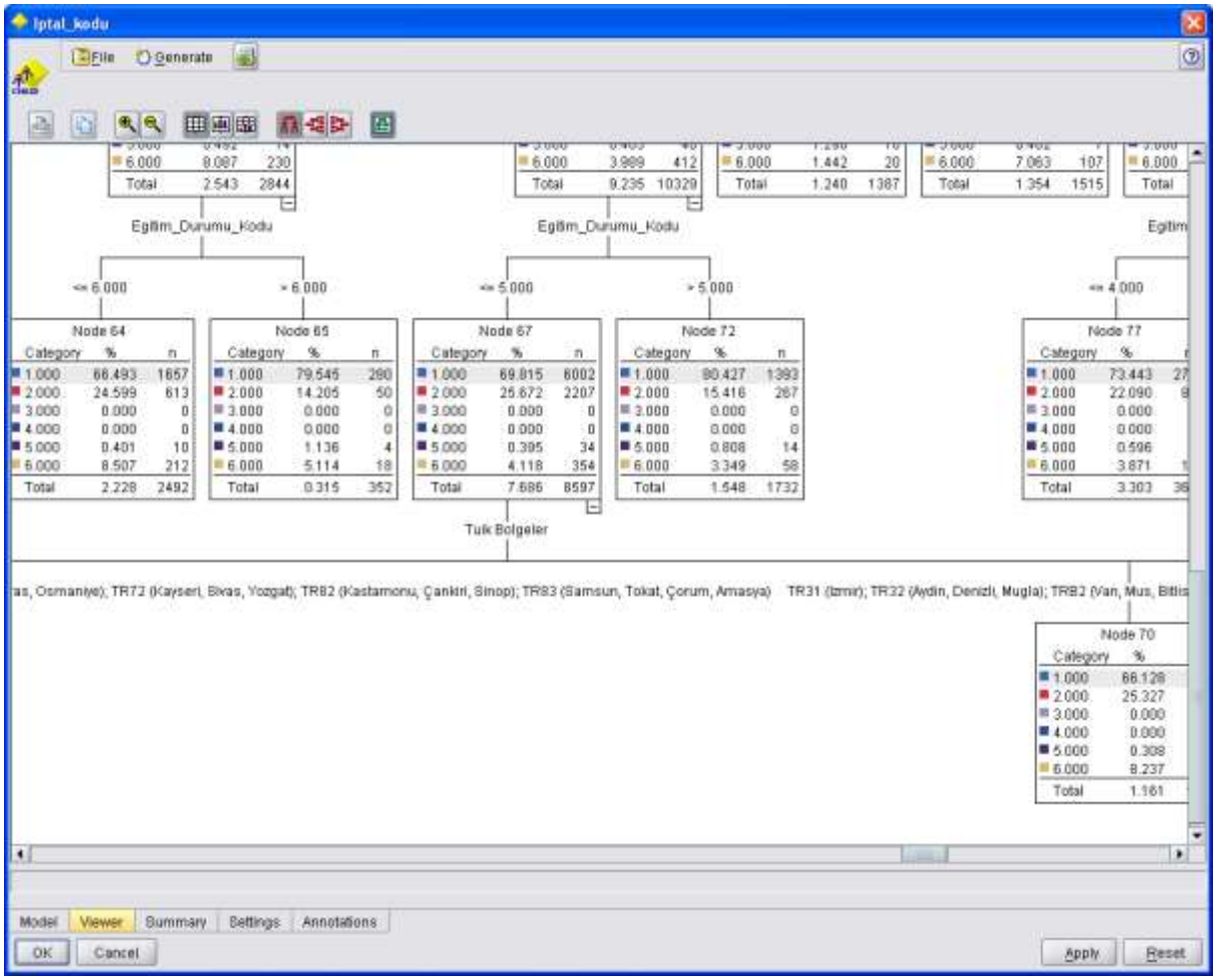

Şekil 4.22. Nod 27 (yıllık dilimi= 2 için başlangıç düğümü)

Karar ağacında yılın ikinci yarısını içeren bölümde ağaç yapısının meslek kodu- TÜİK bölgeler - eğitim kodu- paket kodu olarak dallara ayrıldığı görülmektedir.

İkinci 6 aylık dilimi içeren nod 27'nin alt düzeyleri incelendiğinde ;

Nod 48 için: Bireysel nedenlerden iptal oranının %90 'a yükseldiği gözlenmiştir. Yani yılın ikinci 6 aylık diliminde meslek kodu= 4 (emekli) olanların aboneliklerini bireysel nedenlerden dolayı iptal ettirdikleri görülmektedir. (Şekil 4.22)

Nod 70: İlk başlangıç düğümü nod 0'da iptal nedeni 6 (telefon iptali sebebiyle) oranı % 2.87, yıllık dilimlere göre ayrılınca ikinci 6 aylık dönem için(nod 27) %3.57 olarak değişmişti. Finans meslek grubuna ait olan ticari limitli paket koduna sahip, eğitim durumları lise ve daha az düzeyde olan TR31,TR32,TRC1,TRB2 bölgelerinde yaşayanları temsil eden nod 70'de bu oranın %8.2'ye yükseldiği görülmektedir. (Şekil 4.23)



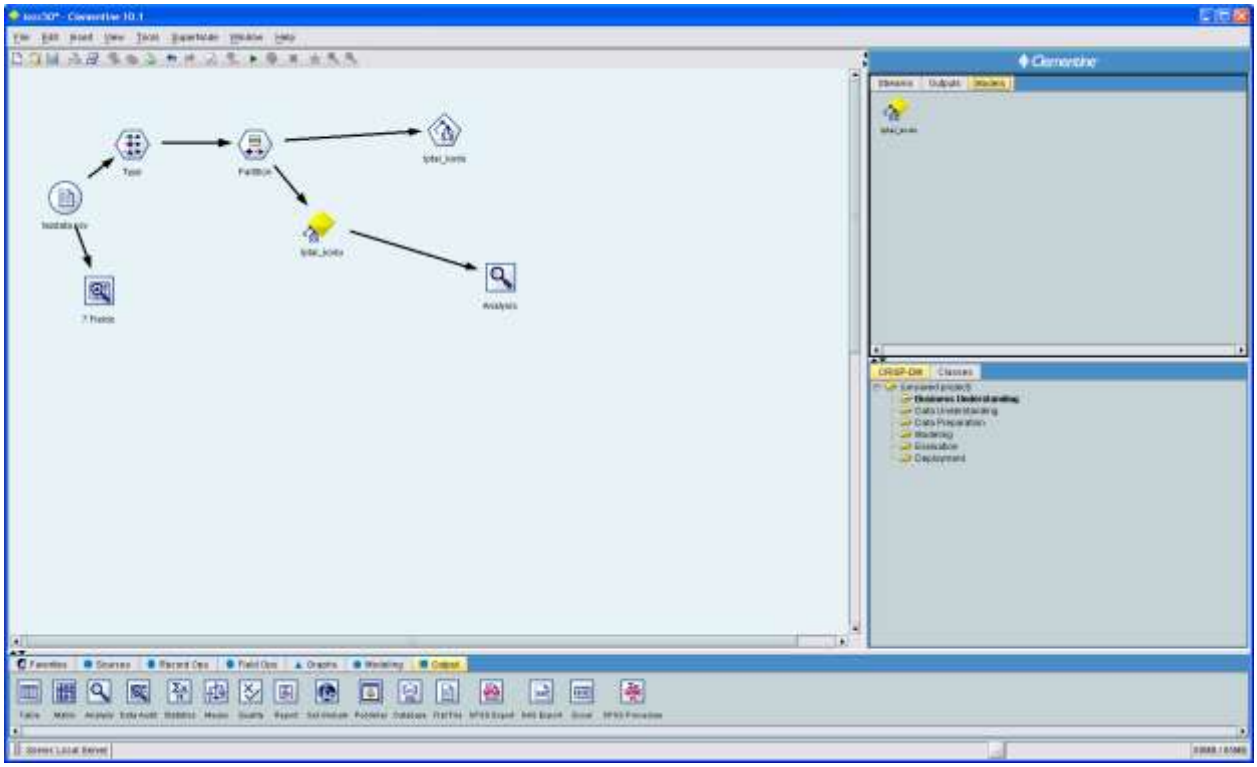

Şekil 4.23. Nod 70

### 4.3.2 C5.0 Algoritması

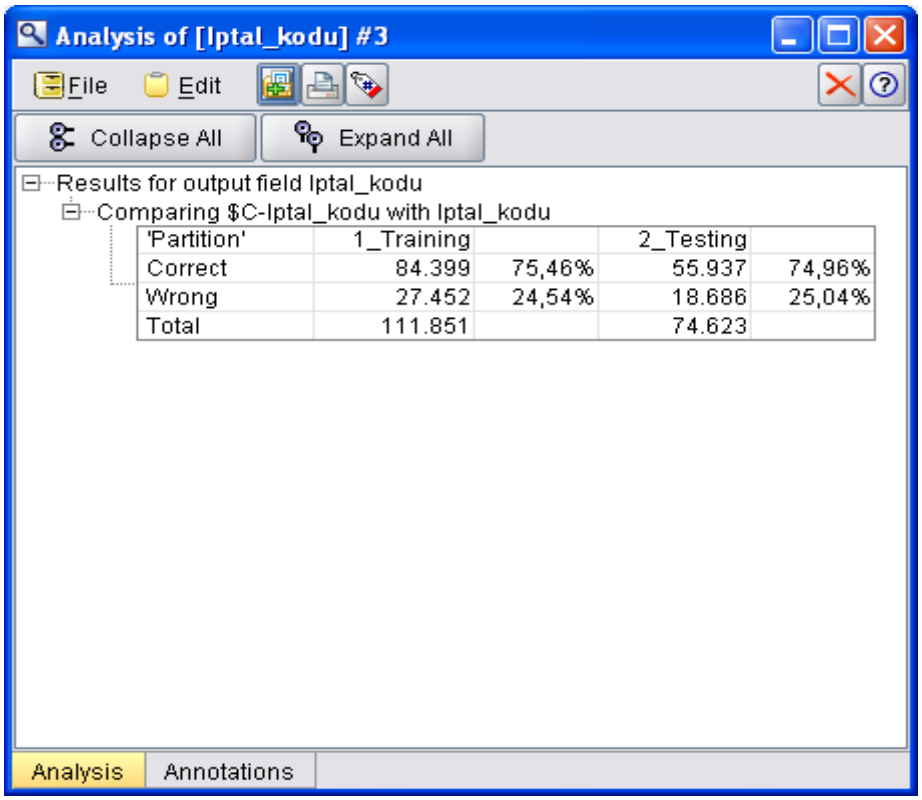
C5.0 algoritması ölçülebilir özellikteki veriler için daha yaygın kullanılmakla birlikte bu çalışmada da uygulanmıştır.

C5.0 algoritmasını çalıştırmak için partition noduna C5.0 modeli bağlanır ve çalıştırılır. C5.0 modeline ait sonuçlar analiz nodu eklenerek incelenir. (Şekil 4.24)

Analiz sonucuna göre modelin test ve training sonuçlarının birbirine yakın çıktığı ve %75 doğru skorlama yaptığı görülmektedir. Bu chaid modeli ile paralellik göstermektedir. (Şekil 4.25)

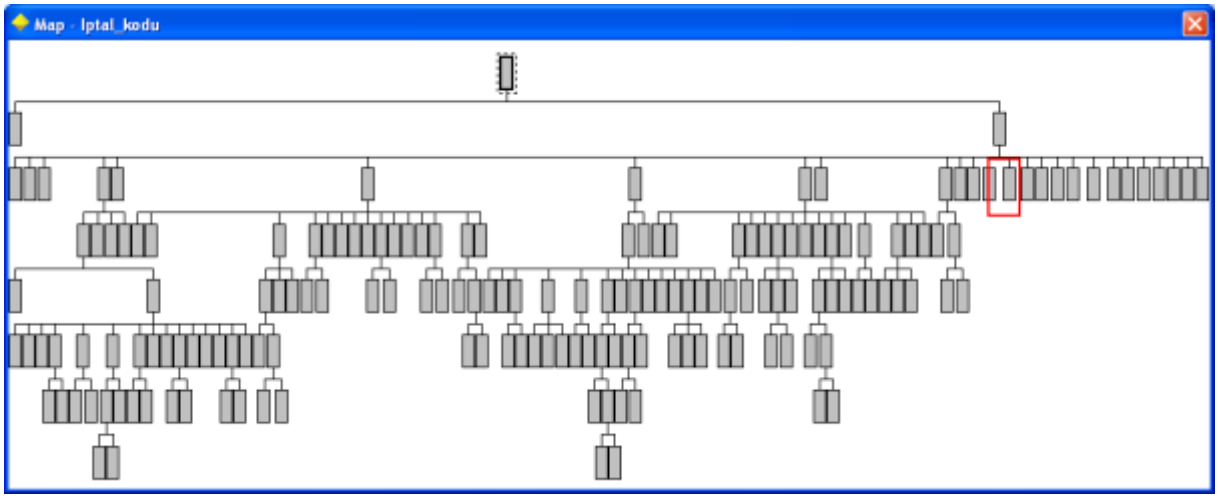


Şekil 4.24. C5.0 modeli ve analizi




Şekil 4.25. C5.0 analiz sonucu

C5.0 modeli sonucunda oluşan karar ağacının genel yapısı Şekil 4.26'da görülmektedir.



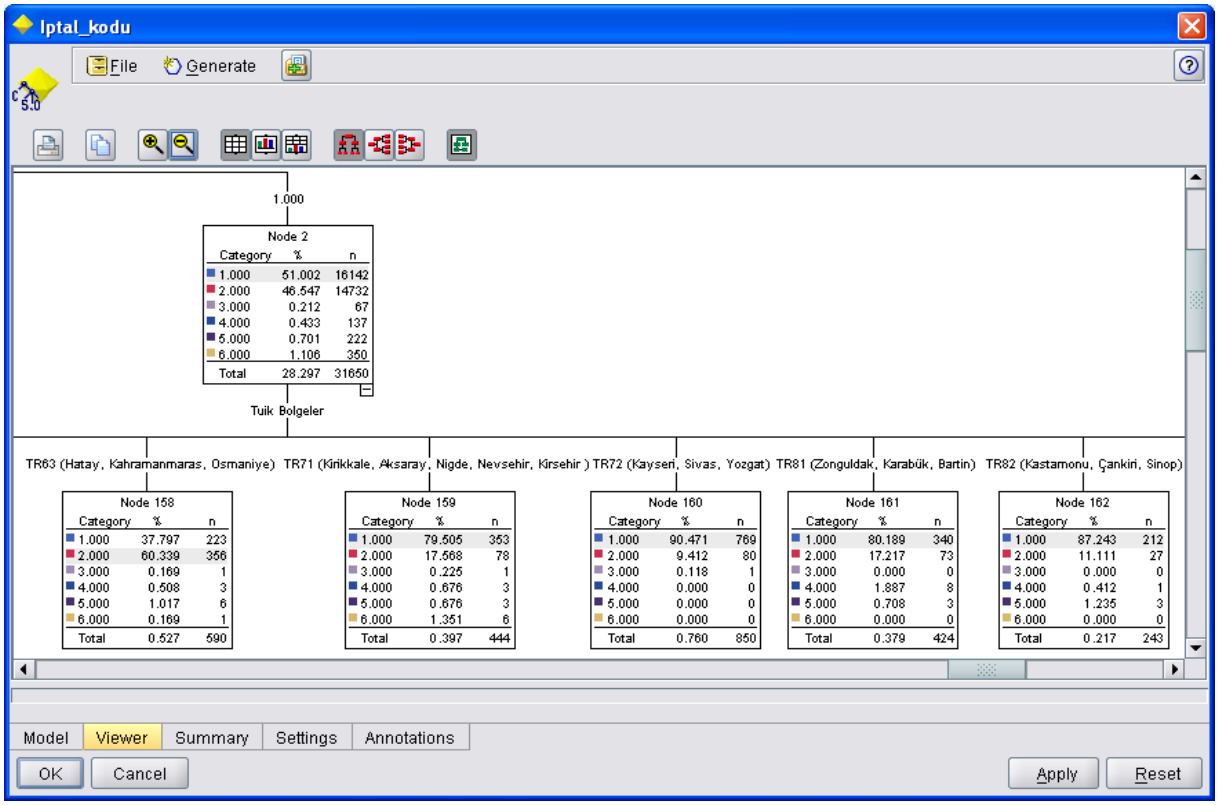

Şekil 4.26. C5.0 karar ağacı genel yapısı

C5.0 karar ağacı incelendiğinde yıllık dilim ayrımında 2. dilim için herhangi bir ağaç yapısı oluşmadığı görülmüştür. İlk 6 ayı kapsayan yıllık dilimde ise büyük değişim nodlar:

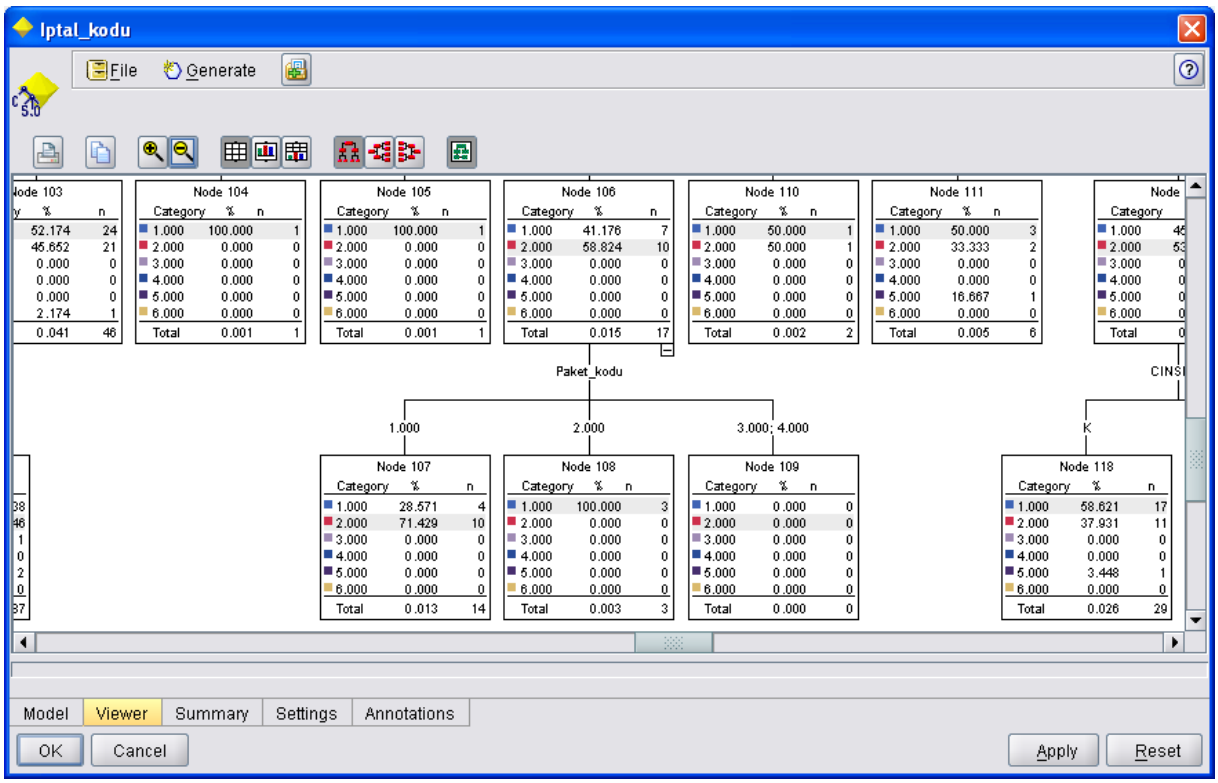
Nod 160 için: Nod 2'de bireysel tercihlerden iptal oranı %51 iken nod 160'da bu oran %90'a yükselmiştir. Bu nod TR72 (Kayseri, Sivas, Yozgat) illerini içermektedir. (Şekil 4.27)

Nod 108 için: Bir üst düğüm nod 106'da bireysel tercihlerden iptal oranı %41 iken nod 108 paket kodu 2 (Bireysel limitsiz) seçeneği için bu oran %100'e yükselmiştir. (Şekil 4.28)

Genel olarak değerlendirildiğinde C5.0 diğer yöntemler kadar sağlıklı bir sonuç vermemiştir. Bunun nedeni olarak ele alınan veri yapısının C5.0'a uygun olmadığı söylenebilir.




Şekil 4.27. Nod 160




Şekil 4.28. Nod 108

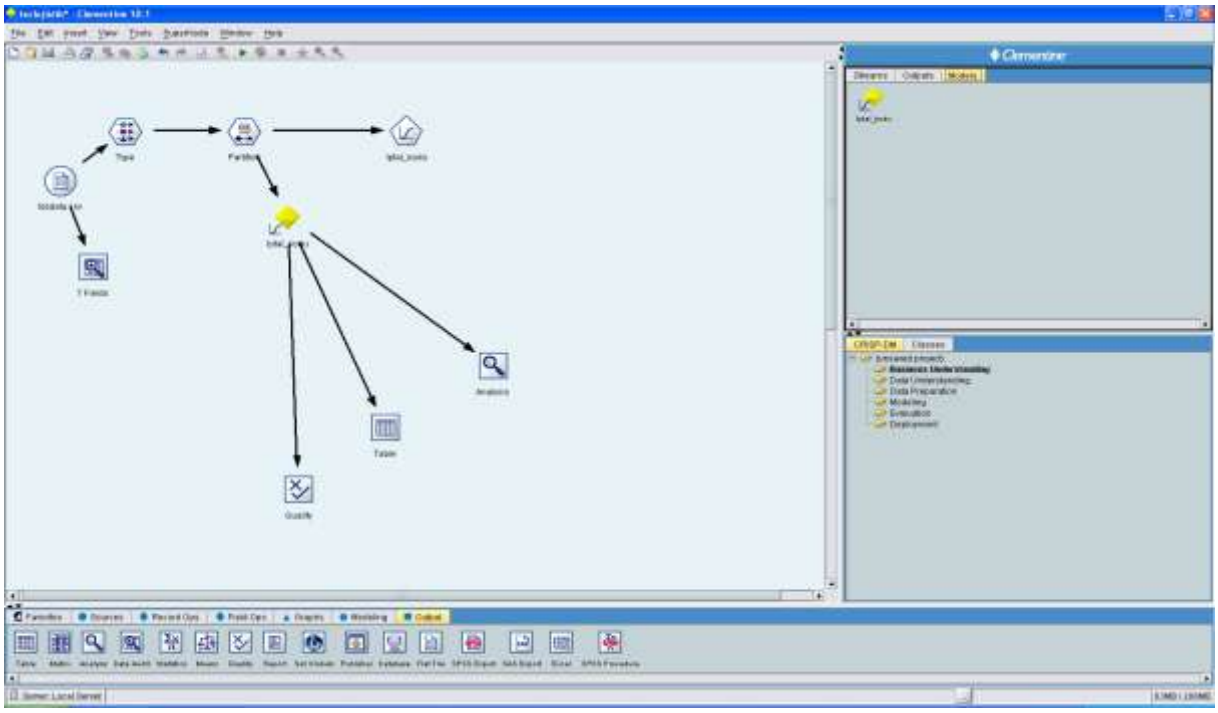
### 4.3.3 Lojistik Regresyon

Lojistik regresyon hedef ve tahminleyici alan bilgileri Çizelge 4.1'de gösterilmektedir.

Çizelge 4.1. Logistik regresyon bilgileri

| | |
|---|---|
| Hedef Alan : | İptal Nedeni |
| Tahminleyici Alanlar: | Yıllık dilim |
| | Cinsiyet |
| | Paket Kodu |
| | TÜİK Bölgeler |
| | Meslek Kodu |
| | Eğitim Durumu |
| | Paket Kodu |

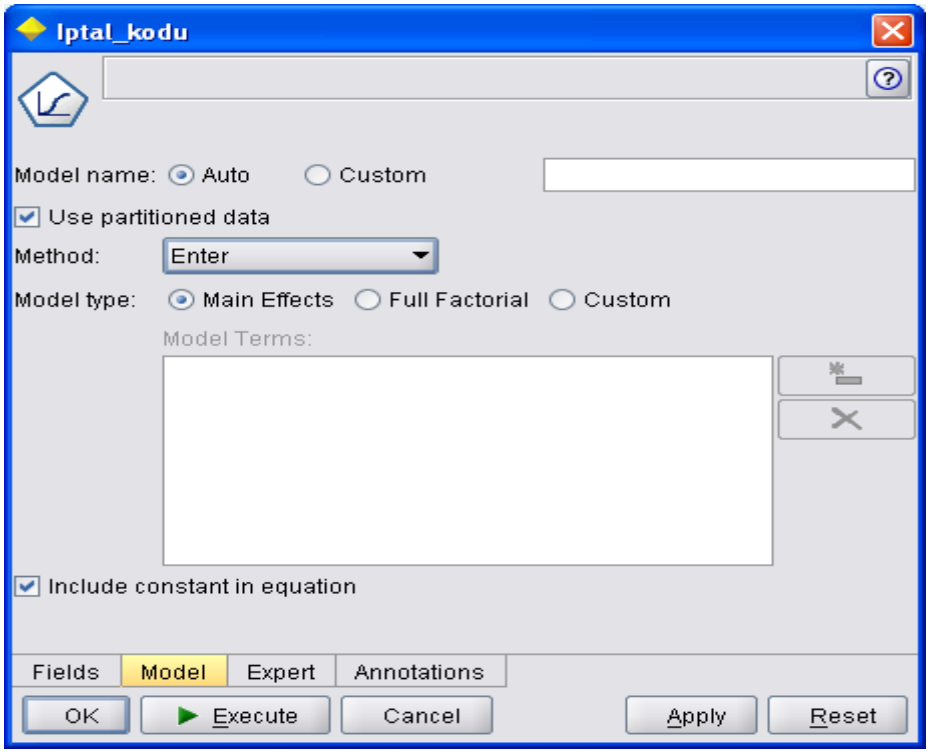
Lojistik regresyon çalıştırmak için "partition nodu"'na lojistik nodu bağlanır ve çalıştırılır. Lojistik regresyona ait sonuçlar analiz nodu eklenerek incelenir. (Şekil 4.29)



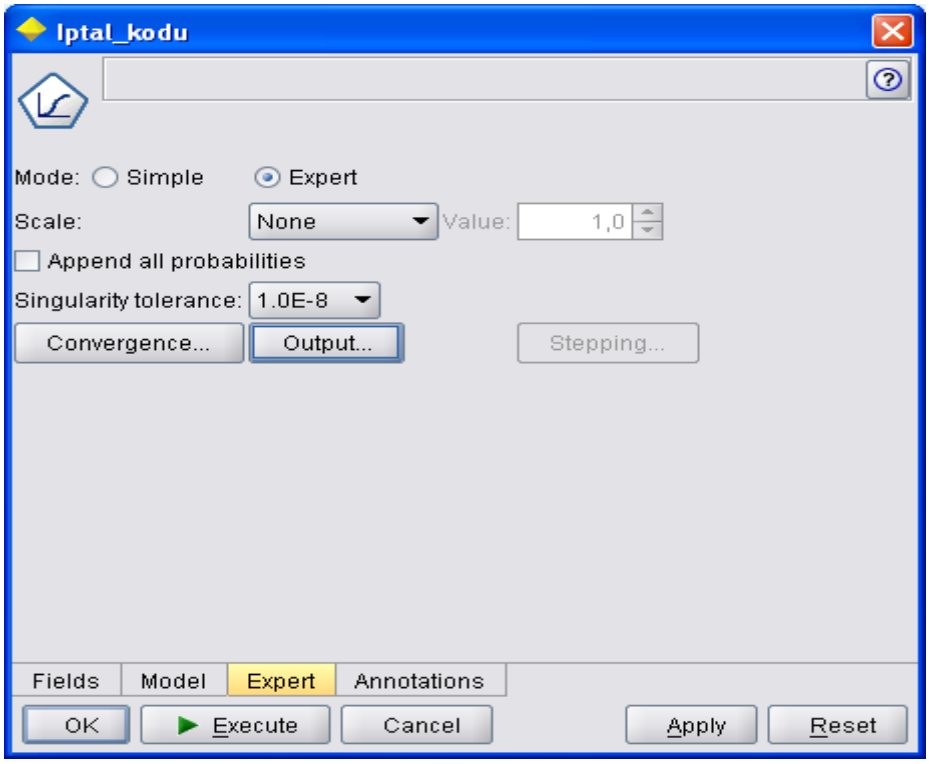
Şekil 4.29. Lojistik regresyon ve analizi

Lojistik regresyon nodunda “model tab”’ı ile denklemde sabit değer bulunup bulunmadığı belirlenir. Method alanında enter metod seçildiğinde anlamlı olsun olmasın tüm değişkenleri dikkate almaktadır. “Model type” alanında “main effect” (temel etkiler) seçilerek model faktörlerin ve sayısal tahmin edicilerin temel etkileri incelenir. (Şekil 4.30)

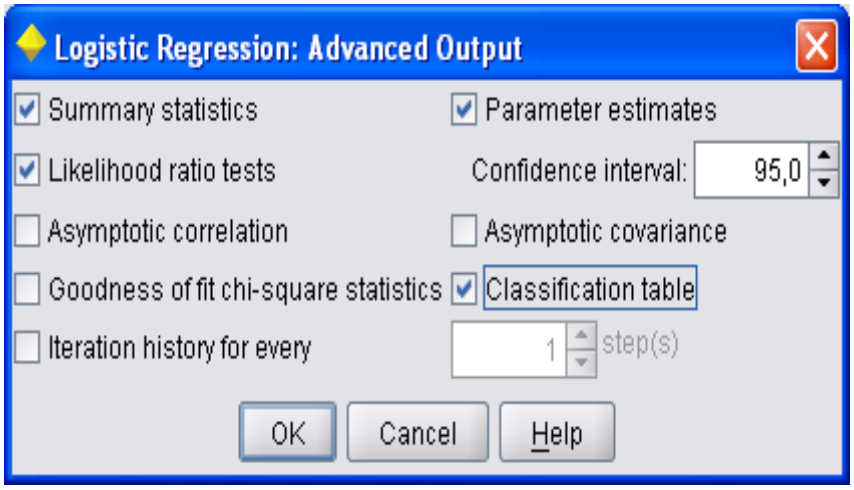
Lojistik regresyon nodunda “expert tabı”’nda “ append all probabilities” seçilirse her kayıtın yanına output alanının bütün kategorileri için tahmin edilen olasılıklar kayıt edilecektir. Seçilmezse tahmin edilen kategori için bir tane olasılık alanı eklenecektir.(Şekil 4.31) Exper butonunda output seçilerek default olarak gelen “summary statistics”- özet istatistikleri, “parameter estimates”- parametre tahminleri ve “likelihood ratio test”- olabilirlik oran testi alanlarıyla beraber “clasification table” alanı da seçilir. (Şekil 4.32)




Şekil 4.30. Lojistik regresyon nodu model tabı




Şekil 4.31. Lojistik regresyon nodu expert tabı




Şekil 4.32 Lojistik regresyon nodu expert tabı output butonu

Logistik regresyon sonucunda elde edilen kayıt durum özeti tablosu Çizelge 4.2'de yer almaktadır. Bu tabloda alanların kategorilere göre sıklıkları ve yüzdeleri yer almaktadır. Bir kayıtın analizde yer alabilmesi için burada yer alan tüm alanların dolu olması gerekmektedir. Yüzdelik dilimler incelendiğinde kayıtların %72'sinin bireysel nedenlerden dolayı aboneliklerini iptal ettirdikleri, TÜİK bölgeler incelendiğinde %25'lik oranla TR10 (İstanbul) 'un ilk sırada yer aldığı, iptal eden abonelerin

%76'sının erkek olduğu, %26'sının lise mezunu olduğu, %37'sinin ticari limitli pakete sahip olduğu ve %71'inin yılın 2. diliminde aboneliklerini iptal ettirdiği görülmektedir. Analizde yaklaşık 112.000 kayıt yer almaktadır.

Çizelge 4.2. Kayıt durum özeti

| | Nominal Regression | | |
|---|---|---|---|
| Case Processing Summary | | | |
| | | N | Marginal Percentage |
| Iptal_kodu | 1 | 80963 | 72.4% |
| | 2 | 26755 | 23.9% |
| | 3 | 67 | .1% |
| | 4 | 137 | .1% |
| | 5 | 714 | .6% |
| | 6 | 3215 | 2.9% |
| Tuik Bolgeler | TR10 (Istanbul) | 28426 | 25.4% |
| | TR21 (Tekirdag, Edirne, Kirklareli) | 2844 | 2.5% |
| | TR22 (Balikesir, Çanakkale) | 3156 | 2.8% |
| | TR31 (Izmir) | 9541 | 8.5% |
| | TR32 (Aydin, Denizli, Mugla) | 5218 | 4.7% |
| | TR33 (Manisa, Afyon, Kütahya, Usak) | 3185 | 2.8% |
| | TR41 (Bursa, Eskisehir, Bilecik) | 6511 | 5.8% |
| | TR42 (Kocaeli, Sakarya, Düzce, Bolu, Yalova) | 5399 | 4.8% |
| | TR51 (Ankara) | 10600 | 9.5% |
| | TR52 (Konya, Karaman) | 2736 | 2.4% |
| | TR61 (Antalya, Isparta, Burdur) | 4866 | 4.4% |
| | TR62 (Adana, Mersin) | 5276 | 4.7% |
| | TR63 (Hatay, Kahramanmaras, Osmaniye) | 2263 | 2.0% |
| | TR71 (Kirikkale, Aksaray, Nigde, Nevsehir, Kirsehir ) | 1729 | 1.5% |
| | TR72 (Kayseri, Sivas, Yozgat) | 3124 | 2.8% |
| | TR81 (Zonguldak, Karabük, Bartin) | 1603 | 1.4% |
| | TR82 (Kastamonu, Çankiri, Sinop) | 822 | .7% |
| | TR83 (Samsun, Tokat, Çorum, Amasya) | 3295 | 2.9% |
| | TR90 (Trabzon, Ordu, Giresun, Rize, Artvin, Gümüshane) | 2997 | 2.7% |
| | TRA1 (Erzurum, Erzincan, Bayburt) | 956 | .9% |
| | TRA2 (Agri, Kars, Igdir, Ardahan) | 583 | .5% |
| | TRB1 (Malatya, Elazig, Bingöl, Tunceli) | 1638 | 1.5% |
| | TRB2 (Van, Mus, Bitlis, Hakkari) | 830 | .7% |
| | TRC1 (Gaziantep, Adiyaman, Kilis) | 1859 | 1.7% |
| | TRC2 (Sanliurfa, Diyarbakir) | 1471 | 1.3% |
| | TRC3 (Mardin, Batman, Sirnak, Siirt) | 923 | .8% |
| Cinsiyeti | E | 85408 | 76.4% |
| | K | 26443 | 23.6% |

| | | | |
|---|---|---|---|
| **Meslek kodu** | **1** | 483 | .4% |
| | **10** | 8818 | 7.9% |
| | **11** | 116 | .1% |
| | **12** | 2078 | 1.9% |
| | **13** | 3584 | 3.2% |
| | **14** | 4591 | 4.1% |
| | **15** | 1727 | 1.5% |
| | **2** | 501 | .4% |
| | **3** | 29998 | 26.8% |
| | **4** | 11920 | 10.7% |
| | **5** | 13173 | 11.8% |
| | **6** | 21681 | 19.4% |
| | **7** | 1485 | 1.3% |
| | **8** | 480 | .4% |
| | **9** | 11216 | 10.0% |
| **Egitim_Durumu_Kodu** | **1** | 683 | .6% |
| | **10** | 262 | .2% |
| | **2** | 418 | .4% |
| | **3** | 28209 | 25.2% |
| | **4** | 10835 | 9.7% |
| | **5** | 40942 | 36.6% |
| | **6** | 7432 | 6.6% |
| | **7** | 20959 | 18.7% |
| | **8** | 1480 | 1.3% |
| | **9** | 631 | .6% |
| **Paket_kodu** | **1** | 27010 | 24.1% |
| | **2** | 19177 | 17.1% |
| | **3** | 53291 | 47.6% |
| | **4** | 12373 | 11.1% |
| **Yillik_Dilim** | **1** | 31650 | 28.3% |
| | **2** | 80201 | 71.7% |
| **Valid** | | 111851 | 100.0% |
| **Missing** | | 0 | |
| **Total** | | 111851 | |
| **Subpopulation** | | 9844(a) | |

a. The dependent variable has only one value observed in 5739 (58.3%) subpopulations.

Modelin modelin anlamlılığını test etmek için Ki kare istatistiği kullanılır.

$H_0$ : Model anlamsızdır. (4.1)

$H_1$ : Model anlamlıdır. (4.2)

Çizelge 4.3 ‘e göre p <0.05 olduğu için model anlamlıdır. Pseudo $R^2$ model tarafından açıklanan değişkenliği veren uyum iyiliği ölçütüdür. Modelimizde değişkenlik %16 oranında açıklanmaktadır.

Çizelge 4.3. Model uyum tablosu

| Model Fitting Information | | | | |
|---|---|---|---|---|
| Model | Model Fitting Criteria | Likelihood Ratio Tests | | |
| | -2 Log Likelihood | Chi-Square | df | Sig. |
| Intercept Only | 55.060.214 | | | |
| Final | 35.495.414 | 19.564.800 | 265 | .000 |
| Pseudo R-Square | | | | |
| Cox and Snell | .160 | | | |
| Nagelkerke | .210 | | | |
| McFadden | .121 | | | |

Model uyum tablosu omnibus testi ile modelin anlamlılığını test etmektedir. Olabilirlik oran testi (Likelihood ratio tests) ise modeldeki bütün etkilerin önemliliğini test eder. Çizelge 4.4’e göre bütün etkiler önemlidir. Bu tabloya göre sabit değerler test edilememektedir ancak sabitlerin anlamlılığı için parametre tahminleri tablosu yer almaktadır. (Çizelge 4.5)

Çizelge 4.4 Olabilirlik oran testi tablosu

| Likelihood Ratio Tests | | | | |
|---|---|---|---|---|
| Effect | Model Fitting Criteria | Likelihood Ratio Tests | | |
| | -2 Log Likelihood of Reduced Model | Chi-Square | df | Sig. |
| **Intercept** | 35495.414(a) | .000 | 0 | . |
| **Tuik Bolgeler** | 38057.242(b) | 2.561.828 | 125 | .000 |
| **Cinsiyeti** | 35541.198(b) | 45.784 | 5 | .000 |
| **Meslek kodu** | 36.673.925 | 1.178.511 | 70 | .000 |
| **Egitim_Durumu_Kodu** | 35.942.498 | 447.084 | 45 | .000 |
| **Paket_kodu** | 37448.609(b) | 1.953.195 | 15 | .000 |
| **Yillik_Dilim** | 37.954.801 | 2.459.387 | 5 | .000 |

Parametre tahminleri tablosunda bütün değişkenlerin denklemlerdeki katsayıları ve bu katsayıların anlamlarını test etmek için hesaplanmış istatistikler bulunmaktadır. Hedef değişkenimizin 6 düzeyi olduğu için lojistik regresyon beş ayrı denklem vermektedir. Örnek olarak iptal nedeni 1 (Bireysel tercih) incelendiğinde paket kodu (1,2,3) cinsiyet (erkek) ve yıllık dilim(1) katsayılarının anlamlı olduğu görülmektedir. (Çizelge 4.5)

Çizelge 4.6 ‘da yer alan sınıflandırma tablosu gözlemlerin ne kadarının doğru, ne kadarının hatalı sınıflandırıldığını göstermektedir. Toplamda modelin %74.5 oranında doğru tahminde bulunduğu görülmektedir. Tabloda satırlar gerçek düzeyleri, sütunlar ise tahmin edilen düzeyleri vermektedir. En iyi tahmin edilen grubun iptal nedeni 1 (bireysel tercih) olduğu görülmektedir.

Çizelge 4.5. Parametre tahminleri tablosundan bir kesit

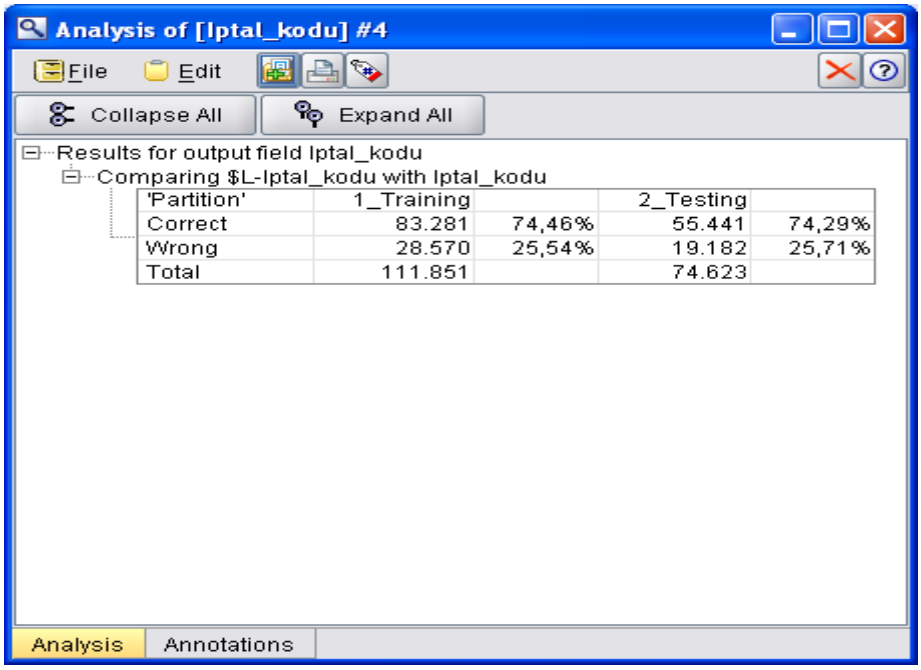
| Parameter Estimates | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Iptal kodu(a) | | B | Std. Error | Wald | df | Sig. | Exp(B) | 95.0% Confidence Interval for Exp(B) | |
| | | | | | | | | Lower Bound | Upper Bound |
| 1 | Intercept | 3.162 | .320 | 97.379 | 1 | .000 | | | |
| | [Tuik Bolgeler=TR10 (Istanbul)] | .791 | .140 | 32.045 | 1 | .000 | 2.205 | 1.677 | 2.899 |
| | [CINSIYETI=E] | .261 | .049 | 28.745 | 1 | .000 | 1.298 | 1.180 | 1.429 |
| | [CINSIYETI=K] | 0(b) | . | . | 0 | . | . | . | . |
| | [Meslek kodu=1] | -.687 | .203 | 11.482 | 1 | .001 | .503 | .338 | .749 |
| | [Meslek kodu=10] | .379 | .092 | 16.792 | 1 | .000 | 1.460 | 1.218 | 1.750 |
| | [Meslek kodu=11] | -.100 | .520 | .037 | 1 | .847 | .905 | .326 | 2.507 |
| | [Paket_kodu=1] | -.998 | .135 | 54.728 | 1 | .000 | .369 | .283 | .480 |
| | [Paket_kodu=2] | -1.327 | .080 | 277.197 | 1 | .000 | .265 | .227 | .310 |
| | [Paket_kodu=3] | -.797 | .075 | 112.516 | 1 | .000 | .451 | .389 | .522 |
| | [Paket_kodu=4] | 0(b) | . | . | 0 | . | . | . | . |
| | [Yillik_Dilim=1] | .928 | .107 | 74.935 | 1 | .000 | 2.530 | 2.051 | 3.122 |
| | [Yillik_Dilim=2] | 0(b) | . | . | 0 | . | . | . | . |

Çizelge 4.6. Sınıflandırma tablosu

| Classification | | | | | | | |
|---|---|---|---|---|---|---|---|
| | Predicted | | | | | | |
| Observed | 1 | 2 | 3 | 4 | 5 | 6 | Percent Correct |
| 1 | 74524 | 6439 | 0 | 0 | 0 | 0 | 92.0% |
| 2 | 17998 | 8757 | 0 | 0 | 0 | 0 | 32.7% |
| 3 | 25 | 42 | 0 | 0 | 0 | 0 | .0% |
| 4 | 77 | 60 | 0 | 0 | 0 | 0 | .0% |
| 5 | 612 | 102 | 0 | 0 | 0 | 0 | .0% |
| 6 | 3032 | 183 | 0 | 0 | 0 | 0 | .0% |
| Overall Percentage | 86.1% | 13.9% | .0% | .0% | .0% | .0% | 74.5% |

Lojistik regresyon sonuçları yorumlandıktan sonra Clementine ‘de sonuçları genel olarak değerlendirebilmek için “analysis”, “table” ve “quality “ nodları çalıştırılır. (Şekil 4.29)

Analiz sonucuna göre modelin “testing” ve “training” sonuçlarının birbirine yakın çıktığı ve %74 doğru skorlama yaptığı görülmektedir. Bu chaid modeli ile paralellik göstermektedir. (Şekil 4.33)

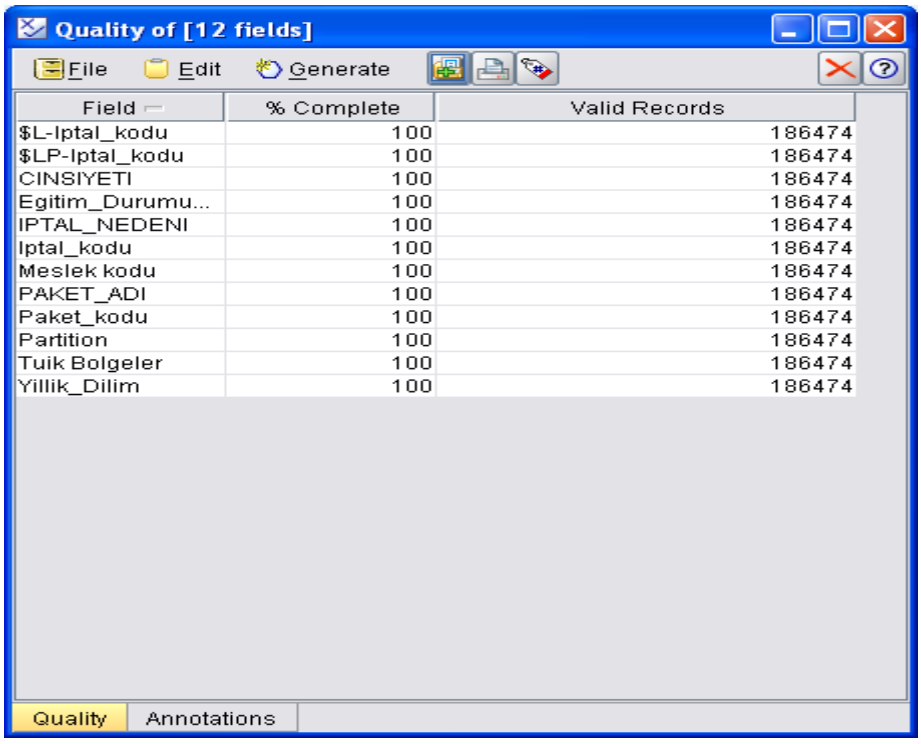



| 'Partition' | 1_Training | | 2_Testing | |
|---|---|---|---|---|
| Correct | 83.281 | 74,46% | 55.441 | 74,29% |
| Wrong | 28.570 | 25,54% | 19.182 | 25,71% |
| Total | 111.851 | | 74.623 | |

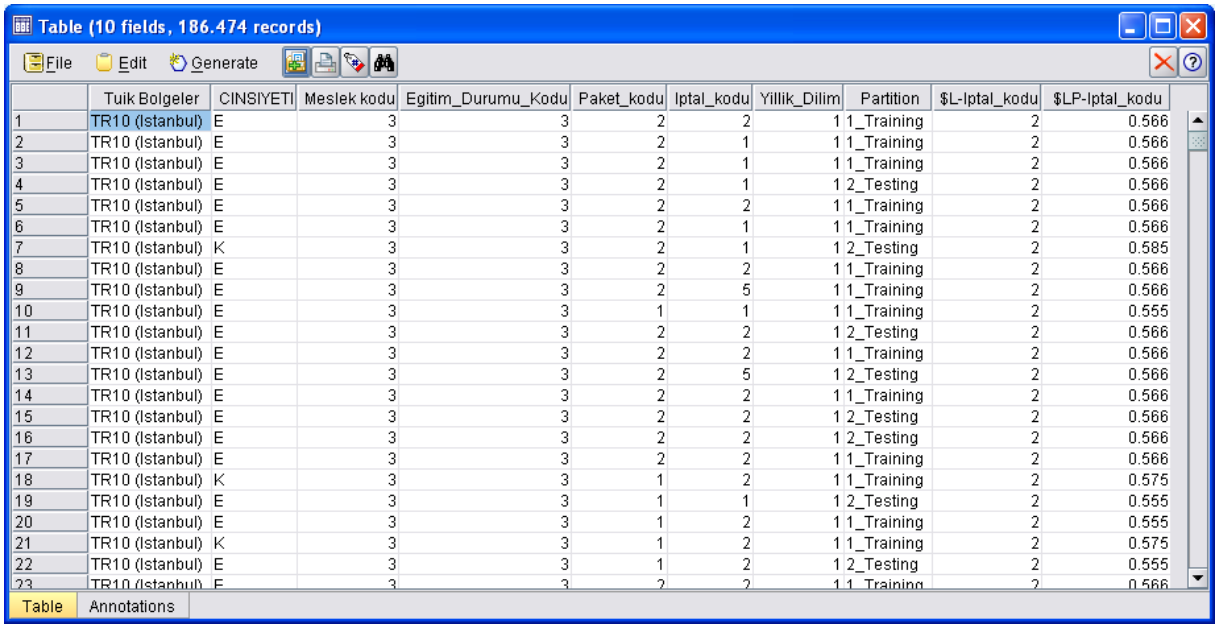
Şekil 4.33. Lojistik regresyon analiz sonucu

“Quality nodu” bir veri akımında eksik değerlerin raporlanmasını sağlar. Tüm değişkenlerimiz için eksik bilgi olmadığı görülmektedir. (Şekil 4.34)

“Table nodu”’nda değişkenlerin gerçek değerleri ile oluşturulan model sonucunda çıkan değer bilgileri görüntülenmektedir. (Şekil 4.35)




Şekil 4.34. Quality nodu




Şekil 4.35. Table nodu

# 5 SONUÇ VE ÖNERİLER

Çalışmada veri madenciliği kavramını, gelişim süreci, uygulama metodları ve bunların istatistikteki yerini anlatmak amaçlanmıştır. Bu doğrultuda öncelikle veri madenciliği sürecinin tarihçesi, OLAP ve istatistik gibi disiplinlerle ilişkisi ele alınmış olup süreç örneklerle ayrıntılı olarak ele alınmış ve veri madenciliği yöntemleri ayrıntılı olarak incelenmiştir. Son bölümde de SPSS Clementine'in veri madenciliği sürecinde uygulaması adım adım verilmiştir.

Günümüzde bilgisayarların çok yaygın kullanılmasıyla birlikte her türlü veri sayısal ortamda kayıt altına alınmaya başlanmıştır. Süpermarkette yapılan alışverişten, bankacılık işlemlerine, telekomünikasyon hizmetlerinden sağlık sektörüne kadar her bir işlem veritabanlarında birer kayıt olarak karşılık bulmaktadır. Sonuç olarak terabaytlar düzeyinde veritabanları oluşmakta ve bu verinin miktarı günden güne artmaktadır. Dolayısıyla büyük hacimli veriler arasında stratejik öneme sahip bilgi elde etmenin yolu Veri Madenciliğidir.

SPSS Clementine kullanılarak yapılan uygulama 2GB RAM ve 160 GB sabit diske sahip bir diz üstü bilgisayarda yapılmıştır. Bilgisayarın yeterli özelliklere sahip olması, Clementine programının kurulum ve çalıştırma adımlarının kısa sürede gerçekleştirilmesini sağlamıştır. Clementine'nin veri büyüklüğünde herhangi bir kısıtlaması olmaması özelliğinden dolayı veri incelemesinde bir sorunla karşılaşılmamıştır.

Çalışmada yapılan uygulama sonucunda 7 boyutlu 186474 veriden bir servis sağlayıcı firmanın 2009 yılına ait ADSL hizmetini iptal eden abonelerin iptal nedenleri ve bunu etkileyen değişkenleri incelenmiş ve yorumlanmıştır.

Elde edilen sonuçlar konunun uzmanlarıyla paylaşılmış ve ortaya çıkan sonuçlar neticesinde yılın 2 dönemi arasındaki farklılaşmanın alternatif servis sağlayıcı firmalarının olması, fiyat kırma stratejileri ve ekonomik nedenlerden kaynaklanabileceği fikrine varılmıştır.

Verileri anlamlı bilgiye çevirmek için bir çok istatistik programı bulunmaktadır. Ancak bu programların çoğu büyük hacimlerdeki veriler ile çalışamamaktadırlar. Clementine bu yönüyle araştırmaların daha kapsamlı ve kesin sonuçlar alınmasında etkili olmuş

ve ön plana çıkmıştır. İstatistiğin daha anlaşılabilir ve uygulanabilir olmasına yol açmıştır.

Uygulamada veriler, clementine programı kullanılarak daha kolay analiz edilebilir ve yorumlanabilir hale getirilmiştir.

Çalışmada servis sağlayıcı firmanın abonelik iptal edilirken almış olduğu bilgiler kullanılmıştır. İncelemelerde iptal nedenlerinin bireysel tercih ve ekonomik nedenler üzerine baskın olduğu gözlemlenmiştir. Daha ayrıntılı bilgi içeren bir verinin toplanarak incelenmesi ile sonuçların yeniden değerlendirilmesi önerilebilir. Bu, iptal işlemi sırasında abonelere yöneltilen soru sayısının artırılması ve iptal sebebinin ayrıntılı olarak incelenmesi ile mümkün olabilir.

**SÖZLÜK**

Terabayt : Bilgisayarlarda kullanılan, 1024 gigabayt büyüklüğündeki ölçü birimidir.

Gigabayt : Bilgisayarlarda kullanılan, 1024 megabayt anlamına gelen bir ölçü birimidir.

Rasyo : İşletmenin yapısı ve işletme faaliyetlerinin verimliliği hakkında bilgi veren rakamlardır.

**KAYNAKLAR**


1. Ahola, J., Rinta-Runsala, E., 2001, Data Mining Case Studies in Customer Profiling, VTT İnformation Technology, TTE1-2001-29
2. Akpınar, H., 2000, Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği, İ.Ü. İşletme Fakültesi Dergisi, C:29, S: 1/Nisan 2000, s: 1-22
3. Alkan A., 2007, Cozum Veri Madenciliginde, LinkPlus
4. Alpaydın, E., 2000, Zeki Veri Madenciliği: Ham Veriden Altın Bilgiye Ulaşma Yöntemleri, Bilişim 2000 Eğitim Semineri
5. Apte, C., Liu, B., Pednault, E.P.D., Symth., P., 2002, Business Application of Data Mining, Communications of The Acm August 2002/Vol. 45, No. 8
6. Ballard, C., Herreman, D., Schau, D., Bell, R., Kim, E., Valencic, A., 1998, Data Modeling Techniques for Data Warehousing, International Technical Support Organization, http://www.redbooks.ibm.com
7. Bounsaythip,C. and Runsala,E.R., 2001, Overview of Data Mining for Customer Behavior Modeling, VTT Technology, Research Report TTE1-2001-18
8. Cabena, P., Hadjinian, P., Stadler, R., Verhees, J., and Zanasi, A., 1998, Discovering Data Mining: From Concept to Implementation, Prentice Hall, Upper Saddle River,NJ
9. Doğan, N., Özdamar, K., 2003, Cahid Anlizi ve Aile Planlaması ile İlgili Bir Uygulama,T Klin Tıp Bilimleri 2003-23
10. Feelders, A., Daniels, H., Holsheimer, M., 2000, Metohodological and Practical Aspects of Data Mining, Information & Management 37 (2000) 271-281
11. Forcht, K., and Cochran, K., 1999, Using Data Mining and Datawarehousing Techniques
12. Ganguly, A.R., Steinhaeuser, K., 2008, Data Mining for Climate Change and Impacts, 2008 IEEE International Conference on Data Mining Workshops
13. Han, J. and Kamber M., 2006, Data Mining Concepts and Techniques, Morgan Kaufmann Publishers
14. Hand, D., and Manilla, H., Smyth, P., 2001, Principles of Data Mining, MIT Press
15. Hand, D.J., 1999 ,Statistics and Data Mining: Intersecting Disciplines, Department of Mathematics, SIGKKD Explorations, Volume 1, Issue1, Page 19, Imperial College, London, UK

16. Hand, D.J.,2000, Data Mining New Challenges for Statisticans

17. Hegland, M., 2001, Data Mining Techniques, Acta Numerica (2001), pp. 313–355 c Cambridge University Press

18. Hirji, K.K., 2001, Exploring Data Minnig İmplementation, Communications of The Acm July 2001/Vol. 44, No. 7

19. Honga, T., Wang, C., and Taoc, Y., 2001, A New İncremental Data Mining Algorithm, Intelligent Data Analysis 5 (2001) 111–129 111, IOS Press

20. Hornick, F.M., Marcadé, E. and Venkayala, S., 2007, Java Data Mining: Strategy, Standard and Practice a Practical Guide for Architecture, Design and Implementation, Morgan Kaufman

21. Jayagopal, B., 2008, Applying Data Mining Techniques To Credit Scoring, Amadeus Software Limited

22. Larose, D.T.,2005, Discovering Knowledge in Data: An Introduction to Data Mining, Wiley Publishing

23. Lee, S.J., Siau. K., Statistical Models for Data Mining Techniques, Industrial Management & Data Systems 101/1 [2001] 41-46, MCB University Press

24. Luan, J., 2002, Data Mining and Its Applications in Higher Education

25. Luo, Q., 2008, Advancing Knowledge Discovery and Data Mining, 2008 IEEE International Conference on Data Mining Workshops

26. Maimon, O. and Rokach, L., 2005, Data Mining & Knowledge Discovery Handbook, Springer, Tel-Aviv University, Israel

27. Mitchell, T. M., 1997, Machine Learning, McGraw-Hill

28. Olson, D.L. and Delen, D., 2008, Advanced Data Mining Techniques, Springer

29. Özkan, Y.,2008, Veri Madenciliği Yöntemleri

30. Özmen, Ş., 2001, İş Hayatı Veri Madenciliği ile İstatistik Uygulamaları, Marmara Universitesi İ.İ.B.F.

31. Quinlan, J.R., 1986, Induction of Decision Trees, Kluwer Academic Publishers, Boston, Machine Learning 1: 81-106

32. Quyang, J., Patel, N., Sethi, I.K., 2008, Chi-Square Test Based Decision Trees Induction in Distributed Environment, 2008 IEEE International Conference on Data Mining Workshops

33. Ratner, B., 2010, Chaid Its Original Intent, DM Stat-1 Article,

http://www.dmstat1.com/res/OriginalCHAIDintent.html

34. Rud, O., 2001, Data Mining Cookbook: Modeling Data for Marketing, Risk and Customer Relationship Management, Wiley Publishing

35. Rule Quest, 2010, http://www.rulequest.com/see5-comparison.html

36. Rygielski, C., Wang J.C., Yen, D.C., 2002, Data Mining Techniques for CRM, Technology in Society 24 (2002) 483–502

37. Silahtaroğlu, G., 2008, Kavram ve Algoritmalarıyla Temel Veri Madenciliği

38. Soares, C., Peng, Y., Meng, J., Washio, T. and Zhou, Z., 2008, Applications.of.Data.Mining.in.E-Business.and.Finance, IOS.Press

39. Spangler, W.E., Gal-Or, M., May, J.H., 2003 ,Data Mining to Profile TV Viewers, Communications of The Acm December 2003/Vol. 46, No. 12

40. SPSS Clementine 10.1 Handbook, 2009

41. SPSS, 1999, Field-Tested Data Mining 10 Essential Strategies and Tips

42. Statistics Solutions, 2010, CHAID, http://www.statisticssolutions.com/methods-chapter/statistical-tests/chaid/

43. Taniar, D., 2008, Data Mining and Knowledge Discovery Technologies, Monash University, Australia, IGI Publishing

44. The Gartner Group, 2010, www.gartner.com

45. Two Crows Corporation,1999, Introduction to Data Mining and Knowledge Discovery Third Edition by

46. Yeniay, Ö. ve Kadılar, C., 2001, Veri madenciliği ve İstatistik, Hacettepe Üniversitesi, İstatistik Bölümü

**EKLER DİZİNİ**

## EK 1. TÜİK BÖLGELER

| | |
|---|---|
| 1 | TR10 (İstanbul) |
| 2 | TR21 (Tekirdağ, Edirne, Kırklareli) |
| 3 | TR22 (Balıkesir, Çanakkale) |
| 4 | TR31 (İzmir) |
| 5 | TR32 (Aydın, Denizli, Muğla) |
| 6 | TR33 (Manisa, Afyon, Kütahya, Uşak) |
| 7 | TR41 (Bursa, Eskişehir, Bilecik) |
| 8 | TR42 (Kocaeli, Sakarya, Düzce, Bolu, Yalova) |
| 9 | TR51 (Ankara) |
| 11 | TR52 (Konya, Karaman) |
| 12 | TR61 (Antalya, Isparta, Burdur) |
| 13 | TR62 (Adana, Mersin) |
| 14 | TR63 (Hatay, Kahramanmaraş, Osmaniye) |
| 15 | TR71 (Kırıkkale, Aksaray, Niğde, Nevşehir, Kırşehir ) |
| 16 | TR72 (Kayseri, Sivas, Yozgat) |
| 17 | TR81 (Zonguldak, Karabük, Bartın) |
| 18 | TR82 (Kastamonu, Çankırı, Sinop) |
| 19 | TR83 (Samsun, Tokat, Çorum, Amasya) |
| 20 | TR90 (Trabzon, Ordu, Giresun, Rize, Artvin, Gümüşhane) |
| 21 | TRA1 (Erzurum, Erzincan, Bayburt) |
| 22 | TRA2 (Ağrı, Kars, Iğdır, Ardahan) |
| 23 | TRB1 (Malatya, Elazığ, Bingöl, Tunceli) |
| 24 | TRB2 (Van, Muş, Bitlis, Hakkari) |
| 25 | TRC1 (Gaziantep, Adıyaman, Kilis) |
| 26 | TRC2 (Şanlıurfa, Diyarbakır) |

**EK 2. CİNSİYET**

| | |
|---|---|
| 1 | KADIN |
| 2 | ERKEK |

**EK 3. YILLIK DİLİM**

| | |
|---|---|
| 1 | İLK 6 AYLIK DİLİM |
| 2 | İKİNCİ 6 AYLIK DİLİM |

**EK 4. PAKET KODU**

| | |
|---|---|
| 1 | BİREYSEL LİMİTLİ |
| 2 | BİREYSEL LİMİTSİZ |
| 3 | TİCARİ LİMİTLİ |
| 4 | TİCARİ LİMİTSİZ |

**EK 5. MESLEK KODU**

| | |
|---|---|
| 1 | BASIN YAYIN/ REKLAM/ HALKLA İLİŞKİLER |
| 2 | ÇALIŞMAYAN |
| 3 | DİĞER |
| 4 | EMEKLİ |
| 5 | EV HANIMI |
| 6 | FİNANS |
| 7 | HİZMET SEKTÖRÜ |
| 8 | HUKUK |
| 9 | İŞÇİ |
| 10 | KAMU |
| 11 | KÜLTÜR-SANAT-SPOR |
| 12 | MÜHENDİS/MİMAR |
| 13 | ÖĞRENCİ |
| 14 | ÖĞRETMEN/ ÖĞRETİM GÖREVLİSİ/ AKADEMİSYEN |
| 15 | SAĞLIK |

## EK 6. EĞİTİM DURUMU

| | |
|---|---|
| 1 | EĞİTİMSİZ |
| 2 | OKUL ÖNCESİ |
| 3 | ILKOKUL |
| 4 | ORTAOKUL |
| 5 | LİSE |
| 6 | YÜKSEK OKUL |
| 7 | ÜNİVERSİTE |
| 8 | YÜKSEK LİSANS |
| 9 | DOKTORA |
| 10 | TANIMSIZ |

## EK 7. İPTAL NEDENİ

| | |
|---|---|
| 1 | BİREYSEL TERCİH |
| 2 | EKONOMİK NEDENLER |
| 3 | MÜŞTERİ MEMNUNİYETSİZLİĞİ |
| 4 | SERVİS SAĞLAYICI DEĞİŞİKLİĞİ |
| 5 | TEKNİK ENGELLER |
| 6 | TELEFON İPTALİ SEBEBİYLE |

## ÖZGEÇMİŞ

Adı Soyadı : Fatma Meltem Kocabaş

Doğum Yeri : Ankara

Doğum Yılı : 29.06.1977

Medeni Hali : Bekar

Eğitim ve Akademik Durumu:

Lise 1991-1998 Kanuni Lisesi

Lisans 1994-1998 Hacettepe Üniversitesi, İstatistik Bölümü

Yabancı Dil: İngilizce

İş Tecrübesi:

1998-2008 Logo Business Solutions

2008-2009 Siemens

2009-... İnnova Bilişim Çözümleri A.Ş.