T.C.
SAKARYA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

KAÇAK SU KULLANIMININ TESPİTİNDE VERİ MADENCİLİĞİ YAKLAŞIMI

YÜKSEK LİSANS TEZİ

End. Müh. Muhammed Ali YAVUZ

Enstitü Anabilim Dalı : ENDÜSTRİ MÜHENDİSLİĞİ

Tez Danışmanı : Yrd. Doç. Dr. Bayram TOPAL

Eylül 2009

T.C.
SAKARYA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

KAÇAK SU KULLANIMININ TESPİTİNDE VERİ MADENCİLİĞİ YAKLAŞIMI

YÜKSEK LİSANS TEZİ

End. Müh. Muhammed Ali YAVUZ

Enstitü Anabilim Dalı : ENDÜSTRİ MÜHENDİSLİĞİ

Bu tez 18/09/2009 tarihinde aşağıdaki jüri tarafından Oybirliği ile kabul edilmiştir.

Yrd. Doç. Dr. Bayram TOPAL	Yrd. Doç. Dr. İsmail GÜMÜŞ	Yrd. Doç. Dr. Gültekin ÇAĞIL
Jüri Başkanı	Üye	Üye

ÖNSÖZ

Günümüz işletmeleri acımasız rekabet şartlarında ayakta kalmak için birçok yöntem kullanmaktadırlar. Kullanılan yöntem ne olursa olsun yöntemin en önemli bileşeni bilgidir. Bilgi ise artık çok pahalıdır. İşletmeler hem maliyetleri düşürmek, hem de ihtiyaç duyduğu bilgi türü materyalleri temin etmek zorundadırlar. Hem bu durumu aşmak hem de en doğru bilgiye ulaşmak için işletmeler her türlü veriyi veri tabanları ve veri ambarlarına kaydetmeyi; ihtiyaç duyduklarında ise onları etkin bir şekilde kullanmaya imkan sağlayan veri madenciliği yöntemini kullanmaya başlamışlardır. Temeli mevcut verilerden anlamlı ve kullanılabilir bilgiler çıkarmak olduğu için veri madenciliği müşteri ilişkileri yönetiminden hilekarlık tespitine kadar çok geniş bir yelpazede kullanım alanı bulmaktadır.

Bu çalışmanın hazırlanmasında bana maddi ve manevi desteğini esirgemeyen dostlarım End. Yük. Müh. Turgay ÖZTÜRK'e, End. Yük. Müh. Muhammed ÇETİN'e, Sayın Zülküf YILDIZ'a, Sayın Muhammed ASLIBAY'a, bana destek olan mesai arkadaşlarıma, uygulama sırasında yardımlarını esirgemeyen değerli büyüğüm Sayın İdris KAYMAK'a, çalışmada kullanılan verilerin temininde bana yardımcı olan ADASU yöneticilerine ve çalışanlarına, çalışmalarım sırasında beni destekleyen daire başkanım Sayın End. Yük. Müh. Metin BAYRAM'a, beni bilgi ve tecrübesiyle yönlendiren değerli hocam Sayın Yrd. Doç. Dr. Bayram TOPAL'a ve bugünlere gelmeme sebep olan Anneme ve Babama teşekkürü bir borç bilirim.

İÇİNDEKİLER

SİMGELER VE KISALTMALAR LİSTESİ

AID	:	Automatic Interaction Detector
CART	:	Classification and Regression Trees
CRISP-DM	:	Veri Madenciliği için Sektörler Arası Standart Süreci
CHAID	:	Chi-Squared Automatic Interaction Detector
C&RT	:	Classification and Regression Trees
DR	:	Doğrusal Regresyon
ID3	:	Induction of Decision Trees
GA	:	Genetik Algoritmalar
KA	:	Karar Ağacı
KPSS	:	Kamu Personeli Seçme Sınavı
LR	:	Lojistik Regresyon
MARS	:	Multivariate Adaptive Regression Splines
MR	:	Manyetik Rezonans
QUEST	:	Quick, Unbiased, Efficient, Statistical Tree
RBFN	:	Radial Bases Function Network
SLIQ	:	Supervised Learning in Quest
SPRINT	:	Scalable Parallelizable Induction of Decision Trees
VA	:	Veri Ambarı
VM	:	Veri Madenciliği
VT	:	Veri Tabanı
VTBK	:	Veri Tabanlarında Bilgi Keşfi
YSA	:	Yapay Sinir Ağları
WWW	:	World Wide Web

ŞEKİLLER LİSTESİ

TABLOLAR LİSTESİ

ÖZET

Anahtar kelimeler: Veri Madenciliği, Hilekârlık Tespiti, Kaçak Su Kullanımı

Bilginin temel yapısını oluşturan veri, son dönemde gelişen veri madenciliği kavramı ile daha bir önem kazanmıştır. Dünyada ve Türkiye'de veri madenciliğine olan ilgi ve yatırım büyük miktarlara ulaşmıştır. Dünyada perakendecilik, e-ticaret, bankacılık, sigortacılık, telekomünikasyon, sağlık ve eğitim alanlarında yaygın olarak kullanılan veri madenciliği, son dönemde Türkiye'de de özellikle marketçilik, banka ve sigortacılık, dolandırıcılık ve hilekarlık tespiti ile e-devlet alanlarında kullanılmaya başlanmıştır.

Bu çalışmada, veri madenciliğinin tanımı, kullanım alanları, model ve algoritmaları ayrıntılı olarak ele alınmıştır. Uygulama kısmında ise, kaçak su kullanımı engellemek için il yerel yönetiminin ilgili biriminin gerçek verileri kullanılmıştır. Birinci aşamada veriler düzenlenerek bir veri seti oluşturulmuş, daha sonra bu veri seti uygun model kurularak analiz edilmiştir. Elde edilen sonuçlar istatistik yöntemler kullanılarak test edilip, işletmenin ileride kaçak su kullanması muhtemel abonelerini tespit etmesine yönelik bir model oluşturulmaya çalışılmıştır.

DATA MINING APPROACH FOR DETECTION ILLEGAL USAGE OF WATER

SUMMARY

Key Words: Data Mining, Fraud Detection, Illegal Usage of Water

Being the basic structure of knowledge, data has gained considerable importance with the emergence of the concept of data mining. Investment and interest in data mining has been growing and already reached big sums in the world as well as in Turkey. Data mining is used worldwide in various social and industrial areas such as retail marketing, e-commerce, banking, insurance, telecommunications, health and education. In Turkey, in recent years it is being utilized especially in the areas of retail marketing, banking, insurance, fraud detection and e-state.

In this research, the definition of data mining, the areas of its application, the models and the algorithms have been examined intensively. In the implementation stage, real data taken from city government department that work about usage of water. In the first stage, all data have been restored for creating a data-set then this set has been analyzed by using an appropriate model. The results obtained, have been tested using statistical methods and results making good sense and affecting the relations between the company and members about illegal usage of water.

BÖLÜM 1: GİRİŞ

Dünyada ekonomik sınırların kaldırılıp "Küreselleşme" adı altında dünya küçük bir köy mertebesinde erişime imkân tanınması işletmeler arası rekabet, ticaret ve ilişkiler de muazzam boyutlara ulaştırmıştır. Bu durumdan çok karlı çıkan işletmeler olduğu gibi zararlı çıkan hatta ömrünü tamamlamak zorunda kalan işletmeler de olmuştur. Dünyada oluşan bu yeni oyun alanında var olabilme yarışına giren işletmeler oyunu kuralına göre oynamak için çağın en önemli kaynağı olan bilgiyi elde etmenin, bilgiyi saklamanın, etkin bir biçimde onu kullanmanın yollarını aramaya başlamışlardır. Mevcut birçok veri analiz teknikleri olduğu halde incelenecek verilerin devasa boyutlarda olması hem zaman hem de insan kaynağı açısından kısıtlayıcı bir faktör olmuştur. Fakat teknolojik gelişmeler sonucu ucuzlayan, hızlanan, birçok farlı işlemi aynı anda tam ve doğru olarak yapabilen bilgisayarlar bu noktada işletmelerin imdadına yetişmiştir. Bilgisayarlara ve işletme taleplerine göre oluşturulmuş veri analiz tekniklerini kullanan sürece veri madenciliği denmektedir. Süreç sonunda elde edilen veriler işletmeleri o kadar tatmin etmiştir ki kamudan finans sektörüne, müşteri ilişkileri yönetiminden hilekârlık tespitine kadar geniş bir alanda tercih edilmekte ve uygulanmaktadır.

1.1. Tezin Amacı

Bu tezin hazırlanma amacı;

- Temel veri kavramlarının açıklanması,
- Veri Madenciliği uygulaması sırasında oluşabilecek problemler, VM kullanım alanları ve amaçları gibi VM ile ilgili temel kavramların belirtilmesi,
- VM ile ilgili olarak literatürdeki çalışmaların bir kısmının derlenmesi,
- VM sürecinin açıklanması,
- Süreç adımlarının kullanılarak kaçak su kullanımının tespitine yönelik bir model oluşturulmasıdır.

1.2. Tezin Kapsamı

Tez çalışması beş bölümden oluşmaktadır.

Bölüm II' de veri ile ilgili kavramların tanımı, veri madenciliğinin tanımı, amaçları, kullanım alanları, gelişme nedenleri, karşılaştığı temel problemler, bu konuda yapılmış çalışmalar, veri madenciliği modelleri ve sürecin aşamaları açıklanmaktadır.

Bölüm III' de süreç aşamalarından modelleme aşamasına kadar olan kısmına yer verilmiştir.

Bölüm IV' de modelleme aşamasına yer verilmiş ve son bölümde ise modelden elde edilen sonuçlar değerlendirilmiş ve çalışmada yer alan modelin daha sonra kullanılabilirliğini artırmak amacıyla bazı önerilerde bulunulmuştur.

BÖLÜM 2: VERİ MADENCİLİĞİ

2.1. Veri İle İlgili Temel Kavramlar

2.1.1. Veri kavramı

Veri; kendi başına değersiz, istediğimiz amaç doğrultusunda bilgidir. Bilgi ise bir amaca yönelik işlenmiş veridir. Bir diğer ifade ile bilgi, bir soruya yanıt vermek için veriden çıkardığımız sonuç olarak tanımlanabilir[1].

Veri bir kişinin formülleştirmeye veya kayıt etmeye değer bulduğu her şey olarak da tarif edilebilir. Veriyi tanımlamak için çok farklı kavram seçeneği mevcuttur. Bu kavramlar aşağıdaki gibi sıralanabilir[2]:

– Veri (Data): Herhangi bir özel anlam içermeyen, kayıt edilebilen, sınıflandırılabilen, depolanabilen, bir bilgi sistemine girilen, yapısal olmayan, işlenmemiş girdiler, nesneler, aktiviteler, işlemlerin tümüne denir. Veri; sayılar, harfler ve onların anlamıdır. Veri hakkındaki bu veriye 'meta data' denir.

– Byte: En küçük adreslenebilir birim olan "bit"in 8 adedinin oluşturduğu bütündür.

– Veri Parçası: Alan veya veri elementi olarak da tanımlanabilecek veri parçası bir veya birden fazla byte'dan oluşan en küçük kimliklendirilmiş veridir.

– Veri Toplamı: Veri toplamı bir kayıt içerisindeki veri parçalarının birleşiminden oluşan bir bütündür.

– Kayıt: Kayıt, veri toplamlarının oluşturduğu bir bütündür.

– Kısım: Kısım terimi kayıt ve veri toplamı gibi veri bölümünü tarif eden iki tanımın gereksiz olduğuna inanan IBM gibi firmaların geliştirdiği bir kavramdır. Bu kavram kayıt ve veri toplamını kapsamaktadır.

– Dosya: Dosya, kayıtlar bütünüdür.

– Veri Tabanı: Veri parçaları, veri kayıtları ve bu kayıtlar arasındaki ilişkileri içeren bir bütündür.

– Bilgi (Information): Herhangi birine söylendiğinde bireyin kafasında söylenen bu ifadeye ait bir anlam uyandıran, karar alma aşamalarında verilerin işlenip anlamlı hale getirilerek kullanıcıya sunulmuş halidir. Veri bilginin hammaddesidir. Veriyi bilgiye çevirmeye "veri analizi" denir.
– Kurumsal Bilgi-Çıkarımı (Knowledge): Belirli bir amaca yönelik olarak bilginin çeşitli analiz, sınıflama ve gruplama işlemlerinden geçirilerek, gerektiği zamanlarda potansiyel olarak kullanıma hazır hale getirilmesidir. Türkçede günlük kullanımda bilgi sözcüğü ile hem 'Information', hem de 'Knowledge' ifade edilmekte olduğundan ve henüz kurumsal bilginin (çıkarımın) örgüt içinde kullanımı yaygınlaşmadığından, kavramların ifade edilmesi sırasında güçlükler yaşanmaktadır.

2.1.2. Veri kaynakları

– İçsel veri: Bu tip veriler insanlar, ürünler, servisler ve prosesler ile ilgilidir. Örneğin işçilere ait ödemeler muhasebe bölümünde, malzeme ve makineler ile ilgili veriler imalat bölümünde tutulmaktadır[2].
– Dışsal veri: Bu tip veriler uydular ve algılayıcılardan toplanan ticari verilerdir. Cd sürücülerden, internetten, film müzik veya seslerden, resimlerden, televizyondan, grafik ve diyagramlardan elde edilen veriler bu kategoriye girer. Hükümet raporları, yerel bankalar, enstitüler, özel şirketler de önemli dışsal veri kaynaklarıdır[2].
– Personel Verisi: Nesnel satış tahminleri, rakiplerin neler yapabileceği ile ilgili fikirler, şirkete özgü haber portalları gibi işletmenin kendi uzmanlık bilgileriyle bir araya getirdikleri verilerdir[2].

2.1.3. Veri modelleri

Veri modeli, veriyi bir kurala göre yapılandırma şeklidir. Bu yapılandırma içerisinde iki unsur bulunur. Bu unsurlar; yapı ve işlemlerdir. Yapı; sistemin veriyi yapılandırma şeklidir. İşlemler ise kullanıcıların veri tabanındaki veriyi düzenleme imkânlarıdır. Tüm özellikler bir model tarafından yansıtılamaz. Eğer bir model uygun olarak formülleştirilmişse kullanıcıların ihtiyaçlarını karşılayabilir. Modellerin eksiklikleri iki grup altında toplanabilir. Birincisi, veri yapısının bir bölümünün temsil edilmemesi ve ikincisi çeşitli yollarla veri yapısı üzerinde değişiklik

yapılamamasıdır. Bir veri modeli, verinin hangi kurallara göre yapılandırılacağını belirler. Fakat yapılar verinin anlamı ve nasıl kullanılacağı hakkında tam bir fikir vermezler. Veri modeli veri tabanında bulunan verilerin mantıksal organizasyonunu belirleyen kurallar kümesi olarak tanımlanabilir. Veri modelleri ikiye ayrılır[2];

2.1.3.1. Basit veri modelleri

Basit veri modellerindeki amaç, verinin basit, anlaşılabilir bir yapıya sokulmasıdır. Bunlar genel yapılardır. Basit veri modelleri daha çok programlamaya dayalı bir veri modelidir. Dosyalama sistemleri oluşturmak amacıyla kullanılmaya başlanan veri modelidir. Aynı zamanda bilgisayarlarda veri işleme ihtiyacının ortaya çıkması ile dosyalama sistemleri oluşturmak amacı ile kullanılmaya başlanan veri modelleridir. Basit veri modelleri; hiyerarşik veri modeli ve ağ veri modeli olmak üzere ikiye ayrılmaktadır[2].

1. Hiyerarşik veri modeli: Hiyerarşik veri modeli bir ağaç yapısı şeklindedir. Ayrıca hiyerarşi sıralamasında üstteki varlıklar ebeveyn, alttakiler ise çocuklar olarak isimlendirilir. Hiyerarşik modelleme tekniği varlıklar arasında bire çoklu ilişki tiplerinin bulunduğu verilerin modellenmesi esnasında kullanılır. Bu teknikteki 1 kısmındaki kayıt tiplerine baba, n kısmındaki kayıt tiplerine oğul adı verilir. Oğullarında oğulları tanımlanabiliyorsa düğüm adını alır.

2. Ağ veri modeli: Hiyerarşik veri modelinin basit yapılı olmasına rağmen tek bir kökün olmadığı durumlarda modellemede sorunlar çıkmaktadır. Aynı zamanda ilişki tipleri ikili, yani varlık arasında kurulan birebir ilişki söz konusudur. Ağ veri modeli iki varlık arasında bire çoklu ilişkiden oluşan küme kavramını kullanır. "Bir" tarafında olan varlık kümenin sahibi, "Çok" tarafında olan varlık ise kümenin üyesidir. Bir üye başka bir kümenin sahibi olabilir. Fakat bir varlık aynı tipte iki kümeye birden üye olamaz. Buna karşılık bir üye aynı tipte olmayan veya daha fazla kümeye sahip olabilir. Ağ veri modelleri, tablo ve grafik temellidir. Grafikteki düğümler varlık tiplerine karşılık gelir ve tablolar şeklinde temsil edilir. Grafiğin okları, ilişkileri temsil eder ve tabloda bağlantılar olarak temsil edilir.

2.1.3.2. Geliştirilmiş veri modelleri

Var olan bir verinin üzerinde bilgisayar kullanarak işlem yapabilmek için o verinin bilgisayarda işlenmesi yeterli değildir. Burada aynı zamanda kullanıcıların ve veri üzerindeki işlem yapacak analist1erin bakış açıları da çok önemlidir. Tüm kullanıcıların farklı bakış açılarının bütün1eşik bir model ile veri tabanına yansıtılması veri modeli oluşturmaktadır. Geliştirilmiş veri modelleri; varlık-ilişki veri modelleri, ilişkisel veri modelleri ve nesne yönelimli veri modelleri şeklinde sıralanabilir[2].

Varlık-ilişki veri modeli: Varlık-ilişki işlemi, analizler ve şemalandırma için önemli bir tekniktir. Organizasyonun veri ve gereksinimlerinin yukarıdan aşağıya planlamasında kullanılır. Bu şema, işletme açısından önemli olan iş varlıklarının gösterildiği bir grafiktir. Varlık gerçek veya soyut, kesin, görülebilir veya görülemez olabilir. Görülebilir varlıklara müşteri, çalışan, fatura ve bölüm örnek olarak verilebilir. Görülemez varlıklara ise olay, iş adı, zaman periyodu ve kazanç merkezi örnek verilebilir. Kayıt etmek istenilen bir varlık, renk, boyut, maddi değer, yüzdelik değerlendirme, adres, maaş, tarih, kod veya cinsiyet gibi özniteliklere sahip olabilir. Varlıklar arasındaki ilişkilerin üç önemli çeşidi vardır. Bunlar[2];
Bire bir ilişki: Bir varlıktan diğerine bire bir ilişkiler, birinci varlığın her bir değeri ikinci varlığın sadece bir değeri ile eşleşir.
Bire çoklu ilişki: A varlığından B varlığına bire çoklu ilişki, A varlığının bir değeri B varlığının sıfır bir veya birçok değerleriyle herhangi bir zamanda ilişkilendirilmiş olduğu anlamına gelir.
Çoklu ilişki: Bazı durumlarda, bir varlık-ilişki şemasında çoklu ilişkilere ihtiyaç duyulur.

İlişkisel veri modeli: İlişkisel veri modeli tablolardan oluşur. Tablolar ilişki olarak isimlendirilir. Tablolar arasında ortak olan sütunlar ile ilişkiler sağlanmış olur. Tablolar iki boyutludur, satır ve sütunlardan oluşur. Tablolarla ilgili bir takım kurallar vardır; her sütunun kendine özgü bir ismi olmalıdır ve o sütundaki veriler sütun ismi ile uyumlu olmalıdır. Aynı şekilde her satırda bir diğerinden farklı olmalıdır. İlişkisel modelde her şey özellikleri tanımlayan sütunlar ve nesneleri veya kişileri tanımlayan bilgilerin yer aldığı satırlardan oluşan basit bir tablodur[2].

Nesne yönelimli veri modeli: Nesne yönelimli veri modeli ilişkisel modelle karşılaştırıldığında yüksek seviyeli bir modeldir. Çünkü nesne yönelimli veri modeli ilişkisel modelde zor olan hiyerarşiler gibi yapılandırılmaları hızlandırmaktadır. Nesne yönelimli veri modelini önemli kılan bir başka özellik ise verilerin harmanlanması için özel bir yapı sunmasıdır. Nesne yönelimli veri modelinde her şey bir nesnedir. Nesne yönelimli sistemler farklı sistemler ve metodolojiler için kullanılmıştır. Genel olarak bu sistemler gerçek dünyadaki objeleri nesne denilen varlık şeklinde modellemeyi temel almaktadır. Nesneler ortak karakteristikler içeren nesnelerin bulunduğu sınıflar içerisinde gruplandırılırlar[2].

2.1.4. Veri tabanları

Veri tabanı sistemleri, bir veya daha fazla uygulamaya hizmet vermek için bir araya toplanmış birbirleriyle ilişkili veriler toplamıdır[5]. Veri tabanı, sistematik erişim imkânı olan, yönetilebilir, güncellenebilir, taşınabilir, birbirleri arasında tanımlı ilişkiler bulunabilen bilgiler kümesidir[3].

Veri tabanı (VT) sadece verinin alınması değil aynı zamanda o veri üzerinde değişiklik yapılmasına da imkân vermektedir. Veri tabanı bilgisayarda veri depolamak ve işlemek amacıyla kullanılmaktadır. VT, çeşitli tiplerdeki varlıklara, bu varlıkların özniteliklerine ve bunlar arasındaki ilişkilere ev sahipliği yapan bir yapıdır. Bir veri tabanında soyutlama katmaları kullanılarak gerçek dünyanın kavramları bilgisayar ortamına adapte edilebilmektedir. Fiziksel veri tabanı, disk üzerinde bulunan dosya ve indeks koleksiyonu ve bunlara ulaşmak için kullanılan depolama yapılarıdır. Kavramsal veri tabanı, gerçek hayatın bir soyutlamasıdır. Bu soyutlamayı gerçekleştirmek için veri tabanı yönetim sistemi, bir veri tabanı tanımlama dili kullanır. Veri tanımlama dili kavramsal veri tabanını veri modeli olarak tanımlayabilmemizi sağlar. Kavramsal veri tabanı, organizasyon tarafından kullanılan verinin bütünü temsil eder[2].

Bununla birlikte veri tabanı sisteminin kurulum ve bakımının zor ve pahalı olması ve bütünleşik sistemdeki bir bölüm veriye ulaşılamamasının tüm sistemin çalışmamasına sebep olması gibi dezavantajları da bulunmaktadır. İşletmeleri veri

tabanı yaklaşımına götüren pek çok problem mevcuttur. Bunlardan bazıları şunlardır[2]:

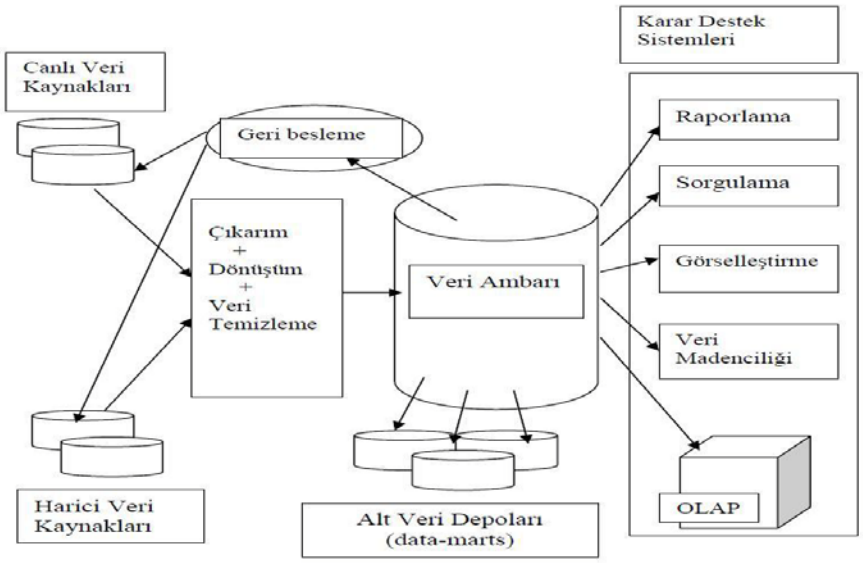
- Basit ihtiyaçlara çabuk yanıtlar alınamaması.
- Düşük veri kalitesi ve doğruluğu
- Değişime hızlı ayak uyduramama
- Yüksek gelişim maliyetleri
- Gerçek dünya için geçersiz veri modeli kullanımı.

Veri tabanı sistemlerinin başlıca üç özelliği vardır:
Özerklik: Bir veri tabanı diğer veri tabanlarıyla etkileşimde olmak için kendi kontrol politikasını oluşturabilir.
Heterojenlik: Veri modelleri, sorgulama dilleri, veri tabanından veri tabanına farklılık gösterebilir.
Dağıtım: Fiziksel olarak farklı ortamlarda yerleşmiş bulunan veri tabanları.

VT kullandıkları veri modellerine göre hiyerarşik veri modellerini kullanan hiyerarşik veri tabanları, ilişkisel veri modellerini kullanan ilişkisel veri tabanları ve nesne yönelimli veri modellerini kullanan nesne yönelimli veri tabanları olmak üzere üç kısma ayrılır[2].

2.1.5. Veri ambarları

Günümüz yöneticileri çok değişken olan iş dünyasında satışlardan rakiplerine, müşterilerden yürütülen projelere kadar her türlü bilgiye her zamankinden daha hızlı ve doğru olarak ihtiyaç duymaktadırlar. Hiçbir şey tam zamanında elde edilmiş verinin yerini tutamamaktadır. Bu bilgiyi elinde tutan, güce de sahip demektir. İşte veri ambarları bilginin güce dönüştürülmesinde bir aracıdır. Veri Madenciliği sık sık veri ambarlarıyla karıştırılmaktadır. En basit anlamda veri madenciliği ve veri ambarları, birbirlerinin tamamlayıcısıdır. Veri ambarları verinin belli bir yapıda saklanması için kurulurken, veri madenciliği bu saklanan verinin bilgiye dönüştürülmesini sağlar. Kısaca veri ambarları, veri madenciliğinin omurgası gibidir[4]. Bu ilişki Şekil 2.1.'de görselleştirilmiştir.



Şekil 2.1. Veri Ambarı (VA) mimarisi[7]

Veri ambarcılığı çeşitli şekillerde tanımlanmıştır. Veri ambarcılığının babası sayılan Bill Inmon veri ambarını 1992'de "Veri ambarı (VA), yönetimin karar sürecini desteklemede kullanılan, konuya yönelik, entegre, zamana bağlı, kalıcı veri topluluğudur." şeklinde tanımlamıştır. Başka bir tanıma göre ise veri ambarı, basitleştirilmiş biçimde hareket sistemlerinden özetlenen ve kümelenen verinin saklandığı yerdir[2,4].

VA, iş dünyasında bilgiye hemen ulaşmak amacıyla karar vericiler için tasarlanmış bir bilgisayar sistemidir[3].

VA, operasyonel, kalıcı, entegre ve tarihsel derinliği olan verilerin, karar destek sisteminin işlevini desteklemek, verilerden anlamlı ilişkiler kurarak sonuçlar çıkarmak üzere modellenmiş süreçlerin toplamıdır. Böylelikle veriler, organizasyondaki karar vericilerin faydalanmaları için saklanarak veriye hızlı ve tek kaynaktan ulaşmaları imkânı sağlanmaktır. En basit tanımıyla veri ambarı, OLTP (Online Transaction Processing - Çevrimiçi İşlem Süreci) veri tabanından çıkarılan operasyonel verinin depolandığı merkezdir[5].

Bir başka tanıma göre VA; operasyonel veri tabanından, içsel ve dışsal kaynaklardan gelen, entegre edilmiş, temiz, arşivlenmiş, büyük hacimli verilerin yönetim tarafından karar vermeyi destekleyecek ve kullanıma olanak sağlayacak şekilde

derlendiği depolama alanlarıdır[5].

VA, bir işletmenin veya kamusal bir kurumun değişik bölümleri tarafından toplanan bilgilerin, gelecekte değerlendirilmek üzere arka plandaki sistemlerde birleştirilmesinden oluşan geniş ölçekli veri deposudur. Veri ambarları müşteri, tedarikçi, ürün bilgisi, stok, alış ve satış verisi gibi önemli özneler üzerine kurulur ve veriler veri ambarlarına tarihi bir bakış açısından bilgi sağlamak için depolanır[6].

Teknoloji boyutu öne çıkarılarak yapılan bir tanımda VA; bilgiyi kullananların daha iyi ve daha hızlı karar vermelerini amaçlayan teknolojilerin bütünü olarak tanımlanmıştır. Bir başka tanımda ise VA; bir kurumda gerçekleşen tüm operasyonel işlemlerin, en alt düzeydeki verilerine kadar inebilen, etkili analizler yapabilmesi amacıyla özel olarak modellenen, tarihsel derinliği olan ve operasyonel sistemlerden fiziksel olarak farklı ortamlardaki yapılar üzerinde gerçekleşen süreçler toplamı olarak tanımlanmıştır[4].

Veri ambarı, bir işletmenin ya da kurumun çeşitli birimleri tarafından canlı sistemler aracılığı ile toplanan verilerin, ileride değerlendirmeye alınabilecek olanlarının geri planda yer alan bir sistemde birleştirilmesinden oluşan büyük ölçekli bir veri deposudur. Günümüzün ticari işletmelerinde bilgi sistemleri iki ayrı başlık altında toplanmaktadır. Bunlar[1];

- Canlı Sistemler: Bu sistemlerde güncel veriler bulunur. Günlük yapılan işleri ve işlemleri gerçekleştirmek, sonuçları saklamak bu sistemlerin görevidir. Bu sistemler, marketlerde ya da mağazalarda stok takibi, üye borçları, satış işlemleri, ödeme kayıtları gibi bilgilerin işlendiği ve tutulduğu bilgi sistemleri olabilir[1].
- Karar Destek Sistemleri: İşletmelerde yer alan ikinci tür bilgi sistemleri ise karar destek sistemleridir. Bu sistemlerde yer alan bilgiler, çeşitli incelemelerden ve araştırmalardan geçerek, işletmelerin ileride karını ya da verimliliğini arttırması, gelecekte izlenecek politikalarının belirlenmesi gibi yönetimsel kararların alınmasına yardımcı olur ve bu kararların daha doğru verilmesini kolaylaştırır. Bu sistemlerde verilerin erişimi asıl amaç değildir. Karar destek sistemlerinin önceliği performanstır. Karar destek sistemlerinde veriler, canlı sistemlere oranla çok daha büyük boyutlardadır. Verilerin büyük

boyutundan dolayı, verilerin incelenmesi ve incelemelerden sonuçlar çıkartılması, sistem kaynaklarını aşırı kullanmakta ve uzun süre almaktadır. Veri ambarı, karar destek sistemi olarak nitelendirilebilir[1].

İşletmelerde kullanılan üç çeşit veri ambarı vardır[2]:

- Tüm kuruma hizmet eden kurumsal (geleneksel) veri ambarı,
- İşletmedeki belirli bir iş birimini veya bölümü desteklemek üzere tasarlanmış minyatür bir veri ambarı olan veri pazarı (data mart),
- Veri ambarı tekniklerinin hareket sistemlerine uyarlandığı operasyonel veri deposu.

2.1.5.1. Veri ambarının karakteristik özellikleri

- Konuya Yönelik Olma: Operasyonel veri ihtiyacı, uygulamanın anlık ihtiyaçları ile ilgilidir ve o anda geçerli iş kurallarına dayanır. Veri ambarı dünyası ise müşteri, mal veren, ürün ve etkinlik gibi temel konular etrafında organize olur. Veri ambarındaki veri karar vermeye yöneliktir ve zaman derinliği çok daha fazla olduğundan daha karmaşık ilişkilere imkân tanır[2].
- Bütünleşik yapı: Sitemlerden veri ambarına veri aktarılırken veri entegre edilir ve hepsi aynı formata getirilir. Böylece değişik kaynaklardan gelen veri, veri ambarında tek ve genel olarak üzerinde anlaşmaya varılmış bir şekilde yer alır. Veri ambarındaki veri, temiz, geçerliliği onaylanmış ve uygun biçimde kümelenmiş olmalıdır[2].
- Kalıcı Ortam: Operasyonel veri tabanlarından gelen veriler güncellenmeden veri ambarına giremezler, güncellenip veri ambarına girdikten sonra ise eski verinin güncellendiği anlamını taşımazlar sadece veri ambarında kronolojik olarak yerlerini alırlar. Eski veriler ise yerlerini muhafaza etmeye devam ederler[5].
- Zamana Bağlı Olma: Veri ambarındaki veri referans alınan zaman birimi ile birlikte kaydedilir ve veri bir kez doğru biçimde kaydedildikten sonra kullanıcılar tarafından güncellenemez. Veri ambarındaki veri tipik olarak 3-10 yıllık bir zaman dilimini kapsar[2].

2.1.5.2. Veri ambarının yapısı ve hedefleri

Veri ambarları farklı tipte verilerden ve Tablo 2.1.'de belirtilen hedeflerden oluşmaktadır.

Geçerli Detay Veri: Geçerli (güncel) detay veri, en çok ilgilenilen en son olayları gösterir. Bu veri en düşük atomiklik seviyesinde depolandığından oldukça büyük hacimlidir. Geçerli detay veri çoğu zaman erişimi oldukça hızlı fakat pahalı ve yönetimi oldukça karışık olan disk depolarında depolanmaktadır. Geçerli detay veri genellikle operasyonel sistemlerde şu anda mevcut olan operasyon verilerinin uygun biçimde veri ambarına aktarılmış halidir[4].

Eski Detay Veri: Eski detay veri, aynı seviyede depolanan geçerli detay veri ile tutarlı fakat daha az erişilen veridir. Veri ambarlarının çoğunda, tutulan detaylı veriler belli bir yaşa ulaştıkları zaman diskten daha büyük bir veri saklama ortamına gönderilmesini öngören kurallar bulunmaktadır[4].

Az Özetlenmiş Veri: Veri ambarı kullanıcılarının yapabileceği bazı analiz ve sorgular için istenebilecek standart değerleri önceden özetlemek veri ambarından daha hızlı cevap alınmasını ve performansın iyileşmesi ile birlikte daha fazla kullanılmasını sağlamaktadır[4].

Çok Özetlenmiş Veri: Çok özetlenmiş veri yoğundur ve kolayca erişilebilir. Karar vermek için gerekli veri çoğunlukla çok özetlenmiş veriler kullanılarak elde edilmektedir. Üst düzey yöneticilerin ihtiyaç duydukları bazı bilgiler yoğun ve kolayca erişilebilir olmalıdır[4].

Meta data (Veri Bilgisi):Veri hakkında veri anlamına gelen meta data; belirli bir grup verinin, kim tarafından, ne zaman, nasıl toplandığını ve verinin nasıl biçimlendirildiğini tanımlar. VA' da toplanan bilginin anlaşılabilmesi için meta veri gereklidir. Veri ambarının en önemli bileşenlerinden birisidir ve veri ambarını tanımlayan veridir. Meta data şu şekilde sınıflandırılabilir[4]:

- Teknik meta veri: Veri ambarı tasarımcılarının ve yöneticilerinin işlemlerini yerine getirirken kullandıkları veridir.
- Ticari meta veri: Kullanıcıya veri ambarındaki verinin kullanılmasında kolaylık sağlayan veridir.

– Veri ambarının kendi işlemleriyle ilgili meta veri: Bunlar veri ambarı versiyonları, denetim işlemleri, yedekleme ile elde edilen verilerdir.

Tablo 2.1. Veri Ambarının Hedefleri[4]

Uygulama Hedefleri	**Bilgi Hedefleri**	**Meta Data Hedefleri**
Karar Destek	Erişebilirlik	İş tanımlarının yapılması
Tahmin Modelleme	Tutarlılık	İş kurallarının tanımlanması
Planlama	Güvenlik	Bilgi uyumunun yürütülmesi
	Şartlara ve çevreye uyma yeteneği	

2.1.5.3. Veri ambarı ihtiyacı

Bir işletmenin büyüklüğü veri ambarı ihtiyacının bir ölçüsü değildir. İşletmenin bir veri ambarına ihtiyacı olup olmadığına karar verirken ise bazı anahtar göstergelere bakarak başlanabilir. Bu göstergelerden bazıları şunlardır[2]:

– İşletme değişken ve rekabetin çok yoğun olduğu bir pazarda faaliyet göstermesi,
– Müşteriler hakkında sağlıklı bilgi elde etme ihtiyacının olması,
– Kazanç sağlayacak ve/veya verimliliği arttıracak bilgiye dayalı ürünler veya hizmetler oluşturma fırsatlarının olması,
– Sık kullanılan ve birbiriyle ilişkili kurumsal verinin birçok değişik yerde ve farklı sistemlerde bulunması,
– "Aynı veri ama farklı sonuç" şeklindeki sorunun işletmede sürekli bir rahatsızlık haline gelmiş olması,
– Gerçek karar destek sistemlerine ihtiyacın olması,
– Kullanıcıların daha etkili ve anlık sorgulama ve raporlama istemeleri,
– Bir bilgi dağıtımı alt yapısına ihtiyaç olması.

VA finans(bankalar, sigorta şirketleri, leasing, factoring ve borsa şirketleri), üretim, ulaşım, iletişim, perakendecilik ve kamu (vergi dairelerinde) sektörü gibi pek çok sektörde kullanılmaktadır[2].

Gelecek kuşak VA uygulamalarında ise her düzeyde müşteri ilişkisini düzenlemek için gerçek zamanlı analiz yöntemleri gerekecektir. Bugünün rekabetçi ortamındaki müşteri ilişkileri yönetimi veri ambarı uygulamalarını bire-bir ilişkileri düzenlemek

için yapılanma yönüne kaydıracaktır. Müşteriyle etkileşim analitik karar destek sistemleriyle birleşerek 'etkin veri ambarı' çözümlerine olgunluk kazandıracaktır[2].

2.1.5.4. Veri ambarı yönetimi

Veri ambarları geçmişe yönelik birçok yıllık veriyi kapsadıkları için işlevsel veri tabanlarından yaklaşık olarak 4 kat daha büyüktür. Bu yüzden gerçek zamanlı olarak güncellenmesi çok zordur. Ancak üzerinde çalışılan uygulamaların durumuna göre günde en az bir kez yenilenmelidir. Bir veri ambarının yönetimi şunları kapsamaktadır[4];

- Güvenlik ve önceliklerin belirlenmesi
- Çeşitli kaynaklardan gelen verinin incelenmesi
- Veri kalitesinin kontrolü
- Meta verinin yönetimi ve veri güncellenmesi
- Veri ambarının durumunun kontrol edilmesi ve raporlanması
- Verinin düzenlenmesi
- Verinin yedeklenmesi (backup) ve başlangıca döndürme (restore) işlemleri
- Veri ambarının depolama yönetimi

2.1.6. Veri ambarı ile veri tabanının karşılaştırılması

Veritabanı içerisindeki bilgiler genelde anlık bilgilerdir. Yani belirli bir süre sonunda güncelliğini kaybedecek olan bilgilerdir[1].

Veri tabanlarından beslenen veri ambarları ise, verileri depolamaktadır. Depolanan veriler güncel olmasalar dahi geçerlilikleri daha uzun sürmektedir. Veri tabanları ile veri ambarlarını tutukları kayıt sayısına göre değerlendirmek gerekirse, veri ambarlarında ne kadar çok veri tutulursa yapılan analizler o kadar gerçeğe yakın çıkacaktır[7].

Diğer taraftan, veri tabanındaki kaydın artması canlı sistemlerin kullanımını etkileyecek ve verilere erişim yavaşlayacaktır. Canlı sistemlerin yavaşlaması hiçbir işletmenin istemediği bir durumdur[1,7].

2.2. Veri Madenciliğinin Tanımı

İşletmelerde ve devlet kurumlarında 90'lı yılların başından itibaren bilgisayar sistemlerinin yaygınlaşması ile her türlü veri farklı depolama alanları içinde hızla büyüyen boyutlarda saklanmaya başlamıştır. Zamanla kurum ve işletmelerin mevcut veriler üzerinde yaptıkları çalışmalar ile elde edilen sonuçlar geleceğe yönelik planlamada kullanılarak kazanç elde etme çalışmaları artmıştır. Daha fazla değerli veriyi toplama çalışmaları önde gelen amaçlardan biri olmuştur[6].

Önceden istatiksel veriler ile devasa büyüklükteki veri tabanlarından işe yarayacak örüntüler bulmak için istatiksel metot ve yöntemler kullanılırdı ki bunların sonucunda oluşan verilerin incelenmesi için uzman kişilere gerek vardı. İstatistiksel yaklaşımların kullanımında bu paketlerin dezavantajları ortaya çıkmaktaydı. Başka bir dezavantajı ise her farklı ihtiyaç için bu işlemlerin tekrarlanmasıydı[8].

Veri Madenciliğini (VM) istatistik yöntemlere üstün kılan özelliği, çok fazla miktarda veriyle çalışabilir olmasıdır. İstatistikte, ana kütleden seçilen bir örneklem üzerinde çalışarak genelleştirme yapılmaya çalışılır. Fakat bu durumun gelecekteki isletme ihtiyaçlarını tam olarak karşılayamama, iş çevresindeki gelişmelere ve değişimlere cevap verememe gibi eksik yönleri vardır. Bazen veri madenciliği teknikleriyle daha basit, ayrıntılı ve uygulanabilir kararlar alınabilmektedir[3].

Otomatik veri toplama araçları ve veri tabanı teknolojilerindeki gelişme, veritabanlarında, veri ambarlarında ve diğer bilgi depolarında çok miktarda bilgi depolanmasına sebep olmaktadır. Büyük miktarlardaki veri içindeki gizli örüntülere, değerli bilgilere geleneksel çözümleme araçlarıyla ulaşmak oldukça zordur. Dolayısıyla toplanan veri miktarı büyüdükçe ve toplanan verilerdeki karmaşıklık arttıkça, daha iyi çözümleme tekniklerine olan gereksinim artmakta ve veri madenciliği uygulamaları alternatif bir çözüm olarak karsımıza çıkmaktadır. Veri madenciliği uygulamaları; ilişkisel veritabanları, veri ambarları, gelişmiş veritabanları ve bilgi depoları (nesne kaynaklı, nesne ilişkili, uzamsal, metin, çoklu ortam, heterojen veritabanları, zamansal veriler ve WWW) üzerindeki veriler

üzerinde gerçekleştirilmektedir. Bu bağlamda veri madenciliğinin üç türünden söz etmek mümkündür[9]:

- Doğrudan veri madenciliği: Veri ambarındaki verilerin doğrudan kullanıldığı yöntemdir. Bir doktorun hastasının kapalı damarlarını bulmak için görüntüleme cihazlarından faydalanması doğrudan veri madenciliğine bir örnektir.
- Varsayım deneme ve varsayımı daha iyi hale getirme: Bu yöntemle kullanıcı çalıştığı konuya ilişkin bazı varsayımlar üretir ve bu varsayımların sistem tarafından doğrulanmasını, değiştirilmesini veya daha uygun hale getirilmesini amaçlar.
- Dolaylı ya da saf veri madenciliği: Veri madenciliği türlerinin içerisindeki en genel yöntemdir. Hiçbir kısıtlama ve kullanıcıların bulacağı bilginin türü hakkında belli bir beklenti yoktur. Bu aynı zamanda en güç yöntemdir.

Veri madenciliği; verideki trendleri, ilişkileri ve profilleri belirlemek için veriyi sınıflandıran bir analitik araç ve bilgisayar yazılım paketidir. Spesifik veri madenciliği yazılımları; kümeleme, doğrusal regresyon, sinir ağları, Bayes ağları, görselleştirme ve ağaç tabanlı modeller gibi pek çok modeli içerir. Veri madenciliği uygulamalarında yıllar boyu istatistiksel yöntemler kullanılmıştır. Bununla birlikte, bugünün veri madenciliği teknolojisinde eski yöntemlerin tersine büyük veri kümelerindeki eğilim ve ilişkileri kısa zamanda saptayabilmek için yüksek hızlı bilgisayarlar kullanılmaktadır. Böylece veri madenciliği, gizli trendleri minimum çaba ve emekle ortaya çıkarmaktadır[10].

Farklı çalışmalarda Veri madenciliğine (VM) ilişkin temelde aynı olmakla birlikte farklı tanımlar yapılmıştır. Aşağıda veri madenciliğinin farklı tanımlarından bazıları belirtilmiştir.

Hand (1998), veri madenciliğini istatistik, veritabanı teknolojisi, örüntü tanıma, makine öğrenme ile etkileşimli yeni bir disiplin ve geniş veritabanlarında önceden tahmin edilemeyen ilişkilerin ikincil analizi olarak tanımlamıştır[10].

Veri Madenciliği; geniş veritabanlarından bilgi çıkartabilmek amacıyla makine öğrenmesi, örüntü tanıma, istatistik, görselleştirme gibi alanların tekniklerini bir araya getiren disiplinler arası bir alandır[11].

Veri Madenciliği(VM), büyük miktarlardaki verinin içinden geleceğin tahmin edilmesinde yardımcı olacak anlamlı ve yararlı bağlantı ve kuralların bilgisayar programlarının aracılığıyla aranması ve analizidir[12].

Gartner Grup tarafından yapılan tanımda ise veri madenciliği, istatistik ve matematik tekniklerle birlikte ilişki tanıma teknolojilerini kullanarak, depolama ortamlarında saklanmış bulunan veri yığınlarının elenmesi ile anlamlı yeni ilişki ve eğilimlerin keşfedilmesi sürecidir[11,12].

Diğer bir tanımlama ise "Veri ambarlarında tutulan çok çeşitli ve çok miktarda veriye dayanarak daha önce keşfedilmemiş bilgileri ortaya çıkarmak, bunları karar verme ve eylem planını gerçekleştirmek için kullanma sürecidir"[2].

Veri madenciliği, temel olarak bilgisayar destekli bir bilgi çözümleme işlemidir. VM, ayrı sorgular vererek büyük miktarda olan veriden yararlı bilgi, desenler ve eğilimler çıkarabilmektir. VM, verinin sahibine anlamlı ve yararlı olacak şekilde veri kümesinin içinde şüphe uyandırmayan ilişkileri bulmak ve veriyi yeni bir şekilde özetlemek için veri kümelerinin incelenmesidir[12].

Jacobs (1999), veri madenciliğini, ham datanın tek başına sunamadığı bilgiyi çıkaran veri analizi süreci olarak tanımlamıştır. David (1999), veri madenciliğinin büyük hacimli datalardaki örüntüleri araştıran matematiksel algoritmaları kullandığını söylemiştir. DuMouchel (1999), veri madenciliğinin geniş veri tabanlarındaki birliktelikleri araştırdığını belirtmiştir. Kitler ve Wang (1998), veri madenciliğini oldukça tahminci anahtar değişkenlerin binlerce potansiyel değişkenden izole edilmesini sağlama yeteneği olarak tanımlamışlardır. Bransten (1999), veri madenciliğinin insanın asla bulmayı hayal bile edemeyeceği trendlerin keşfedilmesini sağladığını belirtmiştir[10].

Tüm bu tanımlardan sonra veri madenciliğini istatistiksel yöntemler serisi olarak görmek mümkün olabilir. Ancak veri madenciliği, geleneksel istatistikten birkaç yönde farklılık gösterir. Veri madenciliğinde amaç, kolaylıkla mantıksal kurallara ya da görsel sunumlara çevrilebilecek nitel modellerin çıkarılmasıdır. Bu bağlamda, veri madenciliği insan merkezlidir ve bazen insan–bilgisayar ara yüzünü birleştirir. Veri madenciliği sahası, istatistik, makine bilgisi, veri tabanları ve yüksek performanslı işlemler gibi temelleri de içerir[2].

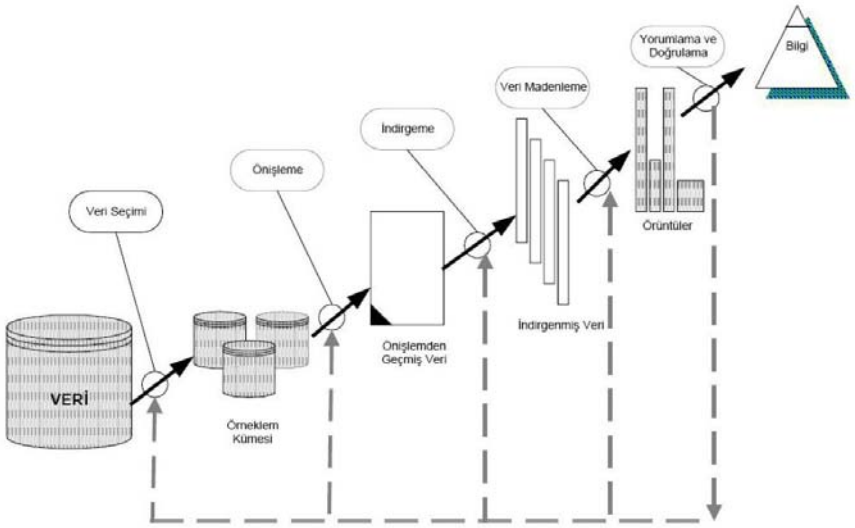
Genel olarak veri madenciliği, bir veri ambarına ve bir yazılım paketine gereksinim duyar. Diğer temel gereksinimleri şu şekilde sıralanabilir:

- Veriye erişilebilirlik
- Etkin erişim yöntemleri
- Veri problemlerinde dinamiklik
- Etkin algoritmalar
- Yüksek performanslı uygulama sunucusu (server)
- Sonuç dağıtımında esneklik
- Verinin temizlenmesi[10,13].
- Çok miktarda, güvenilir veri ön şarttır. Çözümün kalitesi öncelikle verinin kalitesine bağlıdır.
- Uygulama ile ilgili ve yararlı olabilecek her tür bilginin öğrenmeye yardım için sisteme verilmesi gerekmektedir.
- Sonuçların tutarlılığının uzmanlar tarafından denetlenmesi gerekir.
- Veri madenciliği tek aşamalı bir çalışma değildir, tekrarlıdır. Sistem ayarlanana dek birçok deneme gerektirir[5].

Etkin bir veri madenciliği uygulayabilmek için dikkat edilmesi gereken noktalar aşağıdaki gibi özetlenebilir[2];

- Farklı tipteki verileri ele alma
- Veri madenciliği algoritmasının etkinliği ve ölçeklenebilirliği
- Sonuçların yararlılık, kesinlik ve anlamlılık kıstaslarını sağlaması
- Keşfedilen kuralların çeşitli biçimlerde gösterimi
- Farklı bir kaç soyutlama düzeyi ve etkileşimli veri madenciliği
- Farklı ortamlarda yer alan veri üzerinde işlem yapabilme
- Gizlilik ve veri güvenliğinin sağlanması

Sonuç olarak veri madenciliği, işletmelerdeki mevcut veri ambarlarının kullanılarak farklı disiplinlerdeki yöntemler yardımıyla VM süreci öncesinde görülemeyen hatta öngörülemeyen bilgi ve ilişkilerin karar vericiler tarafından kullanılmak üzere keşfidir. Şekil 2.2.'de gösterildiği üzere aslında veri madenciliği Veri Tabanlarında Bilgi Keşfi (VTBK) sürecinde bir adımdır. Fakat VTBK sürecinin en önemli işlevini görmesinden olsa gerek birçok çalışmada tüm süreci belirtmek için kullanılmıştır. Bu tez boyunca da Veri Madenciliği kavramı sürecin tamamını belirtmek için kullanılacaktır.



Şekil 2.2. Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği [6]

2.3. Literatürde Veri Madenciliği

Çoban, VM'yi yaygın kullanım alanlarının dışında bir alanda imalat sanayinde tedarikçi seçimi sürecinde kullanmıştır[2].
Özçınar, KPSS sonuçlarının tahmin edilmesinde VM tekniklerinden olan regresyon analizi kullanılmıştır[11].

Kalıkov, çalışmasında Veri Madenciliği tekniklerini kullanarak e-ticaret amaçlı kurulan bir yayınevi web sitesinin veri tabanında tutulan verilerin analizlerini yapmıştır. Bu tekniklerin uygulanması sonucunda, veri tabanında bulunan sanal

ürünlerin(kitapların) kategorilerine göre doğru yerleştirilmesinde yardımcı olacak bilgiler keşfetmiştir[12].

Altıntaş, çalışmasında veri madenciliği yöntemlerinden olan kümeleme algoritmalarını bir bankanın müşteri bilgilerini barındıran bir veri tabanı üzerinde uygulayarak bankanın müşterilerini kredilerini ödeme durumlarına göre kümelere ayırmasını sağlamıştır[14].

Gazi, çalışmasında GSM operatörleri tarafından yapılan kampanyaların, cep telefonu kullanıcıları üzerindeki etkisini analiz etmek amacıyla VM tekniklerini kullanmıştır[7].

Akbulut, yaptığı çalışmada bir kozmetik markasının müşteri gruplarını ve ayrılma eğilimi gösteren müşteri profilini belirleyerek; bu müşterilere özel pazarlama stratejileri geliştirilmesini hedeflemiş ve segmantasyon için kümeleme teknikleri, ayrılacak müşteri profilini belirlemek için ise sınıflama teknikleri kullanmıştır[10].

Aydın, çalışmasında asenkron motorların stator, rotor ve mil yatağı gibi bileşenlerinde oluşan arızaları yumuşak hesaplama ve veri madenciliği teknikleri ile teşhis etmiştir. VM tekniklerinden olan yapay sinir ağları gibi yumuşak hesaplama teknikleri kullanarak kırık rotor, sarım, mil yatağı sürtünmesi ve eksantriklik arızaları başarılı bir şekilde teşhis etmiştir[15].

Tiryaki, VM tekniklerinden olan sınıflandırma algoritmasını bir lojistik firmasının verilerine uygulamıştır[16].

Dolgun, çalışmasında birliktelik kuralları yöntemi ile pazar sepeti analizi yapmıştır[1].

Göral, çalışmasında kredi kartı başvuru aşamasında sahtecilik tespiti için VM yardımıyla öngörüsel bir model oluşturmuştur. Tüm başvuruları skorlamakta olan modelin sonucunda ortaya çıkan rapor, tüm başvurular için bir sahtekârlık skoru içermektedir[17].

Güntürkün, VM'yi kalite iyileştirme çalışmaları üzerinde kullanmıştır[18].

Tezcanlar, VM'yi Petro-kimya sektöründe bir işletmede bir yıllık dönemdeki müşteri profili ve satış verilerini inceleyip pazarlama stratejilerine tavsiye niteliğinde sonuçlara ulaşma amacıyla kullanmıştır[4].

Yılmaz, Kütahya İlinde sosyal sınıfların belirlenmesi ve tüketici profilinin çıkarılması amacına yönelik uygulamasında VM kullanmıştır[19].

Kasap, sigortacılık sektöründe müşteri ilişkileri yönetimi yaklaşımıyla veri madenciliği teknikleri birlikte kullanmıştır[20].

Özbay, internet bankacılığında yapılan dolandırıcılık işlemlerinin, veri madenciliği teknik ve metotları kullanılarak belli ölçüde önlenmesini temin eden bir model geliştirmiştir[8].

Yılmaz Koltan, çalışmasında İstanbul Menkul Kıymetler Borsası Ulusal 100 endeksinde sanayi ve hizmet sektörlerinde faaliyet gösteren 173 işletmenin 2004-2006 yıllarına ait yıllık finansal göstergelerinden yararlanarak veri madenciliği tekniklerinden birisi olan karar ağaçları tekniği uygulamıştır[3].

Ceran, VM'yi esnek akış tipi çizelgeleme problemlerinin çözümünde genetik algoritma ile birlikte kullanmıştır[21].

Çalışkan, soğuk hava tesislerinde optimum soğutma grubu seçiminde VM kullanmıştır[22].

Tosun, VM teknikleriyle kredi kartlarında müşteri kaybetme analizi yapmıştır[23].

Aktürk, borsa ile ilgilenen kişiler üzerindeki risk düzeyini aşağı yönde indirgeyebilmek amacıyla yaptığı çalışmasında VM teknikleri kullanmıştır[24].

Baysal ise bayi değerlendirmesi amacıyla yaptığı çalışmasında VM uygulamıştır[25].

Martens ve arkadaşları lojistik regresyon, C4.5 karar ağacı ve yapay sinir ağı kullanarak şirketlerin geleceğe yönelik endişelerini gidermek için bir model oluşturmaya çalışmışlar[29].

Sinha ve arkadaşları veri madenciliği sınıflandırma algoritmaları ile tanım kümesi birleştirmesine yönelik çalışmalarını dolaylı borç verme yani kredi üzerinde uygulamışlar. Uygulamada lojistik regresyon, karar ağacı, k-en yakın komşu algoritması ve yapay sinir ağı gibi VM yöntemleri kullanılmıştır. Kredi verilecek müşterilerin bilgileri sınıflandırılarak risk durumuna göre gruplar oluşturan bir model üzerinde çalışmışlardır[30].

Jie Sun ve Hui Li finansal tehlikelerin tahmini üzerine yaptıkları çalışmalarında Çin Hisse Senedi Piyasası ve muhasebe araştırmaları veri tabanı verilerine karar ağacı uygulayarak bir model oluşturmuşlar[31].

Shah ve Zhong kötü niyetli kişilerden mahremiyeti koruma amacıyla veri madenciliği tekniklerinden k-en yakın komşu kümeleme algoritmasını kullanarak bir model oluşturmaya çalışmışlar[32].

Chu ve arkadaşları var olan müşterileri kaybetmemek için hibrit bir veri madenciliği çalışması yapmışlar. Çalışmalarında C5.0, ve bir yapay sinir ağı algoritmasından oluşan bir model geliştirmeye çalışmışlar[33].

Hung ve arkadaşları telekomünikasyon şirketlerinde müşteri ilişkileri yönetimi veri madenciliği uygulaması için VM algoritmalarından C5.0 karar ağacı ve yapay sinir ağı kullanarak bir model uygulaması yapmışlar[34].

Hsu pazarlama, üretim artırma ve endüstriyel standartları geliştirme amacıyla veri madenciliğini giysi endüstrisinden bir işletmede uygulamış ve çalışmasında kümeleme algoritmalarını kullanmıştır[35].

Sugumaran ve arkadaşları titreşen prizmatik gövde üzerinde güvenlik analizini veri madenciliği tekniklerinden C5.0 karar ağacı algoritmasını kullanarak yapmışlardır[36].

Huang ve arkadaşları tedarikçi müşteri değer analizinde VM tekniklerinden k-ortalamalar algoritmasını kullanarak bir model üzerinde çalışmışlar[37].

Wu ve Yen izinsiz girmeleri tespit için VM tabanlı çalışmasında C4.5 karar ağacı algoritmasını kullanmışlardır[38].

Delen ve arkadaşları sağlık hizmetleri sigorta kapsamı analizinde yapay sinir ağı ve karar ağacı tekniklerini kullanmışlardır[39].

Lu ve Chen Tayvan borsa yatırımcıları için bilgi ifşası için VM uygulaması yapmışlar ve çalışmalarında karar ağaçlarını kullanmışlardır[40].

Chang ve Shyue Tayvan nüfus sayımında mağdur sosyal sınıflarını inceledikleri çalışmalarında karar ağacı ve kümeleme algoritmalarını kullanmışlardır[41].

Turhan ve arkadaşları yazılım virüslerini tespit için VM kaynak kodu çalışmalarını telekomünikasyon sektöründe naive bayes algoritması kullanarak uygulamışlar[42].
Chien ve Chen ileri teknoloji endüstrisinde personel seçimi ve insan sermayesini geliştirmeye yönelik bir VM çalışması yapmışlar ve çalışmalarında farklı karar ağacı algoritmalarını denemiş nihai modelde ise CHAID karar ağacı kullanmışlardır[43].

Chang geç gelişim gösteren çocuklara erken müdahalede bulunma amacıyla bir VM uygulaması yapmıştır. Çalışmada farklı karar ağacı algoritmaları denenmiştir[44].

Chien ve arkadaşları yarı iletken üretiminde verim artırma amacıyla yaptıkları çalışmalarında VM tekniklerinden k-ortalamalar algoritması ve karar ağacı algoritmalarını kullanmışlardır[45].

Chen ve Lin çalışmalarında ürün çeşitliliği ve boş raf dağıtımı probleminde VM yaklaşımını kullanmışlardır[46].

Kirkos ve arkadaşları sahte finansal beyanları tespit etmek için VM tekniklerinden yapay sinir ağı, Bayes güven ağları ve karar ağacı kullanmışlardır[47].

Yen ve Lee müşteri işlemlerinden ilginç bilgiler keşfetmek için etkin bir VM yaklaşımı geliştirmeye çalışmışlar[48].

Enke ve Thawornwong yapay sinir ağı ve VM kullanarak borsada hisse senedi iadelerini tahmin için bir model üzerinde çalışmışlardır[49].

Bayam ve arkadaşları yaşlı sürücüler ve kazalar arasındaki ilişkiyi tespit etmek için karar ağacı ve yapay sinir ağı tekniklerini kullanan bir model üzerinde çalışmışlar[50].

2.4. Veri Madenciliğinin Amaçları

Veri madencisinin geleneksel yöntemlerde olduğunun aksine başlangıçta herhangi bir amacı ya da varmak istediği bir kavram yoktur. Yapılacak analizlerden sonra elde edilen verilerin bir istatistikçi gözü ile incelenip daha önceden düşünülmemiş kavramların ortaya çıkarılması, başarılı bir VM süreci olarak kabul edilmektedir[3]. Buradaki temel amaç, değişkenler arasındaki ilişkilerden çok, geleceğe yönelik sağlıklı öngörülerin üretilmesidir. Bu anlamda VM, öz bilginin keşfedilmesi anlamında bir "kara kutu" bulma yaklaşımı olarak kabul edilmektedir ve bu doğrultuda yalnızca keşifsel veri analizi tekniklerini değil, sinir ağı tekniklerinden hareketle geçerli öngörüler yapmak ve öngörülen değişkenler arasındaki ilişkilerin belirlenmesi mümkün olduğu için aynı zamanda sinir ağı tekniklerini de kullanmaktadır[14].

Biraz daha detaylandırmak gerekirse veri madenciliğinin amaçlarını aşağıdaki başlıklar altında toplamak mümkündür;

- Öngörü: Hangi ürünlerin, hangi dönemlerde, hangi şartlarda, hangi miktarlarda satılacağına ilişkin öngörülerde bulunmak[2,13]
- Tanıma: Aldığı ürünlerden bir müşterinin tanınması, kullandığı programlar ve yaptığı işlemlerden bir kullanıcının tanınması[2,13]
- Sınıflandırma: Birçok parametrenin birleşimi kullanılarak ürünlerin, müşterilerin vb. sınıflandırılması[2,13]

- En iyileme: Belirli kısıtlamalar çerçevesinde zaman, yer, para ya da ham madde gibi sınırlı kaynakların kullanımını en iyilime ve üretim miktarı, satış miktarı ya da kazanç gibi değerleri büyütme de veri madenciliği amaçlarındandır[2,13].
- Ön tahmin
- Benzer gruplama
- Kümeleme
- Tanımlama[26]

2.5. Neden Veri Madenciliği

Otomatik veri toplama araçları ve veri tabanı teknolojilerindeki gelişme, veritabanlarında, veri ambarlarında ve diğer bilgi depolarında çok miktarda bilgi depolanması sonucunu doğurmuştur. Çok fazla veri var, ancak bilgi yok… Veri ambarları ve veri madenciliği büyük miktarlardaki veriler içindeki gizli örüntüler, geleneksel çözümleme araçlarıyla bulunamaz. Toplanan veri miktarı büyüdükçe ve toplanan verilerdeki karmaşıklık arttıkça, daha iyi çözümleme tekniklerine olan gereksinim de artmaktadır. Bu tür bilgiler, bilgi bulma/keşfetme (knowledge discovery) ya da veri madenciliği (data mining) olarak bilinen teknikler yardımıyla çözümlenebilir[27].

Veri madenciliği aşağıdaki karakteristiklere sahip problemlerin çözümünde daha çok tercih edilir[13]:

- Büyük miktarlarda veriye erişilebildiği zaman,
- Veri birçok değişkene sahipse,
- Veri karmaşık, çok değişkenli ve doğrusal değil ise,
- Çıktıları ya da davranışı tahmin etmek gerekiyorsa,
- Henüz anlaşılmayan birliktelik ve ilişkileri bulmak gerekiyorsa,

Veri tabanı hacimlerinin veri madenciliği gerektirecek düzeye ulaşması, pazarlama, reklam ve imalatta küçük müşteri gruplarına ve bireylere kadar ulaşılması gerekiyorsa[2].

2.6. Veri Madenciliğinin Kullanım Alanları

Günümüzde veri madenciliği teknikleri başta işletmeler olmak üzere çeşitli alanlarda başarı ile kullanılmaktadır. Veri madenciliğinin asıl amacı, veri yığınlarından anlamlı bilgiler elde etmek ve bunu eyleme dönüştürecek kararlar için kullanmaktır[27]. Son yıllarda Ülkemizde de geniş bir kullanım alanı bulan veri madenciliğinin kullanıldığı alanlar ve kullanım amaçları aşağıda belirtilmiştir:

Pazarlama alanında;

- Müşterilerin satın alma örüntülerinin belirlenmesi
- Müşterilerin demografik özellikleri arasındaki bağlantıların bulunması
- Posta kampanyalarında cevap verme oranının artırılması
- Pazar sepeti analizi (Market Basket Analysis)
- Müşteri ilişkileri yönetimi (Customer Relationship Management)
- Müşteri değerlendirme (Customer Value Analysis)
- Satış tahmini (Sales Forecasting)
- Müşteri dağılımı
- Çeşitli pazarlama kampanyaları
- Mevcut müşterilerin elde tutulması için geliştirilecek pazarlama stratejilerinin oluşturulması
- Çapraz satış analizleri
- Çeşitli müşteri analizleri[12]
- Müşteri şikâyetlerinin incelenmesi,
- Satış kampanyalarının verimlilik analizlerinin yapılması[14]

Bankacılık alanında;

- Farklı finansal göstergeler arasında gizli korelâsyonların bulunması
- Kredi kartı dolandırıcılıklarının tespiti
- Kredi kartı harcamalarına göre müşteri gruplarının belirlenmesi
- Kredi taleplerinin değerlendirilmesi
- Müşteri dağılımı
- Usulsüzlük tespiti
- Risk analizleri
- Risk yönetimi[12]

Sigortacılık alanında;

- Yeni poliçe talep edecek müşterilerin tahmin edilmesi
- Sigorta dolandırıcılıklarının tespiti
- Riskli müşteri örüntülerinin belirlenmesi[12]

Perakendecilik alanında;

- Satış noktası veri analizleri
- Alış-veriş sepeti analizleri
- Tedarik ve mağaza yerleşim optimizasyonu[12]

Borsa alanında;

- Hisse senedi fiyat tahmini,
- Genel piyasa analizleri,
- Hisse tespitleri,
- Alım-satım stratejilerinin optimizasyonu[11]

Telekomünikasyon alanında;

- Kalite ve iyileştirme analizleri
- Hatların yoğunluk tahminleri[12]
- Çağrı ayrıntı analizleri[2].
- Müşteri bağlılığı[2].

Sağlık ve İlaç Sektöründe;

- Test sonuçlarının tahmini
- Ürün geliştirme
- Tıbbi teşhiste
- Tedavi sürecinin belirlenmesi[12]
- MR verileri ile sinir sistemi bölge ilişkilerinin belirlenmesinde[27].

Endüstri alanında;

- Kalite kontrol analizleri
- Lojistik
- Üretim süreçlerinin optimizasyonu[12]
- Operasyonel süreçte oluşabilecek olası kayıpların veya suiistimallerin tespiti

- Kurum teknik kaynaklarının optimal şekilde kullanılmasını sağlamak
- Firmanın finansal yapısının, makro ekonomik değişmeler karşısındaki duyarlılığı
- Geçmiş ve mevcut yapı analiz edilerek geleceğe yönelik tahminler[14].

Eğitim alanında;

- Öğrenci davranışlarının öngörülmesi
- Öğrencilerin ders seçme eğilimlerinin belirlenmesi[11].

Hilekârlık Tespitinde; geçmişe ait veriler kullanılarak, geçmişte hilekârlık yapmış kişilere ait veriler incelenebilir ve bunlara ait bir model kurulabilir. Geliştirilen bu model kullanılarak hilekârlığa meyilli olanlar tespit edilebilir. Hilekârlık belirlemenin en yaygın kullanım alanları sigortacılık sektörü, finans sektöründe kredi kartı servisleri, perakendecilik sektörü ve telekomünikasyon sektörüdür[2]. Bunun dışında ev veya işyerlerinde kullanılan elektrik, su ve telefon gibi abonelik gerektiren durumlarda mevcut kullanımı düşük gösterme ya da tamamen kaçak kullanma gibi yasa dışı durumların tespitinde de VM kullanılmaktadır.

Web uygulamaları alanında;

- Kullanıcı taraflı bilgiler (tarayıcı, dil vb..) ışığında altyapı düzenlemeleri.
- Kullanıcı profillerine uygun ürünlerin reklam kampanyaları en çok ziyaret ettikleri sayfalara koyulabilir.
- Farklı web şablonları, temaları arasında kullanıcı istekleri değerlendirilebilir.
- Kötü niyetli kullanıcı istekleri belirlenip bunlara karşı alınması gereken önlemler belirlenebilir[27].

Kamu uygulamaları alanında;

- Kaynakların doğru olarak kullanımını sağlama ve planlama.
- Kamu güvenliğini sağlama amacı ile güvenlik problemlerini önceden tahmin etmek.
- Rastlantısal olaylardaki sorunların çözümüne dair izleri keşfetme ve olası güvenlik sorunlarını es zamanlı olarak tespit edebilme ve çözüm üretebilme.

- Vergi ile ilgili yolsuzlukları ve izlerini belirleme, yolsuzlukları es zamanlı olarak belirleme.
- Sağlık ödemeleri.
- Kamu kurumlarında programların uygulanması gibi konularda şüpheli durumların tespiti, suiistimal ve israfları belirleme ve milyonlarca dolarlık zararı engelleme.
- Emniyet birimleri için suç istatistiklerine dair online raporlama, hangi profildeki insanların ne tür suçlara meyilli olduklarını belirleme, es zamanlı suç engelleme politikaları oluşturmak[9].

Tablo 2.2.'de 2003 yılında yapılan bir araştırma sonucuna göre veri madenciliğinin sektörler bazında kullanımına ilişkin sonuçlar yer almaktadır[27].

Tablo 2.2. Veri Madenciliği Uygulama Alanları[27]

131 KİŞİDEN TOPLAM 279 OY	
Bankacılık (37)	**13%**
Bioteknoloji / Genetik (27)	**10%**
Pazarlama / Organizasyon (29)	**10%**
Web (15)	**5%**
Eğlence / Haber (4)	**1%**
Sahtekârlık Tespiti (24)	**9%**
Sigortacılık (23)	**8%**
Yatırım / Hisse Senedi (8)	**3%**
İmalat (5)	**2%**
Medikal (16)	**6%**
Perakende (17)	**6%**
Bilimsel Çalışmalar (24)	**9%**
Güvenlik (6)	**2%**
Tedarik Zinciri Analizi (3)	**1%**
Telekomünikasyon (21)	**8%**
Seyahat (5)	**2%**
Diğer (12)	**4%**
Bilinmeyen (3)	**1%**

2.7. Veri Madenciliği Sürecinde Karşılaşılan Problemler

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

Diğer sorunlar da verinin konu ile uyumsuzluğundan doğabilir[10]. Küçük veri kümelerinde hızlı ve doğru bir biçimde çalışan bir sistem, çok büyük veri tabanlarına uygulandığında tamamen farklı davranabilir. Veri madenciliği sürecinde karşılaşılan başlıca problemler aşağıda belirtilmiştir[2].

– Veri tabanı boyutu: Veritabanları genel olarak veri madenciliği dışındaki amaçlar için tasarlanmışlardır. Bu yüzden, öğrenme görevini kolaylaştıracak bazı özellikler bulunmayabilir[10]. Günümüzde kullanılan veritabanı boyutları hızla arttığından, bu veritabanları üzerinde çalıştırılacak makine öğrenmesi algoritmaları çok yavaş çalışmaktadır. Küçük veri kümeleri üzerinde hızlı çalışan algoritmalar yüz binlerce hatta milyonlarca veri üzerinde uygulandığında yetersiz sonuçlar verirler. Dolayısıyla veri madenciliği yöntemleri ya sezgisel bir yaklaşımla arama uzayını taramalıdır ya da örneklemi yatay/dikey olarak indirgemelidir[15]. Yatay indirgeme, nitelik değerlerinin önceden belirlenmiş genelleme sıra düzenine göre, bir üst nitelik değeri ile değiştirilme işlemi yapıldıktan sonra aynı olan satırların çıkarılması işlemidir. Dikey indirgeme, artık niteliklerin indirgenmesi işlemidir. Özellik seçimi yöntemleri ya da nitelik bağımlılık çizelgesi uygulanarak yapılır[15, 27].

Veri tabanlarında tutulan verilerin iki boyutu vardır[27]:

1. Yatay boyut: VT'lerde tutulan bilgilerin özelliklerini ifade eden satırların sütunsal detaylarıdır.
2. Dikey boyut: VT'lerde tutulan kayıt sayısını ifade etmektedir.

– Gürültülü veri: Büyük veri tabanlarında pek çok niteliğin değeri yanlış olabilir. Bu hata, veri girişi sırasında yapılan insan hataları veya girilen değerin yanlış ölçülmesinden kaynaklanır. Veri girişi ya da veri toplanması sırasında oluşan sistem dışı hatalara "gürültü" adı verilir. Ancak günümüzde kullanılan ticari ilişkisel veri tabanları veri girişi sırasında oluşan hataları otomatik biçimde gidermek konusunda az bir destek sağlamaktadır. Hatalı veri gerçek dünya veri tabanlarında ciddi problem oluşturabilir. Bu durum, bir veri madenciliği yönteminin kullanılan veri kümesinde

bulunan gürültülü verilere karsı daha az duyarlı olmasını gerektirir. Gürültülü verinin yol açtığı problemler tümevarımsal karar ağaçlarında uygulanan metotlar balgamında kapsamlı bir biçimde araştırılmıştır. Eğer veri kümesi gürültülü ise sistem bozuk veriyi tanımalı ve ihmal etmelidir[2,13,16]. Quinlan, gürültünün sınıflama üzerindeki etkisini araştırmak için bir dizi deney yapmıştır. Deneysel sonuçlar, etiketli öğrenmede etiket üzerindeki gürültü öğrenme algoritmasının performansını doğrudan etkileyerek düşmesine sebep olmuştur. Buna karsın eğitim kümesindeki nesnelerin özellikleri/nitelikleri üzerindeki en çok % 10'luk gürültü miktarı ayıklanabilmektedir[2,13].

– Eksik veri: Örneklem kümesindeki kayıtların eksik olması ya da bazı kayıtlar için bazı niteliklerin veya nitelik değerlerinin olmamasıdır. Bu eksiklik; hatalı ölçüm araçlarından, veri toplama sürecinde deneyin tasarımında yapılan değişiklikten ya da birbirine benzer ancak özdeş olmayan veri kümelerinin birleştirilmesinden kaynaklanıyor olabilir[13].
– Boş (null) değerler: Veritabanlarında geçersiz veri, değeri birincil anahtarda yer almayan herhangi bir niteliğin değeri olabilir. Bir kayıtta eğer bir nitelik değeri geçersiz ise o nitelik bilinmeyen ve uygulanamaz bir değere sahiptir. Bu durum, ilişkisel veritabanlarında sıkça karşımıza çıkmaktadır. Bir ilişkide yer alan tüm kayıtlar aynı sayıda niteliğe, niteliğin değeri geçersiz olsa bile, sahip olmalıdır. Örneğin kişisel bilgisayarların özelliklerini tutan bir ilişkide bazı model bilgisayarlar için ses kartı modeli niteliğinin değeri geçersiz olabilir[15].

Veri kümelerinde yer alan boş değerler için çeşitli çözümler söz konusudur. Bunlar; boş değerli kayıtlar tamamıyla ihmal edilebilir, boş değerler yerine olası bir değer atanabilir. Bu değerler o nitelikteki en fazla frekansa sahip bir değer veya ortalama bir değer olabilir, varsayılan bir değer olabilir, boş değerin kendisine en yakın değer olabilir[13].

– Artık veri: Verilen veri kümesi, eldeki probleme uygun olmayan veya artık nitelikler içerebilir. Bu durum pek çok işlem sırasında karşımıza çıkabilir. Artık nitelikleri elemek için geliştirilmiş algoritmalar özellik seçimi olarak adlandırılır. Özellik seçimi, tümevarıma dayalı öğrenmede budama öncesi yapılan bir işlemdir. Başka bir deyişle, özellik seçimi, verilen bir ilişkinin içsel tanımını, dışsal tanımın

taşıdığı (veya içerdiği) bilgiyi bozmadan onu eldeki niteliklerden daha az sayıdaki niteliklerle (yeterli ve gerekli) ifadeleyebilmektir. Özellik seçimi yalnızca arama uzayını küçültmekle kalmayıp, sınıflama işleminin kalitesini de arttırır[2,16].

– Dinamik veri: Veri tabanlarındaki bilgiler, veri eklendikçe ya da silindikçe değişebilir. Veri madenciliği perspektifinden bakıldığında, kuralların hala aynı kalıp kalmadığı ve istikrarlılığı problemi ortaya çıkar[10].

Kurumsal çevrim-içi veri tabanları dinamiktir, yani içeriği sürekli olarak değişir. Bu durum, bilgi keşfi metotları için önemli sakıncalar doğurmaktadır. İlk olarak sadece okuma yapan ve uzun süre çalışan bilgi keşfi metodu bir veri tabanı uygulaması olarak mevcut veri tabanı ile birlikte çalıştırıldığında mevcut uygulamanın da performansı ciddi ölçüde düşer. Diğer bir sakınca ise, veri tabanında bulunan verilerin kalıcı olduğu varsayılıp, çevrimdışı veri üzerinde bilgi kesif metodu çalıştırıldığında, değişen verinin elde edilen örüntü1ere yansıması gerekmektedir. Bu işlem, bilgi keşfi metodunun ürettiği örüntü1eri zaman içinde değişen veriye göre sadece ilgili örüntü1eri yığmalı olarak günleme yeteneğine sahip olmasını gerektirir. Aktif veri tabanları tetikleme mekanizmalarına sahiptir ve bu özellik bilgi kesif metotları ile birlikte kullanılabilir[2].

– Farklı tipteki verilerin ele alınması: Gerçek hayattaki uygulamalar makine öğreniminde olduğu gibi yalnızca sembolik veya kategorik veri türleri değil, fakat aynı zamanda tamsayı, kesirli sayılar, çoklu ortam verisi, coğrafi bilgi içeren veri gibi farklı tipteki veriler üzerinde işlem yapılmasını gerektirir[16]. Kullanılan verinin saklandığı ortam, düz bir kütük veya ilişkisel veri tabanında yer alan tablolar olacağı gibi, nesneye yönelik veri tabanları, çoklu ortam veri tabanları, coğrafik veri tabanları vb. olabilir. Saklandığı ortama göre veri, basit tipte olabileceği gibi karmaşık veri tipleri (çoklu ortam verisi, zaman içeren veri, yardımlı metin, coğrafi, vb.) de olabilir. Bununla birlikte veri tipi çeşitliliğinin fazla olması bir VM algoritmasının tüm veri tiplerini ele alabilmesini olanaksızlaştırmaktadır. Bu yüzden veri tipine özgü adanmış VM algoritmaları geliştirilmektedir[16,27].

– Sınırlı Bilgi: Veri tabanları genel olarak veri madenciliği dışındaki amaçlar için tasarlanmışlardır. Bu yüzden, öğrenme görevini kolaylaştıracak bazı özellikler bulunmayabilir[16].

– Belirsizlik: Yanlışlıkların şiddeti ve verideki gürültünün derecesi ile ilgilidir. Veri tahmini bir keşif sisteminde önemli bir husustur[16].

2.8. Veri Madenciliğinde Kullanılan Modeller

Gerek tanımlayıcı gerekse tahmin edici modellerde yoğun olarak kullanılan belli başlı teknikler; Hipotez Testi Sorgusu, Sınıflama ve Regresyon Sorgusu, Kümeleme Sorgusu, Ardışık Örüntüler, Birliktelik Kurulları olarak sıralanabilir. Sınıflama ve Hipotez Testi modelleri tahmin edici, kümeleme, birliktelik kuralları ve ardışık örüntü modelleri tanımlayıcı modellerdir [27]. Veri madenciliği modelleri verilerde var olan gizli bilgiyi ortaya çıkartmaya yarayan metotlar olup genel olarak iki ana gruba ayrılır:

2.8.1. Tanımlayıcı (descriptive) modeller

Doğrulamaya dayalı modellerde; kullanıcı tarafından ispatlanmak istenen bir hipotez ortaya sürülür ve VM algoritmalarıyla bu hipotez ispatlanmaya çalışılır. Çok boyutlu analizlerde ve istatistiksel analizlerde tercih edilen metottur. Hipotez testi buna güzel bir örnektir[8].

Tanımlayıcı modeller analiste daha önceden bir hipoteze sahip olmaksızın, veri kümesinin içinde ne tür ilişkiler olduğunu anlama imkânı sunar. Analizcinin çok geniş veri tabanlarındaki bilgileri incelemek, örüntüleri keşfetmek için doğru soruları sorup hipotezler geliştirmesi pratikte zor olduğundan, ilginç örüntüleri keşfetme önceliği veri madenciliği programına bırakılır. Keşfedilen bilginin kalitesi ve zenginliği, uygulamanın kullanışlılığını ve gücünü oluşturur. Kümeleme, birliktelik kuralları, çok kullanılan tanımlayıcı modellerdir[27].

Tahmin edici modeller kümeleme modelini, homojen veri grupları oluşturması için veri ön isleme aşaması olarak ta kullanmaktadırlar. Birliktelik kuralları, bir arada olan olayların ya da özelliklerin keşfedilmesi sürecidir, ilişki analizi ya da pazar sepet analizi olarak da adlandırılır. Birliktelik kuralları genellikle "eğer şu olursa daha sonra bu olur" seklindedir. Birliktelik kuralları oluşturmada en çok kullanılan algoritmalar Apriori ve GRI'dir. Özetleme tanımlayıcı istatistikleri kullanarak verinin betimlenmesidir, genellikle açıklayıcı veri analizi için uygulanır. Görselleştirme, verinin grafik öğeleri yardımıyla betimlenmesidir, genellikle ayrık değerleri tespit etmede, veri ön islemede, trend ve ilişkilerin bulunmasında kullanılır[11].

Tanımlayıcı modellerde amaç, büyük veri kümelerindeki desen ve ilişkileri tespit ederek, incelenen sistemin anlamını kavramaktır. "25 yas altı bekâr kişiler ile 25 yas üstü evli kişiler üzerinde yapılan ve ödeme performanslarını gösteren bir analiz tanımlayıcı modellere örnek olarak verilebilir"[12].

VM' de kullanılan bazı algoritma ve teknikler hem tahmin edici hem de tanımlayıcı modellerde kullanım alanı bulduğundan bu çalışmada ya tanımlayıcı model grubunun içinde ya da tahmin edici model grubunun içinde belirtilecektir.

2.8.1.1. Kümeleme analizi

Kümeleme analizi denetimsiz öğrenme kategorisine giren bir algoritmadır. Sınıflama algoritmasında olduğu gibi ortak özellikleri olan veriler bir kümeye girer. Alt kümelere ayrılmak için keşfedilen kurallar yardımıyla bir kaydın hangi alt kümeye girdiği kümeleme algoritması sayesinde bulunur. Kümeleme algoritması genelde astronomi, nüfus bilimi, bankacılık uygulamaları gibi uygulamalarda kullanılır[8].
Kümeleme modellerinde amaç, üyelerinin birbirlerine çok benzediği, ancak özellikleri birbirlerinden çok farklı olan kümelerin bulunması ve veri tabanındaki kayıtların bu farklı kümelere bölünmesidir. Sınıflandırmaya benzemekle birlikte grupların önceden belirlenmesi bakımından ondan farklıdır. Temel özellikleri oluşacak küme sayısının belirsiz olması, küme sonuçlarının dinamik olması ve kümelerle ilgili bir ön bilgi olmayabileceğidir. Kümeleme algoritması veritabanını alt kümelere ayırır. Her bir kümede yer alan elemanlar dâhil oldukları grubu diğer

gruplardan ayıran ortak özelliklere sahiptir. Kümeleme modellerinde amaç, küme üyelerinin birbirlerine çok benzediği, ancak özellikleri birbirlerinden çok farklı olan kümelerin bulunması ve veri tabanındaki kayıtların bu farklı kümelere bölünmesidir. Kümeleme analizinde; veri tabanındaki kayıtların hangi kümelere ayrılacağı veya kümelemenin hangi değişken özelliklerine göre yapılacağı konunun uzmanı olan bir kişi tarafından belirtilebileceği gibi veri tabanındaki kayıtların hangi kümelere ayrılacağını, geliştirilen bilgisayar programları da yapabilmektedir[16].

Kümelemede, genellikle k-ortalamalar algoritması ya da Kohonen şebekesi gibi istatistiksel yöntemler kullanılmaktadır. Hangi yöntem kullanılırsa kullanılsın süreç aynı şekilde işler. Her kayıt var olan kümelerle karşılaştırılır. Bir kayıt kendisine en yakın kümeye atanır ve bu kümeyi tanımlayan değeri değiştirir. Optimum çözüm bulununcaya kadar kayıtlar yeniden atanır ve küme merkezleri ayarlanır. En yaygın kullanılan kümeleme algoritması "k ortalamalar algoritması"dır[10].

Kümeleme işlemi, heterojen yapıya sahip bir kitleyi daha homojen birkaç alt gruba ya da kümeye bölme işlemidir. Sınıflama ile kümelemeyi birbirinden ayıran en önemli fark, kümeleme işleminin sınıflama işleminde olduğu gibi önceden belirlenmiş bir takım sınıflara göre bölme yapmamasıdır. Sınıflamada her bir veri, önceden sınıflandırılmış bir takım sınıflar üzerinde yapılan bir eğitim neticesinde ortaya çıkan bir modele göre önceden belirlenmiş olan bir sınıfa atanmaktadır. Kümeleme işleminde ise önceden tanımlanmış sınıflar ya da örnek sınıflar bulunmamaktadır. Verilerin kümelenmesi işlemi, verilerin birbirlerine olan benzerliklerine göre yapılmaktadır. Oluşan sınıfların hangi anlamları taşıdığının belirlenmesi tamamen çözümlemeyi yapan kişiye kalmıştır. Kümeleme işlemi çoğunlukla bir başka VM uygulaması için bir ilk işlem olarak kullanılır. Kümeleme modellerinde amaç, küme üyelerinin birbirlerine çok benzediği, ancak özellikleri birbirlerinden çok farklı olan kümelerin bulunması ve veritabanındaki kayıtların bu farklı kümelere bölünmesidir. Literatürde birçok kümeleme algoritması bulunmaktadır. Kullanılacak olan kümeleme algoritmasının seçimi, veri tipine ve çalışmanın amacına bağlıdır. Kümeleme analizinde genel olarak bölme, hiyerarşik, yoğunluk tabanlı, ızgara tabanlı ve model tabanlı yöntemler kullanılmaktadır[1].

2.8.1.2. Birliktelik kuralları

İlişki analizi ya da birliktelik kuralları, bir veri kümesinde kendiliğinden, sıklıkla gerçekleşen, birlikte ya da aynı süre içinde alınma, yapılma, oluşma gibi etkileri keşfetme temeline dayanır. Bu yöntem bankacılık işlemlerinin analizinde ya da sepet analizi tekniğinde yaygın olarak kullanılır. Sepet analizi, bir alışveriş sırasında veya birbirini izleyen alışverişlerde müşterinin hangi mal veya hizmetleri satın alma eğiliminde olduğunun belirlenmesiyle müşteriye daha fazla ürün satılması yollarından biridir[16].

Birliktelik kuralları, bir arada olan olayların ya da özelliklerin keşfedilmesi sürecidir. Birliktelik kuralları genellikle "eğer şu olursa daha sonra bu olur" şeklindedir. Genellikle açıklayıcı veri analizinde, ayrık değerleri tespit etmede, veri ön işlemede, eğilim ve ilişkilerin bulunmasında kullanılır. Bir alışveriş sırasında veya birbirini izleyen alışverişlerde müşterinin hangi mal veya hizmetleri satın almaya eğilimli olduğunun belirlenmesi, müşteriye daha fazla ürünün satılmasını sağlama yollarından biridir. Bununla birlikte bu teknikler, tıp, finans ve farklı olayların birbirleri ile ilişkili olduğunun belirlenmesi sonucunda değerli bilgi kazanımının söz konusu olduğu ortamlarda da önem taşımaktadır. Bir birliktelik algoritması oluşturmadan önce kurallar belirlenmelidir. Büyük veri tabanında ilişkileri bulacak algoritmalar geliştirmek çok zor değildir. Fakat geliştirilen algoritmalar önemli ilişkileri ortaya çıkaracağı gibi önemsiz birçok ilişkiyi de ortaya çıkarır. Bu yüzden, büyük veri tabanlarında küçük alt kümeler bulunmalıdır. Büyük veri tabanlarında birliktelik kuralları bulunurken, şu iki işlem basamağı takip edilir[27]:

1- Sık tekrarlanan öğeler bulunur. Bu öğelerin her biri en az, önceden belirlenen minimum destek sayısı kadar sık tekrarlanırlar.

2- Sık tekrarlanan öğelerden güçlü birliktelik kuralları oluşturulur. Bu kurallar minimum destek ve minimum güven değerlerini karşılamalıdır.

Ayrıca, büyük veri tabanlarında çok sayıda ilişki bulunabileceğinden, birliktelik kuralları sayısı da sınırsız olabilir. Dolayısıyla ilginç ilişkilerle önemsiz ilişkilerin ayrılması gerekir. Birliktelik kuralları oluşturmada en çok kullanılan algoritmalar Apriori, GRI, AIS ve SETM'dir[27].

Birliktelik kuralları; ticaret, mühendislik, fen ve sağlık sektörlerinin içinde bulunduğu birçok alanda uygulanmaktadır. Birliktelik kuralları, VM araştırmalarında çok büyük yatırımlar yapılan, VM'nin özel bir uygulama alanıdır. Birliktelik kuralları aynı işlem içinde çoğunlukla beraber görülen nesneleri içeren kurallardır. Birliktelik kurallarının bulunmasında birçok yöntem vardır. Büyük veritabanlarında birliktelik kuralları bulmak için algoritma geliştirmek çok zor değildir, buradaki zorluk bu tür algoritmaların çok küçük değerli diğer birçok birliktelik kuralını da meydana çıkarmasıdır. Bulabileceğimiz olası birliktelik kuralları sayısı sonsuzdur. Birliktelik kurallarıyla ilgili problem, birliktelik kurallarını bulmada bir eşik değeri bulmaktır. Önemsiz gürültüden, değerli bilgiyi ayırabilmek ve bu eşik değerini bulabilmek çok zordur. Bu yüzden ilginç birliktelik kurallarından ilginç olmayanları ayırt edebilmek için bazı ölçütlerin belirlenmesi gereklidir. Bu ölçütler destek ve güven değerleridir. Birliktelik kuralı madenciliğin amacı, kullanıcı tarafından belirlenen minimum destek ve güven değerlerini sağlayan kuralların bulunmasıdır. Anlamlılığı destek ve güven değerleri ile ölçülen birliktelik kuralları, "X nesnesini alan bir müşterinin muhtemelen Y nesnesini de alması" tipindeki kuralların tanımlanmasını amaçlamaktadır[1].

Bu kriterler şu şekilde hesaplanmaktadır. Burada X ve Y ürünleri arasındaki ilişki incelenmektedir[20].

$P(X \cap Y)$ = X ve Y ürünlerini almış müşteri sayısı / toplam müşteri sayısı

$P(X \cap Y)$, destek kriteri adı verilmektedir. Destek kriteri X ürününü alan bir müşterinin Y ürününü alma olasılığını yani X ürününü alıp sonra Y ürününü alma olasılığını gösteren bir değerdir. Bu değer bire yaklaştıkça güçlenmektedir.

$P(X/Y) = P(X \cap Y) / P(Y)$

$P(X/Y)$, güven kriteri olarak tanımlanmaktadır. Bu kriter Y ürününü alan bir müşterinin X ürününü alma olasılığını göstermektedir. Aynı destek kriteri gibi, güven kriteri de bire yaklaştıkça güçlenmektedir. Birliktelik kurallarına örnek vermek gerekirse, tatil için uçak bileti alan bir kimsenin, belli bir olasılıkla araba kiralaması verilebilir[20]. Ya da "Düşük yağlı peynir ve yağsız süt alan müşteriler %85 olasılıkla diyet süt alırlar"[10].

2.8.1.3. Ardışık zamanlı örüntüler

Ardışık örüntü keşfi, bir zaman aralığında sıklıkla gerçeklesen olaylar kümelerini bulmayı amaçlar. Bir ardışık örüntü örneği söyle olabilir: Bir yıl içinde A yazarının "X" romanını satın alan insanların %70'i B yazarının "Y" adlı kitabını da satın almıştır. Bu tip örüntüler perakende satış, telekomünikasyon ve tıp alanlarında yoğun bicimde kullanılmaktadır[8].

Ardışık örüntü keşfi, bir zaman aralığında sıklıkla gerçekleşen olaylar kümelerini bulmayı amaçlar. Bu tip örüntüler perakende satış, telekomünikasyon ve tıp alanlarında yararlıdır[16].

Ardışık analiz ise birbiriyle ilişkisi olan ancak birbirini izleyen dönemlerde gerçekleşen ilişkilerin tanımlanmasında kullanılır[10].

Belli bir dönem boyunca nesneler arasındaki birlikteliklerin incelenmesi olan ardışık örüntü keşfi bazı kaynaklarda "ardışık zamanlı örüntü çözümlemesi" olarak da isimlendirilmiştir[1].

2.8.2. Tahmin edici (predictive) modeller

Keşfe dayalı modellerde herhangi ispatlanmak istenen bir hipotez yoktur. VT keşfedici olarak araştırılarak, gizli olan bilgiler açığa çıkarılır. Doğrulamaya dayalı algoritmaların tersine bu algoritmalarda ortada ispatlanması istenen hipotezler yoktur. Tam tersine bu algoritmalar otomatik keşfe dayanmaktadır. İstisnai durumların keşfi, karar ağacı, kümeleme gibi algoritmalar bu yaklaşıma göre kurulmuştur[8].

Tahmin edici modellerin amacı sonuçları bilinen verilerden hareket edilerek bir model geliştirilmesi ve kurulan bu modelden yararlanılarak sonuçları bilinmeyen veri kümeleri için sonuç değerlerin tahmin edilmesidir. Burada tahmin etme, yargıya varma, sınıflandırma v.s benzer işlevleri görecek şekilde ve çalıştırılabilir kod olacak bir model üretme amaçlanmaktadır. "Örneğin bir banka önceki dönemlerde vermiş

olduğu kredilere ilişkin gerekli tüm verilere sahip olabilir. Bu verilerde bağımsız değişkenler kredi alan müşterinin özellikleri, bağımlı değişken değeri ise kredinin geri ödenip ödenmediğidir. Bu verilere uygun olarak kurulan model, daha sonraki kredi taleplerinde müşteri özelliklerine göre verilecek olan kredinin geri ödenip ödenmeyeceğinin tahmininde kullanılmaktadır"[12].

Tahmin, geçmiş tecrübelerden elde edilen bilgiler ve mantık kullanılarak, gelecekte olması muhtemel durumlar hakkında öngörüde bulunmaktır. Tahmin edici modeller karar alma süreçlerinde önemli bir rol oynar. Tahmin edici modellerde sonuçları bilinen verilerden hareket edilerek bir model geliştirilmesi ve kurulan bu modelden yararlanılarak sonuçları bilinmeyen veri kümeleri için sonuç değerlerinin tahmin edilmesi amaçlanır. Tahmin edici modellerin temel iki türü sınıflandırma ve regresyondur. Sınıflandırma, veri nesnesini daha önceden belirlenen sınıflardan biriyle esleştirme sürecidir. Verileri ve karsı gelen sınıfları içeren eğitim kümesi ile eğitilen sistem, sonraki aşamalarda sınıf bilgisine sahip olunmayan verilerin ait olduğu sınıfların bulunması için kullanılır. Müşteri Segmentasyonu, kredi analizi, is modellemesi ve benzeri birçok alanda kullanılan sınıflandırma yöntemi günümüzde en çok kullanılan veri madenciliği yöntemidir. Regresyon, sürekli sayısal bir değişkenin, aralarında doğrusal ya da doğrusal olmayan bir ilişki bulunduğu varsayılan diğer değişkenler yardımıyla tahmin edilmesi yöntemidir. Regresyon modeli, sayısal değerleri tahmin etmeye yönelik olması dışında sınıflandırma yöntemine benzetilebilir. Çok terimli lojistik regresyon gibi kategorik değerlerin de tahmin edilmesine olanak sağlayan tekniklerin geliştirilmesi ile sınıflandırma ve regresyon modelleri giderek birbirine yaklaşmakta ve dolayısıyla aynı tekniklerden yararlanılması mümkün olmaktadır[11].

2.8.2.1. Sınıflandırma ve regresyon modelleri

İstenilen bir değişken bağımlı değişken ve diğerleri tahmin edici (bağımsız) değişkenler olarak adlandırılır. Amaç, girdi olarak tahmin edici değişkenlerin yer aldığı modelde, çıktının bağımlı değişkenin değerinin bulunduğu anlamlı bir model kurmaktır. Bağımlı değişken sayısal değil ise problem sınıflama problemidir. Eğer bağımlı değişken sayısal ise problem regresyon problemi olarak adlandırılır. Mevcut verilerden hareket ederek geleceğin tahmin edilmesinde faydalanılan ve VM

yöntemleri içerisinde en yaygın kullanıma sahip olan sınıflama ve regresyon modelleri arasındaki temel fark, tahmin edilen bağımlı değişkenin kategorik veya süreklilik gösteren bir değere sahip olmasıdır[1].

Sınıf olmak için her kaydın belli ortak özellikleri olması gerekir. Ortak özelliklere sahip olan kayıtların hangi özellikleriyle bu sınıfa girdiğini belirleyen algoritma, sınıflama algoritmasıdır. Sınıflama algoritması, denetimli öğrenme kategorisine giren bir öğrenme biçimidir. Denetimli öğrenme, öğrenme ve test verilerinin hem girdi hem de çıktıyı içerecek şekilde olan verileri kullanmasıdır. Sınıflama sorgusuyla, bir kaydın önceden belirlenmiş bir sınıfa girmesi amaçlanmaktadır. Bir kaydın önceden belirlenmiş bir gruba girebilmesi için sınıflama algoritması ile öğrenme verileri kullanılarak hangi sınıfların var olduğu ve bu sınıflara girmek için bir kaydın hangi özelliklere sahip olması gerektiği otomatik olarak keşfedilir. Test verileriyle de bu öğrenmenin testi yapılarak ortaya çıkan kurallar optimum sayısına getirilir. Sınıflama algoritmasının kullanım alanları sigorta risk analizi, banka kredi kartı sınıflaması, sahtecilik tespiti gibi alanlardır[8].

Sınıflama sorgusu yeni bir veri elemanını daha önceden belirlenmiş sınıflara atamayı amaçlar. Veri tabanında yer alan çoklular, bir sınıflama fonksiyonu yardımıyla kullanıcı tarafından belirlenir veya karar niteliğinin bazı değerlerine göre anlamlı ayrık alt sınıflara ayrılır. Bu yüzden sınıflama, denetimli öğrenmeye (supervised learning) girer. Sınıflama algoritması bir sınıfı diğerinden ayıran örüntüleri keşfeder ve iki şekilde kullanılır[16]:

- Karar Değişkeni ile Sınıflama: Seçilen bir niteliğin aldığı değerlere göre sınıflama işlemi yapılır. Seçilen nitelik karar değişkeni adını alır ve veri tabanındaki çoklular karar değişkeninin değerlerine göre sınıflara ayrılır. Bir sınıfta yer alan çoklular, karar değişkeninin değeri açısından özdeştir.
- Örnek ile Sınıflama: Bu biçimdeki sınıflamada veri tabanındaki çoklular iki kümeye ayrılır. Kümelerden biri pozitif, diğeri negatif çokluları içerir.

Sınıflama, verinin önceden belirlenen çıktılara uygun olarak ayrıştırılmasını sağlayan bir tekniktir. Çıktılar, önceden bilindiği için sınıflama, veri kümesini denetimli (supervised) olarak öğrenir. Örneğin; A finans hizmetleri şirketi; müşterilerinin yeni bir yatırım fırsatıyla ilgilenip ilgilenmediğini öğrenmek istemektedir. Daha önceden

benzer bir ürün satmıştır ve geçmiş veriler hangi müşterilerin önceki teklife cevap verdiğini göstermektedir. Amaç; bu teklife cevap veren müşterilerin özelliklerini belirlemek ve böylece pazarlama ve satış çalışmalarını daha etkin yürütmektir. Müşteri kayıtlarında müşterinin önceki teklife cevap verip vermediğini gösteren "evet"/ "hayır" şeklinde bir alan bulunur. Bu alan "hedef " ya da "bağımlı" değişken olarak adlandırılır. Amaç, müşterilerin diğer niteliklerinin (gelir düzeyi, iş türü, yaş, medeni durum, kaç yıldır müşteri olduğu, satın aldığı diğer ürün ve yatırım türleri) hedef değişken üzerindeki etkilerini analiz etmektir. Analizde yer alan diğer nitelikler "bağımsız" ya da " tahminci" değişken adını alır. Regresyon, sürekli sayısal bir değişkenin, aralarında doğrusal ya da doğrusal olmayan bir ilişki bulunduğu varsayılan diğer değişkenler yardımıyla tahmin edilmesi yöntemidir[16].

Regresyon modeli, sayısal değerleri tahmin etmeye yönelik olması dışında sınıflandırma yöntemine benzetilebilir. Sınıflama gruplanacak verileri tahmin ederken, regresyon süreklilik gösteren değerlerin tahmin edilmesinde kullanılır. Çok terimli lojistik regresyon gibi kategorik değerlerin de tahmin edilmesine olanaklı tekniklerin geliştirilmesi ile sınıflandırma ve regresyon modelleri giderek birbirine yaklaşmakta ve dolayısıyla aynı tekniklerden yararlanılması mümkün olmaktadır[27].

1. Diskriminant (ayrımlama) analizi: Veri setini tanımlama sürecinde amaç, veri hakkında özet bir bilgi elde etmektedir. Ayrımlama ise, veri setindeki farklılıkları ortaya koymak için yapılan bir işlemden ibarettir. Ayrımlama işleminde kullanılan en önemli yöntemlerden birisi Diskriminant analizidir[20].

Diskriminant analizi, bir dizi gözlemi önceden tanımlanmış sınıflara atayan bir tekniktir. Model, ait oldukları sınıf bilinen gözlem kümesi üzerine kurulur. Bu küme, öğrenme kümesi olarak da adlandırılır. Öğrenme kümesine dayalı olarak, diskriminant fonksiyonu olarak bilinen doğrusal fonksiyonların bir kümesi oluşturulur. Diskriminant fonksiyonu, yeni gözlemlerin ait olduğu sınıfı belirlemek için kullanılır. Yeni bir gözlem söz konusu olduğunda için tüm diskriminant fonksiyonları hesaplanır ve yeni gözlem diskriminant fonksiyonunun değerinin en yüksek olduğu sınıfa atanır[10].

Genel olarak birimlerin gruplamasında bazı matematiksel eşitliklerden faydalanılır. Diskriminant fonksiyonu olarak adlandırılan bu eşitlikler birbirine en çok benzeyen Grupları belirlemeye olanak sağlayacak şekilde grupların ortak özelliklerini belirlemek amacıyla kullanılmaktadır. Grupları ayırmak amacıyla kullanılan karakteristikler ise diskriminant değişkenleri olarak adlandırılmaktadır. Kısaca diskriminant analizi, iki veya daha fazla sayıdaki grubun farklılıklarının diskriminant değişkenleri vasıtasıyla ortaya konması işlemidir. Araştırıcının, p tane özelliği bilinen gözlemleri belli özelliklerine göre bazı gruplara ayırmak istemesi, elde edilecek somut ve özetleyici bilgiler açısından istatistiksel değerlendirmede önemli bir konudur. Diskriminant analizinin amaçlarını dört grupta toplanabilir[20]:

- Analiz öncesi tanımlanmış iki ya da daha fazla grubun (örneğin, mali açıdan başarılı ve başarısız işletmeler) ortalama özellikleri arasında önemli farklar olup olmadığının, bağımsız değişkenlere (açıklayıcı değişken) bağlı olarak istatistiksel olarak test edilmesi,
- Her bir değişkenin, gruplar arasındaki farka katkısının saptanması,
- Grup içi değişime oranla, gruplar arasındaki ayırımı maksimize eden tahmin değişkenleri kombinasyonunun belirlenmesi ve bu sayede başlangıçtaki açıklayıcı değişken sayısından daha az sayıda değişken ile gruplar arasındaki önemli farklılıkların açıklanması,
- Analiz öncesi tanımlanmış grupların atanması ile ilgili yöntemlerin geliştirilmesi, yeni bireylerin hangi gruba ait olduklarının saptanmasıdır.

Tüm istatistiksel ve matematiksel modellerde olduğu gibi, diskriminant analizi de bazı varsayımlara dayanmaktadır. Bunlar[20]:

- Ana kütle belli özelliklere göre gruplanabilir. Birbirinden farklı iki veya daha fazla grup söz konusu olmalıdır.
- Veriler ana kütleden rastsal olarak seçilmiştir.
- Bağımsız değişkenler çok boyutlu normal dağılıma sahiptirler.
- Gruplara ait ortalamalar ve kovaryans matrisi önceden bilinir. Grupların kovaryans (sapma) matrisleri eşittir. Bu varsayımın sağlanamadığı durumlarda, diskriminant analizinin karesel formu kullanılabilir.
- Grupların eşit sayıda birimden oluşmadığı durumlarda, üyelerin önsel olasılıklarının bilindiği varsayılır.
- Herhangi bir birimin yanlış sınıflandırmanın maliyeti önceden bellidir.

Diskriminant (ayırma) analizi, iki veya daha fazla sayıdaki grubun ayrımı ile ilgilenen çok değişkenli bir istatistik analiz tekniğidir. Diskriminant analizi bağımlı değişkenin nominal (metrik olmayan veya kategorik), bağımsız değişkenlerin ise metrik olduğu hallerde kullanılan en uygun tekniktir. İki gruplu bağımlı değişken söz konusu olduğunda analiz diskriminant analizi olarak ifade edilirken; grup sayısı üç veya daha fazla olduğunda analiz çoklu diskriminant analizi adını almaktadır. Diskriminant analizinde amaç, çok değişkenli problemin tek değişkenli biçime dönüştürülmesidir. Yani tüm değişkenlerin uygun ağırlıklarla katılacağı tek bir değişkenin (fonksiyon) elde edilmesidir. Pek çok analizde bağımlı değişken iki gruptan oluşur. Kadın-erkek, yüksek-düşük gibi. Diskriminant analizi ile iki veya daha çok (üç veya daha fazla) grup analiz edilebilir. Yüksek dereceli terimleri kapsayan fonksiyonlarda olduğu gibi değişkenlerden bazılarının sürekli, bazılarının kesikli olması durumlarında diskriminant analizine alternatif olarak lojistik regresyon analizi önerilmektedir[28].

2. Naive-Bayes algoritması: Naive-Bayes; hem tahmin edici hem de tanımlayıcı bir sınıflama tekniğidir. Her ilişkide koşullu bir olasılık türetmek için bağımlı ve bağımsız değişkenler arasındaki ilişkiyi analiz eder[13].
Naive Bayes, sürekli veri ile çalışmaz. Bu nedenle sürekli değerleri içeren bağımlı ya da bağımsız değişkenler kategorik hale getirilmelidir. Örneğin; bağımsız değişkenlerden biri yaş ise, sürekli değerler "<20" "21-30", "31-40" gibi yaş aralıklarına dönüştürülmelidir. Naive Bayes, modelin öğrenilmesi esnasında, her çıktının öğrenme kümesinde kaç kere meydana geldiğini hesaplar. Bulunan bu değer, öncelikli olasılık olarak adlandırılır. Naive Bayes aynı zamanda her bağımsız değişken/bağımlı değişken kombinasyonunun meydana gelme sıklığını bulur. Bu sıklıklar öncelikli olasılıklarla birleştirilmek suretiyle tahminde kullanılır[10].

3. Karar ağaçları: Karar verici, türlü seçeneklerin gerçekleşmesinin belirli ya da belirsiz olduğu bir problemle ilgili en iyi karara ulaşmak amacıyla, bazı işlemlerin yerine getirilmesi için birtakım yöntemlere veya araçlara ihtiyaç duyar. Seçenek sayısının fazla olduğu ve/veya ardışık aşamalarda karar almanın söz konusu olduğu problemlerin analizi, modellerin kurulması ve çözümlenmesi işlemlerinde karar vericiler bu araçlardan birisi olan "Karar Ağacı Analizi"ni kullanabilirler[28].

Karar ağaçları (KA), yaygın olarak kullanılan sınıflama algoritmalarından biridir. Karar ağacı yapılarında, her düğüm bir nitelik üzerinde gerçekleştirilen testi, her dal bu testin çıktısını, her yaprak düğüm ise sınıfları temsil eder. En üstteki düğüm kök düğüm olarak adlandırılır. Karar ağaçları, kök düğümden yaprak düğüme doğru çalışır[10].

Karar Ağacı olası tüm eylemlerin yönlerini, eylemlerin yönlerine etkisi olabilecek tüm olası faktörleri ve tüm bu faktörlere dayanan her bir olası sonucu, verilere bağlı olarak değerlendiren, çizgi, kare, daire gibi geometrik semboller kullanımı yoluyla karar vericiye sorunu anlamada kolaylık sağlayan düzenleme biçiminde tanımlanabilir. Bu tanıma göre karar ağacının türlü eylem seçeneklerini, farklı olası etkenlerin ve eylemlerin sonuçlarını içerdiği söylenebilir[28].

Karar ağacı yöntemini kullanarak verinin sınıflanması iki basamaklı bir işlemdir. İlk basamak öğrenme basamağıdır. Öğrenme basamağında önceden bilinen bir eğitim verisi, model oluşturmak amacı ile sınıflama algoritması tarafından çözümlenir. Öğrenilen model, sınıflama kuralları veya karar ağacı olarak gösterilir. İkinci basamak ise sınıflama basamağıdır. Sınıflama basamağında test verisi, sınıflama kurallarının veya karar ağacının doğruluğunu belirlemek amacıyla kullanılır. Eğer doğruluk kabul edilebilir oranda ise, kurallar yeni verilerin sınıflanması amacıyla kullanılır[1].

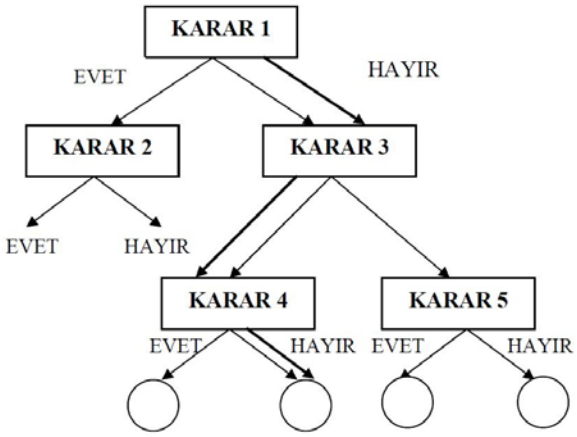
İstatistiksel yöntemlerde veya yapay sinir ağlarında veriden bir fonksiyon öğrenildikten sonra bu fonksiyonun insanlar tarafından anlaşılabilecek bir kural olarak yorumlanması zordur. Karar ağaçları ise veriden oluşturulduktan sonra ağaç kökten yaprağa doğru inilerek kurallar yazılabilir. Bu şekilde kural çıkarma veri madenciliği çalışmasının sonucunun doğrulanmasını sağlar. Bu kurallar uygulama konusunda uzman bir kişiye gösterilerek sonucun anlamlı olup olmadığı denetlenebilir. Sonradan başka bir teknik kullanılacak bile olsa karar ağacı ile önce bir kısa çalışma yapmak, önemli değişkenler ve yaklaşık kurallar konusunda analiste bilgi verir ve daha sonraki analizler için yol gösterici olabilir[16].

Karar ağacı, çoklu regresyondaki sınırlılıkları aşmak amacıyla geliştirilmiştir. Bu yöntemde karar ağaçları kullanılarak veri kümesi sonlu sayıda sınıfa ayrılır. Karar ağacındaki düğümler nitelik isimleriyle, dallar nitelik değerleriyle, yapraklar da farklı sınıf isimleriyle etiketlenir. Kök düğüm olarak da adlandırılan ilk eleman en yüksek karar düğümüdür, kullanılan algoritmaya bağlı olarak her düğüm iki veya daha fazla dala sahip olur. İki dala sahip olan karar ağaçları ikili ağaç, daha fazla dala sahip olanlar ise çok yollu ağaç olarak adlandırılır. Her dal bir başka karar düğümüyle, ya da ağacın sonuyla yani yaprak düğümle sonlanır. Karar düğümlerinde gerçekleştirilen her bölünmede oluşturulan gruplar arasındaki mesafenin maksimum olması bir başka deyişle elde edilen grupların mümkün olduğu kadar saf olması istenir. Kategorik değerleri sınıflandırmak için oluşturulan karar ağaçlarına sınıflandırma ağacı, sürekli sayısal değişkenleri tahmin etmek için kullanılan karar ağaçlarına ise regresyon ağacı denilmektedir[11].

Karar ağacı, karar vericinin en iyi karara ulaşılabilmesi için yapılan gerek olasılık gerekse maksimum fayda esas alınarak düzenlenen bir tekniktir. Karar ağacı analizi, genellikle seçenekler üzerinde yapılan bir analiz türüdür. Bu analizin veri madenciliğinde kullanılma sebepleri şunlardır[20].

- Maliyeti azdır.
- Anlaşılması ve yorumlanması kolaydır.
- Veri tabanına kolay entegre edilebilmektedir.
- Güvenirliliği yüksektir.

Bu analizin uygulamasında veri seti iki kısma ayrılır. İlk veri seti karar ağacını oluşturmak, ikinci kısım ise karar ağacını kontrol etmek amaçlı kullanılmaktadır[20]. Ağaç meydana getirilirken kurulan sistemin çalışıp çalışmadığı belirlenir. Eğer ağaç istenen düzeyde çalışıyorsa, dallanma sonlandırılır. Bu durum "durdurma" olarak adlandırılır. Durdurma kriteri ağacın hassasiyetini gösterir. Geç durdurulan bir ağaç daha fazla dallanacak, bu da istenmeyen sonuçların ortaya çıkmasına sebep olacaktır. Erken durdurulan ağaç ise tam öğrenmenin gerçekleşmemesi olasılığını taşıyacaktır. Şekil 2.3.'de basit bir karar ağacı yapısı görülmektedir.



Şekil 2.3. Örnek Bir Karar Ağacı[28]

Karar ağacı analizinde kullanılan algoritmalar şunlardır:

- CART veya C&RT**:** CART veya C&RT (Classification and Regression Trees) Breiman, Friedman, Olshen ve Stone tarafından 1984 yılında geliştirilmiş ikili ağaç olarak büyüyen bir algoritmadır. C&RT veriyi iki alt kümeye ayırır. Böylece bir sonraki adımda oluşacak olan alt küme, bir öncekinden daha homojen olacaktır. Bu süreç sonuç bulunana kadar devam eden, kendini tekrarlayan bir süreçtir. C&RT algoritması karmaşık bir algoritmadır. Büyük verilerle çalışıldığında sonuç bulması uzun sürmektedir. C&RT sınıflandırma ve regresyon analizi için kullanılan bir algoritmadır[28].
- CHAID: CHAID (Chi-Squared Automatic Interaction Detector) algoritması adından da anlaşılacağı gibi ayırma kriteri olarak Ki-kareyi kullanır. CHAID algoritması, tahmin edici değişkenin tüm değerlerini dikkate alarak analiz yapar. Hedef değişkeni dikkate alarak istatistik olarak benzer olan değişkenleri birleştirir ve farklı olan değişkenlerle işlemi sürdürür. Daha sonra karar ağacının ilk dalını oluşturmak için en iyi tahmin edici değişkeni seçer. Her bir düğüm, seçilen değişkenin benzer değerlerinden oluşur. Bu süreç ağaç tamamıyla büyüyene kadar tekrarlanarak devam eder. Yapılacak testler hedef değişkenin türüne göre değişmektedir. Eğer değişken sürekli bir değişken ise F testi, kategorik (nominal/ordinal) bir değişken ise Ki- kare testi kullanılır[20,28]. CHAID en popüler karar ağacı metotlarından biridir. CHAID algoritması ikili bir algoritma değildir, ki bu ağaçta herhangi bir seviyede ikiden çok kategori üretmesi anlamına gelir. Bu nedenle daha geniş ağaç oluşturmaya eğilimlidir. Her tür değişken için kullanılabilen bir tekniktir[28].

- C4.5: C4.5 algoritması 1993 yılında Quinlan tarafından ortaya atılmıştır. Karar ağacı oluşturulurken kayıp veriler hesaba katılmaz. C4.5 algoritması, kalitatif değişkenleri dikkate alır. Ayrıca kayıp verileri diğer veri ve değişkenler yardımı ile tahmin ederek, daha hassas ve daha anlamlı kurallar çıkartabilen bir ağaç üretebilir[28].
- C5.0: C5.0 algoritması örnek üzerindeki alan bölümlendirmelerinden en fazla bilgi sağlayacak şekilde çalışır ve alanlar bölünemeyecek duruma gelene kadar her bir alt örnek bölünmeye devam eder[19].
- QUEST: QUEST(Quick, Unbiased, Efficient, Statistical Tree) 1997 yılında Loh ve Shih tarafından geliştirilmiş olan yeni bir tekniktir. İkili büyüyen bir ağaç algoritmasıdır. Ayrı ayrı değişken seçimi ve ayırım noktası seçimi ile ilgilenir. QUEST'deki birim değişken ayıracı, tahmini olarak tarafsız değişken seçimini gerçekleştirir. QUEST algoritmasının C&RT algoritmasına benzer avantajları vardır, ancak ağaçlar yavaş büyüyebilir ve ikili olduğu için karar ağacı çok geniş olabilir[28].
- ID3(Induction of Decision Trees)[1]
- Exhaustive CHAID[1]
- MARS (Multivariate Adaptive Regression Splines)[1]
- SLIQ (Supervised Learning in Quest)[1]
- SPRINT (Scalable Parallelizable Induction of Decision Trees)[1].

4. Kaba kümeler: Kaba küme teorisi 1970'li yıllarda Pawlak tarafından geliştirilmiştir. Kaba küme teorisinde bir yaklaştırma uzayı ve bir kümenin alt ve üst yaklaştırmaları vardır. Yaklaştırma uzayı, ilgilenilen alanı ayrı kategorilerde sınıflandırır. Alt yakınlaştırma belirli bir altkümeye ait olduğu kesin olarak bilinen nesnelerin tanımıdır. Üst yakınlaştırma ise alt kümeye ait olması olası nesnelerin tanımıdır. Alt ve üst sınırlar arasında tanımlanan herhangi bir nesne ise "kaba küme" olarak adlandırılır[10].

Kaba küme teorisi sınıflandırmada, kesin olmayan ya da gürültülü veri içindeki yapısal ilişkileri keşfetmek için kullanılmaktadır. Kesikli değerli niteliklere uygulanır. Sürekli değerli nitelikler, kaba küme uygulanmadan önce kesiklileştirilmelidir[13].

5. Genetik algoritmalar: Genetik algoritmaların (GA) literatürde çok geniş bir kullanım alanı vardır, ancak teknik burada kısaca ele alınacaktır. Genetik algoritmalar çok değişkenli fonksiyonların optimizasyonu amacıyla kullanılan nümerik araştırma araçlarıdır. Olasılık kurallarına göre çalışan genetik algoritmalar, yalnızca amaç fonksiyonuna ihtiyaç duyarlar. Geleneksel optimizasyon yöntemlerine göre farklılıkları olan genetik algoritmalar, parametre kümesini değil kodlanmış biçimlerini kullanırlar. Genetik algoritmalarda, başlangıç olarak bir çözüm seti oluşturulur ve bu çözümü geliştirmek için biyolojik evrimi esas alan bir süreç kullanılır. Bu sürecin sonunda adaen iyi ikromozoma ulaşmak amaçtır. Çözüm uzayının tamamı değil bir kısmı taranır. Etkin arama yapılarak çok daha kısa bir zamanda çözüme ulaşılır[28].

GA, Darwin tarafından geliştirilen "evrim teorisini"ne dayalıdır. Algoritma ilk olarak popülasyon adı verilen bir çözüm kümesi (öğrenme veri kümesi) ile başlatılır. Bir popülasyondan alınan sonuçlar bir öncekinden daha iyi olacağı beklenen yeni bir popülasyon oluşturmak için kullanılır. Evrim süreci (yeni popülasyonlar yaratma iterasyonu) tamamlandığında bağımlılık kuralları veya sınıf modelleri ortaya konmuş olur[10].

GA, optimizasyon problemlerinde olduğu gibi sınıflandırma için de kullanılabilir. Veri madenciliğinde, diğer algoritmaların uygunluğunu değerlendirmek için kullanılabilirler[12].

GA, biyoloji biliminden yararlanılarak geliştirilmiş önemli makine öğrenimi yöntemlerinden birisidir. İlk olarak 1960 ve 1970'li yıllarda çalışılmış ve 1980'li yılların sonuna doğru kabul edilmiş bir yöntemdir. Genetik algoritmalar iki konuda doğadan esinlenmiştir. Bunlardan birincisi, çok sayıda canlı türlerinden çevreye uyum sağlayanın yasaması ve uyum sağlayamayanların yok olmasıdır. Bu yaklaşım, problem çözümlerine esin kaynağı olmuştur. İkincisi ise, canlıların DNA yapısıdır. Bu yaklaşım ise, kodlama problemleri için esin kaynağı olmuştur. GA açıklanabilir sonuçlar üretirler. Değişik tiplerdeki verileri isleme özelliğine sahip olan genetik algoritmalar en iyileme (optimization) amacı ile kullanılabilirler. Ayrıca genetik algoritmalar yapay sinir ağları ile ortaklasa çalışarak başarılı sonuçlar

üretmektedirler. GA yapay sinir ağlarının eğitilmesi, bellek tabanlı yöntemlerde kombinasyon fonksiyonunun oluşturulması gibi islerde de kullanılmışlardır. Tüm bu olumlu yönlerine rağmen genetik algoritmaların kullanımlarında bazı sıkıntılar da yaşanmaktadır. Bunlardan en belirgin olanı karmaşık sorunların genetik kodlanmasının çok zor olmasıdır. Ayrıca en iyi (optimal) sonucun üretildiğine dair bir garanti de bulunmamaktadır[1].

Genetik ortamın, programlama teknikleri kullanılarak kodlanması genetik algoritma olarak adlandırılır. En iyinin korunumu ve doğal seçilim ilkesinin benzetim yoluyla bilgisayarlara uygulanması ile elde edilen bir arama yöntemidir. Genetik algoritma (GA), bir problemin olası çözümlerinden oluşan sabit büyüklükte bir çözüm grubu içinde tekrarlanarak yürütülen işlemlerden oluşan bir yöntemdir[13].

6. K-en yakın komşu algoritması: Bellek tabanlı yöntemler istatistikte 1950'li yıllarda önerilmiş olmasına rağmen o yıllarda gerektirdiği hesaplama ve bellek yüzünden kullanılamamış ama günümüzde bilgisayarların ucuzlaması ve kapasitelerinin artmasıyla, özellikle de çok işlemcili sistemlerin yaygınlaşmasıyla, kullanılabilir olmuştur. Bu yönteme en iyi örnek, k-en yakın komşu algoritmasıdır (k-nearest neighbor)[16].

K-en yakın komşu algoritması, veri madenciliğinde kümeleme amacıyla kullanılan tekniklerden birisidir. Bu teknik ile kümeleme analizi tahmin yapmak için kullanılabilir. Algoritma, bilinmeyen bir gözlem değerinin hangi sınıfa dahil olduğunu belirlemek için, örüntü uzayını araştırarak bilinmeyen gözlem değerine en yakın olan k adet küme sayısını bulur. Yakınlık öklid uzaklığı ile belirlenir. Daha sonra bilinmeyen değer, k- en yakın komşu içinden en çok benzediği kümeye atanır. K-en yakın komşu algoritması aynı zamanda bilinmeyen gözlem değeri için bir gerçek değerin tahmininde de kullanılabilir[28].

Komşuluk hesaplamaları yapılırken, daha yakın komsulara daha büyük ağırlık değerleri atanabilir. K-en yakın komşuluk yönteminde sınıflandırılmak istenen olay sayısı arttıkça hesaplamalar için gereken sürede hızlı bir şekilde artar, k en yakın komşuluk modelinin işlem hızını artırmak için genellikle bütün veri hafızada tutulur[11].

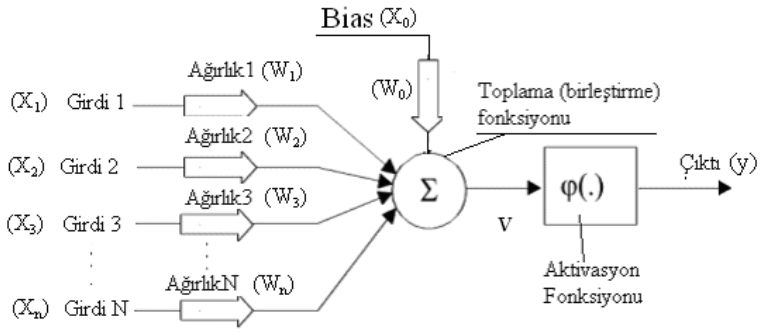
K-ortalamalar algoritması oldukça etkin bir algoritma olmakla birlikte; sadece nümerik veri ile çalışır fakat veri madenciliği uygulamaları sıklıkla kategorik verileri de içermektedir. K-ortalamalar algoritmasının geliştirilmesi ile elde edilen k-modlar algoritması ise kategorik veriler üzerinde çalışabilen bir algoritmadır. K-ortalamalar algoritmasında küme merkezleri, küme ortalaması alınarak hesaplanırken; k-modlar algoritmasında küme merkezlerinin belirlenmesinde kümede en sık tekrarlanan değerler (mod) dikkate alınır [10].

7. Yapay sinir ağları: Sinir ağları, tanımlayıcı ve tahminci veri madenciliği algoritmalarındandır. İnsan beyninin fizyolojisini taklit ederler. Komplike ve belirsiz veriden bilgi üretirler. Keşfettikleri örüntü ve trendler, insanlar ya da bilgisayarlarca kolay keşfedilemez. Bu tür karmaşık problemlerde birbirleriyle etkileşimli yüzlerce değişken bulunur. Bu teknik, veritabanındaki örüntüleri, sınıflandırma ve tahminde kullanılmak üzere genelleştirir. Sinir ağları algoritmaları sayısal veriler üzerinde çalışırlar[10].

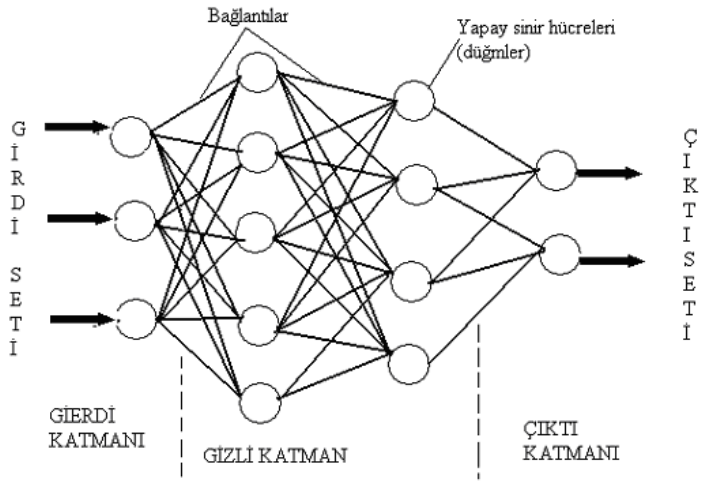
Sinir ağları, insan beyninden esinlenilerek şekillendirilmiştir. İnsan beyni çok karmaşık bir yapıya sahiptir. İnsan beyninin karmaşık yapısı daha karmaşık öğrenme makinelerinin oluşturulmasında model olarak alınmış ve yapay sinir ağları (YSA) olarak adlandırılmıştır. YSA için farklı yapılar vardır ve bunların her biri verilen işleri yapmak için farklı yol ve öğrenme yöntemleri kullanırlar. Yapay sinir ağının, belirli bir işi yapmak için eğitildiği evreye şifreleme evresi, sınıflama ya da kestirim yapma evresine ise şifre çözme evresi adı verilir. YSA, VM için çok kullanışlı bir yöntem olmasının yanında anlaşılmaz modeller de ortaya çıkardığı için uygulaması çok uzun zaman gerektirebilir. YSA, verideki eğilimleri ve örüntüleri belirlemek için çok uygundur. Bu yüzden tahmin yapmak için de kullanılır. YSA, insanların deneyimlerinden bir takım bilgiler çıkartması gibi kendisine verilen örneklerden bir takım bilgiler çıkartma yeteneğine sahiptir. YSA öncelikle sonuçları bilinen belirli bir veri kümesi üzerinde öğrenme algoritmaları çalıştırılarak eğitilir. Bu eğitim neticesinde yapay sinir ağının içerisindeki ağırlıklar belirlenir. Bu ağırlıklar kullanılarak yeni gelen veriler işlenir ve bir sonuç üretilir. Yapay sinir ağlarının en olumsuz tarafı ise bu ağırlıkların neden ilgili değeri aldığının bilinmemesi ya da çıkan sonucun neden geçerli bir sonuç olduğunu açıklayamamasıdır. Yapay sinir ağlarını kullanmak için en iyi yaklaşım onları içi bilinmeyen bir şekilde çalışan kara

kutular olarak düşünmek olacaktır. YSA; sınıflama, kümeleme ve tahmin amaçları ile kolaylıkla kullanılabilecek genel amaçlı ve güçlü araçlardır. Ekonomik alanlardan tıbbi konulara, değerli müşterilerin belirlenmesi için yapılan kümeleme işlemlerinden kredi kartlarında sahtekârlıkların belirlenmesine kadar çok geniş bir alanda uygulanabilmektedir[1].

Şekil 2.4.'de bir yapay sinir hücresinin yapısı yer almaktadır. Şekil 2.5.'de yer aldığı gibi YSA'nın üç katmanı vardır. İlk olarak giriş katmanı, karar vermede kullanılacak verilerin sisteme aktarıldığı katmandır. Bu katman, ağın içinde görünmeyen bir katmana bağlıdır. Bu katmanda birden çok birbirine bağlı "nöronlar" vardır. Bunlar giriş değerine göre bir çıkış değeri üretir ve bir sonraki nörona gönderir. Her bir nöron kendi üzerindeki karmaşık fonksiyonu kullanarak bir sonuç üretir ve sonunda bu sonuçlar çıkış katmanı tarafından toplanır ve modelin sonucu olarak dışarı verilir. Yapı kurulduktan sonra sinir ağının eğitilmesine sıra gelir. Giriş verileri verilir ve çıkış değeri alınır. Bu değer gerçek değer ile karşılaştırılır ve ağın içerisindeki nöron fonksiyonlarının bu sonuçtaki hata miktarına göre ayarlanması sağlanır. Bu şekilde birçok değer ağa verilir ve ağın elimizdeki verinin yapısını öğrenmesi sağlanır. Öğrenme işlemi tamamlandıktan sonra sinir ağımız kullanıma hazır hale gelir[19].



Şekil 2.4. Yapay sinir hücresinin yapısı [51]



Şekil 2.5. Bir yapay sinir ağı örneği [51]

Sınıflama amaçlı kullanılan yapay sinir ağları, geri yayılım algoritması ve RBF (Radial Basis Function) ağlarıdır. Kümeleme amaçlı kullanılan yapay sinir ağları, 80'lerin basında Kohonen tarafından geliştirilen öz düzenlemeli haritadır. Yapay sinir ağlarının VM açısından kuvvetli yönleri şunlardır[1]:

- Çok geniş açıdaki (spektrum) sorunların çözümünde kullanılabilirler,
- Çok karmaşık durumlarda dahi iyi sonuçlar üretmektedirler,
- Hem sayısal hem de kategorik veriler üzerinde işlem yapabilirler.

Yapay sinir ağlarının VM açısından zayıf yönleri de şunlardır[1];

- 0 ile 1 arasında giriş verileri olması zorunludur,
- Ürettikleri sonuçların açıklamasını yapamazlar.

Yapay sinir ağlarında kullanılan öğrenme algoritmaları veriden, üniteler arasındaki bağlantı ağırlıklarını hesaplar. YSA istatistiksel yöntemler gibi veri hakkında parametrik bir model varsaymaz yani uygulama alanı daha geniştir ve bellek tabanlı yöntemler kadar yüksek işlem ve bellek gerektirmez[16].

Yapay sinir ağları, insan beyninin özelliklerinden olan öğrenme yolu ile yeni bilgiler türetebilme, yeni bilgiler oluşturabilme ve keşfedebilme gibi yetenekleri herhangi bir yardım almadan otomatik olarak gerçekleştirmek amacı ile geliştirilen bilgisayar sistemleridir. Bu yetenekleri geleneksel programlama yöntemleri ile gerçekleştirmek oldukça zordur veya mümkün değildir. Bu nedenle yapay sinir ağlarının,

programlanması çok zor veya mümkün olmayan olaylar için geliştirilmiş adaptif bilgi işleme ile ilgilenen bir bilgisayar bilim dalı olduğu söylenebilir[28].

8. İstisna analizi: Bir veri tabanı, tüm veri modelinin davranışını sergilemeyen veriler içeriyor olabilir. Bu tür veriler istisna (outliers) olarak adlandırılırlar. Birçok veri madenciliği tekniği istisnaları gürültü olarak adlandırır. Buna rağmen bazı uygulamalarda örneğin hile tespitinde (fraud detection), daha seyrek oluşmuş olan olaylar sık oluşmuş olanlara göre daha ilginç ve önemli olabilirler. Outlier verinin analizi, outlier analiz olarak adlandırılır[1,16].

9. Faktör analizi: Faktör analizi, veriler arasındaki ilişkilere dayanan, verilerin daha anlamlı ve özet biçimde sunulmasını sağlayan çok değişkenli bir istatistik analiz tekniğidir. Faktör analizinin amacı esas olarak değişkenler arasındaki karşılıklı bağımlılığın kökenini araştırmaktır. Faktör analizi fazla sayıdaki veri setinin azaltılması ve basitleştirilmesi amacıyla yaygın olarak kullanılmaktadır. Faktör analizinin başlıca varsayımları, veri matrisinin analiz öncesi kriter ve tahmin değişkenleri alt matrislerine bölüştürülmemesi ve değişkenler arasındaki ilişkinin doğrusal olduğudur. Başta sosyal bilimler olmak üzere pek çok alanda sıkça kullanılan faktör analizi, çok değişkenli analiz tekniklerinden biridir. Araştırmacı gözlenemeyen bazı özellikler için (imaj, kişilik, zeka, vatanseverlik gibi) bir takım ölçekler geliştirmek durumundadır. Faktör analizi bu durumda kullanılır. Faktör analizi, özellikle tüketici eğilimleri, tüketici tercihleri, tüketici davranışları gibi davranışsal konular başta olmak üzere çeşitli pazarlama sorunlarında da sık sık başvurulan bir tekniktir[28].

10. Doğrusal regresyon, lojistik regresyon: Doğrusal regresyonda, veriler düz bir çizgi kullanılarak modellenir. Doğrusal regresyon (DR), kestirim çeşitlerinden en basit olanıdır. İki değişkenli (bivariate) doğrusal regresyon rastgele değerler üretir[12].

Regresyon analizi, bir ya da daha fazla bağımsız değişken ile hedef değişken arasındaki ilişkiyi matematiksel olarak modelleyen bir yöntemdir. Veri madenciliğinde yaygın olarak kullanılan regresyon modellerinden doğrusal

regresyonda tahmin edilecek olan hedef değişken sürekli değer alırken; lojistik regresyonda hedef değişken kesikli bir değer almaktadır. Doğrusal regresyonda hedef değişkenin değeri; lojistik regresyonda ise hedef değişkenin alabileceği değerlerden birinin gerçekleşme olasılığı tahmin edilmektedir[10].

Lojistik regresyon (LR) son yıllarda ünlenmiş ve çoklu regresyon analizi ile benzeştiği için, analiz yapan kişilerce yoğun bir biçimde kullanılmaya başlanmıştır. Bu yöntem, çeşitli varsayımlardan sapma (normallik, ortak kovaryansa sahip olma gibi) durumunda diskriminant analizi ve çapraz tablolara bir alternatif olmaktadır. Bağımlı değişkenin 0/1 gibi ikili (binary) ya da ikiden çok düzey içeren kesikli değişken olması durumunda, normallik kısıtı olmaması nedeniyle kullanım rahatlığının yanı sıra, çözümlemeden elde edilen modelin matematiksel olarak çok esnek olması ve kolay yorumlanabilir olması sebebiyle gün geçtikçe daha çok ilgi görmektedir. Lojistik regresyon analizinin kullanılabileceği iki örnek aşağıda verilmiştir. Bir kablolu yayın şirketinin pazarlama yöneticisi; özel bir TV programları paketine üye olabilecek kişilerin bu pakete üye olma olasılıkları ile; bu kişilerin geliri, eğitimi, mesleği, yaşı, medeni durumu ve çocuk sayısı arasında bir ilişki olup olmadığını incelemek istemektedir. Bir denetçi, inceleme yaptığı bir firmanın başarısız olma olasılığının, firmanın finansal oranları ve firma büyüklüğüyle ilgili olup olmadığını tespit etmek istemektedir[28].

11. Kanonik korelasyon analizi: Çoklu regresyon analizinde bir bağımlı ve birden fazla bağımsız değişken arasındaki ilişki tahmin edilirken; kanonik korelasyon analizinde birden fazla bağımlı ve birden fazla bağımsız değişken arasındaki ilişki tahmin edilmeye çalışılır. Kanonik korelasyon analizi, bağımlı değişkenler seti ile bağımsız değişkenler seti arasındaki korelasyonu belirlemeyi amaçlayan çok değişkenli bir istatistik tekniğidir[28].

2.8.2.2. Tahmin modelleri

– Bellek tabanlı yöntemler: İnsanlar kararlarını genellikle daha önce yaşadıkları deneyimlere göre verirler. Örneğin doktorlar bir hastayı incelerken, elde ettiği bulguları daha önce tedavi ettiği benzer hastalığa yakalanmış hastalar üzerindeki

deneyimlerini kullanarak değerlendirirler. Bellek tabanlı yöntemler de benzer şekilde deneyimleri kullanmaktadır. Bu yöntemlerde, bilinen kayıtların bulunduğu bir veritabanı oluşturulur ve sistem yeni gelen bir kayda komşu olan diğer kayıtları belirler ve bu kayıtları kullanarak tahminde bulunur ya da bir sınıflama işlemi uygular. Bellek tabanlı yöntemlerin en önemli özelliği veriyi olduğu gibi kullanabilme yeteneğidir. Diğer VM yöntemlerinin aksine bellek tabanlı yöntemler, kayıtların şekli (format) yerine sadece iki işlemin varlığı ile ilgilenir. Bu işlemler, iki kayıt arasındaki uzaklığı belirleyen bir uzaklık fonksiyonu ve komşu kayıtları işleyerek bir sonuç üreten kombinasyon fonksiyonudur. Bellek tabanlı yöntemler sahtekarlık tespiti ve klinik işlemler gibi alanlarda kullanılmaktadır. Bellek tabanlı yöntemlerin güçlü olduğu noktalar[1];

- Kolayca anlaşılabilir sonuçlar üretir,
- Rasgele seçilen, hatta birbiri ile ilgisiz olabilen verilere bile uygulanabilir,
- Çözümleme alanlarının çok olduğu durumlarda dahi etkili olarak çalışabilir,
- Eğitim kümesinin oluşturulması basittir.

Bellek tabanlı yöntemlerin zayıf olduğu noktalar ise[1]:

- Sınıflama ve tahmin işlemleri için kullanıldığında işlem maliyeti yüksektir,
- Eğitim kümesi için büyük miktarlarda alana ihtiyaç vardır.

– Hipotez testi: Hipotez testi algoritmaları doğrulamaya dayalı algoritmalardır. Doğrulanacak hipotez VT üzerindeki verilerle belli doğruluk ve destek değerlerine göre sınanır. Sınama işlemi uzman tarafından bir varsayım kural olarak ortaya çıkarılmak istendiğinde ve ortaya çıkmış bir kuralın budanması ve genişletilmesi durumunda yapılır[8].

Hipotez ya mantıksal bir kural ya da mantıksal bir ifade ile gösterilir. Her iki biçimde de seçilen veritabanındaki nitelik alanları kullanılır[27].

Hipotez testi algoritmasında öne sürülen hipotez genellikle belirli bir örüntünün veri tabanındaki varlığıyla ilgili bir tahmindir. Bu tip bir analiz özellikle keşfedilmiş bilginin genişletilmesi veya damıtılması işlemleri sırasında yararlıdır[16,27].

Etkin bir VM algoritması geliştirebilmek için aşağıdaki hususlara dikkat edilmesi gerekmektedir[16]:

1. Veri gizliliği ve güvenliğinin sağlanması: Bir VTBK sisteminde keşfedilen bilgi pek çok farklı açıdan ve soyutlama düzeyinden izlenebildiği için, gizlilik ve veri güvenliği, VM sistemini kullanan kullanıcının haklarına ve erişim yetkilerine göre sağlanmalıdır.

2. Sonuçların yararlılık, kesinlik ve anlamlılık kıstaslarını sağlaması: Elde edilen sonuçlar analiz için kullanılan VT'yi doğru biçimde yansıtmalıdır. Bunun yanı sıra gürültülü ve aykırı veriler işlenmelidir. Bu işlem elde edilen kuralların kalitesini belirlemede önemli bir rol oynar.

3. Farklı tipteki verileri ele alma: Gerçek hayattaki uygulamalar makine öğreniminde olduğu gibi yalnızca sembolik veya kategorik veri türleri üzerinde değil, aynı zamanda tamsayı, kesirli sayı, çoklu ortam verisi ve coğrafi veri gibi farklı tipteki veriler üzerinde de işlem yapılmasını gerektirir. Kullanılan verinin saklandığı ortam, düz bir kütük veya ilişkisel VT'de yer alan tablolar olabileceği gibi, nesneye yönelik VT'ler, çoklu ortam VT'leri ve coğrafi VT'ler vb. de olabilir. Saklandığı ortama göre veri, basit tipte olabileceği gibi karmaşık veri tipleri (çoklu ortam verisi, zaman boyutlu veri, yardımlı metin, coğrafi veri vb.) de olabilir. Bununla birlikte veri tipi çeşitliliğinin fazla olması bir VM algoritmasının tüm veri tiplerini ele alabilmesini olanaksızlaştırmaktadır. Bu yüzden veri tipine özgü adanmış VM algoritmaları geliştirilmektedir.

4. Farklı ortamlarda yer alan veri üzerinde işlem yapabilme: Kurumlar yerel ağlar üzerinden pek çok dağıtık ve heterojen VT üzerinde işlem yapmaktadır. Bu VM'nin farklı kaynaklarda birikmiş biçimli ya da biçimsiz veriler üzerinde analiz yapabilmesini gerektirir. Veri büyüklüğünün yanı sıra verinin dağıtık olması, yeni araştırma alanlarının ortaya çıkmasına sebep olmuştur. Bunlar, koşut ve dağıtık VM algoritmalarıdır.

5. Veri madenciliği algoritmasının etkinliği ve ölçeklenebilirliği: Çok büyük hacimli veri içinden bilgi elde etmek için kullanılan VM algoritmasının etkin ve ölçeklenebilir olması gerekir. Bu, VM algoritmasının çalışma zamanının tahmin edilebilir ve kabul edilebilir bir süre olmasını gerektirir. Üssel veya çok terimli bir karmaşıklığa sahip bir VM algoritmasının uygulanması kullanışlı değildir.

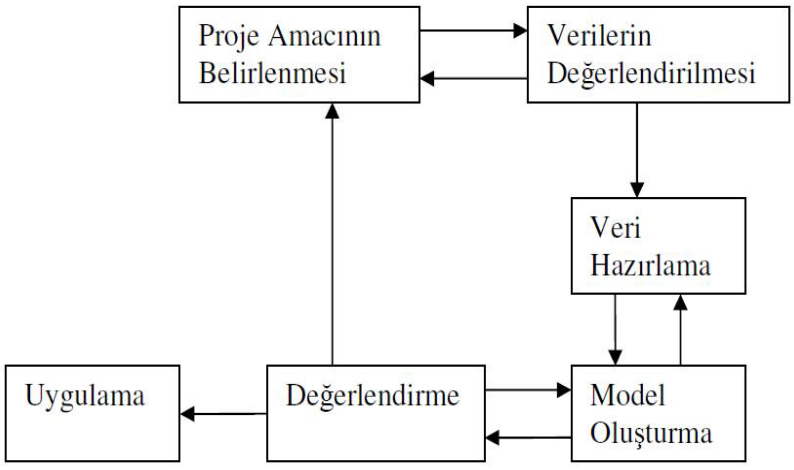
6. Keşfedilen kuralların çeşitli biçimlerde gösterimi: Bu özellik keşfedilen bilginin gösterim biçiminin seçilebilmesini sağlayan yüksek düzeyli bir dil tanımının

yapılmasını ve grafik arayüzünü gerektirir.

7. Farklı bir kaç soyutlama düzeyi ve etkileşimli veri madenciliği: Büyük VT'lerden VM sorgularıyla elde edilecek bilginin edinilmesi güçtür. Bu yüzden VM sorgusu, elde edilen bilgilere göre kullanıcıya etkileşimli olarak sorgusunu değiştirebilmeyi, farklı açılardan ve farklı soyutlama düzeylerinden keşfedilen bilgiyi inceleyebilme esnekliğini sağlamalıdır.

2.9. Veri Madenciliği Sürecinin Aşamaları

Birçok kurum kendi problemlerine, verilerine ve sahip oldukları diğer kaynaklara göre kendi veri madenciliği sürecini oluşturmaktadır ancak veri madenciliği sürecinin oluşturulmasında yapılan yanlışlıklar, sürecin etkinliğine zarar vermektedir. Veri madenciliği sürecinin standartlaştırılması konusunda farklı grup, kurum ve şirketler çeşitli standartlar oluşturmuşlardır bunlardan en çok takip edileni Daimler Chrysler ve SPSS tarafından 1996 yılında oluşturulan Veri Madenciliği için Sektörler Arası Standart Sürecidir (CRISP-DM). CRISP-DM süreci altı aşamadan oluşan etkileşimli ve yinelemeli bir süreçtir. Şekil 2.6'da gösterilen akış semasının herhangi bir aşamasında elde edilen sonuçlara göre sonraki aşamaya ya da önceki bir aşamaya geçilip yeni belirlenen problemlere, ilgi alanlarına göre iyileştirmeler ya da farklı işlemler yapılabilir[11].



Şekil 2.6. CRISP-DM Süreci[11]

2.9.1. Araştırma probleminin tanımlanması

Veri madenciliği çalışmalarında başarılı olmanın en önemli şartı, projenin hangi işletme amacı için yapılacağının ve elde edilecek sonuçların başarı düzeylerinin nasıl ölçüleceğinin tanımlanmasıdır. Ayrıca yanlış tahminlerde katlanılacak olan maliyetlere ve doğru tahminlerde kazanılacak faydalara ilişkin tahminlere de bu aşamada yer verilmelidir. Bu aşamada mevcut iş probleminin nasıl bir sonuç üretilmesi durumunda çözüleceğinin, üretilecek olan sonucun fayda-maliyet analizinin başka bir ifadeyle üretilen bilginin işletme için değerinin doğru analiz edilmesi gerekmektedir[10,16].

2.9.2. Verileri tanıma aşaması

Veri anlama aşaması veri toplamakla başlamaktadır. Daha sonra benzer verileri bir araya getirme, veri niteliklerini tanımlama, verileri keşfetme, gizli bilgileri sınıflandırma ile sürece devam etmektedir. Diğer bir anlatımla bu aşama ilk verilerin toplanması, verinin tanımlanması, verilerin keşfedilmesi ve verilerin kalitesinin incelenmesi faaliyetlerini içerir[3].

2.9.3. Veri hazırlama aşaması

Veri madenciliğinin en önemli aşamalarından biri olan verinin hazırlanması aşaması, analistin toplam zaman ve enerjisinin %50 - %85 ini harcamasına neden olmaktadır. Bu aşamada firmanın mevcut bilgi sistemleri üzerinde ürettiği sayısal bilginin iyi analiz edilmesi, veriler ile mevcut iş problemi arasında ilişki olması gerektiği unutulmamalıdır. Proje kapsamında kullanılacak sayısal verilerin, hangi iş süreçleri ile oluşturulduğu da bu veriler kullanılmadan analiz edilmelidir. Bu sayede analist veri kalitesi hakkında fikir sahibi olabilir[10,16].

Veri hazırlama aşaması kendi içinde veri toplama, değer biçme, birleştirme-temizleme, seçim ve dönüştürme adımlarından oluşur.

- Veri toplama**:** Tanımlanan problem için gerekli olduğu düşünülen verilerin ve bu verilerin toplanacağı veri kaynaklarının belirlenmesi adımıdır. Verilerin toplanmasında kuruluşun kendi veri kaynaklarının dışında, nüfus sayımı,

hava durumu, merkez bankası kara listesi gibi veri tabanlarından veya veri pazarlayan kuruluşların veri tabanlarından faydalanılabilir[10,16].

- Değer biçme: Veri madenciliğinde kullanılacak verilerin farklı kaynaklardan toplanması, doğal olarak veri uyumsuzluklarına neden olacaktır. Bu uyumsuzlukların başlıca olanları farklı zamanlara ait olmaları, kodlama farklılıkları (örneğin bir veri tabanında cinsiyet özelliğinin e/k, diğer bir veri tabanında 0/1 olarak kodlanması), farklı ölçü birimleridir. Ayrıca verilerin nasıl, nerede ve hangi koşullar altında toplandığı da önem taşımaktadır. Bu nedenlerle, iyi sonuç alınacak modeller ancak iyi verilerin üzerine kurulabileceği için, toplanan verilerin ne ölçüde uyumlu oldukları bu adımda incelenerek değerlendirilmelidir[27].
- Birleştirme ve temizleme: Bu adımda farklı kaynaklardan toplanan verilerde bulunan ve bir önceki adımda belirlenen sorunlar mümkün olduğu ölçüde giderilerek veriler tek bir veri tabanında toplanır. Ancak basit yöntemlerle ve baştan savma olarak yapılacak sorun giderme işlemlerinin, ileriki aşamalarda daha büyük sorunların kaynağı olacağı unutulmamalıdır[16,27]. Hatalı veya analizin yanlış yönlenmesine sebep olabilecek verilerin temizlenmesine çalışılır. Genellikle yanlış veri girişinden veya bir kereye özgü bir olayın gerçekleşmesinden kaynaklanan verilerin, önemli bir uyarıcı enformasyon içerip içermediği kontrol edildikten sonra veri kümesinden atılması tercih edilir[10,27].
- Seçim: Bu adımda kurulacak modele bağlı olarak veri seçimi yapılır. Örneğin tahmin edici bir model için, bu adım bağımlı ve bağımsız değişkenlerin ve modelin eğitiminde kullanılacak veri kümesinin seçilmesi anlamını taşımaktadır. Sıra numarası, kimlik numarası gibi anlamlı olmayan ve diğer değişkenlerin modeldeki ağırlığının azalmasına da neden olabilecek değişkenlerin modele girmemesi gerekmektedir. Bazı veri madenciliği algoritmaları konu ile ilgisi olmayan bu tip değişkenleri otomatik olarak elese de, pratikte bu işlemin kullanılan yazılıma bırakılmaması daha akılcı olacaktır. Modelde kullanılan veri tabanının çok büyük olması durumunda tesadüfîliği bozmayacak şekilde örnekleme yapılması uygun olabilir. Günümüzde hesaplama olanakları ne kadar gelişmiş olursa olsun, çok büyük veri tabanları üzerinde çok sayıda modelin denenmesi zaman kısıtı nedeni ile mümkün olamamaktadır. Bu nedenle tüm veri tabanını kullanarak bir kaç

model denemek yerine, tesadüfî olarak örneklenmiş bir veri tabanı parçası üzerinde birçok modelin denenmesi ve bunlar arasından en güvenilir ve güçlü modelin seçilmesi daha uygun olacaktır[27].

- Veri dönüştürme: Kullanılacak model ve algoritma çerçevesinde verilerin tanımlama veya gösterim şeklinde değiştirilmesi gerekebilir. Örneğin kredi riski uygulamasında iş tiplerinin, gelir seviyesi ve yaş gibi değişkenlerin kodlanarak gruplanmasının faydalı olacağı düşünülmektedir [10,16].

2.9.4. Modelin kurulması

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

Tanımlanan problem için en uygun modelin bulunabilmesi, olabildiğince çok sayıda modelin kurularak denenmesi ile mümkündür. Bu nedenle veri hazırlama ve model kurma aşamaları, en iyi olduğu düşünülen modele varılıncaya kadar yinelenen bir süreçtir. Model kuruluş süreci denetimli (supervised) ve denetimsiz (unsupervised) öğrenimin kullanıldığı modellere göre farklılık göstermektedir. Örnekten öğrenme olarak da isimlendirilen denetimli öğrenimde, bir denetçi tarafından ilgili sınıflar önceden belirlenen bir kritere göre ayrılarak, her sınıf için çeşitli örnekler verilir. Sistemin amacı verilen örneklerden hareket ederek her bir sınıfa ilişkin özelliklerin bulunmasıdır. Öğrenme süreci tamamlandığında, tanımlanan kurallar verilen yeni örneklere uygulanır ve yeni örneklerin hangi sınıfa ait olduğu kurulan model tarafından belirlenir. Denetimsiz öğrenmede, kümeleme analizinde olduğu gibi ilgili örneklerin gözlenmesi ve bu örneklerin özellikleri arasındaki benzerliklerden hareket ederek sınıfların tanımlanması amaçlanmaktadır. Denetimli öğrenimde seçilen algoritmaya uygun olarak ilgili veriler hazırlandıktan sonra, ilk aşamada verinin bir

kısmı modelin öğrenimi, diğer kısmı ise modelin geçerliliğinin test edilmesi için ayrılır. Modelin öğrenimi öğrenim kümesi kullanılarak gerçekleştirildikten sonra, test kümesi ile modelin doğruluk derecesi (accuracy) belirlenir[10]. Bir modelin doğruluğunun test edilmesinde kullanılan en basit yöntem basit geçerlilik (simle validation) testidir. Bu yöntemde tipik olarak verilerin %5 ile %33 arasındaki bir kısmı test verileri olarak ayrılır ve kalan kısım üzerinde modelin öğrenimi gerçekleştirildikten sonra, bu veriler üzerinde test işlemi yapılır. Bir sınıflama modelinde yanlış olarak sınıflanan olay sayısının, tüm olay sayısına bölünmesi ile hata oranı, doğru olarak sınıflanan olay sayısının tüm olay sayısına bölünmesi ile ise doğruluk oranı hesaplanır(Doğruluk Oranı=1-Hata Oranı)[10,16].

Sınırlı miktarda veriye sahip olunması durumunda, kullanılabilecek diğer bir yöntem çapraz geçerlilik (cross validation) testidir. Bu yöntemde veri kümesi tesadüfi olarak iki eşit parçaya ayrılır. İlk aşamada a parçası üzerinde model eğitimi ve b parçası üzerinde test işlemi; ikinci aşamada ise b parçası üzerinde model eğitimi ve a parçası üzerinde test işlemi yapılarak elde edilen hata oranlarının ortalaması kullanılır. Bir kaç bin veya daha az satırdan meydana gelen küçük veri tabanlarında, verilerin n gruba ayrıldığı n katlı çapraz geçerlilik (n-fold cross validation) testi tercih edilebilir. Verilerin örneğin 10 gruba ayrıldığı bu yöntemde, ilk aşamada birinci grup test, diğer gruplar öğrenim için kullanılır. Bu süreç her defasında bir grubun test, diğer grupların öğrenim amaçlı kullanılması ile sürdürülür. Sonuçta elde edilen on hata oranının ortalaması, kurulan modelin tahmini hata oranı olacaktır. Model kuruluşu çalışmalarının sonucuna bağlı olarak, aynı teknikle farklı parametrelerin kullanıldığı veya başka algoritma ve araçların denendiği değişik modeller kurulabilir. Model kuruluş çalışmalarına başlamazdan önce, imkânsız olmasa da hangi tekniğin en uygun olduğuna karar verebilmek güçtür. Bu nedenle farklı modeller kurarak, doğruluk derecelerine göre en uygun modeli bulmak üzere sayısız deneme yapılmasında yarar bulunmaktadır. Özellikle sınıflama problemleri için kurulan modellerin doğruluk derecelerinin değerlendirilmesinde basit ancak faydalı bir araç olan risk (yakınsaklık) matrisi kullanılmaktadır. Aşağıda bir örneği görülen bu matriste sütunlarda fiili, satırlarda ise tahmini sınıflama değerleri yer almaktadır. Bazı uygulamalarda doğruluk oranlarındaki küçük artışlar çok önemli olsa da, birçok işletme uygulamasında ilgili kararın niçin verildiğinin yorumlanabilmesi çok daha büyük önem taşıyabilir. Çok ender olarak yorumlanamayacak kadar karmaşıklaşsalar

da, genel olarak karar ağacı ve kural temelli sistemler model tahmininin altında yatan nedenleri çok iyi ortaya koyabilmektedir. Kurulan modelin değerinin belirlenmesinde kullanılan diğer bir ölçü, model tarafından önerilen uygulamadan elde edilecek kazancın bu uygulamanın gerçekleştirilmesi için katlanılacak maliyete bölünmesi ile edilecek olan yatırımın geri dönüş (return on investment) oranıdır. Kurulan modelin doğruluk derecesi ne denli yüksek olursa olsun, gerçek dünyayı tam anlamı ile modellediğini garanti edebilmek mümkün değildir. Yapılan testler sonucunda geçerli bir modelin doğru olmamasındaki başlıca nedenler, model kuruluşunda kabul edilen varsayımlar ve modelde kullanılan verilerin doğru olmamasıdır. Örneğin modelin kurulması sırasında varsayılan enflasyon oranının zaman içerisinde değişmesi, bireyin satın alma davranışını belirgin olarak etkileyecektir[10].

Önemli diğer bir değerlendirme kriteri modelin anlaşılabilirliğidir. Bazı uygulamalarda doğruluk oranlarındaki küçük artışlar çok önemli olsa da, birçok işletme uygulamasında ilgili kararın niçin verildiğinin yorumlanabilmesi çok daha büyük önem taşıyabilir. Çok ender olarak yorumlanamayacak kadar karmaşıklaşsalar da, genel olarak karar ağacı ve kural temelli sistemler model tahmininin altında yatan nedenleri çok iyi ortaya koyabilmektedir[16].

2.9.5. Değerlendirme aşaması

Değerlendirme aşamasında, uygun model ya da modeller kurulduktan sonra, VM sonuçlarının araştırma probleminin amaçlarını gerçekleştirip gerçekleştirmediği değerlendirilir. Bu aşama sonuçların değerlendirilmesi, veri madenciliği sürecinin gözden geçirilmesi ve sonraki adımların ne olacağı hususlarını içermektedir. Bu aşamanın sonunda VM sonuçlarının kullanımı üzerindeki karara varılmaktadır[3].

2.9.6. Uygulama aşaması

Kurulan ve geçerliliği kabul edilen model doğrudan bir uygulama olabileceği gibi, bir başka uygulamanın alt parçası olarak kullanılabilir. Kurulan modeller risk analizi, kredi değerlendirme, dolandırıcılık tespiti gibi işletme uygulamalarında doğrudan kullanılabileceği gibi, promosyon planlaması simülasyonuna entegre edilebilir veya

tahmini envanter düzeyleri yeniden sipariş noktasının altına düştüğünde, otomatik olarak sipariş verilmesini sağlayacak bir uygulamanın içine gömülebilir. Zaman içerisinde bütün sistemlerin özelliklerinde ve dolayısıyla ürettikleri verilerde ortaya çıkan değişiklikler, kurulan modellerin sürekli olarak izlenmesini ve gerekiyorsa yeniden düzenlenmesini gerektirecektir. Tahmin edilen ve gözlenen değişkenler arasındaki farklılığı gösteren grafikler model sonuçlarının izlenmesinde kullanılan yararlı bir yöntemdir [10,16].

BÖLÜM 3: UYGULAMA SÜRECİ

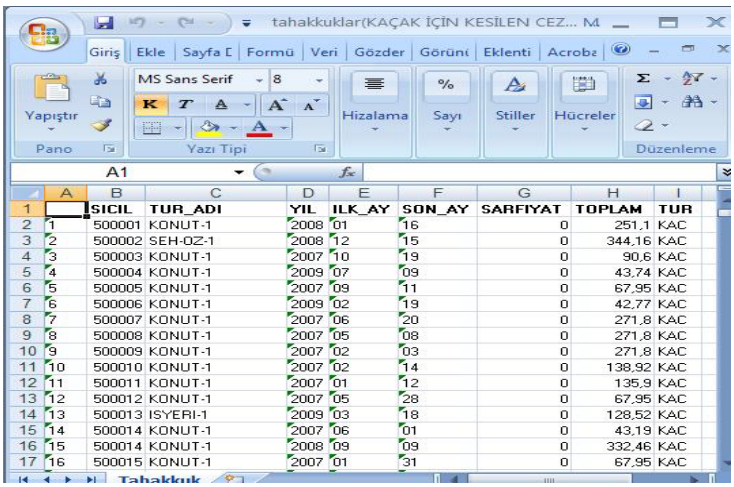
Bu çalışmada Sakarya Büyükşehir Belediyesi Adapazarı Su ve Kanalizasyon İdaresi (ADASU) abonelerinin %10'unun 01.01.2007–20.08.2009 tarihleri arasını kapsayan verilerinden bir kısmı kullanılarak VM uygulaması yapılmıştır. Uygulama sürecinde VTBK süreci takip edilmiş ve uygulama aşamasına kadar olan aşamalar bu bölümde yer almıştır. Yapılan analiz ve modellemelerde Ms-Office Excel 2007, Ms-Office Acces 2007 ve SPSS Clementine 11.1 paket programları kullanılmıştır.

3.1. Araştırma Probleminin Tanımlanması

Topluma abonelik sistemi ile hizmet veren elektrik, doğalgaz, telefon ve su gibi alanlarda hizmet sağlayıcıların en büyük problemlerinden biri kaçak kullanımdır. Bu tip kullanımlar hizmet sağlayıcılara özellikle de çoğunlukla yerel yönetimlerin kontrolündeki içme suyu dağıtım sistemlerini yönetenlere ciddi maddi külfet getirmektedirler. Bu çalışmada ADASU veri tabanındaki abone verilerinden bir kısmı veri madenciliği teknikleriyle analiz edilerek kaçak kullananlar ve kullanmayanların karşılaştırılması amaçlanmıştır.

3.2. Verileri Anlama

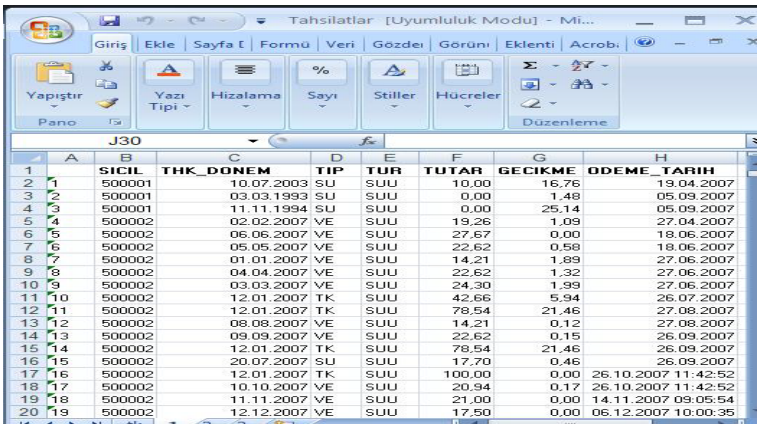
Bu çalışmada kullanılacak olan veriler ADASU veri tabanından alınmıştır. Bu kapsamda analiz edilmek üzere üçü Ms-Office Excel 2003 dosyası ve iki tanesi de text dosyası olmak üzere beş adet dosya alınmıştır.



	A	B	C	D	E	F	G	H	I
1		SICIL	TUR_ADI	YIL	ILK_AY	SON_AY	SARFIYAT	TOPLAM	TUR
2	1	500001	KONUT-1	2008	01	16	0	251,1	KAC
3	2	500002	SEH-OZ-1	2008	12	15	0	344,16	KAC
4	3	500003	KONUT-1	2007	10	19	0	90,6	KAC
5	4	500004	KONUT-1	2009	07	09	0	43,74	KAC
6	5	500005	KONUT-1	2007	09	11	0	67,95	KAC
7	6	500006	KONUT-1	2009	02	19	0	42,77	KAC
8	7	500007	KONUT-1	2007	06	20	0	271,8	KAC
9	8	500008	KONUT-1	2007	05	08	0	271,8	KAC
10	9	500009	KONUT-1	2007	02	03	0	271,8	KAC
11	10	500010	KONUT-1	2007	02	14	0	138,92	KAC
12	11	500011	KONUT-1	2007	01	12	0	135,9	KAC
13	12	500012	KONUT-1	2007	05	28	0	67,95	KAC
14	13	500013	ISYERI-1	2009	03	18	0	128,52	KAC
15	14	500014	KONUT-1	2007	06	01	0	43,19	KAC
16	15	500014	KONUT-1	2008	09	09	0	332,46	KAC
17	16	500015	KONUT-1	2007	01	31	0	67,95	KAC

Şekil 3.1. Kaçak cezaları

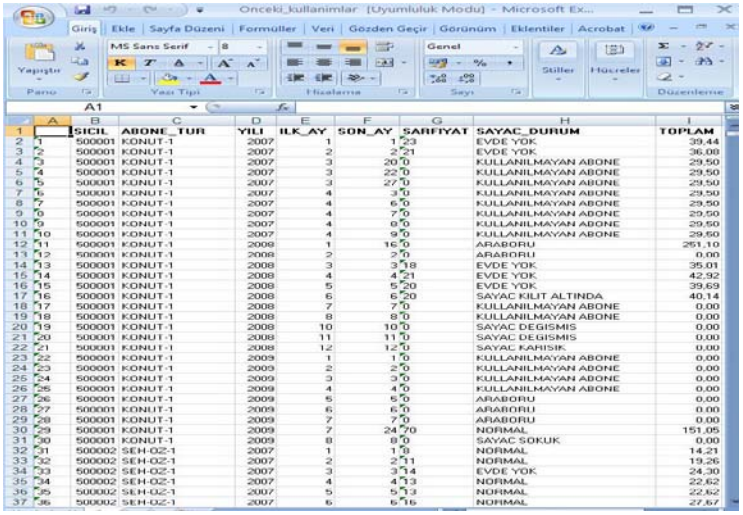
Bu dosyada 01.01.2007 ile 20.08.209 tarihleri arasında kaçak su kullandığı tutanak ile tespit edilen abonelerin maruz kaldıkları ceza tutarları yer almaktadır. Ceza tutarının haricinde abonenin sicil numarası, abonelik türü, cezanın yılı, ayı, günü ve cezanın türü yer almaktadır. Dosyada bir defadan daha fazla ceza yemiş abonelerin her cezası ayrı satırda yer almıştır. Bu tekrarlarla birlikte dosya 7.432 satırdan oluşmaktadır.



	A	B	C	D	E	F	G	H
1		SICIL	THK_DONEM	TIP	TUR	TUTAR	GECIKME	ODEME_TARIH
2	1	500001	10.07.2003	SU	SUU	10,00	16,76	19.04.2007
3	2	500001	03.03.1993	SU	SUU	0,00	1,48	05.09.2007
4	3	500001	11.11.1994	SU	SUU	0,00	25,14	05.09.2007
5	4	500002	02.02.2007	VE	SUU	19,26	1,09	27.04.2007
6	5	500002	06.06.2007	VE	SUU	27,67	0,00	18.06.2007
7	6	500002	05.05.2007	VE	SUU	22,62	0,58	18.06.2007
8	7	500002	01.01.2007	VE	SUU	14,21	1,89	27.06.2007
9	8	500002	04.04.2007	VE	SUU	22,62	1,32	27.06.2007
10	9	500002	03.03.2007	VE	SUU	24,30	1,99	27.06.2007
11	10	500002	12.01.2007	TK	SUU	42,66	5,94	26.07.2007
12	11	500002	12.01.2007	TK	SUU	78,54	21,46	27.08.2007
13	12	500002	08.08.2007	VE	SUU	14,21	0,12	27.08.2007
14	13	500002	09.09.2007	VE	SUU	22,62	0,15	26.09.2007
15	14	500002	12.01.2007	TK	SUU	78,54	21,46	26.09.2007
16	15	500002	20.07.2007	SU	SUU	17,70	0,46	26.09.2007
17	16	500002	12.01.2007	TK	SUU	100,00	0,00	26.10.2007 11:42:52
18	17	500002	10.10.2007	VE	SUU	20,94	0,17	26.10.2007 11:42:52
19	18	500002	11.11.2007	VE	SUU	21,00	0,00	14.11.2007 09:05:54
20	19	500002	12.12.2007	VE	SUU	17,50	0,00	06.12.2007 10:00:35

Şekil 3.2. Kaçak kullananların tahsilatları

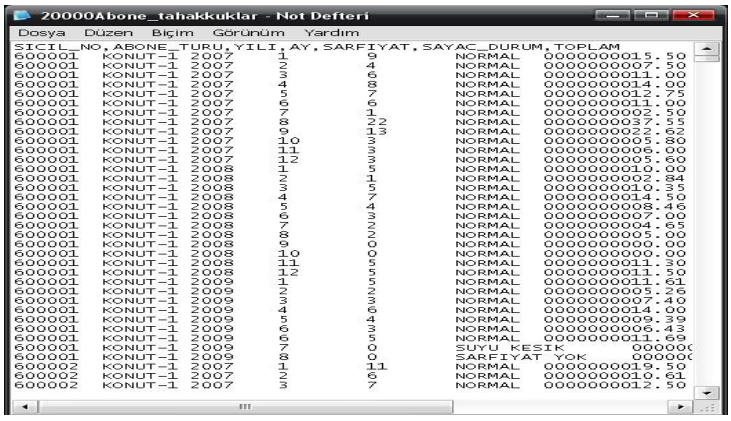
Tahsilatlar dosyasında tutanakla tespit edilmiş kaçak kullanıcılarından ilgili tarih içinde yapılmış tüm tahsilat tutarları ile birlikte abone sicil no, tahsilata ait tahakkuk dönemi (gün, ay, yıl olarak), tip(tahsilat şekli), tür(tahsilat türü), gecikme tutarı ve ödeme tarihi bilgileri yer almaktadır. Her bir tahsilat tutarına ilişkin veri ayrı bir satırda yer almaktadır. Tahsilat dosyası 143.511 satırdan oluşmaktadır.



	A	B	C	D	E	F	G	H	I
1		SICIL	ABONE_TUR	YILI	ILK_AY	SON_AY	SARFIYAT	SAYAC_DURUM	TOPLAM
2	1	500001	KONUT-1	2007	1	1	23	EVDE YOK	39,44
3	2	500001	KONUT-1	2007	2	2	21	EVDE YOK	36,00
4	3	500001	KONUT-1	2007	3	20	0	KULLANILMAYAN ABONE	29,50
5	4	500001	KONUT-1	2007	3	22	0	KULLANILMAYAN ABONE	29,50
6	5	500001	KONUT-1	2007	3	27	0	KULLANILMAYAN ABONE	29,50
7	6	500001	KONUT-1	2007	4	3	0	KULLANILMAYAN ABONE	29,50
8	7	500001	KONUT-1	2007	4	6	0	KULLANILMAYAN ABONE	29,50
9	8	500001	KONUT-1	2007	4	7	0	KULLANILMAYAN ABONE	29,50
10	9	500001	KONUT-1	2007	4	8	0	KULLANILMAYAN ABONE	29,50
11	10	500001	KONUT-1	2007	4	9	0	KULLANILMAYAN ABONE	29,50
12	11	500001	KONUT-1	2008	1	16	0	ARABORU	251,10
13	12	500001	KONUT-1	2008	2	2	0	ARABORU	0,00
14	13	500001	KONUT-1	2008	3	3	18	EVDE YOK	35,01
15	14	500001	KONUT-1	2008	4	4	21	EVDE YOK	42,92
16	15	500001	KONUT-1	2008	5	5	20	EVDE YOK	39,69
17	16	500001	KONUT-1	2008	6	6	20	SAYAC KILIT ALTINDA	40,14
18	17	500001	KONUT-1	2008	7	7	0	KULLANILMAYAN ABONE	0,00
19	18	500001	KONUT-1	2008	8	8	0	KULLANILMAYAN ABONE	0,00
20	19	500001	KONUT-1	2008	10	10	0	SAYAC DEGISMIS	0,00
21	20	500001	KONUT-1	2008	11	11	0	SAYAC DEGISMIS	0,00
22	21	500001	KONUT-1	2008	12	12	0	SAYAC KAHISIK	0,00
23	22	500001	KONUT-1	2009	1	1	0	KULLANILMAYAN ABONE	0,00
24	23	500001	KONUT-1	2009	2	2	0	KULLANILMAYAN ABONE	0,00
25	24	500001	KONUT-1	2009	3	3	0	KULLANILMAYAN ABONE	0,00
26	25	500001	KONUT-1	2009	4	4	0	KULLANILMAYAN ABONE	0,00
27	26	500001	KONUT-1	2009	5	5	0	ARABORU	0,00
28	27	500001	KONUT-1	2009	6	6	0	ARABORU	0,00
29	28	500001	KONUT-1	2009	7	7	0	ARABORU	0,00
30	29	500001	KONUT-1	2009	7	24	70	NORMAL	151,05
31	30	500001	KONUT-1	2009	8	8	0	SAYAC SOKUK	0,00
32	31	500002	SEH-OZ-1	2007	1	1	8	NORMAL	14,21
33	32	500002	SEH-OZ-1	2007	2	2	11	NORMAL	19,26
34	33	500002	SEH-OZ-1	2007	3	3	14	EVDE YOK	24,30
35	34	500002	SEH-OZ-1	2007	4	4	13	NORMAL	22,62
36	35	500002	SEH-OZ-1	2007	5	5	13	NORMAL	22,62
37	36	500002	SEH-OZ-1	2007	6	6	16	NORMAL	27,67

Şekil 3.3. Kaçak kullananların tahakkukları

Tahakkuklar dosyasında ise kaçak kullananların ilgili dönem içindeki tüm tahakkuk tutarları ile birlikte abone sicil numarası, abone türü, yıl, ilk ay (ay), son ay (gün) ve sayaç durumu bilgileri yer almaktadır. Her bir tahakkuk tutarına ilişkin veri ayrı bir satırda yer almaktadır. Tahakkuk dosyası 178.919 satırdan oluşmaktadır.



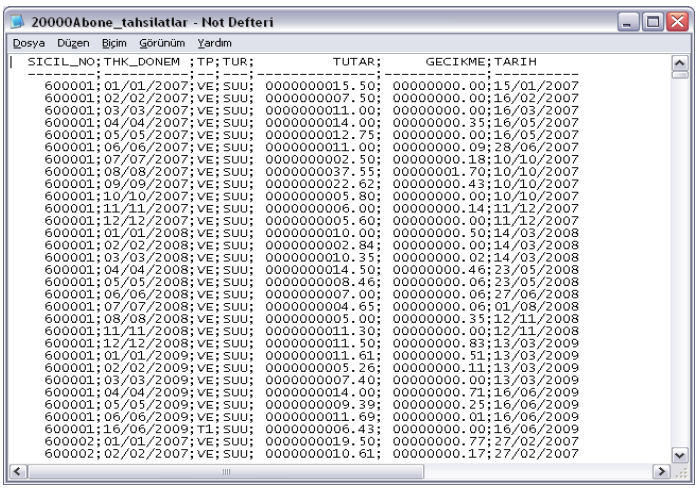
20000Abone_tahakkuklar - Not Defteri

Dosya Düzen Biçim Görünüm Yardım

```
SICIL_NO,ABONE_TURU,YILI,AY,SARFIYAT,SAYAC_DURUM,TOPLAM
600001   KONUT-1 2007     1          9          NORMAL   0000000015.50
600001   KONUT-1 2007     2          4          NORMAL   0000000007.50
600001   KONUT-1 2007     3          6          NORMAL   0000000011.00
600001   KONUT-1 2007     4          8          NORMAL   0000000014.00
600001   KONUT-1 2007     5          7          NORMAL   0000000012.75
600001   KONUT-1 2007     6          6          NORMAL   0000000011.00
600001   KONUT-1 2007     7          1          NORMAL   0000000002.50
600001   KONUT-1 2007     8          22         NORMAL   0000000037.55
600001   KONUT-1 2007     9          13         NORMAL   0000000022.62
600001   KONUT-1 2007     10         3          NORMAL   0000000005.80
600001   KONUT-1 2007     11         3          NORMAL   0000000006.00
600001   KONUT-1 2007     12         3          NORMAL   0000000005.60
600001   KONUT-1 2008     1          5          NORMAL   0000000010.00
600001   KONUT-1 2008     2          1          NORMAL   0000000002.84
600001   KONUT-1 2008     3          5          NORMAL   0000000010.35
600001   KONUT-1 2008     4          7          NORMAL   0000000014.50
600001   KONUT-1 2008     5          4          NORMAL   0000000008.46
600001   KONUT-1 2008     6          3          NORMAL   0000000007.00
600001   KONUT-1 2008     7          2          NORMAL   0000000004.65
600001   KONUT-1 2008     8          2          NORMAL   0000000005.00
600001   KONUT-1 2008     9          0          NORMAL   0000000000.00
600001   KONUT-1 2008     10         0          NORMAL   0000000000.00
600001   KONUT-1 2008     11         5          NORMAL   0000000011.30
600001   KONUT-1 2008     12         5          NORMAL   0000000011.50
600001   KONUT-1 2009     1          5          NORMAL   0000000011.61
600001   KONUT-1 2009     2          2          NORMAL   0000000005.26
600001   KONUT-1 2009     3          3          NORMAL   0000000007.40
600001   KONUT-1 2009     4          6          NORMAL   0000000014.00
600001   KONUT-1 2009     5          4          NORMAL   0000000009.39
600001   KONUT-1 2009     6          3          NORMAL   0000000006.43
600001   KONUT-1 2009     6          5          NORMAL   0000000011.69
600001   KONUT-1 2009     7          0          SUYU KESIK       000000
600001   KONUT-1 2009     8          0          SARFIYAT YOK     000000
600002   KONUT-1 2007     1          11         NORMAL   0000000019.50
600002   KONUT-1 2007     2          6          NORMAL   0000000010.61
600002   KONUT-1 2007     3          7          NORMAL   0000000012.50
```

Şekil 3.4. Kaçak kullanmayanların tahakkukları

Bu dosyada ise tesadüfi olarak seçilen 20.000 abonenin 01.01.2007 tarihinden 20.08.2009 tarihine kadarki süreçte aylık dönemler halinde su sarfiyat miktarı ile birlikte abone sicil numarası, abonelik türü, yılı, ayı, tahakkuk tutarı ve sayaç durumu bilgileri yer almaktadır. Tahakkuk dosyası 626.972 satırdan oluşmaktadır.



Şekil 3.5. Kaçak kullanmayanların tahsilatları

Bu dosyada tesadüfi olarak seçilmiş 20.000 aboneye ait ilgili tarih içinde yapılmış tüm tahsilat tutarları ile birlikte abone sicil no, tahsilata ait tahakkuk dönemi (gün, ay, yıl olarak), tip(tahsilat şekli), tür(tahsilat türü), gecikme tutarı ve ödeme tarihi bilgileri yer almaktadır. Her bir tahsilat tutarına ilişkin veri ayrı bir satırda yer almaktadır. Tahsilatlar dosyası 555.118 satırdan oluşmaktadır. Bu dosyalardaki bazı bilgilerin daha iyi anlaşılabilmesi için ilave olarak üç adet Ms-Office Excel 2003 dosyası daha alınmıştır. Bu dosyalardaki bazı alanlardaki tanımlamaları açıklamak amacıyla ilave olarak alınan dosyalardan birinde sayaç durumu kodları ve tanımlamaları yer almaktadır. Diğer bir dosyada abone türleri ve türlere göre yerleşim yerleri bölge olarak yer almaktadır. Son dosyada ise 04.09.2009 tarihi itibariyle toplam abone sayısı ve 2008 yılı için abone türlerine göre aylık ortalama su sarfiyat miktarları ve tutarları yer almaktadır.

Tablo 3.1. Tüm veri tabanının seçilen veri setiyle karşılaştırılması

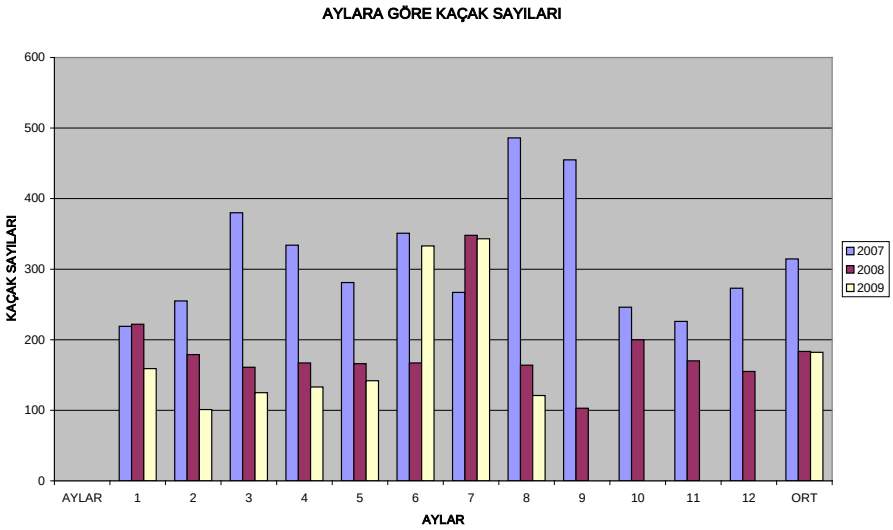
Veri Tabanı Abone Türü	**Sayı**	**%**	**Veri Seti Abone Türü**	**Sayı**	**%**
Konutlar	178.898	88,04	**Konutlar**	24.198	89,77
Dernek ve Vakıflar	976	0,48	**Dernek ve Vakıflar**	65	0,24
İşyerleri	19.097	9,40	**İşyerleri**	2.403	8,91
Park ve Bahçeler	452	0,22	**Park ve Bahçeler**	13	0,05
Sanayi ve Şantiyeler	3.460	1,70	**Sanayi ve Şantiyeler**	262	0,97
Okullar	309	0,15	**Okullar**	13	0,05
TOPLAM	**203.192**	**100,00**	**TOPLAM**	**26.954**	**100,00**

Tablo 3.1.'de ADASU veri tabanından seçilen veri seti ile veri tabanının tümünün abone türüne göre dağılımları birlikte verilmiştir. Her iki durumda da abonelerin yaklaşık %90 ını konut tipi aboneler teşkil etmektedir. Bu durum seçilen veri setinin

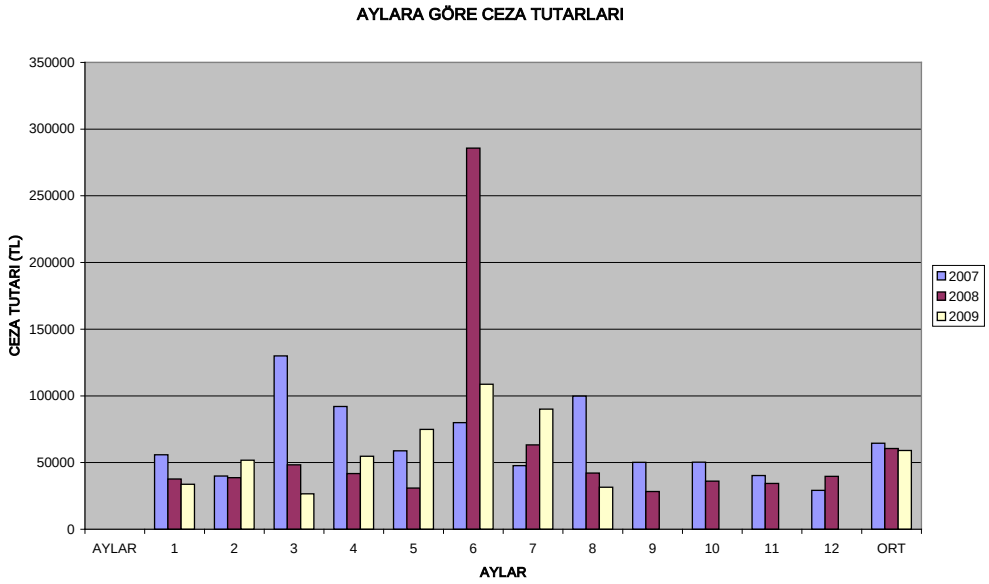
veri tabanının tamamını temsil etme gücünün yüksek olduğunu teyit eder. Tablo 3.2.'de kaçak ceza sayılarının dağılımı yer almaktadır.

Tablo 3.2. Kaçak ceza sayıları

Ceza Sayısı	Sayı	Yüzde (%)
1	5639	94,23
2	283	4,73
3	50	0,85
4	10	0,17
5	2	0,02
TOPLAM	5984	100,00



Şekil 3.6. Aylara göre kaçak sayılarının dağılımı



Şekil 3.7. Aylara göre kaçak ceza tutarlarının dağılımı

Şekil 3.6.'da aylara göre kaçak sayıları her bir yıl için ayrı ayrı olacak şekilde gösterilmiştir. Aynı kaçaklara ilişkin ceza tutarları da aynı düzende Şekil 3.7.'de gösterilmiştir. Grafiklerde 2007 ve 2008 yılının tamamı 2009 yılının ise ilk sekiz ayının verileri gösterilmiştir. Tahakkuk tablolarındaki veri sayıları tahsilat tablolarından fazladır. Bunun sebebi bazı abonelerin su kullandıkları halde ödeme yapmamalarıdır.

3.3. Verinin Hazırlanması

Daha önce ekran çıktıları verilmiş olan tablolar ADASU veri tabanından alınırken süzme işlemine tabi tutulmasına rağmen içlerinde kullanılmasına gerek görülmeyen alan adları tespit edilmiş ve sonraki aşamalarda bu sütunlar çıkarılma sebepleri açıklanarak tablolardan çıkarılacaktır.

Veri seti incelendiğinde bazı bilgilerin ayrı tablolarda ayrı alan adlarıyla verildikleri görülmüş ve bu alan adları tüm tablolarda aynı bilgileri simgeleyecek şekilde düzenlenmiştir.

Tablo 3.3. İlk düzenleme sonrası veri setinde yer alan alan adları

Tahakkuklar	**Tahsilatlar**	**Kaçak Cezaları**
SICIL_NO (Abone No)	SICIL_NO (Abone No)	SICIL_NO (Abone No)
ABONE_TUR (Abone Tipi)	THK_DONEM (Tahakkuk Tarihi)	ABONE_TUR (Abone Tipi)
YIL	TIP (Ödeme Şekli)	YIL
AY	TUR (Ödemenin Türü)	AY
GUN	TUTAR (Tahsilat Tutarı)	GUN
SARFIYAT (Tüketim mik.)	GECIKME (Gecikme tutarı)	SARFIYAT (Tüketim mik.)
SAYAC_DURUM (Okuma kodu)	ODEME_TARIHI (Tahsilat tarihi)	TOPLAM (Ceza tahakkuk tutarı)
TOPLAM (Tahakkuk tutarı)		TUR (Ödemenin Türü)

Tablo 3.1.'de görüldüğü üzere konut tipi abonelerin tüm abonelerin % 90'ını teşkil ettiği için modelde sadece konut tipi abonelere ilişkin veriler kullanılmıştır. Bu

bağlamda veri seti üzerinde bir dizi süzme işlemi yapılmıştır. Bu işlemler sebepleriyle birlikte aşağıda açıklanmıştır.

ADASU veri tabanından tesadüfi olarak seçilen 20.000 abone içinden kaçak kullanmış olan 306 abone veri tekrarını önlemek maksadıyla tahakkuk ve tahsilat tablolarından atılmıştır. Tahsilat tablolarından tür alanından vidanjör ücreti gibi su bedeli dışında kalan tahsilat satırları süzülerek sadece su ile ilgili tahsilat satırları bırakılmıştır. Tahakkuk dosyalarında yer alan gün sütunu model için kullanılmaya değer bulunmadığı için tablodan çıkartılmıştır. Aynı şekilde kaçak cezalarının yer aldığı tablodan da gün sütunu ve tüm tutarlar kaçaktan dolayı oluşan cezalar olduğu için tür alanı çıkartılmıştır. Tahsilat tablolarında ise ödeme şeklini gösteren tür ve ödeme şeklini tip alanı çıkartılmıştır. Böylelikle tahakkuk tabloları yedi tahsilat tabloları beş ve kaçak cezalarının yer aldığı tablo da altı sütuna dönüşmüştür.

Tablolarda her bir abonenin birden fazla satırda verisi bulunmakta ve bazı alanlarda veri tekrarı anlamına gelmekteydi bunun önüne geçebilmek için tablolarda bir dizi birleştirme ve dönüştürme işlemi uygulanmıştır. İlk olarak tahsilat tablolarında tahakkuk dönemi ile ödeme tarihi alanlarının farkı alınarak GECIKME_SURESI (GUN) adında bir alana kaydedilmiştir.

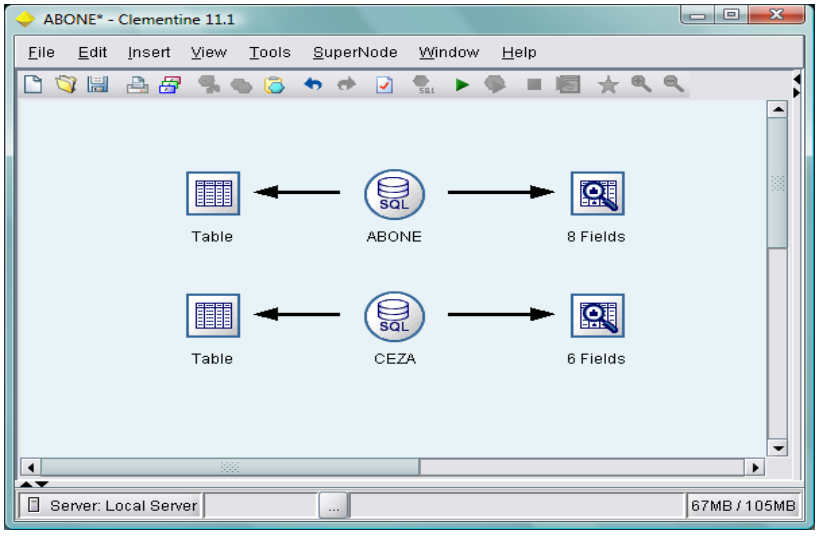
Tahakkuk tablolarında her abonenin her dönemi için Tablo 3.4.'de gösterilen ağırlıklar kullanılarak her bir sayaç durumu yerine tablodaki karşılığı olan sayı yazılmıştır. Her bir abonenin bilgilerinin sadece bir satırda görülebilmesi için tahakkuk dosyasındaki sarfiyat ve tahakkuk tutarını gösteren toplam alanlarının ortalaması, sayaç durumlarını gösteren alan değerlerinin toplamı alınarak sicil numarası ve abone türü alanlarının yanına yerleştirilmiş ve beş alandan oluşan tahakkuk tabloları elde edilmiştir. Aynı şekilde tahsilat dosyasında da tutar, gecikme süresi ve gecikme tutarı alanlarının ortalamaları alınarak sicil numaralarının yer aldığı alanın yanına yerleştirilerek dört alandan oluşan tablo elde edilmiştir. Tahakkuk ve tahsilat tabloları önce kendi aralarında birleştirilerek tek bir tahakkuk ve tahsilat tablosu elde edilmiştir. Daha sonra tahsilat ve tahakkuk tabloları tek sekiz alandan oluşacak şekilde tek bir tabloda birleştirilmiştir. Kaçak cezalarının yer aldığı tabloda birden fazla cezaya maruz kalmış aboneler yer aldığı için bu birleştirmeye tabi tutulmamıştır.

Birleştirilmiş son tabloda sicil no, abone türü, ort. sarfiyat, ort. tahakkuk, toplam sayaç durum (ağırlıklandırılmış), ortalama gecikme tutarı, ortalama tahsilat tutarı ve ortalama gecikme süresi alanları yer almıştır. Süzme işlemleri Ms-Excel 2007 programı; dönüştürme ve birleştirme işlemleri de Ms-Acces 2007 programı kullanılarak yapılmıştır.

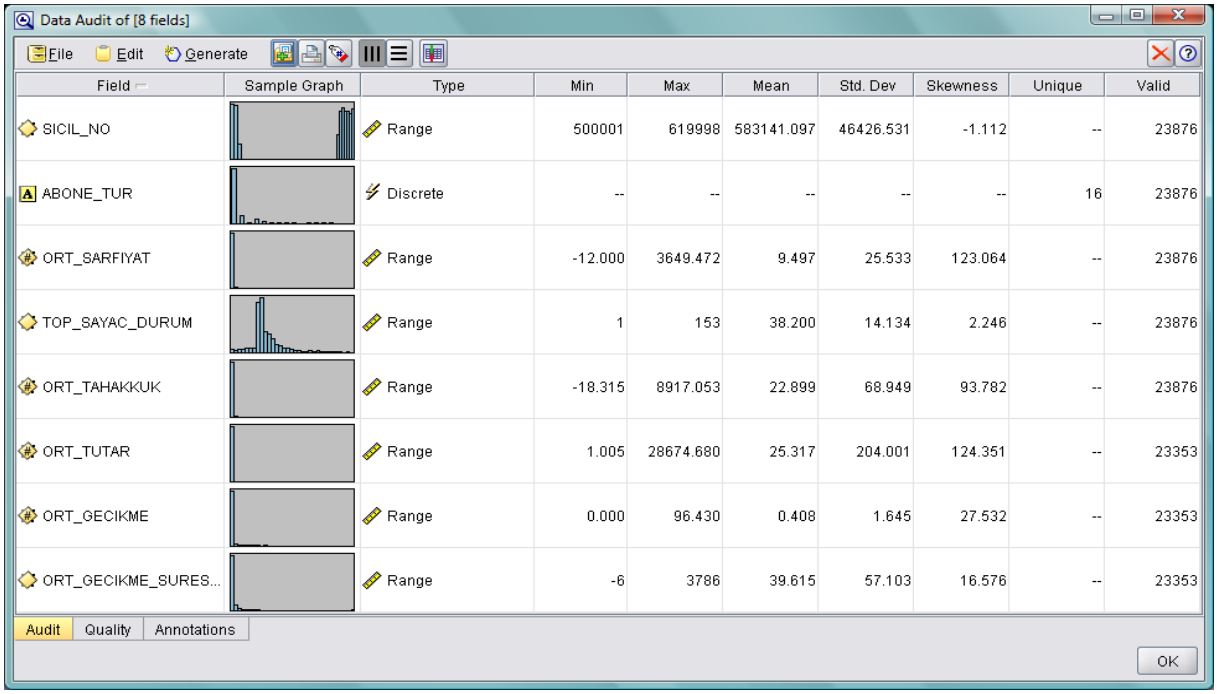
Tablo 3.4. Sayaç okuma kodlarına göre puanlama grupları

Grup-1	**Grup-1**	**Grup-1**	**Grup-2**
Normal Abone	**Normal Abone**	**Normal Abone**	**Kaçak Kul. Abone**
Sayaç Değişmemiş	Tüketim Şüpheli	Normal	Suyu Kesik Kullanım Var
Sayaç Değişmiş	Sayaç Kilit Altında	Devirli	Ara boru
Sayaç Çalışmıyor	Sayaç Üz. Mlz. Var	Camı Kırık	Kaçak ve Usulsüz Kullanım
Damga Müh. Kopuk	Direk Kullanım	Hasarlı Bina	
Yeri Uyg. Değil	Sayaç Ters Tkl.	Yıkık Bina	
Sayaç Gömülü	Sayaç Gövde Tah.	Yazlık	
Kullanılmayan Abone	Sayaç Buğulu	Sayaç Kirli	
İlk Endeks Hatalı	Sayaç Karışık	Evde Yok	
Adres Bulunamadı	Sayaç Sökük	Köpek var	
Kayıtlı Sayaç Bulunamadı	Abone Engeli	Dilekçeli	
Abone Tipi Değişmemiş	Suyu Kesik	Raporlu	
Mükerrer Sözleşme	Kat Alınmış		

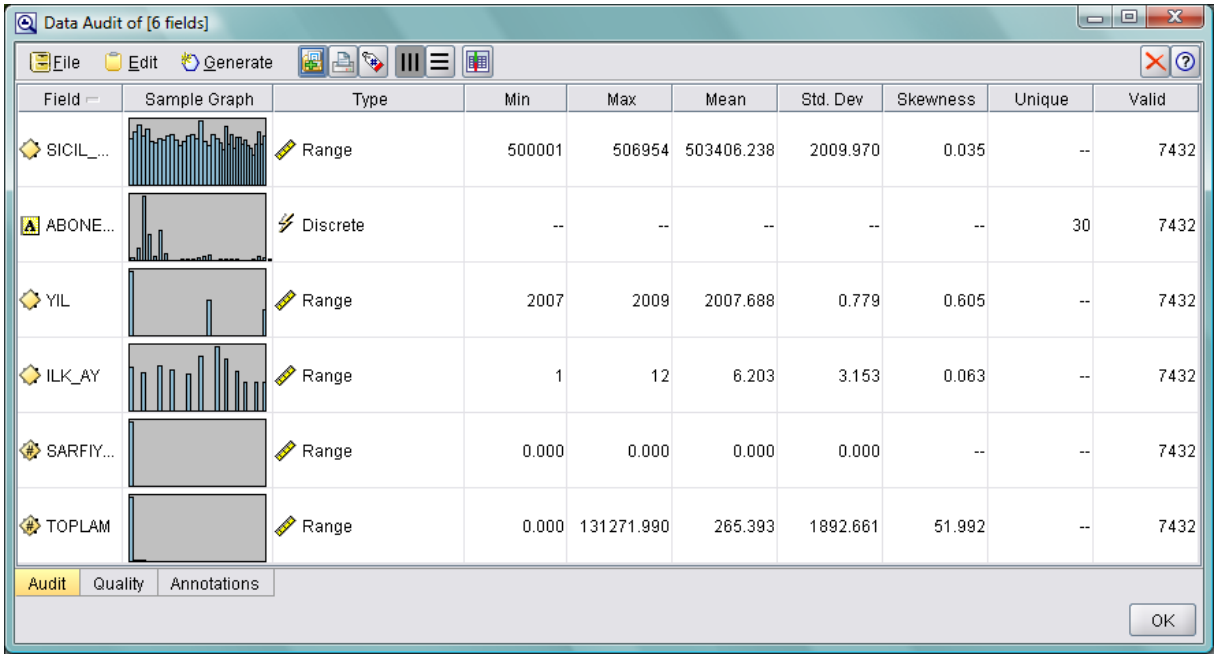
Mdb dosya formatında oluşturulan veri seti ile Clementine programına “Sources” paleti üzerindeki “Database” nodu vasıtası ile bağlantı sağlanmış ve veri setinin kalitesi incelenmiştir. Database nodu clementine’e bir veri tabanından veri tanıtılırken kullanılmaktadır.



Şekil 3.8. Veri kalitesinin incelenmesi clementine ekran çıktısı



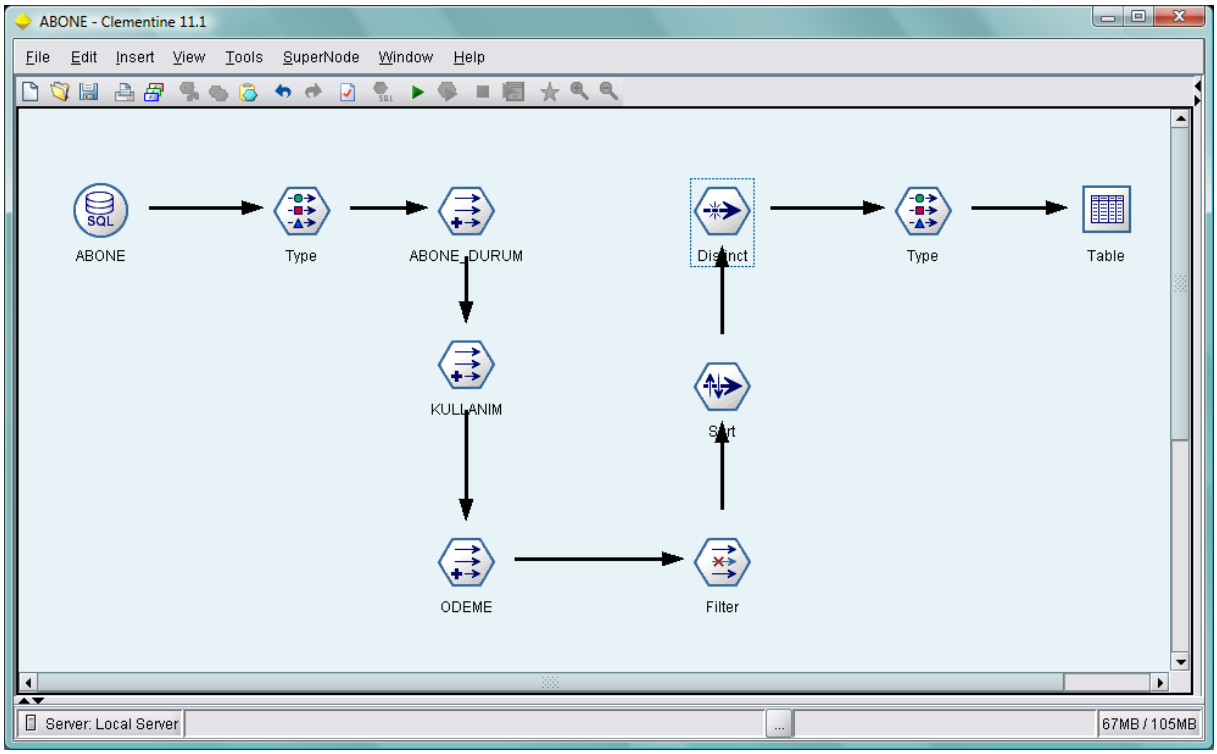
Şekil 3.9. ABONE veri kalitesi inceleme sonuçları



Şekil 3.10. CEZA veri kalitesi inceleme sonuçları

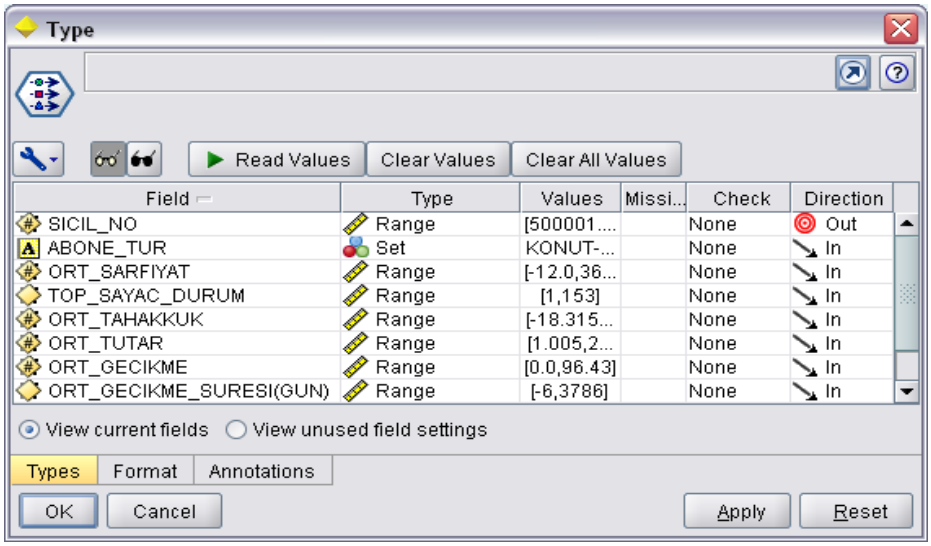
Veri kalitesi incelemeleri sonucunda veri setinde herhangi bir sapma ve kayıp değer tespit edilmemiştir. Fakat bazı alanlarda yanlış veri girişinden kaynaklanan (ortalama sarfiyat miktarının sıfırdan küçük olması gibi) hatalı veriler tespit edilmiş ve bu veriler veri setinden çıkartılmıştır. Aynı şekilde ortalama gecikme süresi bir yıldan

fazla olana veriler de modelin tahmin gücünü etkilemesi için veri setinden çıkartılmıştır. Clementine de tekrar yapılan veri kalitesinin incelenmesinde bu hatalı alan değerlerine rastlanmamıştır. Veri kalitesinin incelenmesinden sonra kurulacak modele uygun olması için veri seti düzenlenmiştir. Bu düzenlemeye ilişkin clementine ekran çıktısı şekil 3.11.'de gösterilmiştir.



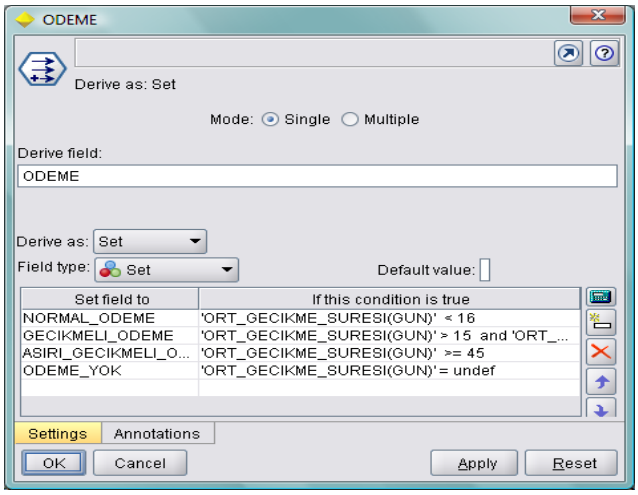
Şekil 3.11. Veri düzenleme clementine ekran çıktısı

Düzenleme ekranında yer alan nodlardan type nodu her alanın tip, yön, eksik değer tanımları gibi özelliklerinin belirlenmesine, derive nodu yeni alanlar oluşturmaya, filter nodu veri setinden istenilen alanların çıkarılmasına, sort nodu veri setinin bir ya da daha fazla alana göre sıralanmasına, distinct nodu kullanıcı tarafından belirlenen alanlar temel alınarak tekrarlanan kayıtlar kontrol edilir ve tekrarlardan ilki ya da ilki hariç tamamını seçmeye ve table nodu ise verilerin oluşturulan son halini tablo halinde görüntülemeye yarar.



Şekil 3.12. Type nodu ekran çıktısı

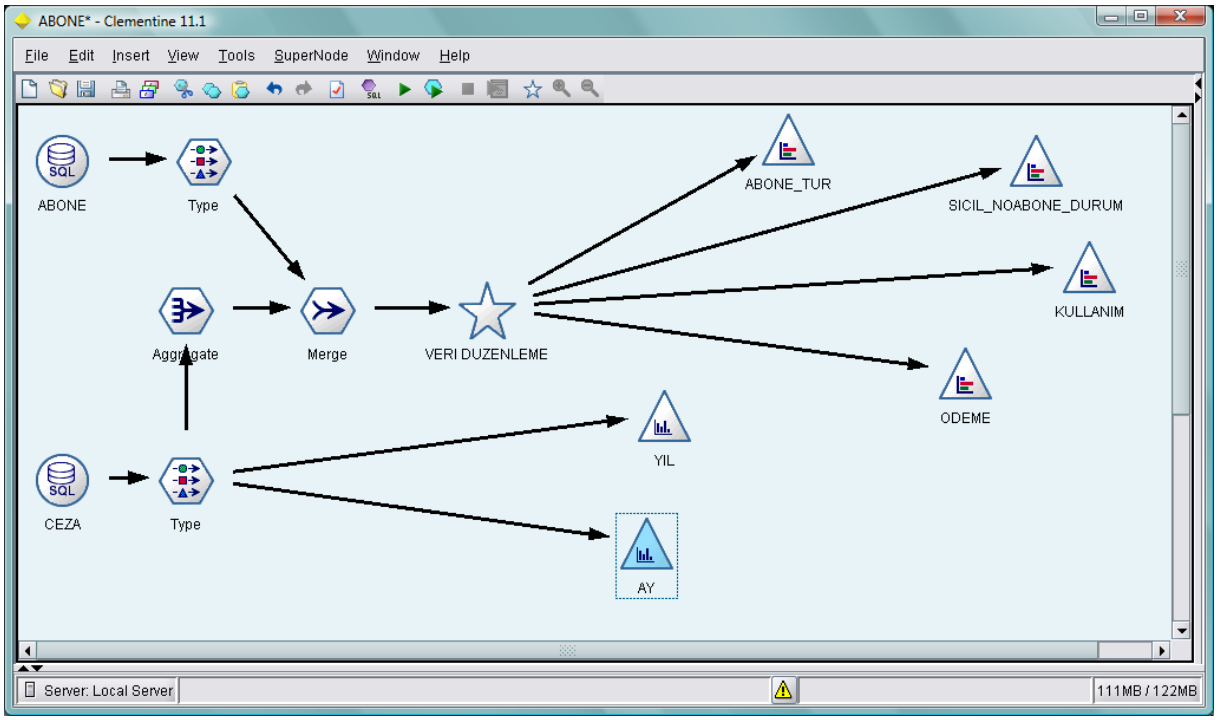
ABONE veri tabanından alınan veriler önce type nodu kullanılarak programa tanıtılmış ve alan tipleri tespit edilmiştir. Daha sonra derive nodu kullanılarak abone durumu, kullanım ve ödeme şeklinde üç yeni alan oluşturulmuştur. İlk derive nodunda aboneleri sicil numaraları 6 ile başlayanlara normal diğerlerine kaçak olacak şekilde etiketleyerek yeni alana bu bilgileri kaydetmiştir. Ortalama sarfiyat miktarı 10 m³ den az olanlar AZ; ortalama sarfiyat miktarı 10-20 m³ arası olanlar normal ve 20 m³ den fazla sarfiyatı olan aboneler de aşırı olarak etiketlenmiş ve yeni oluşan alana kaydedilmiştir. Gecikme sürelerine göre oluşturulan yeni alana ilişkin derive nodu ekran çıktısı Şekil 3.13.'de gösterilmiştir.



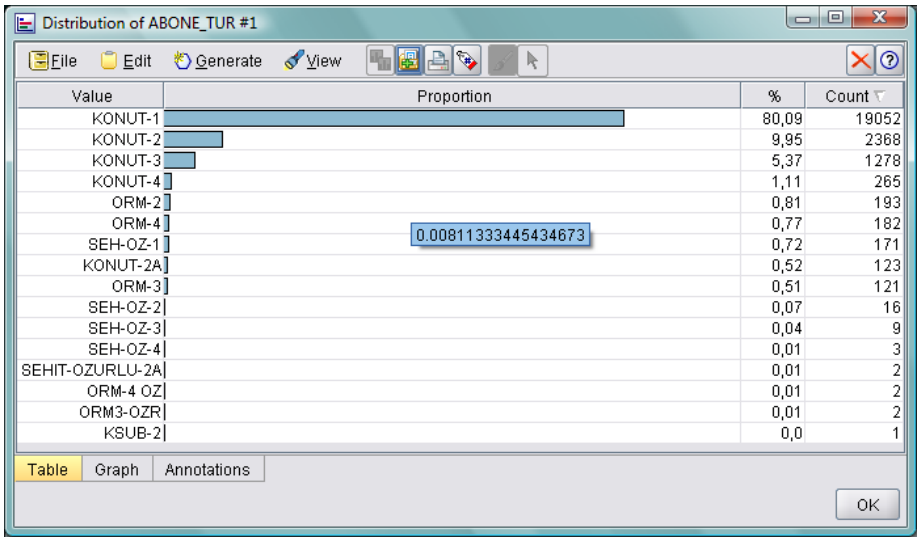
Şekil 3.13. Derive nodu ekran çıktısı

Filter nodu kullanılarak gecikme tutarlarının yer aldığı alan veri setinden çıkartılmıştır. Sort noduyla veri seti yeniden sıralanmış distinct nodu ile sicil

numaralarına göre tekrarlar önlemiştir. Son olarak yine type noduyla veri setindeki alan tipleri belirlenmiştir. Table nodu ile oluşan yeni veri seti tablo halinde incelenmiştir. Veri düzenleme aşamasında oluşturulan nodlar supernod olarak tek bir nod haline getirilerek sonraki aşamalarda görsel sadelik sağlanmaya çalışılmıştır. Oluşan yeni alanlar ile birlikte modelleme aşamasına geçmeden önce veriler arası ilişkilerin daha iyi anlaşılabilmesi için veriler grafiklerle görsel hale getirilmiştir. Şekil 3.14'de ilgili clementine ekran çıktısı yer almaktadır.

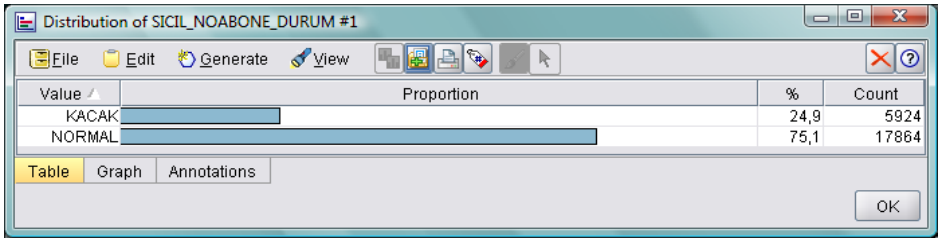


Şekil 3.14. Veri seti ilişki anlama ekran çıktısı



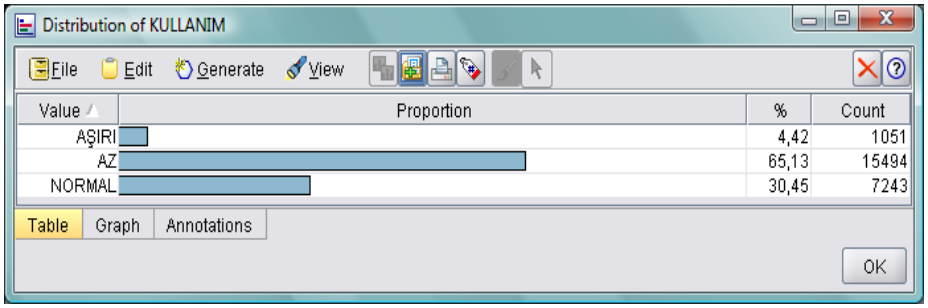
Şekil 3.15. Abone türlerine göre dağılım

Şekil 3.15.'de yer alan abone türlerine göre veri setinin dağılımı incelendiğinde hedef veri setini oluşturan abonelerin %90'lık kısmı Adapazarı merkez, Serdivan ve Erenlerde ikamet etmektedir.

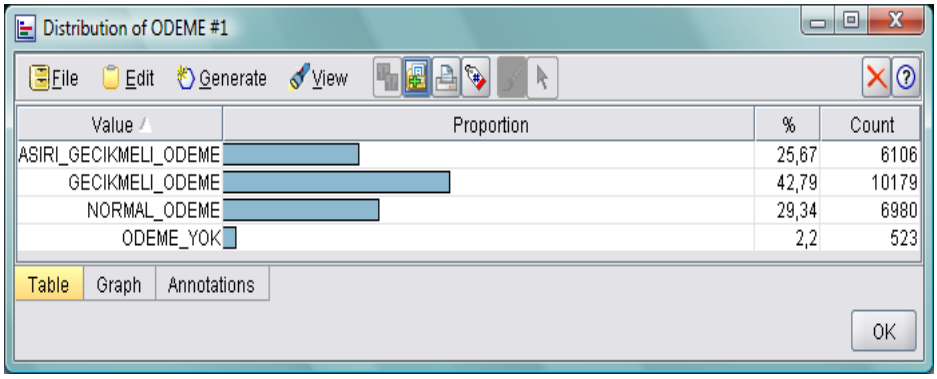


Şekil 3.16. Abone durumuna göre dağılım

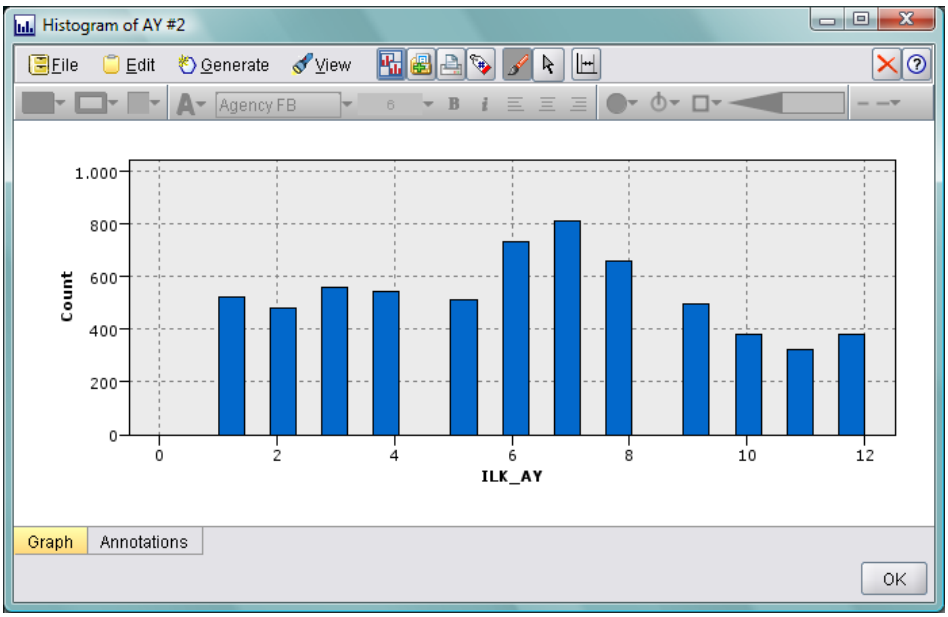
Şekil 3.16.'da yer alan abone durumlarına göre dağılıma göre hedef veri setindeki abonelerin yaklaşık %25'i kaçak kullandığı tespit edilmiş abonelerdir. Şekil 3.17.'de ise abonelerin su kullanım durumuna göre dağılımlarına yer verilmiştir. Bu grafiğe göre veri setinde yer alan abonelerin %65'lik kısmı 10m³'den az su tüketen abonelerdir. Şekil 3.18.'de ise abonelerin ödeme durumuna göre ve Şekil 3.19.'da ise aylara göre kaçak dağılımları yer almaktadır.



Şekil 3.17. Su kullanım durumuna göre dağılım

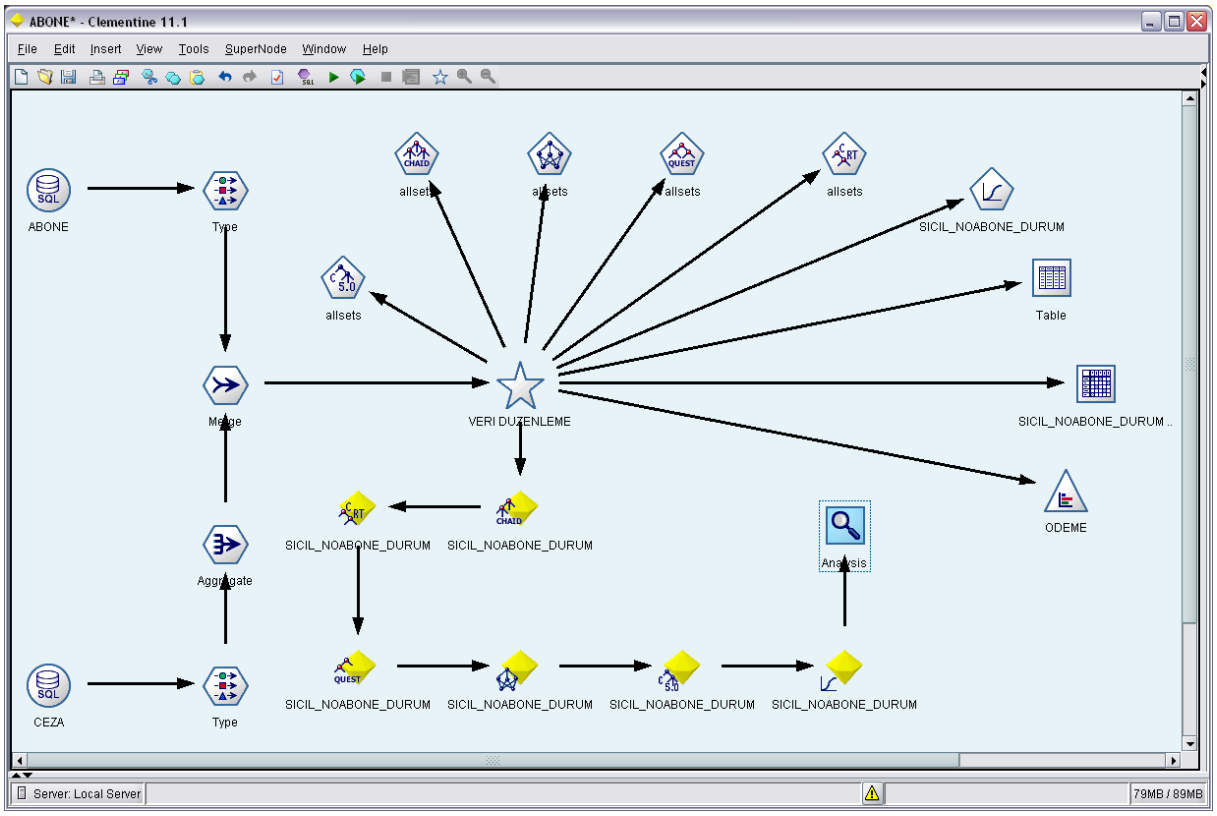


Şekil 3.18. Ödeme durumuna göre dağılım



Şekil 3.19. Aylara göre kaçak kullanım dağılımı

BÖLÜM 4: MODELİN KURULMASI VE ÇALIŞTIRILMASI

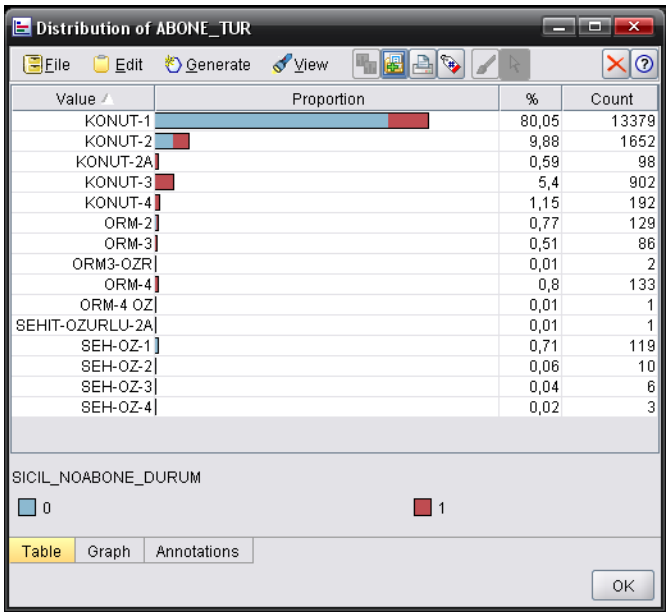


Şekil 4.1. Modelleme clementine ekran çıktısı

Veri hazırlama sonrası kaçak kullanımla ilişkisi tespit edilip anlamlı bulunan değişkenler Şekil 4.1.'de görüldüğü gibi modellenmiştir. Karar değişkeni olarak belirlenen kaçak kullanım ile ilişkili olabilecek tüm değişkenler modelde yer almıştır ve kaçak kullanım için 1 normal kullanım için 0 kodu kullanılmıştır. Modelin öğrenmesi için veri setinin %30'u eğitim için ayrılmış kalan kısım ise test için bırakılmıştır. Bu aşamada analizler karar ağaçları ve yapay sinir ağı kullanarak yapılmış ve her iki algoritmanın tahmin gücü karşılaştırılmıştır. Oluşturulan model eğitim veri seti ile çalıştırılmış ve elde edilen veriler bu bölüm içerisinde yer almıştır. Oluşturulan model içerisinde veri setine uygun olan algoritmalar denenmiş ve en sağlıklı sonucu veren algoritmaların sonucu dikkate alınmıştır. Her bir değişken için denenen algoritmaların tahmin gücü karşılaştırmaları tablolar halinde ilgili

değişkenin yorumlarında verilmiştir. Analizlerde yapay sinir ağı, lojistik regresyon ve karar ağacı algoritmalarından C&R Trees, C5.0, CHAID ve QUEST kullanılmıştır. Veri setindeki veriler %30-%70 şeklinde iki gruba ayrılmıştır. Verilerin %70 ini oluşturan grup eğitim için diğer grup ise test için kullanılmıştır. Her bir değişken ile ilgili olarak önce Clementine ile elde edilen grafik, tablo ya da karar ağacı diyagramı verilmiş daha sonra aynı değişken için kullanılan alternatif algoritma ile karşılaştırılmıştır.

1. Abone türü:

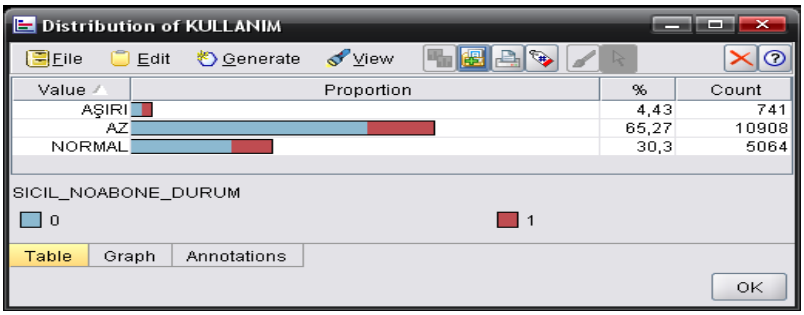


Şekil 4.2. Abone türlerine göre abonelerin dağılımı

Şekil 4.2.'de veri setindeki abonelerin % 80'i konut-1 tipi abone olup merkezde ikamet etmektedirler. % 10'luk kısmı ise konut-2 tipi abone olup Serdivan ve Erenlerde ikamet etmektedirler. Konut-2 tipi abonelerin yaklaşık % 40'ı kaçak kullanmış iken konut-1 tipi abonelerin ise yaklaşık % 15'inin kaçak kullandığı tespit edilmiştir.

2. Kullanım durumu:

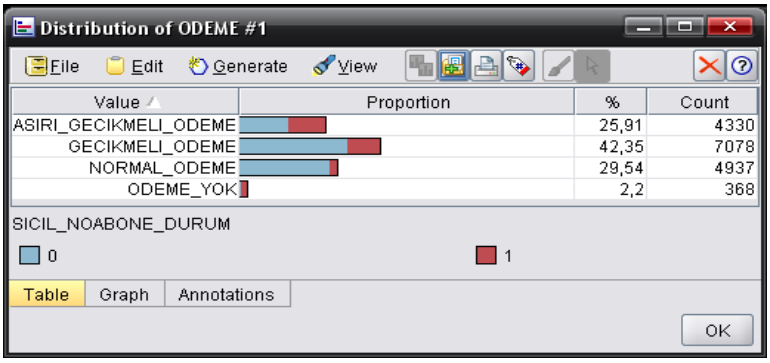
Veri setindeki abonelerin % 65'i 10m³'den daha az, % 30'u ise 10-20m³ arası su tüketen abonelerdir. Aşırı tüketim yapan abonelerin yaklaşık % 45'i, az tüketim yapanların yaklaşık % 22'si ve tüketimi normal olanların ise yaklaşık %25'i kaçak kullandığı tespit edilen abonelerdir. Şekil 4.3.'de tüketim durumuna göre abonelerin dağılımı yer almaktadır.



Şekil 4.3. Kullanım durumuna göre abonelerin dağılımı

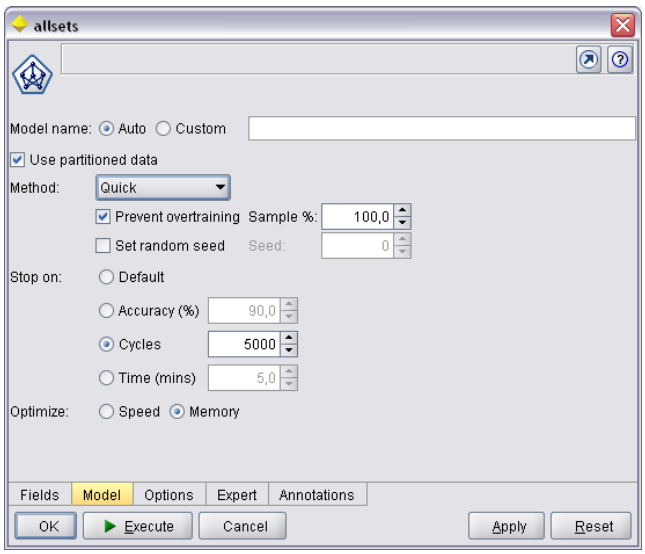
3. Ödeme durumu:

Şekil 4.4.'de yer alan ödeme yapma durumuna göre abonelerin dağılımında aşırı gecikmeli ödeme yapan aboneler tüm aboneler içinde % 25'lik bir çoğunluğa sahip iken bu abonelerin yaklaşık % 45'i kaçak kullandığı tespit edilmiş abonelerdir. Ödemelerini gecikmeli olarak yapanlar % 42, zamanında yapanlar ise yaklaşık olarak % 30'luk bir çoğunluğa sahipler. Benzer şekilde gecikmeli ödeme yapanların dörtte biri, normal ödeme yapanların ise yaklaşık % 10'u kaçak tüketim yapmış olan abonelerdir. Tüm aboneler içinde % 2'lik çoğunluğa sahip olan hiç ödeme yapmayan abonelerin yaklaşık beşte biri normal kullanıcılardır.



Şekil 4.4. Ödeme durumuna göre abonelerin dağılımı

Modelde yer alan üç değişken girdi değişkeni, sicil numaralarına göre abone durumu da çıktı değişkeni olarak kullanılmıştır. Modelde kullanılan alternatif algoritmalar ve metodlar kısaca tanıtıldıktan sonra yapılan deneme sonuçları belirtilmiştir.

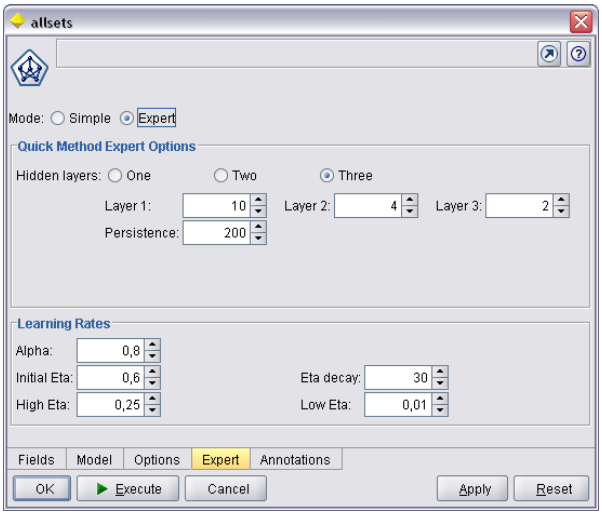


Şekil 4.5. YSA'da model seçenekleri

Şekil 4.5.'de görülen alan YSA oluşturulurken değerlendirilmesi gereken model seçeneklerini içerir. Use partitioned data bölümü işaretlenerek kısımlara ayrılmış veriler kullanılmıştır. Method bölümünde detayları aşağıda belirtilecek olan farklı YSA metodları yer almaktadır. Prevent overtraining sample alanı %100 olarak işaretlenir. Bunun sebebi veri zaten eğitim ve test verisi olarak bölünmüştür ve analiz edilen kısım verinin tamamını içermelidir. Durdurma kriteri seçmek için stop on seçeneği default dışında bir seçenek olarak işaretlenmelidir. Accuracy seçeneğine girilecek % cinsinden değere ulaşana kadar ağ öğrenmeye devam eder. Cycles seçeneğine girilen değer ise ağın öğrenmeyi sonlandırmadan önce geçeceği iterasyon sayısıdır. Time seçeneği ile ise belli bir zaman sonra öğrenme sonlandırılabilir. Fields sekmesinde target alanına çıktı; inputs alanına da girdi değişkenleri atanır. Options sekmesinde duyarlılık analizi gibi seçenekler yer alırken expert sekmesinde daha detaylı analiz seçenekleri mevcuttur. YSA'da kullanılan metodlar ve kısa açıklamaları verilmiş ardından expert seçeneği olanlar detaylandırılmıştır.

– Quick metod: Bu metod ağ için uygun bir şekil (topoloji) seçmek için verinin karakteristiklerini ve başparmak kurallarını kullanır. Şekil 4.6.'de gösterildiği üzere expert sekmesinde hidden layers seçeneğinde gizli (hidden) katman sayısı ve her bir katmanda yer alacak olan düğüm sayısı kullanıcı tarafından belirlenebilir. Gizli katmanda yer alan düğüm sayısının artması karmaşık problemleri çözmeye yardımcı olurken öğrenme zamanının da artırmaktadır. Persistence seçeneğine girilen değer ise gelişme görülmediği halde ağın eğitilmeye devam etmesini sağlayacak devir

sayısını belirler. Learning Rates (Öğrenme oranları) alanında yer alan değerlerden alpha eğitim sırasında ağaırlıkları güncellemede kullanılan bir momentum terimi olup 0-1 arasında bir değere sahip olur. Yüksek değerler momentumu artırır. Eta değerleri her güncellemede kaç tane ağırlığın alıştırıldığını kontrol eden bir öğrenme oranı olup Initial eta başlangıç eta değerini, high eta en yüksek ve low eta da en düşük eta değerini gösterir. Öğrenme başlangıç eta değeri ile başlar en düşük değere iner sonra en yüksek değere çıkar sonra yeniden en düşük değere iner ve son iki adım öğrenme tamamlanana kadar devam eder. eta decay seçeneği ise en yüksek eta değerinden en düşük eta değerine inene kadar oluşacak çevrim sayısını içerir.

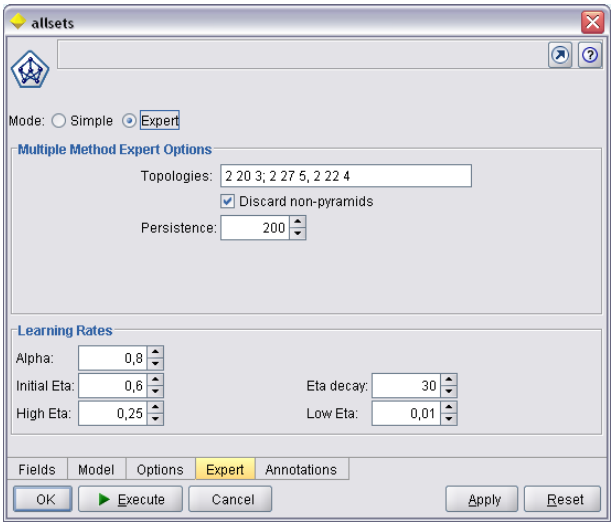


Şekil 4.6. Quick metod expert seçenekleri

– Dynamic metod: Bu metod bir başlangıç topolojisi oluşturur fakat eğitim süreci boyunca her gizli birimin (hidden unit) eklenmesi ve/veya çıkarılmasında bu topolojiyi değiştirir. Bu metodda expert sekmesi yoktur.

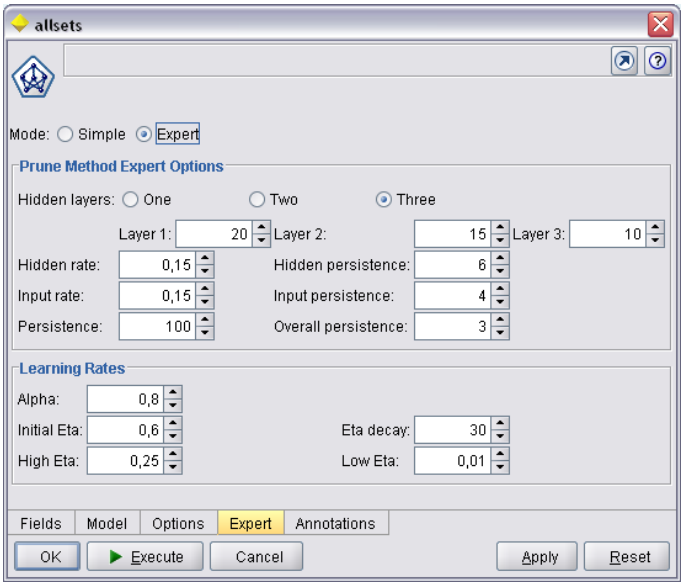
– Multiple metod: Bu metod farklı topolojilere sahip birçok ağ oluşturur (kesin sayı eğitim setine göre değişir). Bu ağlar sahte paralel tarzında eğitilir. Eğitimin sonunda en düşük RMS hatasına sahip olan model son model olarak yer alır. Şekil 4.7.'da gösterilen Expert sekmesinde persistence, discard non-pyramids ve topologies seçenekleri yer alır. Topologies seçeneği gizli katmanda yer alan gizli birim sayısını belirler. Persistence seçeneğine girilen değer ise gelişme görülmediği duruma gelinceye kadar ağın eğitilmeye devam etmesini sağlayacak devir sayısını belirler. Yüksek olması eğitim zamanını artırır. Her katmanın kendinden önceki katmanla

aynı ya da daha az gizli birim içerdiği ağlara piramit denir. Bu tip ağlar piramit olmayanlara göre daha iyi eğitilir bu yüzden discard non-pyramids seçeneği işaretlenir.



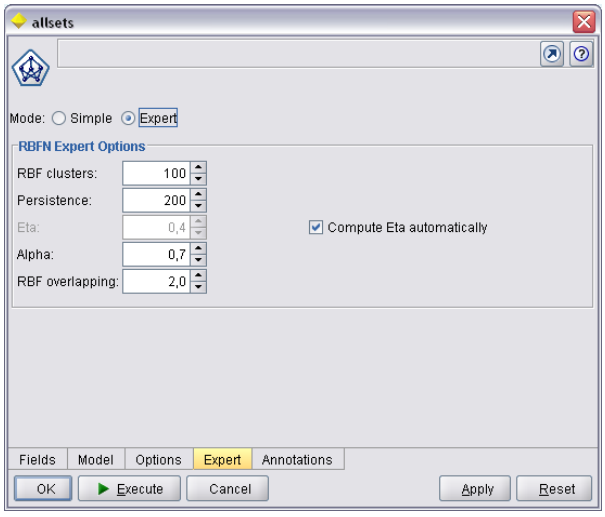
Şekil 4.7. Multiple metod expert seçenekleri

– Prune metod: Bu metod büyük bir ağla eğitime başlar ve eğitim sürecinde gizli ve girdi katmanında yer alan en güçsüz birimleri budar. Genellikle yavaştır ama diğer metodlardan daha iyi sonuçlar verir. Şekil 4.8.'de görüldüğü üzere hidden layers seçeneğinden üç tane gizli katman seçilebilir ve her bir katman için gizli birim sayısı belirlenebilir. Tüm bunlar budama öncesi başlangıç ağı için geçerlidir. Gizli katman sayısının fazla olması daha karmaşık ilişkilerin YSA tarafından öğrenilmesini sağlarken öğrenme süresini uzatır. Hidden rate tek bir gizli birim budamasında budanacak gizli birim sayısını belirler. Hidden persistence hiçbir gelişme görülmediği durumda gizli birim budama operasyon sayısını belirler. Input rate tek bir girdi budamasında budanacak girdi birim sayısını belirler. Input persistence hiçbir gelişme görülmediği durumda girdi birim budama operasyon sayısını belirler. Overall Persistence hiçbir gelişme görülmediği durumda girdi budama döngüsü/gizli birim budama boyunca geçecek zamanı belirler.



Şekil 4.8. Prune metod expert seçenekleri

– RBFN metod: RBFN (Radial Bases Function Network) Radyal tabanlı fonksiyon ağlar metodu hedef alandaki veriyi bölümlendirmek için K-en yakın kümeleme algoritmasına benzer bir teknik kullanır. Şekil 4.9.'de gösterilen expert sekmesinde yer alan RBF Clusters seçeneği kullanılacak radyal tabanlı fonksiyon ya da küme sayısını belirler. Eta değeri sabittir. Kullanıcı belirlemek isterse ilgili seçeneği aktifleştirerek eta değerini kendisi belirleyebilir. RBF overlapping seçeneği veride tanımlanan kümelerin ne kadarının örtüştüğünü kontrol eder.



Şekil 4.9. RBFN metod expert seçenekleri

– Exhaustive Prune metod: Prune metoduyla ilişkili olan bir metod olup ağ eğitme parametreleri en iyiyi bulmak için muhtemel model uzayında tamamen eksiksiz bir

araştırma sağlamak için seçilir. Genelde en yavaş metod olmasına rağmen en iyi sonucu sağlar. Büyük veri setlerinde eğitme çok uzun zaman alabilir. Bu metodda expert sekmesi yoktur.

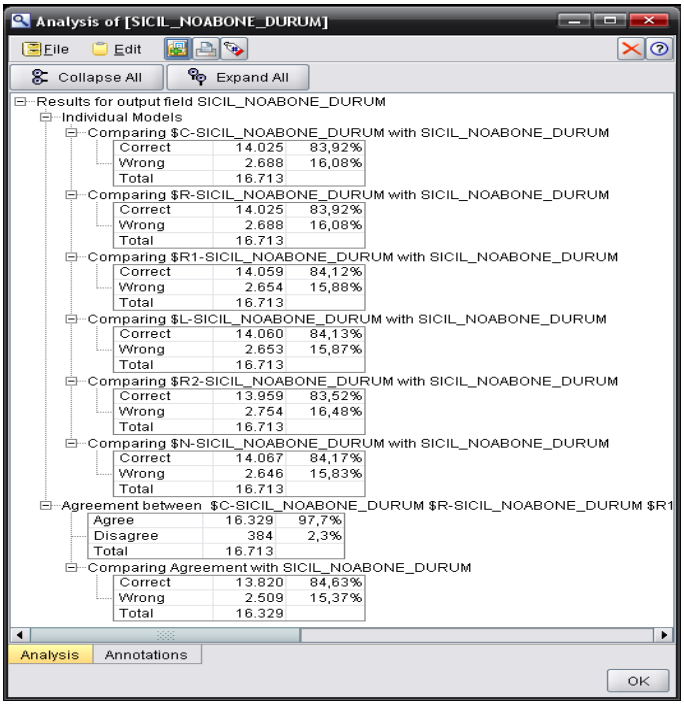
Tablo 4.1. Modelde denenen algoritmalar için eğitim ve test tahmin oranları

YÖNTEM	METOD	AÇIKLAMA	TAHMİN	EĞİTİM		TEST	
Lojistik Regresyon			DOĞRU	14.060	84,13%	6.036	84,27%
			YANLIŞ	2.653	15,87%	1.127	15,73%
Karar Ağacı	C&R Trees		DOĞRU	14.059	84,12%	6.032	84,21%
			YANLIŞ	2.654	15,88%	1.131	15,79%
Karar Ağacı	QUEST		DOĞRU	13.959	83,52%	5.998	83,74%
			YANLIŞ	2.754	16,48%	1.165	16,26%
Karar Ağacı	C5.0		DOĞRU	14.025	83,92%	6.034	84,24%
			YANLIŞ	2.688	16,08%	1.129	15,76%
Karar Ağacı	CHAID		DOĞRU	14.025	83,92%	6.033	84,22%
			YANLIŞ	2.688	16,08%	1.130	15,78%
YSA	Quick	3 katman 20-15-10, eğitim oranları default	DOĞRU	14.068	84,17%	6.036	84,27%
			YANLIŞ	2.645	15,83%	1.127	15,73%
YSA	Quick	3 katman 10-4-2, eğitim oranları default	DOĞRU	14.068	84,17%	6.037	84,28%
			YANLIŞ	2.645	15,83%	1.126	15,72%
YSA	Dynamic		DOĞRU	14.021	83,89%	6.028	84,15%
			YANLIŞ	2.692	16,11%	1.135	15,85%
YSA	Multiple	Eğitim oranları default	DOĞRU	14.068	84,17%	6.036	84,27%
			YANLIŞ	2.645	15,83%	1.127	15,73%
YSA	Multiple	Alp=0,8 initial eta=0,5 high eta=0,3	DOĞRU	14.068	84,17%	6.037	84,28%
			YANLIŞ	2.645	15,83%	1.126	15,72%
YSA	Multiple	Alp=0,7 initial eta=0,7 high eta=0,3	DOĞRU	14.068	84,17%	6.037	84,28%
			YANLIŞ	2.645	15,83%	1.126	15,72%
YSA	Prune	3 katman 20-15-10 kalan seçenekler default	DOĞRU	14.068	84,17%	6.037	84,28%
			YANLIŞ	2.645	15,83%	1.126	15,72%
YSA	RBFN	Clus=20 pers=30 Alp=0,9 Ovrlp=1,0	DOĞRU	13.959	83,52%	5.992	83,65%
			YANLIŞ	2.754	16,48%	1.171	16,35%
YSA	RBFN	Clus=50 pers=100 Alp=0,9 Ovrlp=1,0	DOĞRU	14.057	84,11%	6.035	84,25%
			YANLIŞ	2.656	15,89%	1.128	15,75%
YSA	RBFN	Clus=100 pers=200 Alp=0,7 Ovrlp=2,0	DOĞRU	14.020	83,89%	6.036	84,27%
			YANLIŞ	2.693	16,11%	1.127	15,73%
YSA	Exhaustive Prune		DOĞRU	14.066	84,16%	6.033	84,22%
			YANLIŞ	2.647	15,84%	1.130	15,78%

Yukarıda kısaca açıklanan metodlar içindeki değişik seçeneklerin değerleri değiştirilerek sonucun değişimi izlenmiş ve elde edilen sonuçlar Tablo 4.1.'de belirtilmiştir. Aynı tabloda daha önce belirtilmiş olan algoritmalarla yapılan eğitme

işlemlerinin sonuçları da belirtilmiştir. Tablo 4.1.'de her bir model için eğitim ve test verilerinin tahmin güçlerinin karşılaştırılması yer almaktadır.

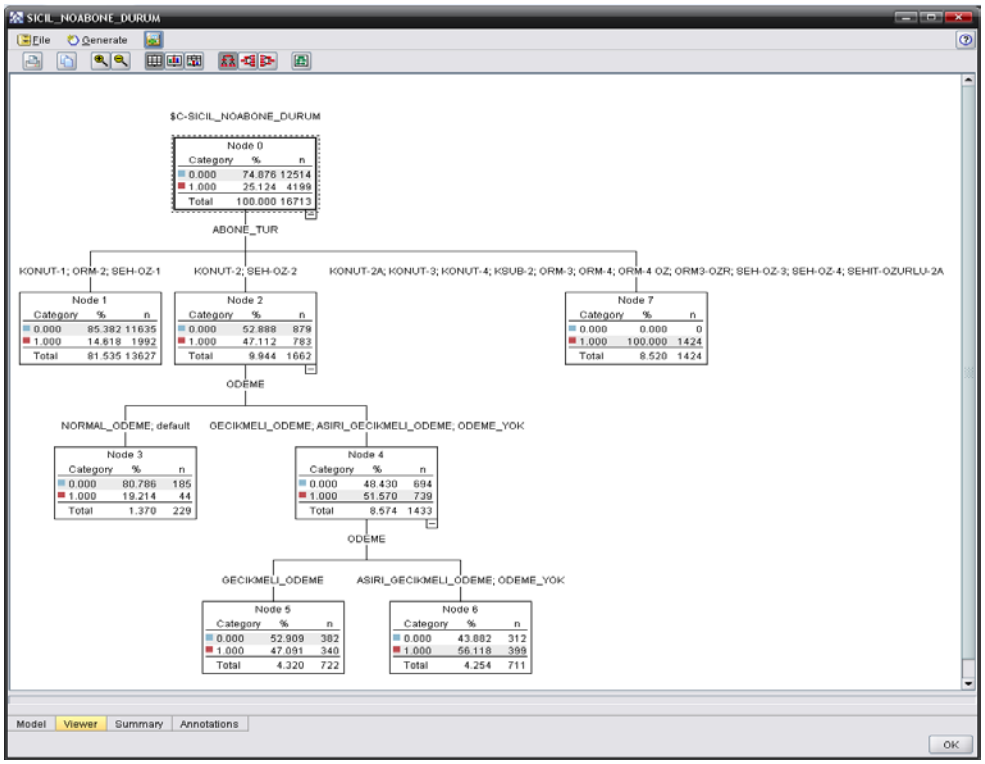
YSA metodları için durdurma kriteri olarak ilk metodda %90 doğruluk değeri kriter olarak girilmiş fakat elde edilen değerler çok fazla değişmemesine rağmen zaman olarak uzun süren bir öğrenme periyodu gerçekleşmiştir. Bu nedenle tüm YSA metodlarında durdurma kriteri beş bin iterasyon olarak belirlenmiştir. Denemeler sonucunda elde edilen değerler göz önüne alındığında YSA metodları içerinde en iyi sonucu Multiple metodunun verdiği görülmektedir. Tablo 4.1.'de sonuçları verilmiş olan algoritma ve metodlardan YSA için seçilmiş olan multiple metodu diğer algoritmalar ile clementine içinde karşılaştırılmış ve sonuç Şekil 4.10.'da gösterilmiştir.



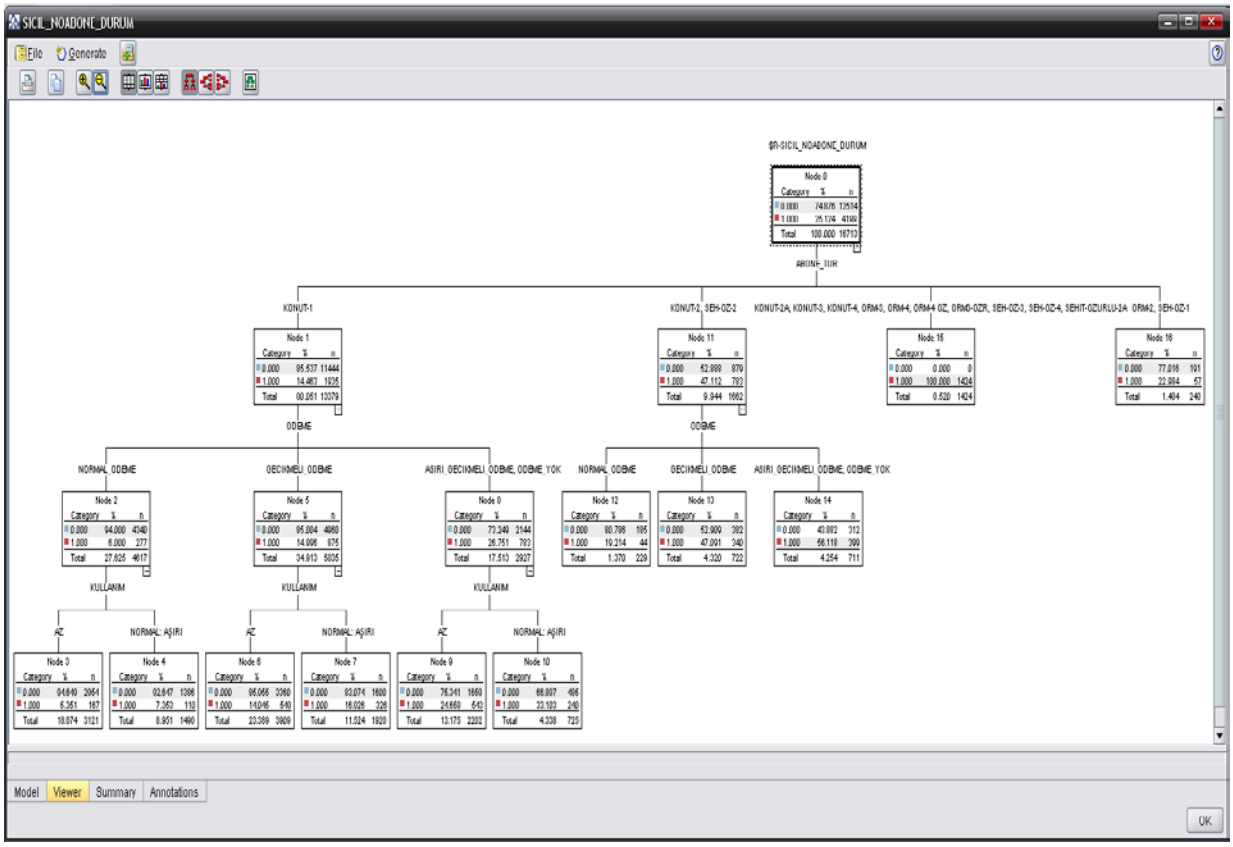
Şekil 4.10. Algoritmaların tahmin gücü karşılaştırması

En yüksek tahmin gücü YSA Multiple metoduyla elde edilmiştir. Sonra sırasıyla lojistik regresyon, C&R Trees, CHAID, C5.0 ve QUEST tahmin gücü yüksekten düşüğe doğru sıralanmıştır. Algoritmaların kendi aralarında % 97 civarında örtüştüğü tespit edilmiş ve bu mutabakatın tahmin gücü de % 84,63 olarak elde edilmiştir. Elde edilen bu sonuçlardan sonra nihai olarak modelde lojistik regresyon, YSA ve karar ağaçlarından en iyi sonucu veren CHAID kullanılacaktır. Denemelerde C5.0

CHAID'den daha iyi sonuç vermesine rağmen Şekil 4.11.'da gösterilen karar ağacında girdi değişkeni olarak abone tipi, ödeme ve kullanım durumu olduğu halde kullanım durumu karar ağacında yer almamıştır. Kullanım durumuna göre yorum yapabilmek için bu karar ağacının yeterli olmadığı gerekçesiyle tahmin gücü bu algoritmaya çok yakın olan CHAID algoritması karar ağacı algoritması olarak tercih edilmiştir.



Şekil 4.11. C5.0 karar ağacı ekran çıktısı

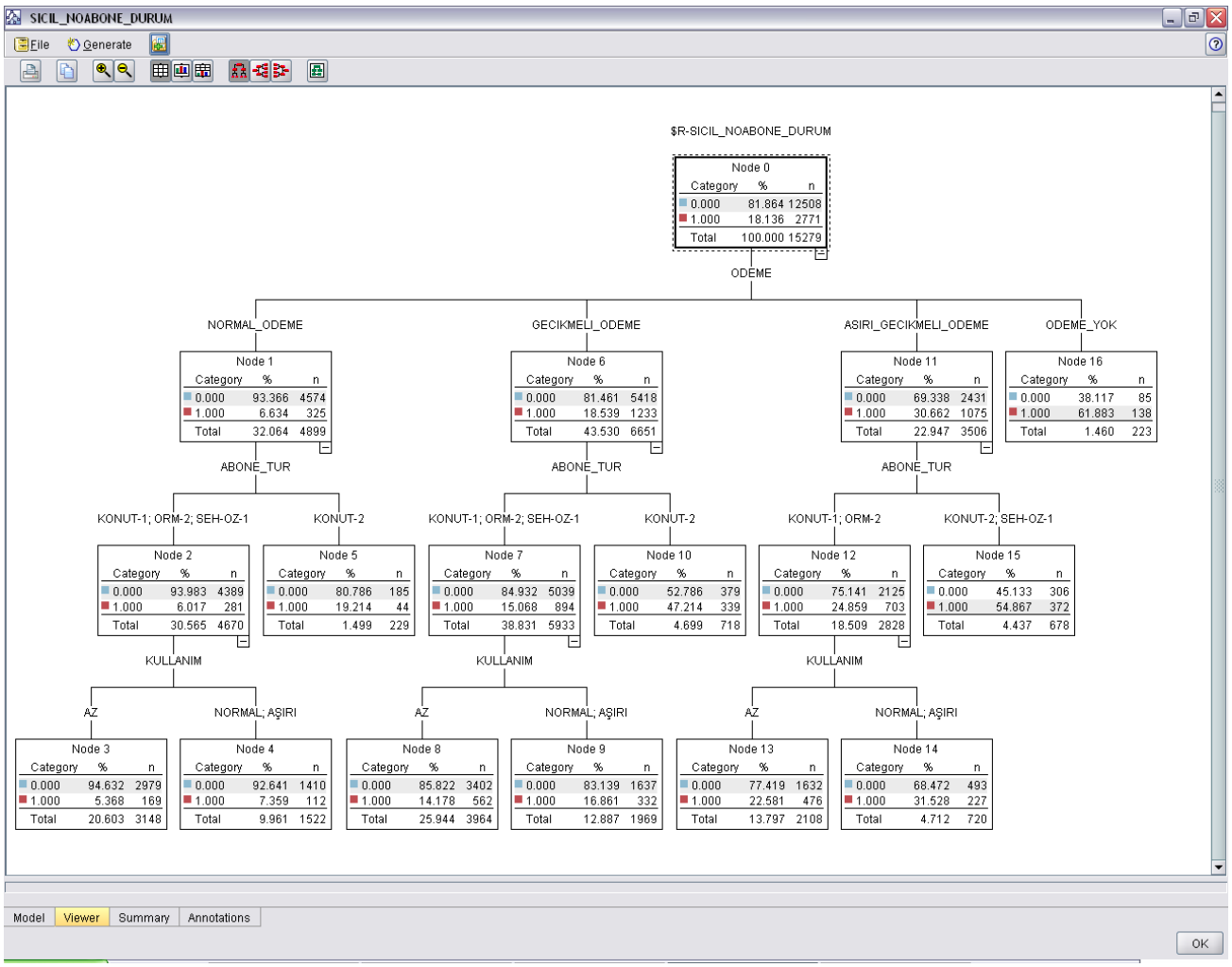


Şekil 4.12. CHAID karar ağacı ekran çıktısı

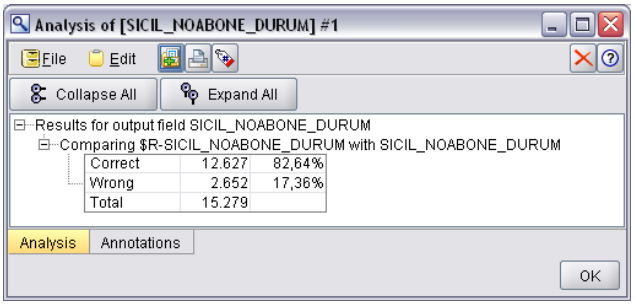
Şekil 4.12.'de yer alan karar ağacında konut-1, konut-2, seh-oz-2, orm-2 ve seh-oz-1 tipi abonelerin dışında kalan abonelerin tamamı kaçak kullanan aboneler olarak görülmektedir. Fakat ADASU ile yapılan görüşmeler sonrası veri setinde yer alan bu konut tipindeki abonelerin kaçak kullanmış olanlardan seçilmiş olduğunu fakat bu sayının da toplam sayının % 2'sini geçmediği anlaşılmıştır. Bu durumda modelin yanıltıcı tahminler yapmasına sebep olabileceği için yukarıda bahsedilen abonelerin dışında kalan abonelerin verileri veri setinden çıkarılmıştır. Yeni veri setinde modeller yeniden çalıştırılarak elde edilen sonuçlar sırasıyla açıklanmıştır.

Şekil 4.13.'de CHAID karar ağacı ekran çıktısı yer almaktadır. Veri setinde yer alan abonelerin yaklaşık % 82'si normal kullanıcı iken kalan miktar kaçak kullanıcıdır. Ödeme durumuna göre yapılan ilk bölünmede dört bölüm oluşmuştur. Bu bölümlendirmeye göre normal ödeme yapan abonelerin (1. düğüm) % 94'e yakın kısmı normal kalan % 6'lık kesim ise kaçak kullanıcıdır. Benzer şekilde gecikmeli ödeme yapanların (6. düğüm) % 81'i ve aşırı gecikmeli ödeme yapanların (11. düğüm) ise yaklaşık % 70'i normal kullanıcı kalan kesimler kaçak kullanıcıdır. Hiç

ödeme yapmamış olan abonelerin (16. düğüm) % 40'a yakını normal kullanıcı iken kalan % 60'lık kısım kaçak kullanıcıdır. Karar ağacı ikinci bölümlendirmeyi abone türüne göre yapmış ve ağaç yapısı hiç ödeme yapmamış olanların yer aldığı düğümün dışındaki düğümlerle yani 1, 6 ve 11. düğümlerle devam etmiştir. Normal ödeme yapan aboneler abone tiplerine göre konut-2 tipinde olanlar bir düğümde (5. düğüm) kalanlar diğer bir düğümde (2. düğüm) toplanmış ve ağaç yapısı bu düğümle devam etmiştir. Çünkü normal ödeme yapan konut-2 tipi aboneler % 2'lik bir çoğunluğu teşkil ederken konut-1, orm-2 ve seh-oz-1 tipi abonelerin oluşturduğu 2. düğüm % 30'luk veriyi içeriyor. Beşinci düğümde yer alan abonelerin % 80'i normal kalan beşte biri ise kaçak kullanıcıdır. Altıncı düğümden devam eden ağaç yapısında abone türüne göre yapılan bölümlendirmede konut-1, orm-2 ve seh-oz-1 tipi aboneler 7. düğümde konut-2 tipi aboneler ise 10. düğümde yer almıştır. Yedinci düğümde normal kullanıcılar % 85'lik bir orana sahip iken 10. düğümde normal kullanıcıların oranı yaklaşık olarak % 53 tespit edilmiştir. Abone tipine göre bölümlendirilmiş son düğüm 11. düğüm olup konut-1 ve orm-2 tipi aboneler 12. ve konut-2 ve seh-oz-1 tipi aboneler 15. düğümde toplanmıştır. On ikinci düğümde yer alan abonelerin yaklaşık dörtte biri kaçak kullanıcısı iken bu oran 15. düğümde % 55'e yükselmiştir. Kullanım durumuna göre devam eden bölümlendirmede ağaç yapısı konut-1 tipi abonelerin bulunduğu düğümler üzerinden devam etmiştir. Kullanım durumuna göre ağaç yapısı az kullanım bir düğümde normal ve aşırı kullanım diğer düğümde olacak şekilde 2, 7 ve 12. düğümler üzerinden devam etmiştir. Üçüncü düğümde yer alan tüketim miktarı az olan abonelerin yaklaşık % 95'i; normal ve aşırı tüketim yapan abonelerin ise % 92'si normal kullanıcıdır. Benzer şekilde sekizinci düğümde de tüketimi az olanların % 85'i normal ve aşırı olanların ise % 83'ü normal kullanıcıdır. Yine on üçüncü düğümde tüketimi az olan abonelerin yaklaşık % 78'i normal ve aşırı tüketim yapanların ise % 68'i normal kullanıcıdır. Kullanım miktarına göre devam eden ağaç yapısında ödeme durumlarına göre ödeme durumu normalden gecikmeli ve aşırı gecikmeli ödemeye doğru gittikçe kullanıma göre normal abonelerin oranları düşmektedir. Konut-2 tipi abonelerden gecikmeli ve aşırı gecikmeli ödeme yapanların kaçak kullanma oranı yaklaşık % 50'dir. Şekil 4.14.'de yer alan sonuç ekran çıktısına göre yeni veri seti için karar ağacının tahmin gücü % 82,64 olup abonelerin 12.627 tanesi doğru, 2.652 tanesi ise yanlış tahmin edilmiştir.

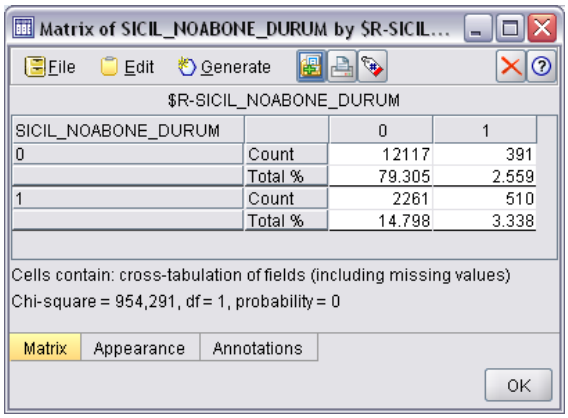


Şekil 4.13. Yeni veri seti için CHAID karar ağacı ekran çıktısı

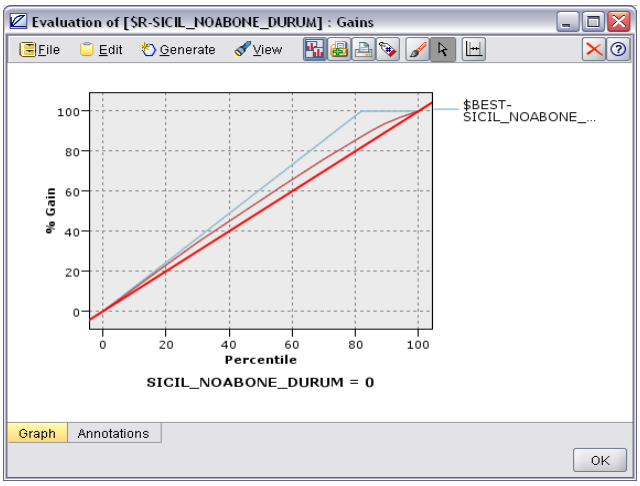


Şekil 4.14. CHAID karar ağacı tahmin gücü ekran çıktısı

Şekil 4.15.'de yer alan çapraz tabloya göre karar ağacı normal kullanıcıların 12.117 tanesini normal olarak tahmin ederken 2.261 tanesini kaçak olarak tahmin etmiştir. Benzer şekilde kaçak kullanan abonelerin de 391 tanesi normal kullanıcı olarak tahmin edilirken 510 tanesi kaçak olarak doğru bir şekilde tahmin edilmiştir. Modelin kaçaklarda tahmin oranı normal kullanıcılardan daha yüksektir. Şekil 4.16.'da modelin tahmin gücünün başarısını gösteren etkinlik grafiği yer almaktadır.

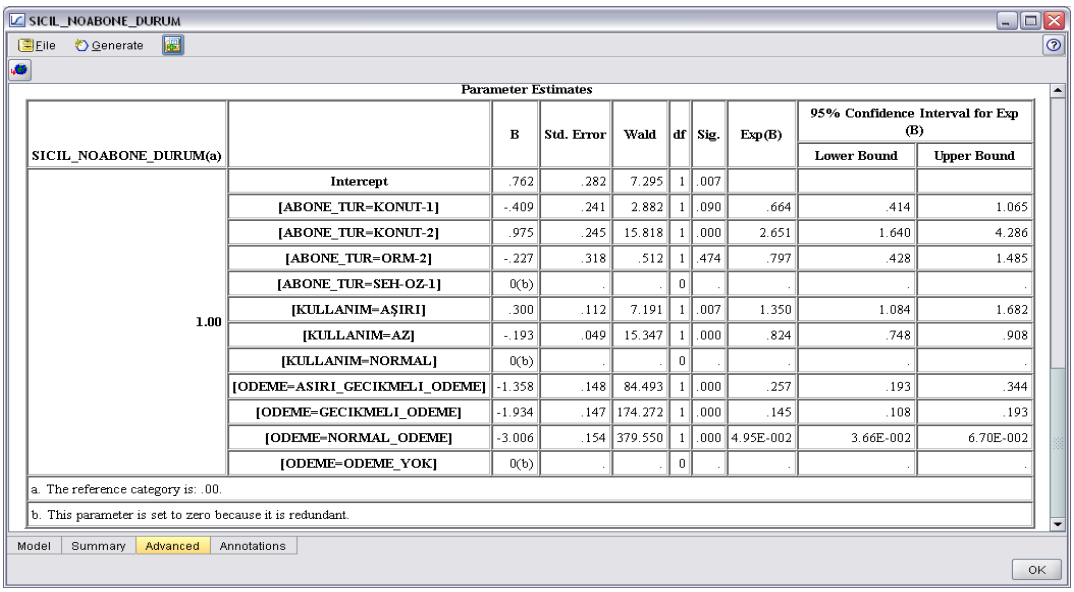


Şekil 4.15. CHAID karar ağacı için çapraz tablo

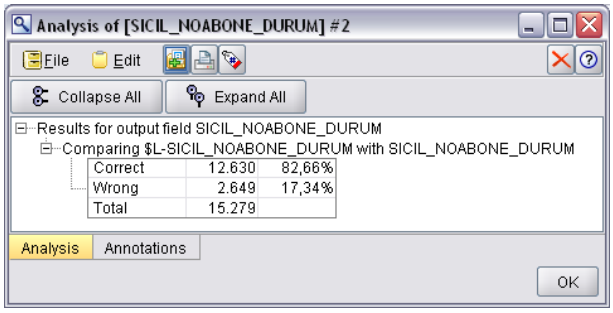


Şekil 4.16. CHAID karar ağacı etkinlik grafiği

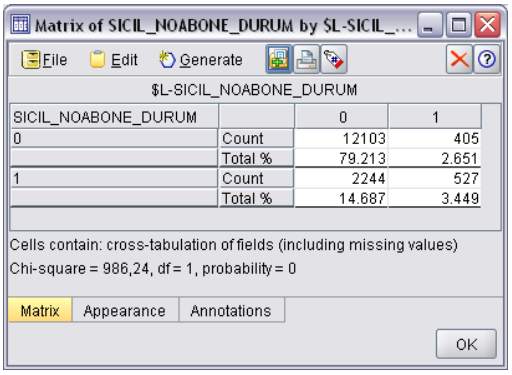
Şekil 4.17.'de yer alan lojistik regresyon ekran çıktısına göre tüm değişkenler anlamlı bulunmuştur.



Şekil 4.17. Lojistik regresyon ekran çıktısı

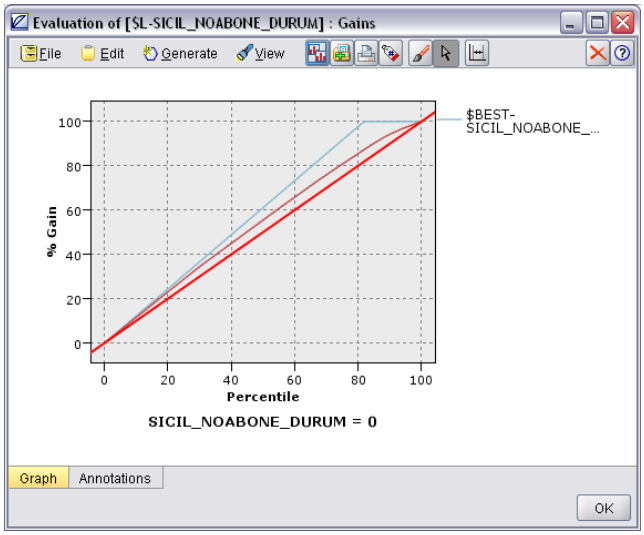


Şekil 4.18. Lojistik regresyon tahmin gücü ekran çıktısı

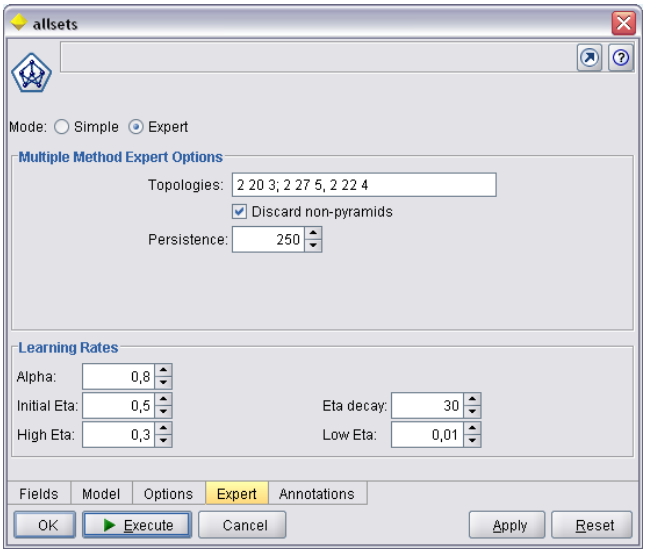


Şekil 4.19. Lojistik regresyon için çapraz tablo

Şekil 4.18.'de yer alan sonuçlara göre lojistik regresyonun tahmin gücü % 82,66 olup model abonelerin 12.630 tanesini doğru, 2.649 tanesini ise yanlış tahmin etmiştir. Şekil 4.19.'da ise aynı model için oluşturulan çapraz sonuç tablosu ve Şekil 4.20.'de ise etkinlik grafiği yer almaktadır.

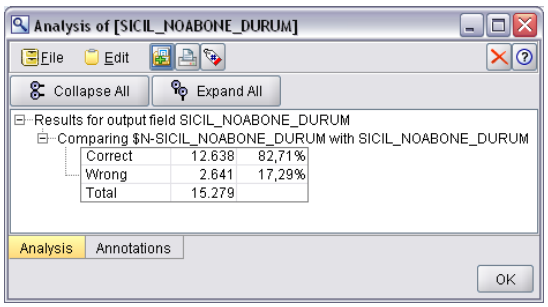


Şekil 4.20.Lojistik regresyon etkinlik grafiği

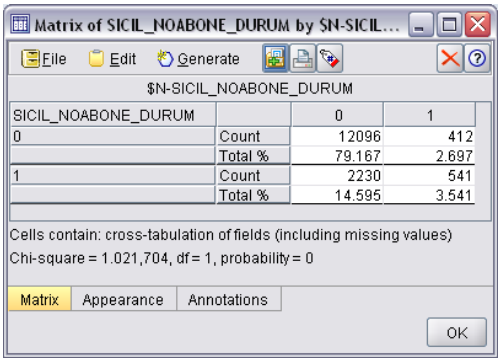


Şekil 4.21. YSA Multiple metodu parametre değerleri

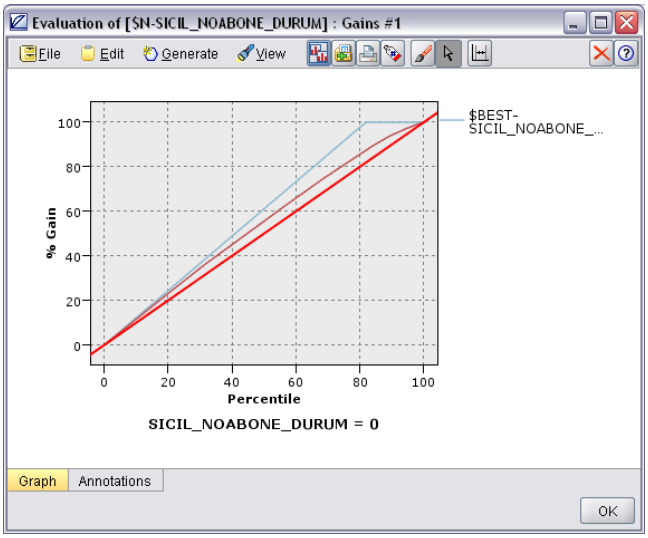
Yeni veri seti için YSA parametreleri Şekil 4.21.'de yer almaktadır. Bu parametrelere göre çalıştırılan model beş bin iterasyon sonra öğrenmeyi durdurmuştur. Şekil 4.22.'de yer alan sonuçlara göre modelin tahmin gücü diğer iki alternatif modelden de yüksek çıkarak % 82,71 olmuştur. Abonelerin 12.638'i doğru; 2.641 tanesi ise yanlış tahmin edilmiştir. Modele ilişkin çapraz sonuç tablosu Şekil 4.23.'de, etkinlik grafiği ise Şekil 4.24.'de yer almaktadır.



Şekil 4.22. YSA tahmin gücü ekran çıktısı



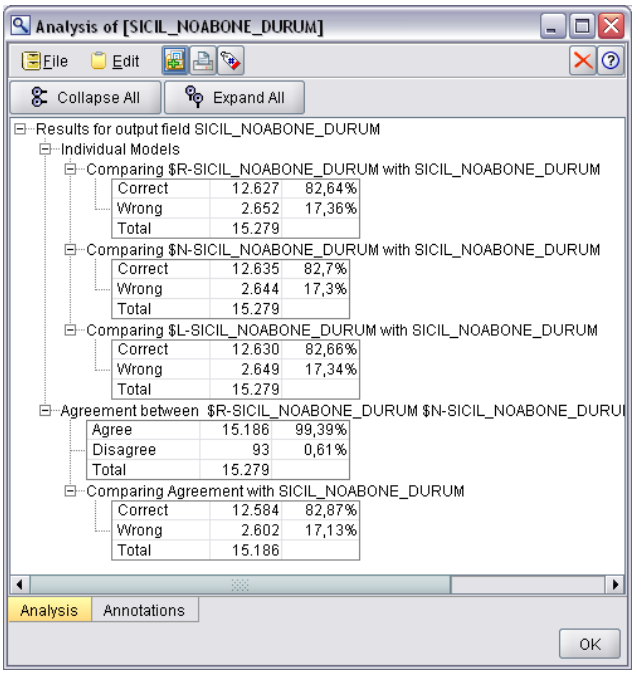
Şekil 4.23. YSA için çapraz tablo



Şekil 4.24. YSA etkinlik grafiği

Bir tahminin başarılı sayılması ideal çizgiye yakın olması ve merkez çizgiye uzak olması ile ilgilidir. Bu durumda tüm modellerin tahminleri başarılı sayılabilir. Etkinlik grafikleri incelendiğinde modeller arasındaki farkın çok az olduğu görülebilir. Modellerin birbirlerine çok yakın tahmin değerlerine sahip oldukları Şekil 4.25.'deki sonuçlardan da çok açık bir şekilde görülebilir. Modelde kullanılan

algoritmaların % 99,39 oranında aynı tahminde bulundukları ve bunun % 82,87 oranında tahmin gücüne sahip olduğu yine aynı şekildeki sonuçlardan elde edilebilir.



Şekil 4.25. YSA, CHAID KA ve LR tahmin değerlerinin karşılaştırılması

BÖLÜM 5: SONUÇ VE ÖNERİLER

Abonelik sistemiyle çalışan organizasyonlar gerçekleşmeden önce yasadışı hizmet alımını tahmin etme noktasında ciddi sıkıntı yaşamaktadırlar. Böyle bir durumun oluşması ise istenmeyen bir durumdur. Fakat oluştuğunda ciddi maddi kayıplar söz konusu olduğundan bu tip durumları önlemek için organizasyonlar farklı çözümler geliştirmişler ve hala daha geliştirmeye devam etmektedirler. Bu çalışmada ADASU için bu amaçla bir model oluşturulmaya çalışılmış, ADASU'dan temin edilen abone verilerinin bir kısmı ile bu modeller eğitilmiş diğer kısmı ile modelin testi tamamlanmış ve bir önceki bölümde yer alan sonuçlar elde edilmiştir. Kaçak kullandığı tutanakla tespit edilen abonelerden çalışma boyunca kaçak kullanıcıları, diğer abonelerden ise normal kullanıcılar olarak bahsedilmiştir.

Bu bölümde önce sonuçlar kısaca değerlendirilecek daha sonra da modelin kullanılabilirliğini artırmak dışında yasadışı kullanımı engellemek için yapılabilecek çalışmalar konusunda önerilerde bulunulacaktır. Modelin sonuçlarından aşağıdaki bilgilere ulaşılmıştır:

- Abonelerden yaklaşık 250 tanesi verilerin ait olduğu 32 aylık dönemde hiç ödeme yapmamış ama birçoğu su kullanmaya devam etmiştir. Bu abonelerin yaklaşık % 40'ı normal kullanıcılardır. Bu aboneler hakkında kaçak su kullandıklarına dair tutanak tutulmadığı halde kullandıkları suyun bedelini ödemedikleri için ADASU'ya verdikleri zarar bakımından kaçak su kullananlarla değerlendirilmelidirler.
- Veri setinin ait olduğu 32 aylık dönem içinde konut tipi abonelerde 6.405 kaçak tutanağı düzenlenmiş ve bunlardan 5.639 tanesi bir sefer kaçak tüketim yapmışken yaklaşık 350 tanesinin iki ve daha fazla defa kaçak su kullandığı tespit edilmiştir. Bu durumun önüne geçilebilmesi için kaçak su tüketenler için daha caydırıcı cezalar verilebilir ve tüketiciler bu konuda broşür ya da cep telefonlarına gönderilecek kısa mesajlarla eğitilebilir.

- Modelin ilk yapılan denemelerinde konut tipi abonelerin oluşturduğu 16 farklı tipte abonenin olduğu veri setiyle modelin tahmin gücü yaklaşık % 84 iken sadece konut-1, konut-2, orm-2 ve seh-öz-1 tipi abonelerin yer aldığı yeni veri setinin tahmin gücü % 2 azalmıştır.
- Model test edilirken bazı sıra dışı aboneler veri setinden çıkarılarak bu verilerin modelin tahmin gücünü etkilemesinin önüne geçilmiştir.
- Modelde denenen algoritmalar arasında en iyi sonucu YSA sağlamış olmasına rağmen sonuçlarının değerlendirilebilir olması bakımından karar ağacı tercihi daha doğru olabilir.
- Bazı abonelerin kaçak kullandıkları birkaç defa tespit edilmesine rağmen sadece bir kez ya da olması gerekenden daha az tutanak tutulmuş olduğu tespit edilmiştir.
- Diğer abone tiplerinde kaçak kullananlar normal kullanıcılardan daha az iken konut-2 ve seh-öz-2 tipi abonelerde kaçak kullanma oranı normal kullanma oranına yakındır.
- Abone tiplerine göre değerlendirme yapıldığında konut-1 tipi abonelerde kaçak kullanma oranı % 14 seviyesinde olup diğer abone tiplerinden daha düşüktür.
- Abonelerin sadece % 30'u zamanında ödeme yaparken bunların çok az bir kısmı kaçak kullanırken yakalanmıştır. Ödeme geciktikçe kaçak kullanma durumunun da arttığı tespit edilmiştir. Aşırı gecikmeli ödeme yapanlarda kaçak kullananlar ile normal kullanıcıların oranları birbirlerine çok yakındır.
- Tüketim miktarı arttıkça abonelerin kaçak kullanma eğilimlerinin de arttığı tespit edilmiştir. Az tüketim yapanlarda kaçak kullananlar normal kullananların dörtte biri iken aşırı kullanan abonelerin yarısı kaçak kullanıcısıdır.
- Abonelerin yaklaşık % 3'ü aşırı gecikmeli ödeme yapan tüketim miktarı az olan konut-1 ya da orm-2 tipinde aboneliğe sahip kaçak kullanıcısıdır.
- Benzer şekilde abonelerin yaklaşık % 4'ü tüketim miktarı az olan ödemesini 1 ay geciktiren konut-1, orm-2 ya da seh-öz-1 tipi aboneliğe sahip olup kaçak kullanan abonedir.

Değerlendirme sonucunda modelin daha etkin çalışabilmesi ve amacına hizmet edebilmesi için aşağıda bazı önerilerde bulunulmuştur:

- Kaçakları önlemek için yapılacak çalışmalar yanında bu hiç ödeme yapmayan abonelerin kullandıkları suyun bedelini ödemelerini sağlayacak çalışmalar da yapılmalıdır.
- Sayaç kodları üçüncü bölümde verilen tablodakine benzer şekilde kodlanarak üç aylık periyotlarla kontrol edilerek kritik değeri aşan aboneler kaçak ekiplerince kontrol edilerek varsa kaçak kullanım tespit edilebilir.
- Ekonomik olarak anlamlı olmak kaydıyla hatlara mahalle ve ana hatlara debimetreler koyularak tahakkuk dönemlerinde debimetrelerden geçen su miktarı ile tahakkuk ettirilen miktar karşılaştırılabilir. Şebekeden olabilecek kayıp toleransı da bırakılarak aradaki farkın hangi debimetrede olduğu tespit edilerek ilgili mahalle ya da sokak incelenir. Debimetre ile tahakkuk miktarları arasındaki fark çıkmaması durumunda da yine debimetreden geçen suyun bedelinin tahsil edilip edilemediği de tespit edilebilir ve bununla ilgili çalışmalar etkin bir biçimde yürütülebilir.
- ADASU'dan alınan verilere göre konut tipi aboneler ayda ortalama 10m³ su tüketmektedirler. Bu miktarın çok altına düşen aboneler süreklilik arz ediyorsa abonelerin durumları incelenebilir evde yaşayan fert sayısına bağlı olarak durum değerlendirmesi yapılabilir.
- Abonelere ait konut türü, evde yaşayan fert sayısı, aylık toplam gelir, yaş ve cinsiyet gibi bilgiler de veri tabanında tutulursa oluşturulan model daha gerçeğe yakın tahminlerde bulunacaktır.
- Bu çalışmada sadece konut tipi aboneler değerlendirilmiştir analiz tüm aboneleri de kapsayacak şekilde genişletilebilir. Değişken sayısının aynı kaldığı durumda bile abone tipi sayısı arttıkça modelin tahmin gücü de artacaktır. Benzer şekilde modelde değerlendirilecek olan değişken sayısının artması da doğru orantılı olarak modelin tahmin gücünü etkileyecektir.
- Sayaç okuma kodları modelde değerlendirilememiştir fakat uygun ve anlamlı bir düzenlemeyle değişken olarak modele dahil edilebilirse modelin tahmin gücü yükselecektir.

KAYNAKLAR

[1] DOLGUN, M.Ö., Büyük Alışveriş Merkezleri İçin Veri Madenciliği Uygulamaları, Yüksek Lisans Tezi, İstatistik ABD, Fen Bilimleri Enstitüsü, Hacettepe Üniversitesi, Ankara, 2006.

[2] ÇOBAN, A., İmalat Sanayinde Veri Madenciliği Destekli Tedarikçi Seçimi Uygulaması, Doktora Tezi, Makine Eğitimi ABD, Fen Bilimleri Enstitüsü, Sakarya Üniversitesi, Sakarya, 2006.

[3] YILMAZ, Ş.K., Veri Madenciliği: İstanbul Menkul Kıymetler Borsası Örneği, Yüksek Lisans Tezi, İşletme ABD, Sosyal Bilimler Enstitüsü, Zonguldak Karaelmas Üniversitesi, Zonguldak, 2008.

[4] TEZCANLAR, P., Müşteri İlişkileri Yönetimi, Veri Madenciliği ve Bir Uygulama, Yüksek Lisans Tezi, İşletme ABD, Sosyal Bilimler Enstitüsü, İstanbul Üniversitesi, İstanbul, 2007.

[5] SEYREK, Ö.S., Müşteri İlişkileri Yönetiminde Veri Madenciliği ve Bir Uygulama, Yüksek Lisans Tezi, İşletme ABD, Sosyal Bilimler Enstitüsü, İstanbul Üniversitesi, İstanbul, 2006.

[6] ÖZÇAKIR, F.C., Müşteri İşlemlerindeki Birlikteliklerin Belirlenmesinde Veri Madenciliği Uygulaması, Yüksek Lisans Tezi, Elektronik-Bilgisayar Eğitimi ABD, Fen Bilimleri Enstitüsü, Marmara Üniversitesi, İstanbul, 2006.

[7] GAZİ, V.E., Veri Madenciliğinde Duyarlılık, Yüksek Lisans Tezi, Bilgisayar Mühendisliği ABD, Fen Bilimleri Enstitüsü, İstanbul Teknik Üniversitesi, İstanbul, 2007.

[8] ÖZBAY, E., Finans Sektöründe Veri Madenciliği İle Dolandırıcılık Tespiti, Yüksek Lisans Tezi, Bilgisayar Mühendisliği ABD, Fen Bilimleri Enstitüsü, Selçuk Üniversitesi, Konya, 2007.

[9] KOCAMAZ, K., Hastane Bilgi Yönetim Sistemlerinde Veri Madenciliği ve Konya Meram Tıp Fakültesindeki Hastane Bilgi Yönetim Sistemi Uygulamasının İncelenmesi, Yüksek Lisans Tezi, İktisat ABD, Sosyal Bilimler Enstitüsü, Selçuk Üniversitesi, Konya, 2007.

[10] AKBULUT, S., Veri Madenciliği Teknikleri İle Bir Kozmetik Markanın Ayrılan Müşteri Analizi ve Müşteri Segmentasyonu, Yüksek Lisans Tezi, Endüstri Mühendisliği ABD, Fen Bilimleri Enstitüsü, Gazi Üniversitesi, Ankara, 2006.

[11] ÖZÇINAR, H., KPSS Sonuçlarının Veri Madenciliği Yöntemleriyle Tahmin Edilmesi, Yüksek Lisans Tezi, Bilgisayar Mühendisliği ABD, Fen Bilimleri Enstitüsü, Pamukkale Üniversitesi, Denizli, 2006.

[12] KALIKOV, A., Veri Madenciliği ve Bir e-Ticaret Uygulaması, Yüksek Lisans Tezi, Elektronik ve Bilgisayar Eğitimi ABD, Fen Bilimleri Enstitüsü, Gazi Üniversitesi, Ankara, 2006.

[13] KIZILKAYA AYDOĞAN, E., Veri Madenciliğinde Sınıflandırma Problemleri İçin Evrimsel Algoritma Tabanlı Bir Yaklaşım: Rough-Mep Algoritması, Doktora Tezi,

Endüstri Mühendisliği ABD, Fen Bilimleri Enstitüsü, Gazi Üniversitesi, Ankara, 2008.

[14] ALTINTAŞ, T., Veri Madenciliği Metotlarından Olan Kümeleme Algoritmalarının Uygulamalı Etkinlik Analizi, Yüksek Lisans Tezi, Endüstri Mühendisliği ABD, Fen Bilimleri Enstitüsü, Sakarya Üniversitesi, Sakarya, 2006.

[15] AYDIN, İ., Arıza Teşhisinde Veri Madenciliği ve Yumuşak Hesaplama Tekniklerinin Kullanımı, Yüksek Lisans Tezi, Bilgisayar Mühendisliği ABD, Fen Bilimleri Enstitüsü, Fırat Üniversitesi, Elazığ, 2006.

[16] TİRYAKİ, S., Lojistik Alanında Bir Veri Madenciliği Uygulaması, Yüksek Lisans Tezi, Fen Bilimleri Enstitüsü, İstanbul Teknik Üniversitesi, İstanbul, 2006.

[17] GÖRAL, M.A., Kredi Kartı Başvuru Aşamasında Sahtecilik Tespiti İçin Bir Veri Madenciliği Modeli, Yüksek Lisans Tezi, Endüstri Mühendisliği ABD, Fen Bilimleri Enstitüsü, İstanbul Teknik Üniversitesi, İstanbul, 2007.

[18] GÜNTÜRKÜN, F., A Comprehensive Review Of Data Mining Applications In Quality Improvement And A Case Study, Yüksek Lisans Tezi, İstatistik ABD, Fen Bilimleri Enstitüsü, Ortadoğu Teknik Üniversitesi, Ankara, 2007.

[19] YILMAZ, E., Kütahya İlinde Sosyal Sınıfların Belirlenmesi ve Veri Madenciliği İle Tüketici Profilinin Çıkarılmasına Yönelik Bir Uygulama, Yüksek Lisans Tezi, İşletme ABD, Sosyal Bilimler Enstitüsü, Dumlupınar Üniversitesi, Kütahya, 2006.

[20] KASAP, E., Sigortacılık Sektöründe Müşteri İlişkileri Yaklaşımıyla Veri Madenciliği Teknikleri ve Bir Uygulama, Yüksek Lisans Tezi, Sigortacılık Bölümü, Bankacılık ve Sigortacılık Enstitüsü, Marmara Üniversitesi, İstanbul, 2007.

[21] CERAN, G., Esnek Akış Tipi Çizelgeleme Problemlerinin Veri Madenciliği ve Genetik Algoritma Kullanılarak Çözülmesi, Yüksek Lisans Tezi, Endüstri Mühendisliği ABD, Fen Bilimleri Enstitüsü, Selçuk Üniversitesi, Konya, 2006.

[22] ÇALIŞKAN, H., Soğuk Hava Tesislerinde Optimum Soğutma Grubu Seçimi İçin Veri Madenciliği Uygulaması, Yüksek Lisans Tezi, Makine Mühendisliği ABD, Fen Bilimleri Enstitüsü, Süleyman Demirel Üniversitesi, Isparta, 2006.

[23] TOSUN, T., Veri Madenciliği Teknikleriyle Kredi Kartlarında Müşteri Kaybetme Analizi, Yüksek Lisans Tezi, Fen Bilimleri Enstitüsü, İstanbul Teknik Üniversitesi, İstanbul, 2006.

[24] AKTÜRK, H., Borsa ve Döviz Verileri Üzerinde Veri Madenciliği Teknolojisini Kullanarak Zarar Riskini Azaltan Bir Uygulama Geliştirimi, Yüksek Lisans Tezi, Bilgisayar Mühendisliği ABD, Fen Bilimleri Enstitüsü, Ege Üniversitesi, İzmir, 2008.

[25] BAYSAL, A.C., Bayi Değerlendirmesi İçin Veri Madenciliği Uygulaması, Yüksek Lisans Tezi, İşletme Mühendisliği ABD, Fen Bilimleri Enstitüsü, İstanbul Teknik Üniversitesi, İstanbul, 2008.

[26] KOLDERE AKIN, Y., Veri Madenciliğinde Kümeleme Algoritmaları ve Kümeleme Analizi, Doktora Tezi, Ekonometri ABD, Sosyal Bilimler Enstitüsü, Marmara Üniversitesi, İstanbul, 2008.

[27] ARSLAN, H., Sakarya Üniversitesi Web Erişim Kayıtlarının Web Madenciliği İle Analizi, Yüksek Lisans Tezi, Elektronik-Bilgisayar Eğitimi ABD, Fen Bilimleri Enstitüsü, Sakarya Üniversitesi, Sakarya, 2008.

[28] ŞİMŞEK, U.T., Veri Madenciliği ve Müşteri İlişkileri Yönetiminde (CRM) Bir Uygulama, Doktora Tezi, İşletme ABD, Sosyal Bilimler Enstitüsü, İstanbul Üniversitesi, İstanbul, 2006.

[29] MARTENS, D., BRUYNSEELS, L., BAESENS, B., WILLEKENS, M., VANTHIENEN,

J., Predicting going concern opinion with data mining, Decision Support Systems, 45, pp. 765-777, 2008

[30] SINHA, P.A., ZHAO, H., Incorporating domain knowledge into data mining classifiers: An application in indirect lending, Decision Support Systems, 46, pp. 287-299, 2008

[31] SUN, J., LI, H., Data mining method for listed companies' financial distress prediction, Knowledge-Based Systems, 21, pp. 1-5, 2008

[32] SHAH, D., ZHONG, S., Two methods for privacy preserving data mining with malicious participants, Information Sciences, 177, pp. 5468-5483, 2007

[33] CHU, B-H., TSAI, M-S., HO, C-S., Toward a hybrid data mining model for customer retention, Knowledge-Based Systems, 20, pp. 703-718, 2007

[34] HUNG, S-Y., YEN, D.C., WANG, H-Y., Applying data mining to telecom churn management, Expert Systems with Applications, 31, pp. 515-524, 2006

[35] HSU, C-H., Data mining to improve industrial standards and enhance production and marketing: An empirical study in apparel industry, Expert Systems with Applications, 36, pp. 4185-4191, 2009

[36] SUGUMARAN, V., KUMAR, R.A., GOWDA, B.H.L., SOHN, C.H., Safety analysis on a vibrating prismatic body: A data-mining approach, Expert Systems with Applications, 36, pp. 6605-6612, 2009

[37] HUANG, S-C., CHANG, E-C., WU, H-H., A case study of applying data mining techniques in an outfitter's customer value analysis, Expert Systems with Applications, 36, pp. 5909-5915, 2009

[38] WU, S-Y., YEN, E., Data mining-based intrusion detectors, Expert Systems with Applications, 36, pp. 5605-5612, 2009

[39] DELEN, D., FULLER, C., McCANN, C., DEEPA, R., Analysis of healthcare coverage: A data mining approach, Expert Systems with Applications, 36, pp. 995-1003, 2009

[40] LU, C-L., CHEN, T-C., A study of applying data mining approach to the information disclosure for Taiwan's stock market investors, Expert Systems with Applications, 36, pp. 3536-3542, 2009

[41] CHANG, C-J., SHYUE, S-W., A study on the application of data mining to disadvantaged social classes in Taiwan's population census, Expert Systems with Applications, 36, pp. 510-518, 2009

[42] TURHAN, B., KOÇAK, G., BENER, A., Data mining source code for locating software bugs: A case study in telecommunication industry, Expert Systems with Applications, 2009

[43] CHIEN, C-F., CHEN, L-F., Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry, Expert Systems with Applications, 34, pp. 280-290, 2008

[44] CHANG, C-L., A study of applying data mining to early intervention for developmentally-delayed children, Expert Systems with Applications, 33, pp. 407-412, 2007

[45] CHIEN, C-F., WANG, W-C., CHENG, J-C., Data mining for yield enhancement in semiconductor manufacturing and an empirical study, Expert Systems with Applications, 33, pp. 192-198, 2007

[46] CHEN, M-C., LIN, C-P., A data mining approach to product assortment and shelf space allocation, Expert Systems with Applications, 32, pp. 976-986, 2007

[47] KIRKOS, E., SPATHIS, C., MANOLOPOULOS, Y., Data mining techniques for the detection of fraudulent financial statements, Expert Systems with Applications, 32, pp. 995-1003, 2007

[48] YEN, S-J., LEE, Y-S., An efficient data mining approach for discovering interesting knowledge from customer transactions, Expert Systems with Applications, 30, pp. 650-657, 2006

[49] ENKE, D., THAWORNWONG, S., The use of data mining and neural networks for forecasting stock market returns, Expert Systems with Applications, 29, pp. 927-940, 2005

[50] BAYAM, E., LIEBOWITZ, J., AGRESTI, W., Older drivers and accidents: A meta analysis and data mining application on traffic accident data, Expert Systems with Applications, 29, pp. 598-629, 2005

[51] ÖZTÜRK. T., Yapay Sinir Ağları İle Talep Tahmini, Lisans Tezi, Endüstri Mühendisliği Bölümü, Kocaeli Üniversitesi, Kocaeli, 2004.

ÖZGEÇMİŞ

Muhammed Ali YAVUZ 1981 yılında Adapazarı'nda doğmuş, ilk ve orta öğretimi farklı şehirlerde okuduktan sonra 2004 yılında Kocaeli Üniversitesi Endüstri Mühendisliği bölümünden mezun olmuştur. Askerlik görevi sonrası bir süre Toprak Mahsulleri Ofisi Konya Şube Müdürlüğünde çalıştıktan sonra 2007 yılında Milli Prodüktivite Merkezi Trabzon Bölge Müdürlüğünde Uzman Yardımcısı olarak göreve başlamış ve bir yıla yakın bir süre bu görevi icra etmiştir. 2009 yılının başından beri ise Türkiye Vagon Sanayi A.Ş. Planlama ve Koordinasyon Daire Başkanlığı bünyesinde çalışma hayatını devam ettirmektedir.