BANKACILIK SEKTÖRÜNDE PERSONEL SEÇİMİ VE PERFORMANS DEĞERLENDİRİLMESİNE İLİŞKİN VERİ MADENCİLİĞİ UYGULAMASI

Hamdi BİLEN

YÜKSEK LİSANS TEZİ
ENDÜSTRİ MÜHENDİSLİĞİ

GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

OCAK 2009
ANKARA

Hamdi BİLEN tarafından hazırlanan BANKACILIK SEKTÖRÜNDE PERSONEL SEÇİMİ VE PERFORMANS DEĞERLENDİRİLMESİNE İLİŞKİN VERİ MADENCİLİĞİ UYGULAMASI adlı bu tezin Yüksek Lisans tezi olarak uygun olduğunu onaylarım.

Prof. Dr. Ertan GÜNER …….
Tez Danışmanı, Endüstri Mühendisliği Anabilim Dalı

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

Prof. Dr. İhsan ALP .....
İstatistik Yöneylem Araştırması ABD, G.Ü.

Prof. Dr. Hadi GÖKÇEN .....
Endüstri Mühendisliği ABD, G.Ü.

Prof. Dr. Ertan GÜNER .....
İstatistik ABD, G.Ü.

Tarih: ...../……/……

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

Prof. Dr. Nail ÜNSAL …….
Fen Bilimleri Enstitüsü Müdürü

TEZ BİLDİRİMİ

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

Hamdi BİLEN

BANKACILIK SEKTÖRÜNDE PERSONEL SEÇİMİ VE PERFORMANS DEĞERLENDİRİLMESİNE İLİŞKİN VERİ MADENCİLİĞİ UYGULAMASI

(Yüksek Lisans Tezi)

Hamdi BİLEN

GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

Ocak 2009

ÖZET

İletişim teknolojilerindeki gelişme ile birlikte "mevcut bilgi"ye ulaşmanın çok kolaylaştığı günümüzde, "bilginin çıkarımı" kavramı giderek önem kazanmaktadır. Çok büyük miktarlardaki verinin yararlı bilgilere dönüştürülmesine duyulan ihtiyaç veri madenciliğinin önemini ortaya koymaktadır. Diğer taraftan personel kalitesi ise günümüzde firmaların rekabet avantajı sağlaması acısından oldukça önemli bir noktaya gelmektedir. Bu çalışmada, veri madenciliği yöntemlerinden sınıflandırma ve kümeleme ile etkili bir personel seçim mekanizması geliştirilerek özellikle personel seçimi sürecinde fayda sağlanması amaçlanmıştır. Çalışmada veri madenciliği yazılımı olarak WEKA kullanılmış ve banka şubelerinde satışa yönelik çalışan personeller için bir uygulama gerçekleştirilmiştir.

Bilim Kodu : **906.2.062**
Anahtar Kelimeler :**Veri madenciliği (VM), kümeleme, sınıflandırma, personel seçimi, Weka**
Sayfa Adedi : **102**
Tez Yöneticisi : **Prof. Dr. Ertan GÜNER**

DATA MINING APPLICATION FOR PERSONNEL SELECTION AND PERFORMANCE EVALUATION IN BANKING SECTOR

(M.Sc. Thesis)

Hamdi BİLEN

GAZİ UNIVERSITY

INSTITUTE OF SCIENCE AND TECHNOLOGY

January 2009

ABSTRACT

As the communication technologies has been progressing, it has been easier to reach "current information" and "knowledge discovery" concept has increasingly become more important. The necessity of turning huge amounts of data into useful information indicates the importance of data mining. On the other hand, personnel quality is an important point for companies in order to maintain competitive advantages. In this study, an effective personnel selection mechanism is improved by classification and clustering in data mining and generating useful decision rules is aimed. WEKA is used as a data mining software and an application for sales employees in banking sector is conducted.

Science Code : 906.2.062
KeyWords : Data Mining, clustring, classification, personnel selection, Weka
Page Number: 102
Adviser : Prof. Dr. Ertan GÜNER

TEŞEKKÜR

Bu tezin hazırlanması aşamasında yardımlarını esirgemeyen, bana çalışmamın her aşamasında yol gösteren Hocam Prof. Dr. Ertan GÜNER' e ve Araş. Gör. Dr. Tahsin Çetinyokuş' a , tezime maddi destek sunan TÜBİTAK' a ve aileme teşekkürü bir borç biliyorum…

İÇİNDEKİLER

Sayfa

Sayfa

Sayfa

ÇİZELGELERİN LİSTESİ

Çizelge **Sayfa**

ŞEKİLLERİN LİSTESİ

Şekil **Sayfa**

RESİMLERİN LİSTESİ

Resim **Sayfa**

SİMGELER VE KISALTMALAR

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

Kısaltmalar	**Açıklama**
BPY	Bireysel Portföy Yöneticisi
ODTÜ	Orta Doğu Teknik Üniversitesi
OLAP	Çevrimiçi Analitik İşleme (OnLine Analytical Processing)
PY	Portföy Yöneticisi
SPK	Sermaye Piyasası Kurulu
TPY	Ticari Portföy Yöneticisi
VM	Veri Madenciliği
VTBK	Veri Tabanı Bilgi Keşfi
YSA	Yapay Sinir Ağları

1. GİRİŞ

Veri miktarı gün geçtikçe artmakta ve artan veri miktarıyla birlikte firmalar bilgi elde etmek adına eldeki verileri etkin bir şekilde kullanmaya çalışmaktadırlar. Artan rekabet koşulları ve gelişen bilgisayar teknolojileri sonucunda firmalar için avantaj sağlayacak bilgiler önem kazanmaktadır. Gerek veri hacmindeki artış gerekse bilişim sektöründe giderek düşen maliyetler veri madenciliğini gittikçe önemli hale getirmektedir.

Veritabanı sistemlerinin artan kullanımı ve hacimlerindeki olağanüstü artış, işletmeleri toplanan verilerden nasıl faydalanılabileceği problemi ile karşı karşıya bırakmıştır. Geleneksel sorgu veya raporlama araçlarının veri yığınları karşısında yetersiz kalması, veri madenciliği (VM) gibi yeni arayışlara neden olmaktadır.

Veri madenciliğinin son dönemlerde bilgi endüstrisinde giderek önem kazanmasında en önemli etken, giderek artan veri ve bu verinin yararlı bilgilere dönüştürülmesine duyulan acil ihtiyaçtır. Veri madenciliği ile ilgili literatürde çok sayıda tanım yapılmış olup yapılan bu tanımlardan çıkan ortak sonuç, veri madenciliğinin büyük veri yığınlarından anlamlı, şaşırtıcı ve fayda sağlayıcı bilgi çıkarımını gerçekleştirmesidir. Küçük ayrıntıların bile büyük rekabet avantajı sağladığı günümüz rekabet koşullarında veri madenciliği önemini giderek arttırmaktadır.

Son dönemlerde başta bankacılık, sigortacılık, finans ve pazarlama sektörü olmak üzere pek çok alanda VM uygulamalarına rastlanılmaktadır. Ancak literatürü incelediğimizde insan kaynakları yönetimine ilişkin çok az sayıda VM uygulamasına rastlanılmaktadır.

Firmaların kendilerine rekabet avantajı sağlaması açısından gün geçtikçe artan rekabet koşulları içerisinde personel kalitesi giderek daha da önemli bir hal almaktadır. Etkili bir personel seçim mekanizması ile doğru insanın, doğru yetenekler ile doğru yerde bulunmasının sağlanması organizasyonlar için kritik bir süreç olmaktadır.

Bankacılık sektörüne yönelik VM uygulamaları incelendiğinde literatürde kredi kartı dolandırıcılıklarının tespiti, kredi kartı harcamalarına göre müşteri gruplarının oluşturulması, kredi taleplerinin değerlendirilmesi gibi uygulamalarla karşılaşılmaktadır. Ancak bankacılık sektöründe personel seçimine yönelik VM uygulamalarına pek rastlanılmamaktadır. Türkiye Bankalar Birliği verilerine göre ülkemizde nüfus ve gelir düzeyindeki gelişmelerle birlikle 1961-2007 döneminde şube sayısı 4 kat, personel sayısı ise 5 kat artarak sırasıyla 7618 ve 158534 olmuştur. Bu derecede çok şubenin ve çalışanın olduğu bir sektörde, etkili bir personel seçimi ve performans değerlendirilmesi oldukça önemli bir konu olmaktadır.

Bu çalışmada, bankacılık sektöründe çalışan ve banka şubelerinde ticari ve/veya bireysel müşterilere hizmet sunan satış personellerinin performans düzeyleri; yaş, cinsiyet, medeni hal, tecrübe, öğrenim durumu, yabancı dil bilgisi gibi kişisel özellikleri; tecrübe, unvan gibi kariyer özellikleri ile çalıştığı şubenin özellikleri dikkate alınarak personellerin değerlendirilmesi ve atanmasına yönelik kriterler ortaya koymak amaçlanmıştır. Çalışanların performans düzeylerine göre gruplara ayrılmasında k-ortalama kümeleme algoritmasından yararlanılmış ve kümeleme sonucu belirlenen performans düzeylerine göre çalışanların şubelere atanmasına yönelik karar kuralları oluşturulmuştur. Bunun için karar ağaçları, yapay sinir ağları (YSA), Bayes sınıflayıcısı gibi sınıflandırma yöntemleri kullanılmış ve sonuçları karşılaştırılmıştır. Çalışmada, ülkemizde faaliyet gösteren bankalardan birine ait insan kaynakları ve performans verisi alınarak VM gerçekleştirilmiştir. Bu süreçte, sınıflandırma ve kümeleme algoritmalarını kolaylıkla uygulayabileceğimiz bir VM yazılımı olan açık kaynak kodlu WEKA kullanılmıştır.

Tez çalışmasının ikinci bölümünde veri madenciliğine genel bir bakış sunulmuş, üçüncü bölümde sınıflandırma, kümeleme ve birliktelik kuralı algoritmalarından bahsedilmiş, sonraki bölümde ise bankacılık sektörü çalışanlarını değerlendirmeye yönelik bir VM uygulaması Weka yazılımı kullanılarak gerçekleştirilmiştir. Son bölümde ise çalışma sonuçlarına yer verilmiş ve genel bir değerlendirme yapılmıştır.

2. VERİ MADENCİLİĞİNE GENEL BAKIŞ

Veri madenciliği, bilgi teknolojilerindeki gelişme ve küresel rekabet dolayısıyla gün geçtikçe büyüyen ve önemi daha da artan bir alan olmaya başlamıştır. Bilimsel çevrelerde uzun yıllardır var olan ancak sektörel ilgiyi çok daha geç bulan bu alanda yapılan çalışmalar giderek çeşitlilik kazanmaktadır [Giudici, 2003].

Son yıllarda bilgi teknolojisinde veri madenciliğinin büyük dikkat çekmesinin en büyük sebebi çok büyük miktarlardaki verinin elde edilebilirliği ve böyle verilerin yararlı bilgilere dönüştürülmesine duyulan ihtiyaçtır. Elde edilen bilgi işletme yönetimi, üretim kontrol ve pazar analizinden, mühendislik tasarımı ve bilimsel keşiflere kadar değişen uygulamalar için kullanılabilmektedir [Han ve Kamber, 2001].

Veri madenciliği büyük veri yığınlarında gizli olan örüntüleri ve ilişkileri ortaya çıkarmak için istatistik ve yapay zeka kökenli çok sayıda ileri veri çözümleme yönteminin tercihen görsel bir programlama ara yüzü üzerinden kullanıldığı bir süreçtir. Veri madenciliği algoritmaları; istatistik kökenli algoritmalar, matematiksel algoritmalar ve yapay zeka algoritmalarını bir arada içerir [Dolgun ve Zor, 2006].

Veri madenciliği keşif odaklıdır. Veri madenciliği, istatistik, karar ağaçları, genetik algoritma, sinir ağları ve görsel teknikler gibi çeşitli teknikleri içermektedir [Chien ve Chen, 2008].

Bir veri madenciliği yöntemini uygulayabilmek, işin gereksinimlerini probleme uyarlayarak bütünleşik bir yöntemin kullanılması demektir. Bunun için, problemin analizi, veritabanı gereksinimlerinin sağlanması ve stratejik kararın alınabileceği son hedef için başarılı, önemli sonuçlar veren, bilgisayarda uygulanabilecek istatistiksel tekniklerin kullanılması gerekmektedir. Stratejik karar kendine özgü yeni ölçülere gerek duyacaktır ve sonuç olarak veri madenciliğinin eyleme geçirdiği bilgilerin faydalı çevrimi olarak adlandırılan yeni iş gereksinimlerini de beraberinde getirecektir [Berry ve Linoff, 1997].

Bilgi teknolojilerinin gelişimi ile birlikte insan kaynakları yönetiminin çıktılarını geliştirmede karar destek sistemleri ve uzman sistemler geliştirilmiştir. Veri madenciliği en çok dikkat çeken başlıklardan biri olarak özellikle göz önüne alınmaktadır. Veri madenciliği, fayda sağlayacak örüntülerin veya kuralların geniş veri tabanlarından otomatik veya yarı otomatik keşfi ve veri analizi ile geçerek elde edilmesidir. Veri madenciliği pazarlama, finans, bankacılık, imalat, sağlık, müşteri ilişkileri yönetimi ve organizasyon öğrenmede sıklıkla uygulanmaktadır. Ancak insan kaynakları yönetimine ilişkin çok az sayıda uygulama yapılmıştır [Chien ve Chen, 2008].

Veri madenciliği için yapılan tanımlardan bazıları ise şöyledir:

Veri madenciliği, veritabanı sahibine anlaşılır ve faydalı sonuçlar vermek amacıyla, büyük miktardaki verilerin daha önceden bilinmeyen ilişki ve kuralların keşfedilebilmesi için modelleme, çıkarım ve seçim sürecidir [Giudici, 2003].

Veri madenciliği, büyük veri kümesi içinde saklı olan genel örüntülerin bulunmasıdır [Holsheimer ve Siebes, 1994].

Veri madenciliği ham verinin tek başına sunamadığı bilgiyi çıkaran veri analizi sürecidir [Jacobs, 1999].

Frawley ve ark. (1991), veri madenciliği önceden bilinmeyen ve potansiyel olarak faydalı olabilecek, veri içindeki gizli bilgilerin çıkarılması olarak tanımlamıştır [Bersone ve ark., 1999].

Fayyad ve arkadaşları (1996), veri madenciliğini geçerli, yeni, potansiyel olarak faydalı ve açıklayıcı örüntülerin veriden keşfedildiği karışık olmayan bir süreç olarak tanımlamıştır [Fayyad ve ark., 1996].

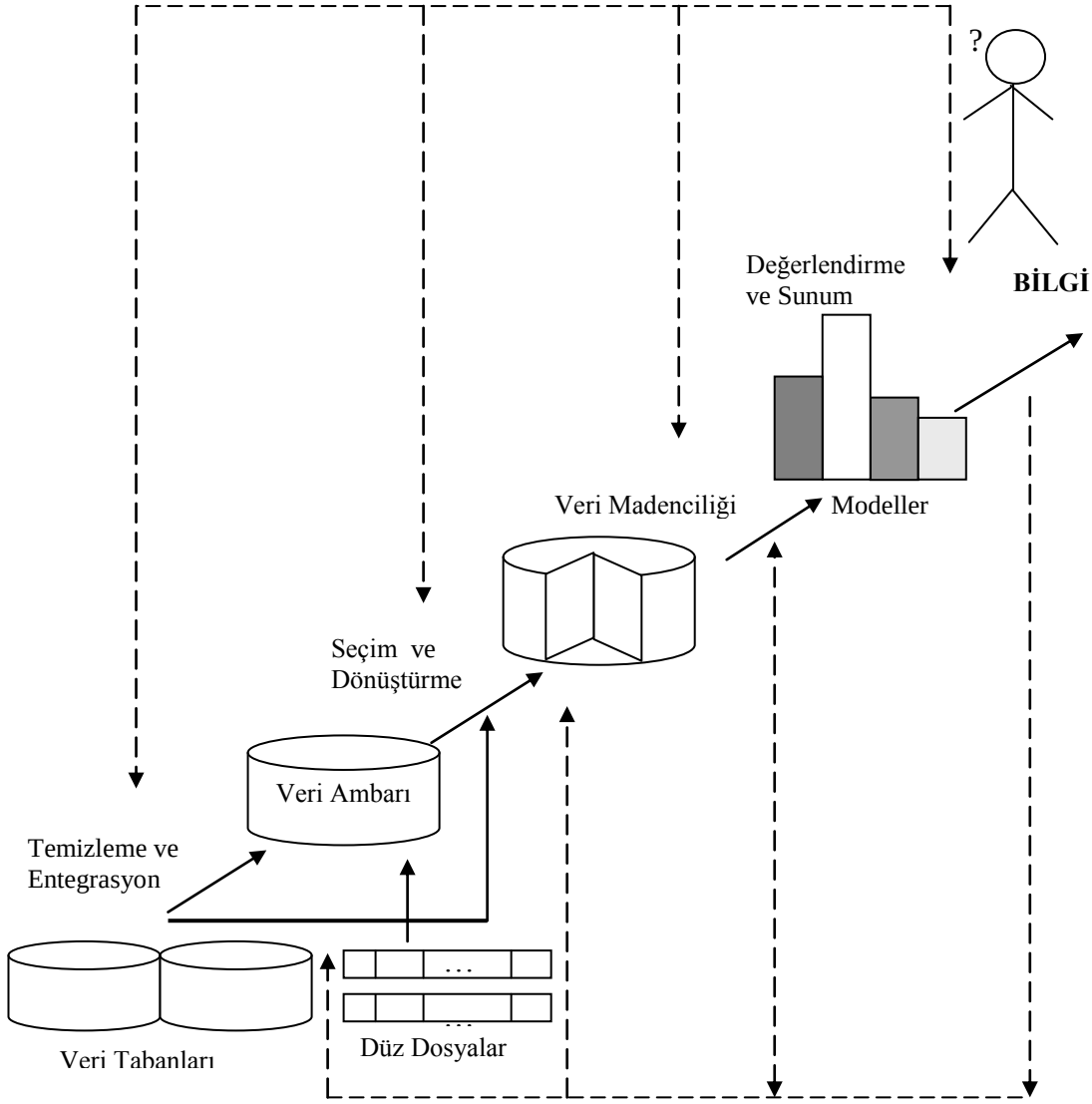
Veri madenciliği istatistik, veritabanı teknolojisi, örüntü tanıma, makine öğrenme ile etkileşimli yeni bir disiplin ve geniş veritabanlarında önceden tahmin edilemeyen ilişkilerin ikincil analizidir [Hand, 1998].

Yapılan tanımlardan da anlaşılacağı gibi veri madenciliği ile büyük veri yığını içindeki verinin anlamlı ve fayda sağlayıcı bilgiye dönüşümü sağlanmaktadır. Bu süreçte ise görsel programlama ara yüzleri kullanılmaktadır.

Waikato Üniversitesi tarafından geliştirilmiş olan Weka ile bu tür çalışmaları tek bir arayüz üzerinden yapmak mümkündür [Witten ve Frank, 2005]. WEKA başta Yeni Zelanda'da tarımsal verinin işlenmesi amacıyla geliştirilmiştir. Bununla birlikte sahip olduğu öğrenen makine metodları ve veri mühendisliği kabiliyeti öyle hızlı ve köklü bir şekilde gelişmiştir ki, veri madenciliği uygulamalarının tüm formlarında yaygın olarak kullanılmaktadır [Frank ve ark., 2004] .

Veri Madenciliği adımı, kullanıcı veya bilgi tabanı ile ilişki halindedir. İlgi çekici modeller kullanıcıya sunulur ve yeni bir bilgi olarak bilgi tabanında depolanabilir. Burada şuna dikkat edilmelidir ki; veri madenciliği, gizli modelleri değerlendirmek için ortaya çıkaran zorunlu bir adım olmasına rağmen, tüm proseste sadece bir adımdır [Fayyad ve ark., 1996; Han ve Kamber, 2001].

Şekil 2.1.' de bilgi keşfi sürecinde tariflenen ve takip eden adımlarda sıralı dizinin bir bileşeni olan veri madenciliği görülmektedir.

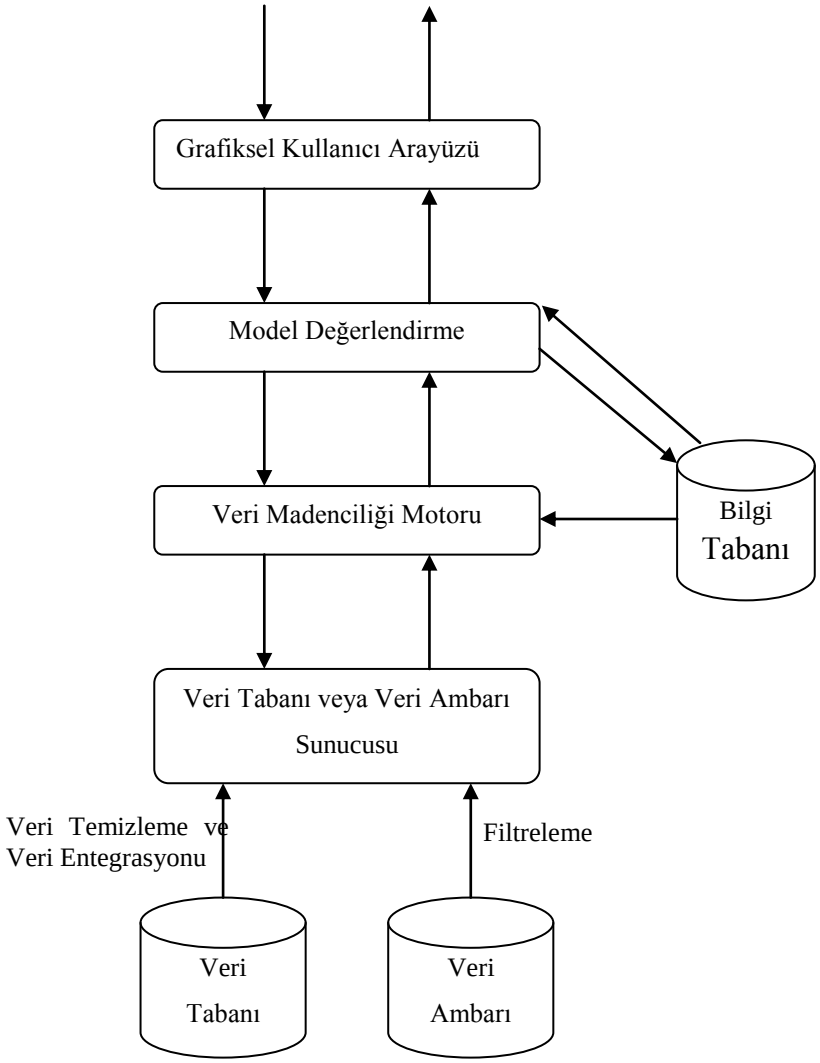


Şekil 2.1. Bilgi keşfi sürecinde veri madenciliği adımı [Han ve Kamber, 2001]

Veri madenciliği, veri tabanı, veri ambarı veya diğer bilgi kaynaklarındaki büyük veri yığınları içerisindeki şaşırtıcı bilgilerin keşfedilmesidir. Bu bakış açısıyla, tipik bir veri madenciliği sisteminin mimarisi Şekil 2.2.' de gösterilen aşağıdaki temel bölümlere sahip olmalıdır [Han ve Kamber, 2001]:

- Veritabanı, veri ambarı veya diğer bilgi kaynakları
- Veritabanı veya veri ambarı sunucusu
- Bilgi tabanı

- Veri madenciliği motoru
- Model değerlendirme modülü
- Grafiksel kullanıcı arayüzü



Şekil 2.2. Tipik bir veri madenciliği sisteminin mimarisi [Han ve Kamber, 2001]

Veri madenciliği, veritabanı teknolojileri, istatistik, makine öğrenme, yüksek performanslı hesaplamalar, model tanıma, sinir ağları, veri görselleştirme, bilgi çıkarımı, görüntü ve sinyal işleme ve uzaysal veri analizi gibi çoklu disiplinlerin tekniklerinin bütünleşmesinden oluşmaktadır. Veri madenciliğinin uygulanmasıyla, veritabanlarından ilgi çekici bilgiler, kurallar veya üst seviye bilgiler elde edilebilir, görüntülenebilir veya farklı açılardan göz atılabilir. Keşfedilen bilgi karar vermede,

proses kontrolünde, bilgi yönetiminde veya sorgu işlemede uygulanılabilir. Bu yüzden veri madenciliği, veri tabanı sistemlerinde en önemli alanlardan biri ve bilgi endüstrisinde umut verici disiplinler arası gelişmelerden birisi olarak değerlendirilmektedir.

2.1. Veri Madenciliğinin Kullanım Alanları

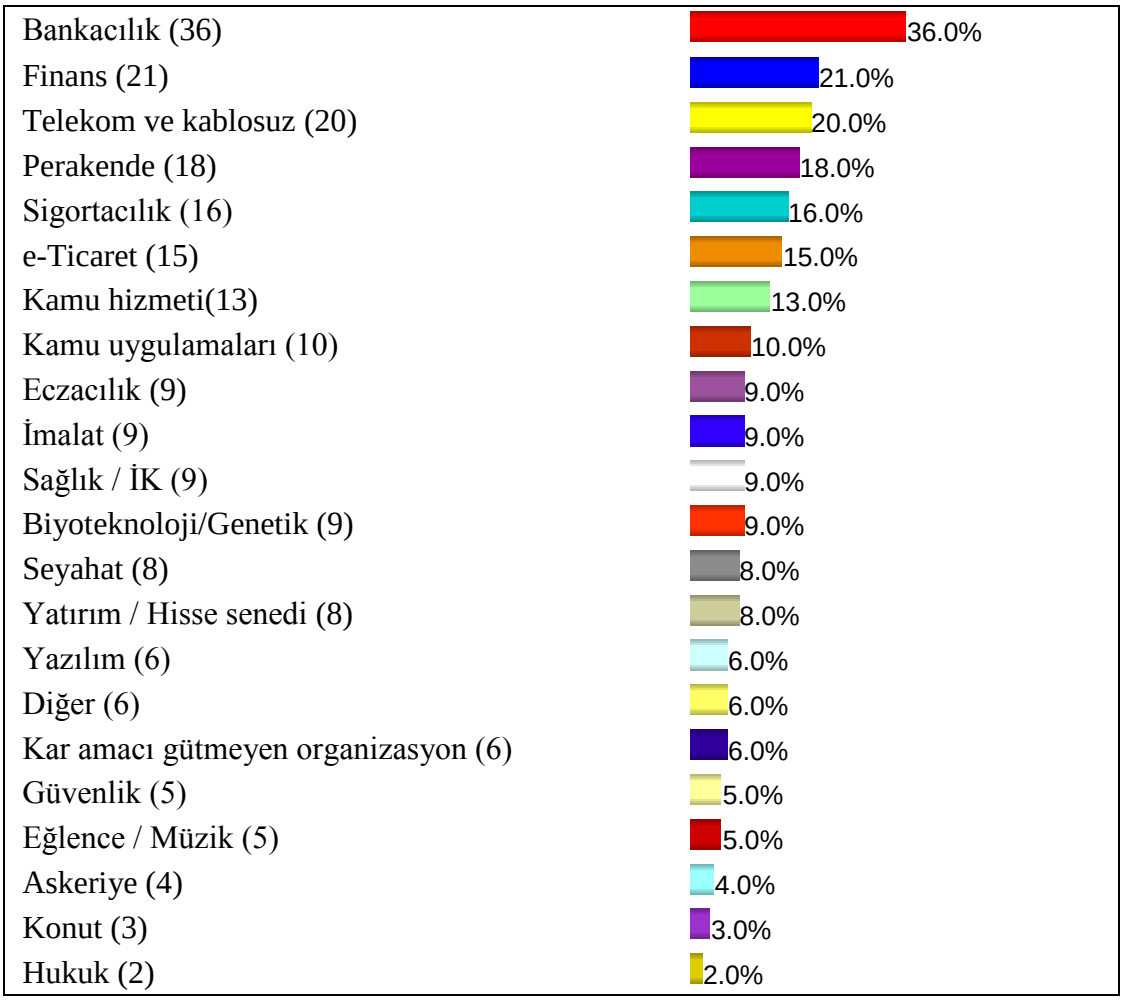
Veri madenciliği, astronomi, biyoloji, finans, pazarlama, bankacılık, sigorta, tıp ve daha bir çok alanda uygulanmaktadır. Son 20 yıldır Amerika Birleşik Devletleri'nde çeşitli VM algoritmalarının gizli dinlemeden, vergi kaçakçılıklarının ortaya çıkartılmasına kadar çeşitli uygulamalarda kullanıldığı bilinmektedir [Çetinyokuş, 2008].

Veri madenciliğinin kullanıldığı alanlardan bazıları şöyledir:

- Perakendecilik – Marketçilik
- Bankacılık
- Sigortacılık
- Taşımacılık / Ulaşım / Konaklama
- Eğitim Öğretim
- Finansal Servisler
- Elektronik Ticaret
- Bilimsel
- Telekomünikasyon
- Mühendislik
- Arama motorları
- Metin madenciliği
- Web sitesi analizleri
- Vergi kaçakçılarının profillerinin çıkartılması

Şekil 2.3.' de 2007-2008 yılında veri madenciliğinin sektörler bazında kullanımına ilişkin bir araştırmanın sonuçları yer almaktadır [KDnuggets**,** 2008]. Bu çizelgede

araştırmaya katılan şirketlerin %36' sı bankacılık alanında veri madenciliğini kullanmaktadır.



Şekil 2.3. Veri madenciliğinin uygulama alanları [KDnuggets, 2008]

- Pazarlama alanında müşteri gruplaması, kampanya ürünleri belirleme, satın alma örüntülerinin belirlenmesi, mevcut müşterileri kaybetmeden yeni müşteriler kazanma, firmaya yarar sağlayacak müşterilerin tespiti, pazar sepeti analizi, satış tahmini.

- Bankacılık ve sigortacılık alanında kredi kartı dolandırıcılıklarının tespiti, kredi taleplerinin değerlendirilmesi, kredi kartı harcamalarına göre müşteri profili

belirlenmesi, sigorta dolandırıcılıklarının tespiti, yeni poliçe talep edeceklerin belirlenmesi.

- Biyoloji, tıp ve genetik alanında gen haritasının çözümlenmesi, genetik hastalıkların ve kanserli hücrelerin tespiti, yeni virüs türlerinin keşfi ve sınıflandırılması.

- Kimya alanında yeni kimyasal moleküllerin keşfi ve sınıflandırılması, yem ve ilaç türlerinin keşfi.

- Yüzey çözümlemesi ve coğrafi bilgi sistemlerinde bölgelerin coğrafi özelliklerine göre sınıflandırılması, kentlerde yerleşim yerleri belirleme, kentlerde suç oranı, kentlere yerleştirilecek posta kutusu, otomatik para makineleri, otobüs durakları gibi hizmetlerin konumlarının tespiti.

- Metin madenciliğinde çok büyük ve anlamsız metin yığınları arasından anlamlı ilişkiler elde etme.

- Web verileri arasında düz metin ve resimden başka akan sayısal veriler de web verileri arasında yer almaktadırlar [Han ve Kamber, 2001]. Bu verilerin çözümlenmesi e-ticaret, web sayfalarının tasarımı ve düzenlenmesi gibi alanlarda VM kullanılmaktadır.

2.2. Veri Madenciliği Örnek Uygulamaları

Veri madenciliğiyle yapılabilecek uygulamalar şu şekilde sınıflandırılmıştır [Alpaydın, 1999]:

- Bağıntı: Müşterilerin beraber satın aldığı malların analizi yapılır. Örneğin "çocuk bezi alan müşterilerin %30'u bira da satın alır." Buradaki amaç ürünler arasındaki pozitif veya negatif ilişkileri bulmaktır.

- Sınıflandırma: Yeni bir nesnenin niteliklerini inceleme ve bu nesneyi önceden tanımlanmış bir sınıfa atamaktır. Burada önemli olan, her bir sınıfın özelliklerinin önceden net bir şekilde belirlenmiş olmasıdır. Örneğin, "eğer yıllık gelir 40.000 YTL' den küçük ve çalışma süresi 5 yıldan az ise kredi riski vardır". Buradaki amaç, kredi verme sürecinde doğru müşterileri bulmaktır.

- Regresyon: Bağımlı ve bağımsız değişkenler arasındaki ilişkinin çıkarımı söz konusudur. "Ev sahibi olan, evli, aynı iş yerinde beş yıldan fazladır çalışan, geçmiş kredilerinde geç ödemesi bir ayı geçmemiş bir erkeğin kredi skoru 825'dir." Başvuru derecelendirmede, bir finans kurumuna kredi için başvuran kişi için bir değer hesaplanır. Bu değer kişinin özellikleri ve geçmiş kredi hareketlerine dayanılarak hesaplanır.

- Zaman İçinde Sıralı İlişkiler: Kredi alan ve kredisinin taksitlerini ödeyen bir müşterinin sonraki taksitlerini ödeme veya geciktirme davranışını değerlendirmek örnek olarak verilebilir. "İlk üç taksitinden en az ikisini geç ödemiş olan müşteriler %60 olasılıkla kanuni takibe gidiyor." gibi sonuçlara ulaşılır.

- Benzer Zaman Sıraları: Zaman içindeki iki hareket serisi arasında bağıntı kurulur. Örneğin, iki farklı şirketin aktif büyüklüklerinin ya da iki farklı ürünün satış miktarlarının zaman içindeki değişimlerini göstermektedir.

- İstisnalar (Fark Saptanması): Buradaki amaç önceki uygulamaların aksine kural bulmak değil, kurala uymayan istisnai hareketleri bulmaktır. Normalden farklı davranış gösteren müşterilerin tespit edilmesi ile örneğin bankacılık sektöründe olası kredi dolandırıcılığının önüne geçilmesi sağlanabilir. Visa kredi kartı için yapılan CRIS sisteminde bir yapay sinir ağı kredi kartı hareketlerini takip ederek müşterinin normal davranışına uymayan hareketler için müşterinin bankası ile temasa geçerek müşteri onayı istenmesini sağlamaktadır.

- Doküman Madenciliği: Dokümanlar arasında ayrıca elle bir tasnif gerekmeden benzerlik hesaplayabilmektir. Doküman veritabanlarındaki büyük miktardaki metin verisinden bilgiyi kavramak, anlamak, yorumlamak ve otomatik olarak süzmek için pek çok disiplinden teknikler kullanır. Son zamanlarda metin madenciliğinden etkilenen bazı alanlar, dizi eşleştirme, metin arama, bilgiye erişme, ana dil işleme, istatistik, bilgi teorisi, hesaplama vb. alanlardır. Değişik metin analiz teknikleri ile birleşen internet arama motorları, çevrim içi doküman madenciliğini kolaylaştırmıştır.

2.3. Veri Madenciliğinin Uygulanabildiği Veri Türleri

Veri madenciliği özünde her tür bilgi kaynağında uygulanabilirdir. Bunlar ilişkisel veritabanları, veri ambarları, işlemsel veritabanları, gelişmiş veritabanı sistemleri, düz dosyalar ve World Wide Web'i içermektedir. Gelişmiş veritabanı sistemleri ise, nesne yönelimli ve nesne ilişkili ve uzaysal veritabanları, zaman serileri veritabanları, yazı veritabanları, çoklu medya veritabanları gibi özel uygulamalara yönelik veritabanlarını kapsamaktadır. Veri madenciliğinin yetenekleri ve teknikleri her bir kaynak sistem için değişebilmektedir[Han ve Kamber, 2001].

2.3.1. İlişkisel veritabanları

İlişkisel bir veritabanı, her birine eşsiz bir isim atanmış tablolar setidir. Her tablo, özelliklerin bir setinden ve büyük bir kayıtlar setinden meydana gelir. İlişkisel bir tablodaki her kayıt, eşsiz bir anahtar tarafından özdeşleştirilen ve bir özellikler seti tarafından tanımlanan bir nesneyi yansıtır. İlişkisel veritabanları için genellikle varlık-ilişki veri modeli gibi anlamsal bir veri modeli oluşturulur.

İlişkisel veritabanlarına, SQL gibi bir ilişkisel sorgulama dilinde yazılmış veritabanı sorguları ile veya grafiksel kullanıcı ara yüzü yardımı ile erişilebilir. Veri tabanında var olan desenler için sorgular çalıştırılırken, veri madenciliğindeki sorgular genelde keşfe dayalı ve ortada olmayan ilişkileri keşfetmeye dayalıdır.

Veri madenciliği sorgularına girdi sağlamak amacıyla veri tabanı kullanılmaktadır. Veri tabanındaki sorgu cümlecikleri VM' nin istediği örneklem kümesini elde etmek amacıyla kullanılmaktadır. Özellikle ilişkilendirme sorgusunda fazla miktarda veri tabanı sorgusu yapmak gerekmektedir.

Büyük miktarlarda verinin veri tabanlarında tutulduğu bilindiğine göre bu verilerin VM teknikleriyle işlenmesine de veri tabanında bilgi keşfi (VTBK) denir. Büyük hacimli olan ve genelde veri ambarlarında tutulan verilerin işlenmesi yeni kuşak araç ve tekniklerle mümkün olabilmektedir. Bundan dolayı bu konularda yapılan çalışmalar güncelliğini korumaktadır. Bazı kaynaklara göre; VTBK daha geniş bir disiplin olarak görülmektedir ve VM terimi sadece bilgi keşfi metotlarıyla uğraşan VTBK sürecinde yer alan bir adımdır [Fayyad ve ark., 1996].

2.3.2. Veri ambarları

Bir veri ambarı, birçok kaynaktan biriktirilen, birleşik bir şema altında depolanan bilgilerin deposudur. Veri ambarları, veri temizleme, veri dönüştürme, veri bütünleştirme, veri yükleme ve periyodik veri yenilemeden oluşan bir süreç yoluyla yapılandırılır.

Veri ambarı kavramı, karar vermede kullanılabilecek yapısal kaliteli bilgiye kolay erişimi sağlama ihtiyacından ortaya çıkmıştır. Karar vermeyi kolaylaştırmak için, bir veri ambarındaki veriler müşteri, ürün, tedarikçi, aktivite gibi temel konular çerçevesinde organize edilir. Veriler, tarihsel perspektifte bilgiler sağlamak için depolanır ve özetlenir.

İş organizasyonlarında bilgi akış mimarisinde veri ambarları iki amaçla oluşturulmaktadır [Kovalerchuk, 2000]:

1. Veri ambarı, hareketsel ve organizasyonel görevler arasındaki depo ve analitik stratejik verilerin birikimini sağlar. Bu veriler daha sonra yeniden kullanılmak

üzere arşivlenir. Veri ambarı, verilerin sorgulanabildiği ve analiz yapılabildiği bir depodur.

2. Veri ambarının pazarda yeni fırsatlar bulmaya, rekabete katkı, yoğun proje çevirimi, iş, envanter, ürün maliyetlerinin azalmasının yanında farklı işlere ait verilerin ilişkilendirilmesi, karar destek ve alınan bilgiye hızlı cevap verebilme gibi birçok katkısı bulunmaktadır.

Veri ambarının gelişmesi ile beraber, verilere daha hızlı şekilde erişme ve çok boyutlu analiz ihtiyaçları ortaya çıkmıştır. Çok boyutlu veri görünümlerine ve özetlenmiş verinin işlenmesine olanak tanıması sebebiyle veri ambarları, Çevrimiçi Analitik İşleme (On-line Analytical Processing, OLAP) için çok uygundur. OLAP çok boyutlu çevrede veri analizini destekleyen sorgu bazlı metottur. OLAP' ta veri, çok boyutlu bir uzay üzerinde tanımlanan ve bir çok boyutları olan küpler biçiminde gösterilir. Her bir boyut bir araya toplanmış bir kümeden oluşmaktadır. OLAP ile çok boyutlu veriler içerisinde derinlemesine farklı boyut analizlerinin yapılması sağlanmaktadır.

Veri madenciliği, sadece istatistiksel tekniklerin tartışıldığı alanlar değildir. Aynı zamanda VM amacıyla, veri ambarlamayı kapsayan çeşitli teknolojileri, teknikleri, çeşitli yazılım paketleri ve dilleri geliştirilmesi ile de ilgilidir. Geleneksel tekniklerin dışında, OLAP' ı içeren çok boyutlu yöntemleri de kapsar. OLAP veri analizlerini kolaylaştıran veri özetleme / bütünleştirme aracı iken veri madenciliği büyük veri topluluğu içinde saklı kalan ilginç verileri keşfeder.

2.3.3. İşlemsel veritabanları

Genel olarak işlemsel veritabanları, her kaydın bir işleme karşılık geldiği bir dosyadan oluşur. Bir işlem tipik olarak benzersiz bir işlem numarasını (trans_ID) ve işlemi oluşturan parçaların listesini içerir. İşlemsel veritabanları, satışlarla ilgili diğer bilgileri de içine alan ek tablolara sahip olabilir.

2.3.4. Gelişmiş veritabanı sistemleri

İlişkisel veritabanı sistemleri işletme uygulamalarında geniş bir yer tutmuştur. Veritabanı teknolojilerinin gelişimi ile birlikte, değişen çeşitlerde gelişmiş veritabanı sistemleri ortaya çıkmış ve yeni veritabanı uygulamalarına olan gereksinime cevap vermek için gelişime uğramıştır.

Bu yeni veritabanı uygulamaları, uzaysal verileri (haritalar gibi), mühendislik tasarım verileri (binaların tasarımı, sistem parçaları veya entegre elektrik devreleri gibi), hipermetin veya çoklu ortam verileri (yazılar, grafikler, video ve ses veriler gibi), zaman ile ilgili verileri (tarihsel kayıtlar veya borsa verileri gibi) ve World Wide Web' i işlemeyi içermektedir. Bu uygulamalar karmaşık nesne yapıları, değişken uzunluktaki kayıtlar, yarı-yapılandırılmış veya yapılandırılmamış veriler, yazı ve çoklu ortam verileri, karmaşık yapılı veritabanı şemaları ve dinamik değişiklikler ile işlem yapabilmek için etkin veri yapılarına ve ölçekli metotlara gerek duymaktadır.
Bu ihtiyaçlara cevap olarak, gelişmiş veritabanı sistemleri ve özel uygulama-yönelimli veri tabanı sistemleri geliştirildi. Bunlar, nesne-yönelimli ve nesne-ilişkili veritabanı sistemleri, uzaysal veritabanı sistemleri, geçici ve zaman serileri veritabanı sistemleri, yazı ve çoklu ortam veritabanı sistemleri, heterojen ve mirasçı veritabanı sistemleri ve web tabanlı evrensel bilgi sistemlerinden meydana gelir.

Böyle veritabanları veya bilgi ambarlarında bilginin etkin bir şekilde depolanması, bulunup işlenmesi, büyük miktarda karmaşık verinin güncelleştirilmesi için karmaşık araçlara ihtiyaç duyulurken, ayrıca bunlar veri madenciliği için verimli zeminler sağlar ve birçok araştırma ve uygulama konusu yetiştirir [Han ve Kamber, 2001].

2.4. Veri Madenciliği Uygulamalarının Artış Sebepleri

Veri madenciliğinin gün geçtikçe artan ilginin nedenleri şu şekilde açıklanabilir [Aktürk ve Korukoğlu, 2008]:

Veri hacmindeki artış

Verilerin sağlıklı bir ortamda saklanması istendiği zaman kolayca erişilebilmesi, sorgulama işlemlerinin insanlara göre daha hızlı yapılması sonucu iş ile ilgili olan tüm veriler artık disklerde saklanmaktadır. Bunun sonucunda ise veriler büyük bir ivme ile artış göstermektedir. Verilerin artması ile birlikte bir takım çıkarsamaların daha güvenilir, daha hızlı ve rekabetçi bir dünyaya ayak uydurması açısından veri madenciliğinin popülaritesi artmaktadır.

İnsanların analiz yeteneğinin kısıtlılığı

Verilerin hızlı bir şekilde işlenmesi bilgisayarlar aracılığı ile yapıldığında insanlara göre çok daha üstünlük sağlamaktadır. İnsanların verileri kendi zekalarını kullanarak analiz etmesinde her zaman objektif olamayışı, bir takım sonuçları bir araya getirip yeni çıkarımları ortaya koymada hızlı ve yeterli olamaması gibi pek çok nedenden ötürü insanlar verilerin analizinde bilgisayarlara göre çok geride kalmaktadır.

Makine öğreniminin düşük maliyetli oluşu

Bir verinin analizi için hem çok sayıda uzman gerekmektedir hem de işin hızlı bir şekilde yapılabilmesi kolay olmamaktadır. Bilgisayarların kullanılmasıyla birlikte işler hem çok daha hızlı hem de çok daha ucuz bir şekilde yapılabilmektedir. Burada insanlara duyulan ihtiyaç, bilgisayarların analizi sonucu ortaya çıkarmış olduğu bilginin yorumlanması aşamasındadır.

2.5. Veri Madenciliği Süreci

Veri madenciliği ile ilgili olarak farklı süreçlerden bahsedilebilmektedir. Ancak, VM uygulamalarını işletme faaliyetlerine uyarlayan kuruluşların oluşturduğu bir konsorsiyum tarafından geliştirilen "Çapraz Endüstri Veri Madenciliği Standart Süreci" yaygın olarak kabul görmektedir. Bilgi keşfi sürecide yer alan VM süreci ise bu yaklaşıma göre 6 adımdan oluşmaktadır [Springer, 2007]:

1. İşin anlaşılması,
2. Verinin anlaşılması,
3. Verinin hazırlanması,
4. Modelleme,
5. Değerlendirme,
6. Modelin ve sonuçların kullanımı.

2.5.1. İşin anlaşılması

Bu adım, amaçların ve gereksinimlerin anlaşılması üzerine odaklanmaktadır. VM çalışmalarında başarılı olmanın öncelikli şartı, uygulamanın amacının açık bir şekilde belirtilmesidir. İşin amacı net bir şekilde ortaya konulmalı, durum değerlendirmesi yapılmalı, VM amaçlarına karar verilmeli ve proje planları yapılmalıdır.

2.5.2. Verinin anlaşılması

Bu adım, başlangıç verisinin toplanması ve tanınmasıyla başlar. Ardından, veri hakkında daha fazla bilgi sahibi olmak için yapılan faaliyetler, veri kalitesiyle ilgili problemlerin belirlenmesi, veri hakkındaki ilk anlayışın ve şaşırtıcı veri altkümelerinin ortaya çıkarılması ile ilerler ve veri kalitesinin doğrulanması ile son bulur.

2.5.3. Verilerin hazırlanması (veri ön işleme)

Örneklem kümesi elde edildikten sonra, örneklem kümesinde yer alan hatalı kayıtların çıkarıldığı ve eksik nitelik değerlerinin değiştirildiği aşamadır. Bu aşama seçilen veri madenciliği sorgusunun çalışma zamanını iyileştirir. Veri madenciliğinin en önemli aşamalarından biri olan verinin hazırlanması aşaması, analistin toplam zaman ve enerjisinin %50 - %85' ini harcamasına neden olmaktadır [Piramuthu, 1998]. Modelin kurulması aşamasında ortaya çıkacak sorunlar, bu aşamaya sık sık geri dönülmesine ve verilerin yeniden düzenlenmesine neden olmaktadır.

Hatalı veya analizin yanlış yönlenmesine neden olabilecek veriler temizlenir. Veri farklı kaynaklardan toplanmışsa ve aralarında farklılıklar varsa gerekli dönüşümler yapılarak bu farklılıklar ortadan kaldırılır. Eksik verilerin bulunduğu kayıtlar proje için fazla enformasyon taşımıyor ise silinir ya da eksik veriler çeşitli yöntemler kullanılarak tahmin edilmeye çalışılır.

Veri Temizleme

Gerçek dünya verileri eksik, yanlış ve tutarsız olma eğilimindedir. Veri temizleme rutinleri verideki eksik değerleri doldurmaya, uç değerleri belirleyerek yanlış değerleri düzeltmeye ve tutarsızlıkları düzeltmeye çalışır.

Veri bütünleştirme

Bir veri analizi görevinde, farklı kaynaklardan gelen verilerin, tek bir veri ambarında birleştiren veri bütünleştirmeyi içermesi büyük olasılıktır. Bu kaynaklar bir çok veritabanı, veri küpleri veya düz dosyaları içerebilir. Bu bir çok kaynaktaki verilerin dikkatli bütünleştirilmesi, sonuç veri setinde gereksiz ve tutarsız verilerin azaltılmasına ve hatta sakınılmasına yardım edebilecektir. Bu da sonraki madencilik sürecinin hızını ve doğruluğunun gelişmesine yardım edebilir.

Veri dönüştürme

Veri dönüştürmede, veriler madencilik için uygun olan formlara dönüştürülür veya birleştirilir. Veri dönüştürme aşağıdakileri içerebilir:

Düzleştirme: Veriden hatalı uç değerlerin silinmesi (atılması) için çalışır.
Bütünleştirme: Özetleme veya bütünleştirme işlemlerinin veriye uygulanmasıdır.
Genelleştirme : Verilerin genelleştirilmesinde alt seviye veri veya ham veri, kavram hiyerarşilerinin kullanılmasıyla daha yüksek seviyelerle değiştirilir.
Normalizasyon: Bir özelliğe ait veri normalizasyonla küçük tanımlanmış bir aralığa düşecek şekilde ölçeklenir.

Alan Yapılandırma: Madencilik sürecine yardım etmek için verilen alanlar setinden yeni alanlar yapılandırılır ve eklenir.

Veri İndirgeme

Büyük miktardaki veri üzerindeki karmaşık veri analizi ve madenciliği, işlemleri uygulanamaz veya imkansız kılacak kadar çok uzun zaman alabilir.

Veri indirgeme teknikleri, hacimce daha küçük indirgenmiş veri setlerini elde etmek için uygulanır. Ama orijinal verinin bütünlüğü de korunmaktadır. Yani, indirgenmiş veri seti üzerindeki madencilik, aynı analitik sonucu üretecek kadar etkin olmalıdır.

Kesiklileştirme

Kesiklileştirme teknikleri, sürekli bir alan için, alanın değişken aralığını aralıklara bölerek değerlerin sayısını düşürmek için kullanılır. Aralık etiketleri daha sonra gerçek veri değerlerini yerleştirmek için kullanılır. Bir alan için değerlerin sayısının düşürülmesi, özellikle işlenmiş veriye sınıflama madenciliğinin karar ağacı tabanlı metotları uygulandığında yararlıdır. Bu metotlar genellikle, her adımda verinin sıralanmasına yüksek miktarda zaman harcanılan yinelemeli yapıdadırlar. Bu yüzden, sıralamak için az sayıda farklı değer olması, bu metotları daha hızlı yapmaktadır.

2.5.4. Modelleme

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 yenilenen bir süreçtir.

Bir veri madenciliği problemi için birden fazla teknik kullanılabilir, problem için uygun olan teknik veya tekniklerin bulunabilmesi için birçok teknik oluşturulup

bunların içinden en uygun olanlar seçilir. Model oluşturulduktan sonra kullanılan tekniğin gereksinimlerine uygun olarak veri hazırlanması aşamasına tekrar dönülüp gerekli değişikliklerin yapılması gerekebilmektedir.

Bir modelin doğruluğunun test edilmesinde pek çok farklı yöntem kullanılabilmektedir. Kullanılan en basit yöntemlerden birisi basit geçerlilik 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).

Değerlendirme aşamasında, daha önce oluşturulmuş olan model, uygulamaya koyulmadan önce son kez tüm yönleriyle değerlendirilir, kalitesi ve etkinliği ölçülür. Modelin ilk aşamada oluşturulan proje amacına ulaşmada etkin olup olmadığı ve problemin tüm yönleri için bir çözüm sağlayıp sağlamadığı karara bağlanır [Two Crows Corporation, 2005].

2.5.5. Modelin değerlendirilmesi

Modelin kurulup, geçerliliğine karar verildikten sonra, modelin iş amacına uygunluğu değerlendirilir. Sonuçlar elde edildikten sonra VM sorgularından ortaya çıkan sonuçların yorumlanma kesimidir. Burada geçerlilik, yenilik, yararlılık ve basitlik açılarından üretilen sonuçlar yorumlanır. Bu aşamanın sonunda ise, ulaşılan VM sonuçlarının kullanılıp kullanılmayacağına karar verilir.

2.5.6. Modelin ve sonuçların kullanımı

Veri madenciliği modeli kurulup geçerliliği kabul edildikten sonra sonuçlar kullanılır. İhtiyaçlara bağlı olarak bu adım, sonuç raporların oluşturulması sağlanır.

Zaman içerisinde koşullarda ve verilerde ortaya çıkan değişiklikler, kurulan modellerin sürekli olarak izlenmesini ve gerekiyorsa yeniden düzenlenmesini gerektirecektir

2.6. Veri Madenciliği İle İlgili Literatür Çalışmaları

Veri madenciliği pazarlama, finans, bankacılık, imalat, sağlık, müşteri ilişkileri yönetimi ve organizasyon öğrenmede sıklıkla uygulanmaktadır. Veri madenciliği ile ilgili literatür çalışmaları incelendiğinde insan kaynaklarına yönelik çok az çalışma yer almaktadır. Veri madenciliği ile personel seçimine yönelik literatürde sadece Chien ve Chen tarafından 2007 ve 2008 yıllarında yayınlanan makalelere rastlanılmıştır.

Chien ve Chen (2008), personel seçimi, personel karakteristikleri ve iş davranışları arasında iş performansı ve işten ayrılmaları içeren ilişki kuralları geliştirmek için veri madenciliği çerçevesi sunmayı amaçlamışlardır. Çalışmalarında, karar ağaçları ve birliktelik kuralları üzerine odaklanmış, veri madenciliği ile personel seçimindeki boşluğun doldurulması ve personel seçimi sürecinde fayda sağlanmasını amaçlamışlardır. Özellikle, personel seçimi kararı için karar ağacı analizi ile kurallar oluşturulmuştur. Pek çok personel verisinin kategorik veri olmasından dolayı sınıflandırma için CHAID karar ağacı oluşturmada kullanılmıştır. Sınıflandırma metodunun performansının değerlendirilmesinde ve yararlı kuralların elde edilmesinde lift (kaldıraç) kriter olarak kullanılmıştır. Çalışmalarını bir firmanın farklı iş fonksiyonlarına sahip mühendis ve yöneticileri içeren endirekt işcilerin işe alımı için gerçekleştirmişlerdir. Sonuçlar personellerin performansı ve işten ayrılmaları ile ilgili karar kuralları sağlamıştır [Chien ve Chen, 2008].

Chien ve Chen, 2007 yılında yayınladıkları makalede, yüksek yetenekli kişileri işe almanın ve elde tutmanın rekabet avantajı elde etmek adına yarı iletken şirketler için kritik bir süreç olduğunu belirtmişlerdir. Geleneksel personel seçim yöntembiliminin statik iş analizleri üzerine odaklandığını ve bunun ileri teknoloji şirketleri için

yeterince uygun olmadığını belirterek çalışmalarında kaba küme teorisi üzerine odaklanmışlardır. Veri madenciliği yaklaşımı ile yeni yeteneklerin performanslarının değerlendirilmesi ve elde tutulabilmesi amacıyla personel seçiminde faydalı ve etkili kurallar oluşturmaya çalışmışlardır. Tayvan' da yer alan bir yarı iletken firma için önerilen metodun geçerliliğini performans ve işten ayrılma davranışlarını içeren iş davranışları üzerinde test etmişlerdir. Kaba küme teorisi bu çalışmada anlaşılması kolay bir yöntem olduğu için kullanılmış ve elde edilen sonuçlar bu uygulamanın pratik sonuçlarını göstermiştir. Çalışma sonucunda, insan kaynakları yönetimine ilişkin kurallar oluşturulmuş ve iş stratejileri geliştirmiştir. Alternatif veri madenciliği yöntemlerinin de bu çalışmalarda uygulanabileceğini belirtilmiş, uygulanan yöntembiliminin operasyonel ya da yönetsel seviyedeki diğer işlere de uygulanabileceğini belirtmişleridir.

Veri madenciliği ile ilgili olarak son birkaç yıldaki sınıflandırma, kümeleme ve ilişki kuralları ile ilgili yapılan çalışmaların bir kısmı aşağıda özetlenmiştir:

Hsia ve arkadaşları (2008), çalışmalarında Tayvan'daki bir üniversitede, kurs tercihleri ve kurs tamamlama oranlarının analizi için veri madenciliği tekniğini kullanmışlardır. 2000-2005 yıllarına ait öğrenci kayıtları karar ağacı, bağlantı analizi ve karar ormanı olmak üzere üç veri madenciliği algoritmasıyla araştırılmıştır. Çalışmalarının amacı, öğrencilerin ders tercihlerinin ve eğitimlerine devam eden öğrencilerin ileriki dönemlerdeki ders tercihlerinin belirlenmesinde veri madenciliği tekniğinin kullanılmasıdır. Karar ağaçları, öğrencilerin kurs tercihlerini bulmada, bağlantı analizi kurs kategorisi ve katılımcı mesleği arasındaki korelasyonun belirlenmesinde, karar ormanı ise katılımcıların tercih ettikleri kursu tamamlama olasılıklarının bulunmasında kullanılmıştır.

Çalışmada karar ağacı olarak CHAID kullanılmıştır. CHAID için amaç değişkenlerin ve tahmin edilecek değişkenlerin tanımlanması gerektiği belirtilmiştir. Kurs kategorisi ve katılımcı mesleği tahmin değişkenleri olarak alınırken katılımcının statüsü amaç değişken olarak alınmıştır. Tercih edilen kursların bulunması için yapılandırılan karar ağacından sonra bağlantı analizi kurs kategorisi ve katılımcı

mesleği arasındaki ilişkinin bulunmasında kullanılmıştır. Son olarak, karar ormanı ile farklı sektörlerden katılımcıların tercih ettiği kurslar belirlenmiştir [Hsia ve ark., 2008].

Hsu (2008), çevrim-içi kişisel İngilizce öğrenimini destekleyecek bir sistem geliştirmiştir. Çalışmasında, hoşnutluk tabanlı analiz, işbirliği filtreleme ve veri madenciliği tekniklerini kullanarak öğrencilerin kendilerine uygun dersleri seçmelerine yardımcı olmayı amaçlamışlardır.

Önerilen sistemde iki veri madenciliği tekniğini kullanmışlardır: kümeleme ve ilişki kuralı. Önerilen İngilizce öğrenme sisteminde öncelikle öğrencileri farklı gruplara ayırmak ve her kümedeki öğrencilerin benzer çalışma davranışları göstermelerini sağlamak için kümeleme algoritması kullanılmıştır. Daha sonra ise, her gruptaki ders ilişkilerini analiz etmek için ilişki kuralı algoritması uygulanmıştır [Hsu, 2008].

Baykasoğlu ve Özbakır (2007), çalışmalarında kural oluşturma için çoklu ifade programlama (MEP) tabanlı yeni bir kromozom temsili ve çözüm tekniği olan birliktelik kuralı için çok ifadeli programlamayı (MEPAR) önermiştir. Yenilenmiş MEP algoritması olan MEPAR madenciliği C/C++ dilinde uygulanmış ve 9 uygun ikili ve n-li medikal veri kümesini sınıflandırmada test edilmiştir. Çalışma sonuçları PART, C4.5, Karar tablosu ve Basit Bayes algoritmaları ile karşılaştırılmış, sonuçlar tahminin kesinliği açısından değerlendirildiğinde dokuz veri kümesinin sekizinde MEPAR daha iyi sonuç vermiş, p-değeri açısından bakıldığında ise sadece bir veri kümesinde PART algoritması MEPAR' dan daha iyi sonuç vermiştir. Ayrıca etkili gen kodlama yapısının mantıksal EĞER-SONRA kurallarının tahmin doğruluğunu doğrudan arttırdığını göstermektedir [Baykasoğlu ve Özbakır, 2007].

Liao ve Wen (2007), çalışmalarında son 10 yılda yapay sinir ağları üzerine yapılmış ve anahtar kelimeleri birliktelik kuralı ve kümeleme olan 10120 makaleyi incelemiştir. 4 karar değişkeni olarak; anahtar sözcük, yazarın milleti, araştırma kategorisi, yayınlanma yılı alınmış ve 110800 veri incelenmiştir. Araştırma sonuçları, bazı özel yapay sinir ağı metodolojisi ve uygulamalarının veri

madeniliğinden çıkarıldığını göstermiştir. 110800 veri MS Access 2002 üzerindeki ilişkisel tablolar üzerinde oluşturulmuş ve MS SQL üzerine veriler transfer edilmiştir. İlişki kuralları ve kümeleme uygulanırken, SPSS Clementine veri madenciliği aracı olarak kullanılmıştır. İlişki kuralları bulmada Apriori, kümelemede ise K-ortalama algoritmaları kullanılmıştır [Liao ve Wen, 2007].

Fu ve arkadaşları (2007), iki farklı ülkedeki kadın ve erkekleri kültür, davranış ve sosyal bağlılık acısından araştırmayı, yaşam kalitelerini belirleyen faktörleri tahmin etmeyi amaçlamışlardır. 278 Avustralya'lı ve 398 Tayvan'lı kadın ve erkeğin yaşam kalitelerini belirlemede CART kullanılmıştır. Çalışmadaki 4 farklı bağımlı değişken olarak; fiziksel, psikolojik, sosyal ve çevresel sağlık çok boyutlu yaşam kalitesi için ölçülmüştür. Bağımsız değişkenler ise, kültür, davranış ve sosyal bağlılık ile sosyodemografik statüsü, dinsel ve ruhsal karakteristiklerdir. Sosyodemografik değişkenler yaş, medeni hal, eğitim düzeyi, mevcut çalışma durumu ve yıllık hane gelirleridir. "Yaş", bu çalışmada sürekli değişken olarak göz önüne alınırken diğer değişkenler (medeni hal, eğitim ve çalışma durumu) çoklu regresyon analizinde kukla olarak kullanılmıştır.

Çalışmanın sonucunda, CART algoritmasının parametrik veri ile veri dönüşümüne gerek kalmadan kullanılabildiği, CART' ın en büyük avantajlarından birinin bağımsız değişkenler arasındaki hiyerarşik ilişkileri ortaya çıkarması olduğu belirtilmiştir [Fu ve ark., 2007].

Plasse ve arkadaşları (2007), geniş dağınık veri seti içinde ikili özellikler arasındaki linkleri analiz edecek bir metot önermiştir. İlk olarak, değişkenler homojen özellik kümeleri sağlayacak şekilde kümelenmiştir. Daha sonra ilişki kuralları her kümeye uygulanmıştır. Önerilen metodoloji, 80000' den fazla araç ve her araçta 3000' den fazla özelliğin mevcut olduğu otomotiv endüstrisinde uygulanmıştır. Her bir özellik 0-1 ikili değerine sahiptir. Çalışılan veri üzerinde çok sayıda kümeleme metodu kullanılmış ve sonuçlar karşılaştırılmıştır. Çalışma, ilişki kuralları ile sınıflandırma metotlarının kombinasyonunun daha uygun olduğunu göstermiştir.

Bu çalışmada, ilişki kuralı olarak Apriori ve Eclat, üzerinde çalışılan veriye uygun en hızlı algoritmalar olduğu için kullanılmıştır. Hangi kümeleme metodu kullanılırsa kullanılsın, kaç tane küme kullanılacağına karar verilmesi gerektiği belirtilmiştir. Denenen tüm farklı sayıdaki kümelerde, her zaman yüksek oranda değişken içeren geniş bir küme olmuştur. 10'dan 100'e kadar farklı sayıdaki kümelerdeki değişken sayıları hesaplanmış, sonuçlar korelasyon katsayısı ve Russel-Rao katsayısı ile birlikte Ward stratejisi kullanılarak hesaplanmıştır. Ward stratejisinin benzerlik katsayısı ne olursa olsun en iyi sonuçları önerdiği sonucuna varılmıştır [Plasse ve ark., 2007].

Hsu ve Chen (2007), veri madenciliğinde karışık veriyi kullanabilen varyans ve entropi odaklı CAVE algoritmasını önermiştir. Varyans, sayısal verinin benzerliğini ölçmede kullanılmıştır. Kategorik verinin benzerliğini ifade etmede uzaklık hiyerarşisi önerilmiştir. Benzer şekilde, kategorik verinin benzerliği, hiyerarşideki uzaklık ağırlıklı entropi ile ölçülmüştür. Yeni bir doğruluk indeksi kümeleme sonuçlarının değerlendirilmesinde kullanılmıştır. CAVE algoritmasının etkinliği sentetik (yapay) ve gerçek veri kümeleri üzerinde test edilmiştir [Hsu ve Chen, 2007].

Seow ve Thomas (2007), banka ve diğer finansal kuruluşların rekabette karşılaştığı iki problem olan müşteri popülasyonun hangi gruplara ayrılacağı ile her grupta hangi teklifin götürüleceği üzerinde durmuşlardır. Çalışmalarında Enterprise Miner 4.3' e sahip SAS 9.1.3 istatistik paket programı kullanılmış ve TAROT uygulanmıştır. Amaç değişkeni olarak teklifin katılımcı tarafından kabul edilip edilmediği alınmıştır. Çalışmada, Southampton Üniversitesindeki öğrencilerin 2001 yılından sonraki hesaplarına ilişin 21 farklı karakteristik ele alınmıştır. Bu karakteristikler içinde, cinsiyet, medeni hal, çocuk sayısı, kredi kartı sayısı, alınan dersler, eğitim bilgileri, hobileri gibi 21 faklı özellik yer almıştır. Çalışma sonucunda TAROT sınıflandırma ağaçları ile her bir kümeye hangi teklifin yapılabileceğine karar verilmiştir. Sonuçlar, TAROT yaklaşımını puanlamadaki uygunluğunu göstermiştir [Seow ve Thomas, 2007].

Kirkos ve arkadaşları (2007), veri madenciliğinde sınıflandırma tekniklerini sahte finansal rapor düzenleyen firmaların belirlenmesinde ve bu faktörlerin tanımlanmasında uygulamıştır. Yapılan örnek, Yunanistan'daki 76 firmanın verilerini içermektedir. Girdi değişkenleri ve sınıflandırma çıktıları arasındaki ilişki modellerde ortaya konulmuştur. Bu çalışmada, karar ağaçları, sinir ağları ve Bayesian Belief Network (BBN) tekniklerinin kullanılırlığı araştırılmıştır. Bu 3 farklı sınıflandırma metodu test edilmiş ve tahmin netliği açısından karşılaştırılmıştır. Çalışmada uygulanan 3 farklı veri madeniliği tekniği tahmin netliği açısından karşılaştırılmıştır. Karar ağacı olarak ID3 uygulanmıştır. Karar ağacı modelinde Sipina Reserach Edition yazılımı kullanılmış ve model 0,05 güven seviyesinde yapılandırılmıştır. Çalışmanın ikinci aşamasında sinir ağları modeli kullanılmış ve Nuclass 7 Non Linear Networks for Classification yazılımı kullanılmıştır. Üçüncü deney aşamasında ise BBN uygulanmış ve yazılım olarak BN Power kullanılmıştır. Modellerin performansları karşılaştırıldığında en iyi performansı BBN metodunun gösterdiği, karar ağaçlarının performansının ise en alt seviyede kaldığı gözlemlenmiştir [Kirkos ve ark., 2007].

Abascal ve arkadaşları (2006), kümeleme ile ilgili çalışmaları adres göstermiş, pozitif geniş değişkenler kümesini tanımlamış, öncelikle kantitatif kriterleri kullanarak değişken değerlerini farklılaştırmış, ardından kalitatif kriterlerle değişkenlerin sıfır değerini alıp almadığına odaklanmıştır. Sıfır değeri, örneğin bir ürünün tüketilmediğini göstermektedir. Genellikle sıfır değerlerinin daha yüksek bir sıklığı mevuttur. Bu verinin analizinde 2 faklı yaklaşım önerilmiştir. Biri, çoklu faktör analizi (MFA), kalitatif ve kantitatif veriyi uzlaştırmaktadır. Diğeri ise fonksiyon ailesi önererek asıl veriyi çevirerek fonksiyonu indekslemek için parametrelerin kullanıldığı ve her bir kriter için ağırlıklı atamanın yapıldığı yaklaşımdır. Tüm prosedürler bir telekomünikasyon firmasının gerçek verisi üzerinde test edilmiş, geniş veri kümelerindeki müşterilerin gruplanması, negatif olmayan tamsayı değişkenleriyle tanımlanması ve önceden tanımlanmamış homojen sınıflar içine yerleştirilmesi yapılmıştır. Yapılan gerçek hayat çalışması için çok değişkenli normal dağılım varsayımı altında 1000 müşteri için 5 farklı tüketim değişkeni alınmıştır [Abascal ve ark., 2006].

He ve arkadaşları (2006), kategorik veriler için yeni ve etkili bir algoritma olan k-ANMI algoritmasını önermişlerdir. Veri kümesindeki sayısal veri için Liu ve arkadaşları (2002) tarafından kullanılan tekniği kullanılmış ve sayısal veri kategorik sınıf etiketine çevrilmiştir. Deney çalışmalarında kümeleme için k-ANMI algoritması, Squeezer algoritması (Z. He ve ark., 2002), GAClust algoritması (Cristofor ve ark., 2002), standart k-mod algoritması (Huang, 1998) ve dByEnsemble (He ve ark., 2005) algoritması olmak üzere 5 farklı algoritma üzerinde çalışılmıştır.

k-ANMI algoritması diğer algoritmalara göre bazı özel avantajlar sağlamıştır. Öncelikle, önerilen algoritma hem kategorik veri kümeleme hem de küme topluluğu için uygundur. İkinci olarak, kategorik veri kümelemeye kolaylıkla yayılabilir. Son olarak, nümerik ve kategorik veri içeren heterojen veriye uygulanabilir [He ve ark., 2006].

Ben-David ve Sterling (2006) tarafından dört farklı veri tabanı için değiştirilmiş en yakın komşu algoritması uygulanmış, daha sonra ise CART ve sinir ağları algoritmaları aynı veri setlerine uygulanarak ve sonuçları karşılaştırılmıştır. En yakın komşu algoritması Matlab ile yazılmışken, CART ve sinir ağları için SPSSs Clementine'nin 7.2 versiyonu kullanılmıştır. Dört veri kümesi için bu üç algoritma test edilmiş ancak hiçbirinde CART veya sinir ağları daha iyi performans gösterememiştir. Ortalama mutlak hatalar dört farklı veri seti için araştırılmış, en iyi sonucu veren en yakın komşu algoritması olmuştur. Sonuçlar, en yakın komşu algoritmasının çok az prototip veya küme ile kesin tahminler yapabildiğini göstermiştir. Çalışmalarından çıkarttıkları ana sonuç, sınıflandırma tekniklerini veri tabanlarında uygularken, karar ağaçlarına veya kurallarına 10' dan fazla dal veya kural için izin vermenin gereksiz olduğudur [Ben-David ve Sterling, 2006].

Questier ve arkadaşları (2005), denetimli ve denetimsiz özellik seçimi için CART ve çok değişkenli regresyon ağacını (MRT) tanımlamıştır. CART metodu denetimli özelliklerin birden çok açıklayıcı değişken x ve bir yanıt değişkeni y ile modellenmesine izin vermektedir. MRT ise CART' tan türetilmiş ve birden yok yanıt

değişkeni y ile işlem yapabilmektedir. Bu da, denetimli özellik seçimine birden çok yanıt değişkeni için izin vermektedir. Hiç bir yanıt değişkeninin uygun olmadığı denetimsiz özellik secimi için, otomatik birleşmeli çok değişkenli regresyon ağacını (AAMRT) önerilmiş, buradaki X sadece açıklayıcı değişken değil aynı zamanda yanıt değişkenidir (X=Y). (AA)MRT açıklayıcı değişkenleri kullanarak benzer yanıt değişkenlerini grupladığından verideki küme yapısı için en sorumlu değişkenleri bulur.

Yapay ve gerçek veri kümelerindeki uygulamalar önerilen metodun özellik seçimi için etkin bir şekilde kullanılabileceğini göstermektedir. Özellik sayısı indirgenirken, en önemli küme yapısı sunulmaktadır. Metot, aynı zamanda küme yapısını gereksiz ve ilişkisiz özellikleri çıkararak geliştirmektedir [Questier ve ark., 2005].

Cho ve Ngai (2003), veri ambarı karakteristiklerine ve veri madenciliği tekniklerinin kullanılması ile uygun sigorta acentesi seçimi üzerine odaklanmıştır. Sigorta ajanslarının servis süresi, satış primi ve sürdürülebilirlik indislerinin tahminlerini de içeren veri ambarıyla bütünleşik üç popüler veri madenciliği tekniği olan diskriminant analizi, karar ağaçları ve yapay sinir ağları üzerine odaklanmıştır. Bu çalışmada, sigorta yöneticilerinin kaliteli ajansları veri ambarı çerçevesinde veri madenciliği kullanılarak karar destek sistemi sunulmaktadır. Veri madenciliği teknikleri arasında sınıflandırma ve tahminde en kolay yolun karar ağaçları olduğu belirtilmiştir. Diğer iki yöntemin aksine karar ağacı analizindeki sonuçların yorumlanabilir olduğu belirtilmiştir. Doğrusal diskriminant analizinin faydası hesaplama kolaylığıdır. Diskriminant analizi satış primi tahmininde daha uygun bulunmuştur. Yapay sinir ağları algoritmasının hesaplama süresi diğer iki algoritmaya göre daha uzun olmasına rağmen üç amacın ikisinde daha göze çarpan tahmin edilebilirlik sağlamıştır.

Guha ve arkadaşları (2000), ikili ve kategorik özellikler üzerinde çalışmıştır. Noktalar arasındaki uzaklığı kullanan kümeleme algoritmasının ikili ve kategorik veri için uygun olmadığını bu çalışmada göstermişlerdir. Bunun yerine noktalar arasındaki link kavramını veri noktaları arasındaki benzerliği ölçmede önermişler,

güçlü bir hiyerarşik kümeleme algoritması olan ROCK algoritmasını sunmuşlardır. Önerilen metot sayısal olmayan benzerlik ölçüsü sunmaktadır. Yapılan deney çalışmaları ile kategorik veride ROCK algoritmasının sadece daha iyi kümeler oluşturduğu değil aynı zamanda iyi ölçeklendirilebilir özellikler sunduğu görülmüştür. Örneğin, mantar veri setinde, sadece yenilebilir ve sadece zehirli mantarları içeren kümeler oluşturulmuştur. Dahası, bulunan kümelerde ciddi oranda farklılıklar yer almaktadır. Bunun aksine, geleneksel merkez tabanlı hiyerarşik algoritmanın bulduğu kümelerin kalitesi ise oldukça düşüktür. Sadece normal büyüklükteki kümeler oluşturmakla kalmamış aynı zamanda da zehirli ve yenilebilir mantarlar aynı kümelerde yer almıştır [Guha ve ark., 2000].

3. VERİ MADENCİLİĞİ MODEL VE TEKNİKLERİ

Veri madenciliğinde uygulamalarındaki yüksek derecede öncelikli iki amaç tahmin ve tanımladır. Tahmin, veri tabanındaki bazı değişkenleri veya alanları kullanarak bilinmeyen ya da ileriki dönemlere ilişkin tahminlerinin yapılmasını içerir. Tanımlama ise, veriyi tanımlayan ve insanların verileri değerlendirebileceği desenleri bulmaları üzerine odaklanır [Fayyad ve ark., 1996].

Veri madenciliğinde tahmin edici modellerde, mevcut verilerden hareket edilerek bir model geliştirilir ve sonuçları bilinmeyen veri kümeleri için sonuç değerlerin tahmin edilmesi amaçlanır.

Tanımlayıcı modellerde ise karar vermeye rehberlik etmede kullanılabilecek mevcut verilerdeki desenlerin tanımlanması sağlanmaktadır. Tanımlayıcı modellerde amaç, büyük veri kümelerindeki desen ve ilişkileri tespit ederek, incelenen sistemin anlamını kavramaktır [Kantardzic, 2002].

Veri madenciliği modellerini gördükleri işlevlere göre:

1- Sınıflama ve Regresyon
2- Kümeleme
3- Birliktelik Kuralları

olmak üzere üç ana başlık altında incelemek mümkündür. Sınıflama ve regresyon modelleri tahmin edici modeller iken kümeleme ve birliktelik kuralları tanımlayıcı modellerdir [Akpınar, 2000] .

3.1. Sınıflama ve Regresyon

Sınıflama kategorik değerleri tahmin ederken, regresyon süreklilik gösteren değerlerin tahmin edilmesinde kullanılır. Sınıflama ve Regresyon modelleri arasındaki temel fark, tahmin edilen bağımlı değişkenin kategorik veya süreklilik

gösteren bir değere sahip olmasıdır. Ancak bazı tekniklerde her iki model giderek birbirine yaklaşmakta ve bunun bir sonucu olarak aynı tekniklerden yararlanılması mümkün olmaktadır [Kalıkov, 2006].

Bu bölümde şu sınıflandırma algoritmalarından bahsedilecektir:

- Karar ağaçları,
- Doğrusal ve çoklu regresyon modelleri,
- Yapay sinir ağları (YSA),
- Saf Bayes sınıflayıcısı,
- K-en yakın komşu algoritması,
- Genetik algoritmalar.

3.1.1. Karar ağaçları ve karar ağacı algoritmaları

Karar ağacı analizi, genellikle seçenekler üzerinde yapılan bir analiz türüdür. Bu analizin veri madenciliğinde kullanılma sebepleri ise şöyledir [Chu, 2005]:

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

Karar ağaçları kolaylıkla sınıflama kurallarına dönüştürülebilmektedir. Bunun için algoritmaya girdi olarak verilerin belirlenen belli nitelikleri, çıktı olarak da verilerin belli bir niteliği verilir ve algoritma bu çıktı niteliğindeki değerlere ulaşmak için hangi girdi nitelik değerlerinin olması gerektiğini ağaç veri yapıları kullanarak keşfeder.

Karar ağaçları genellikle yaprakları ve gövdesi ile ağaç yapısında sunulmaktadır. Gövdeler özelliklerin koşullarını gösterirken yapraklar sınıflandırma sonuçlarını

ortaya koyar. Özellikle karar ağaçlarına yönelik pek çok algoritma ; CART (Breiman ve ark., 1984), CHAID (Kass, 1980), ID3 (Quinlan,1986), C4.5 (Quinlan, 1993) yer almaktadır [Chien ve Chen, 2008].

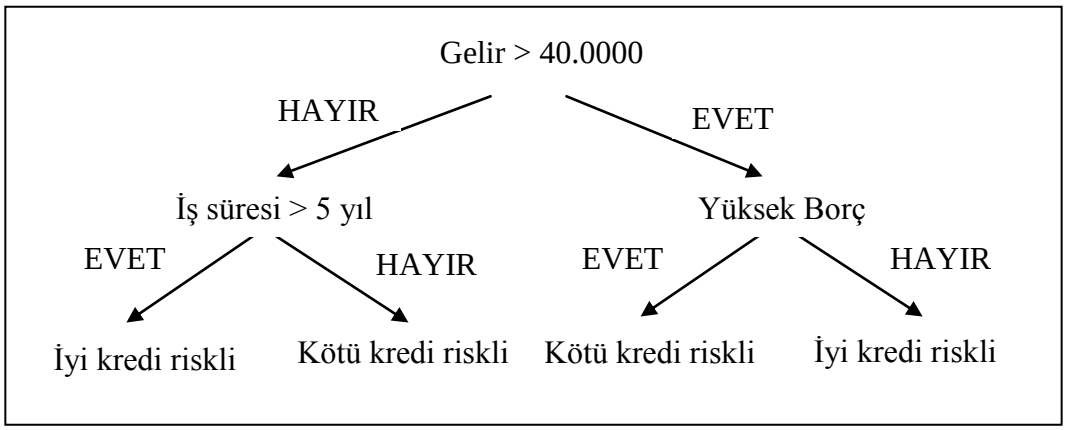
Sınıflandırma ağaçları, Breiman ve arkadaşları tarafından 1984 yılında önerilmiştir. Sınıflandırmada, bir veri seti mevcuttur ve her veri noktası amaç değişkeni ile birlikte karakteristik değerlerden oluşur ve genellikle ikilidir . Amaç, farklı amaç çıktılarına göre özelliklerin kombinasyonunu oluşturmaktır. Sınıflandırma, kredi derecelendirme ve pazarlama dışında pek çok alanda da kullanılmaktadır. Örneğin sağlık alanında Harper ve arkadaşları (2003), CART algoritmasını şeker hastalarının tedavilerine karar vermede kullanmışlardır [Seow ve Thomas, 2007]. Sınıflandırma ağaçlarının gücü, sınıflandırılacak karakteristiklerin etkileşimini incelemesidir. Sınıflandırma ağaçları, verideki en önemli karakteristikleri tanımlar ve amaç değişkenine ulaşmak için en iyi tahmini sağlayan özelliklerin kombinasyonunu belirler. Ağacı yapılandırmak için, öncelikle popülasyon birbirinden mümkün olduğu kadar farklı 2 alt popülasyona ayrılır. Bunu yaparken, her bir karakteristiğe bakılır ve amaç değişkenin çıktılarının farklılaştıracak özelliklerin farklılaştırılması ile optimumu sağlayacak ayırım tanımlanır. Bu süreç, yavru popülasyonlara kadar tekrarlanır [Seow ve Thomas, 2007].

Karar ağaçları sınıflandırma ve tahmin için sıklıkla kullanılan bir veri madenciliği yaklaşımıdır. Sorunla ilgili araştırma alanını alt gruplara ayırmak için kullanılır. Karar ağaçlarında kök ve her düğüm bir soruyla etiketlenir. Düğümlerden ayrılan dallar ise ilgili sorunun olası yanıtlarını belirtir. Her dal düğümü de söz konusu sorunun çözümüne yönelik bir tahmini temsil eder [Altıntaş, 2006]. Sinir ağları gibi diğer metodolojilerin sınıflandırma için kullanılmasına rağmen karar vericiler için kolay tefsir ve anlaşılması karar ağaçlarının avantajlarıdır [Chien ve Chen, 2008].

Karar ağaçlarının en önemli avantajlarından biri EĞER-İSE yapısını kullanması ve bilgi kazanımı sunmada anlaşılabilir olmasıdır [Kirkos ve ark., 2007].

Karar ağaçları, bir sınıf ya da değer oluşturan bir dizi kuralı gösterme yöntemidir. Örneğin, borç uygulamalarını iyi ya da kötü kredi riskine göre sınıflandırmak isteyebilirsiniz.

Şekil 3.1.' de bu problemi çözen basit bir karar ağacı örneği gösterilmektedir, aynı zamanda bu şekil bir karar ağacının tüm basit bileşenlerini de göstermektedir [Two Crows Corporation, 2005].



Şekil 3.1. Karar ağacı örneği [Two Crows Corporation, 2005]

Burada;

EĞER Gelir 40.000 den küçük ve iş süresi 5 yıldan büyük İSE İyi kredi riskli,
EĞER Gelir 40.000 den küçük ve iş süresi 5 yıldan küçük İSE Kötü kredi riskli,
EĞER Gelir 40.000 den büyük ve yüksek borçlu İSE Kötü kredi riskli,
EĞER Gelir 40.000 den büyük ve yüksek borçlu değil İSE İyi kredi risklidir.

Tanımlamalar

Bilgi kazancı ölçütü ağaçtaki her bir düğümde test alanını seçmek için kullanılır. Bu tür bir ölçüt alan seçim ölçütü olarak anılır. En yüksek bilgi kazancı değerine sahip alan ele alınan düğüm için test alanı olarak seçilir. Bu alan sonuç ayrımlarındaki önekleri sınıflamak için gerekli olan bilgiyi en aza indirir ve bu ayrımlarda en az rastsallığı yansıtmaktadır. Böyle bir teorik bilgi yaklaşımı bir nesneyi sınıflamada

ihtiyaç duyulan beklenen test sayısını en küçükler ve basit bir ağacın bulunacağını garantiler.

S, s adet veri örneğini barındıran bir küme olsun. Sınıf etiketi alanının m adet farklı $C_i (i = 1,2,...,m)$ sınıfı tanımlayan m farklı değere sahip olduğunu düşünelim. s_i, C_i sınıfında S' nin örneklerinin sayısı olsun. Verilen örneği sınıflamak için ihtiyaç duyulacak beklenen bilgi Eş. 3.1' de verilmiştir.

$$\textit{Beklenen bilgi}: \; I(s_1, s_2, ..., s_m) = -\sum_{i=1}^{m} p_i \log_2(p_i) \tag{3.1}$$

Burada p_i, keyfi bir örneğin C_i sınıfına ait olması olasılığıdır ve s_i / s ile tahmin edilir.

A alanı v farklı değere sahip olsun $\{a_1, a_2, ..., a_v\}$. A alanı, S' yi v alt sete $\{S_1, S_2, ..., S_v\}$ ayrıştırmada kullanılabilir. Burada S_j, A' nın a_j değerine sahip S' deki örneklerini içermektedir. Eğer A test alanı olarak seçilirse, bu alt setler S setini barındıran düğümden gelişecek dallara karşılık gelecektir. s_{ij}, bir S_j alt setinde C_i sınıfındaki örneklerin sayısı olsun. Entropi**,** ya da A' ya göre alt kümelerine ayrıştırılmasına dayanan beklenen bilgi Eş. 3.2 ' deki gibi hesaplanır:

$$\textit{Entropi}: \; E(A) = \sum_{j=1}^{v} \frac{s_{1j} + ... + s_{mj}}{s} I(s_{1j}, ..., s_{mj}) \tag{3.2}$$

Burada $\frac{s_{1j} + ... + s_{mj}}{s}$ terimi, J alt setinin ağırlığı olarak rol oynar ve alt setteki örnek sayısının, S' deki toplam örnek sayısına bölümüdür. Daha küçük entropi değeri, alt set ayrımlarının saflığının daha büyük olması demektir. S_j alt seti için Eş. 3.3 geçerlidir.

$$I(s_{1j}, s_{2j}, ..., s_{mj}) = -\sum_{i=1}^{m} p_{ij} \log_2(p_{ij}) \quad (3.3)$$

Burada $p_{ij} = \frac{s_{ij}}{|S_j|}$' dir ve S_j'deki bir örneğin C_i sınıfına ait olma olasılığıdır.

A' dan dallanmakla elde edilecek kodlanmış bilgi (kazanç) Eş. 3.4' de gösterilmiştir:

$$\text{Kazanç}(A) = I(s_1, s_2, ..., s_m) - E(A) \quad (3.4)$$

Diğer bir deyişle Kazanç(A), A alanının değerini bilmekten kaynaklanan entropideki beklenen azalmadır.

Algoritma her bir alanın bilgi kazancını hesaplar. En yüksek bilgi kazançlı alan verilen S seti için test alanı olarak seçilir. Bir düğüm yaratılır ve bu alanla etiketlenir. Alanın her bir değeri için dallar yaratılır ve buna göre örnekler ayrıştırılırlar.

Veri madenciliği uygulamalarında yaygın olarak kullanılan karar ağacı algoritmaları ise şöyledir:

- CHAID (Chi-Square Automatic Interaction Detector, Kass, 1980),
- C&RT (Classification and Regression Trees, Breiman ve ark., 1984),
- ID3 (Induction of Decision Trees, Quinlan, 1986),
- C4.5 (Quinlan, 1993).

CHAID

Kass tarafından 1980' de geliştirilen CHAID, özellikle kategorik verilerin analizi için tasarlanmış, ikili olmayan bir karar ağacı tekniğidir [Chien ve Chen, 2008]. CHAID, bölünme kriteri olarak entropi ya da Gini endeksi kullanmak yerine değerlerin tahmininde hangi kategorik tahmin edicinin bağımsızlıktan en uzak olduğunu tanımlamak için Ki-Kare testini kullanmaktadır

C&RT

Breiman ve arkadaşları tarafından 1984 yılında geliştirilen çok sayıdaki açıklayıcı (x) değişkeni ile yanıt (y) değişkenine karar vermede kullanılan istatistiksel bir tekniktir. Kesikli ve sürekli veriler üzerinde çalışabilen her dallanmada iki yeni düğüm oluşturan ikili bir karar ağacıdır ve bölünme kriteri olarak Gini endeksini kullanır [Questier ve ark., 2005]. C&RT kullanılarak kesikli ve sürekli veri tipleri üzerinde regresyon ağaçları oluşturulabilir.

ID3 ve C4.5 Algoritmaları

Karar ağaçları olarak da adlandırılan ID3 ve C4.5 algoritmaları, sınıflandırma modellerini işlemek için Quinlan (1993) tarafından geliştirilmiştir. ID3 yönteminde bölünme kriteri bilgi kazancı değeridir. Buradaki kazanç, bölünme öncesinde ve sonrasında doğru tahmin yapabilmek için ihtiyaç duyulan bilgi miktarındaki farkı anlatmaktadır.

C4.5, ID3'ün geliştirilmiş halidir. C4.5 eksik ve sürekli nitelik değerlerini ele alabilmekte, karar ağacının budanması ve kural çıkarımı gibi işlemleri yapabilmektedir. Karar ağacının kurulması için kullanılacak girdi olarak bir dizi kayıt verilirse bu kayıtlardan her biri aynı yapıda olan birtakım nitelik/değer çiftlerinden oluşur. Bu niteliklerden biri kaydın hedefini belirtir. Problem, hedef-olmayan nitelikler kullanılarak hedef nitelik değerini doğru kestiren bir karar ağacı belirlemektir. Hedef nitelik çoğunlukla ikili değerler alır [Aydoğan, 2003].

Karar ağacı algoritmalarına ilişkin karşılaştırmaya Çizelge 3.1.' de yer verilmiştir.

Çizelge 3.1. CART, CHAID, ID3 ve C4.5 karşılaştırması [Chien ve Chen, 2008]

Algoritma	**Yazar**	**Veri tipi**	**Ağaç budama metodu**	**Her bir düğümdeki dal sayısı**	**Kayıp değer metodu**	**Bölünme kriteri**
CHAID	Kass (1980)	Kesikli	Budama yok	İki veya daha fazla	Kayıp değer dallanması	Ki-Kare testi için P değeri
C&RT	Breiman ve ark. (1984)	Kesikli ve sürekli	Tüm hata oranı	İki	Sıralı/yerine geçen bölünme (alternate /surrogate)	Gini değeri, entropi
ID3	Quinlan (1986)	Kesikli	Budama yok	İki veya daha fazla	Elde edilemeyen	Bilgi kazancı
C4.5	Quinlan (1986)	Kesikli ve sürekli	Tahmini hata oranı	İki veya daha fazla	Olasılıklı ağırlık	Kazanç oranı

3.1.2. Doğrusal ve çoklu regresyon

Regresyon, değerleri bilinen değişkenleri kullanarak diğer değişkenleri tahmin etmek için kullanılır [Two Crows Corporation, 2005]. Regresyon terminolojisinde, tahmin edilecek olan değişken "bağımlı değişken", bağımlı değişkeni tahmin etmek için kullanılan değişken ya da değişkenler ise "bağımsız değişken" olarak adlandırılır.

Doğrusal regresyonda, veri düz bir çizgi kullanılarak modellenir. Doğrusal regresyon, regresyonun en basit halidir. İki değişkenli doğrusal regresyon, rastgele değişken Y'yi bir başka rastgele değişken X'in bir doğrusal fonksiyonu olarak Eş. 3.5' deki gibi modeller.

$$Y = \alpha + \beta X \tag{3.5}$$

Burada Y'nin varyansının sabit olduğu varsayılır ve α ve β sırasıyla doğrunun eksenini kestiği noktayı ve doğrunun eğimini tanımlayan regresyon katsayılarıdır. Bu katsayılar, gerçek veri ve doğrunun tahmini arasındaki hatayı en azaltan en küçük kareler metodu ile çözülebilir. Böylece Eş. 3.6 ve Eş. 3.7 elde edilir:

$$\beta = \frac{\sum_{i=1}^{s}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{s}(x_i - \bar{x})^2} \tag{3.6}$$

$$\alpha = \bar{y} - \beta\,\bar{x} \tag{3.7}$$

Burada $\bar{x}$ $x_1, x_2, ..., x_s$' lerin ortalaması iken, $\bar{y}$ $y_1, y_2, ..., y_s$ 'lerin ortalamasıdır.

Çoklu regresyon, doğrusal regresyonun birden fazla tahminci değişken içeren halidir. Y değişkeninin, çok boyutlu bir özellik vektörünün doğrusal bir fonksiyonu olarak modellenmesine olanak tanır. X_1 *ve* X_2 gibi iki tahminci değişkeni temel alan çoklu regresyon modelinin bir örneği de Eş. 3.8' deki gibidir:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 \tag{3.8}$$

Doğrusal olmayan regresyon

Polinom regresyon, polinom terimleri temel doğrusal modele ekleyerek modellenebilir. Değişkenlere dönüşüm uygulanarak bu doğrusal olmayan modeller, en küçük kareler tekniği ile çözülebilecek doğrusal modellere dönüştürülebilir.

Diğer regresyon modeller

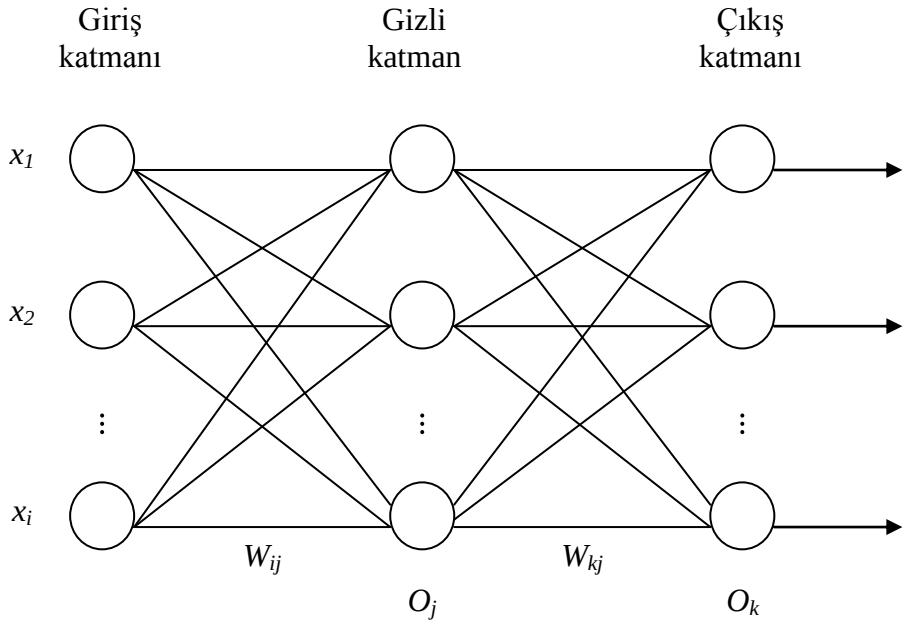
Doğrusal regresyon sürekli değerli fonksiyonları modellemekte de kullanılır. Genelleştirilmiş doğrusal modeller, doğrusal regresyonun kategorik değişkenlerin modellemesinde uygulanabileceğinin teorik esaslarını sunmaktadır. Genelleştirilmiş doğrusal modellerde, Y değişkeninin varyansı, doğrusal regresyondaki sabit değerin tersine Y' nin ortalamasının bir fonksiyonudur. Genelleştirilmiş doğrusal modellerin en bilinen türleri, Lojistik Regresyon ve Poisson Regresyon'dur. Lojistik regresyon tahminci değişkenler setinin bir doğrusal fonksiyonu olarak bazı olayların gerçekleşme olasılıklarını modeller. Sayımlı veriler genellikle poisson dağılım sergiler ve poisson regresyon kullanılarak modellenir.

Lojistik doğrusal modeller yaklaşık olarak, kesikli çok boyutlu olasılık dağılımlarını takip eder. Veri küpü hücreleri ile ilişkili olasılık değerlerinin tahmininde kullanılabilirler.

3.1.3. Yapay sinir ağları

Yapay Sinir Ağları, insanlığın doğayı araştırma ve taklit etme çabalarının en son ürünlerinden bir tanesi olan teknolojidir. 1980'lerden itibaren yaygınlaşan ve Yapay Sinir Ağları adı verilen programlar, basit biyolojik sinir sisteminin çalışma şeklini canlandırmak için tasarlanmışlardır [Yılmaz, 2002].

Bir yapay sinirin öğrenme yeteneği, kullanılan ağırlık oranıyla doğrudan ilişkilidir. Süreçte kullanılan girdiler, dışarıdan elde edilen bilgilerdir.Toplama fonksiyonu bir hücreye gelen net girdi miktarı olarak tanımlanabilir. Aktivasyon fonksiyonu, bu fonksiyon öğrenilme sonucu oluşan değerlerin ortaya çıkarılması için kullanılan bir fonksiyondur. Son olarak çıktı ise, aktivasyon fonksiyonundan elde edilen değer olarak tanımlanabilir [Chu, 2005]. Yapay sinir ağaçlarının katman olarak işleyişi Şekil 3.2.' de gösterilmektedir.



Şekil 3.2. Yapay sinir ağlarının katmanları

3.1.4. Saf Bayes sınıflaması

Bayes sınıflayıcıları istatistiksel sınıflayıcılardır ve bir örneğin belli bir sınıfa ait olma olasılığı gibi sınıf üyelik olasılıklarını tahmin edebilirler. Bayes sınıflaması, bayes teoremine dayanmaktadır.

Saf bayes algoritması sürekli veri ile çalışmadığından değişkenler kategorik hale getirilir. Saf bayes sınıflayıcıları, belli bir sınıf için alan değerlerinin etkisinin diğer alanların değerlerinden bağımsız olduklarını varsayar. Bu varsayım sınıfların şartlı bağımsızlığı olarak adlandırılır. Bu varsayım gereken işlemleri basitleştirmek için yapılmıştır ve bu mantıkla "saf" olarak değerlendirilir.

Saf bayes, modelin öğrenilmesi esnasında , her çıktının öğrenme kümesinde kaç kere meydana geldiğini hesaplar. Bulunan bu değer, öncelikli olasılık olarak adlandırılır. Saf Bayes aynı zamanda her bağımsız değişken / bağımlı değişken kombinasyonunun meydana gelme sıklığını bulur. Bu sıklıklar öncelikli olasılıklarla birleştirilmek suretiyle tahminde kullanılır [Akbulut, 2006].

3.1.5. Diğer sınıflama yöntemleri

Diğer sınıflama yöntemleri, genellikle ticari veri madenciliği sistemlerindeki sınıflamalar için daha az kullanılırlar. Örneğin, en yakın komşu sınıflaması tüm eğitim örneklerini depolar. Bu da çok büyük veri setleri üzerinde yapılan öğrenmede zorluklara neden olabilmektedir.

K-En Yakın Komşu Algoritması

K-En yakın komşuluğunda, K harfi araştırılan komşuların sayısıdır. 5-yakın komşuluğunda, 5 kişiye ve 1-yakın komşuluğunda sadece bir kayıda bakılır [Han ve Kamber, 2001]. Bir veri uzayında, birbirine yakın olan kayıtlar birbirinin yakın komşusu olmaktadır.

En Yakın komşu sınıflayıcıları benzeşme ile öğrenmeyi temel alırlar. Eğitim örnekleri n boyutlu nümerik alanlar olarak tanımlanırlar. Her bir örnek n boyutlu uzayda bir noktaya karşılık gelir. Bu yolla eğitim örneklerinin tamamı n boyutlu uzayda depolanmış olur. Bilinmeyen bir örnek verildiği zaman, k en yakın komşu sınıflayıcısı bu uzayda bilinmeyen örneğe en yakın k eğitim örneğini bulur. Bu k adet eğitim örnekleri, bilinmeyen örneğin "en yakın k komşusu" dur. "Yakınlık" öklid uzaklığı olarak tanımlanır. Buna göre $X = (x_1, x_2, ..., x_n)$ ve $Y = (y_1, y_2, ..., y_n)$ gibi iki nokta arasındaki öklid uzaklığı Eş. 3.9' daki gibi hesaplanır:

$$d(X,Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{3.9}$$

Bilinmeyen örnek, k en yakın komşuları arasındaki en yaygın sınıfa atanır. k=1 olduğunda, bilinmeyen örnek uzayda kendisine en yakın eğitim örneğinin sınıfına atanır.

Genetik Algoritmalar

Genetik algoritmalar , çok değişkenli fonksiyonları optimize etmeyi amaçlayan sayısal bir araçtır. Bu algoritma parametre yerine onların kodlanmış biçimlerini kullanarak en iyiye ulaşmaya çalışır. Yapay zekanın bir uygulaması olan genetik algoritma , kısa sürede çözümleri ortaya çıkarması bakımından önemli bir tekniktir [Kantardzic, 2002].

Genetik algoritmalar, doğal evrim fikrini içermektedir. Genel olarak genetik öğrenme şu şekilde başlar. Bir başlangıç popülasyonu rastgele üretilmiş kuralları içerecek şekilde oluşturulur. Her bir kural bitler katarı şeklinde sunulabilmektedir.

En uygun olanının yaşaması misyonuna uygun olarak, şimdiki popülasyondan en uygun olan kurallar ve bunların çocukları yeni popülasyonu oluşturulur. Tipik olarak bir kuralın uygunluğu bir eğitim örnekleri seti üzerindeki sınıflama doğruluğu tarafından belirlenir.

Çocuklar, çaprazlama ve mutasyon gibi genetik işlemler uygulanarak üretilirler. Çaprazlamada kural çiftlerinin alt katarları yeni kural çiftleri oluşturmak için değiş-tokuş edilirler. Mutasyonda, bir kural katarından rastgele seçilmiş bitler ters çevrilirler.

Önceki kural popülasyonundan yeni popülasyonların üretilmesi süreci bir p popülasyonundaki her bir kural önceden tanımlanmış bir uygunluk eşiğine sahip olana kadar geliştirilmeye devam eder.

Genetik algoritmalar kolaylıkla paralelleştirilebilir ve diğer optimizasyon problemlerinde kullanıldığı gibi sınıflamada kullanılmıştır. Veri madenciliğinde diğer algoritmaların uygunluğunun değerlendirilmesinde kullanılabilir.

3.2. Kümeleme

Kümeleme, veri tabanından ilginç örüntülerin keşfedildiği bir madencilik tekniğidir. Kümelemenin genel düşüncesi, veri tabanını çok sayıda kümeye ayırmak ve aynı kümeye ait verilerin mümkün olduğu kadar yakın ilişkide olmalarının sağlanmasıdır [Hsu, 2008].

Veri madenciliğinde kümeleme yaygın şekilde kullanılan, verileri sınıflar veya kümeler içinde gruplayan, bu sayede aynı küme içindeki verilerin diğer kümedekilere göre daha benzer olduğu bir tekniktir [Han ve Kamber, 2001].

Kümeleme analizi, nesnelerin altdizinlere gruplanmasını yapan işleme denir. Böylece nesneler, örneklenen kitle özelliklerini iyi yansıtan etkili bir temsil gücüne sahip olmuş olur. Kümeleme, bir denetimsiz öğrenme yöntemidir.

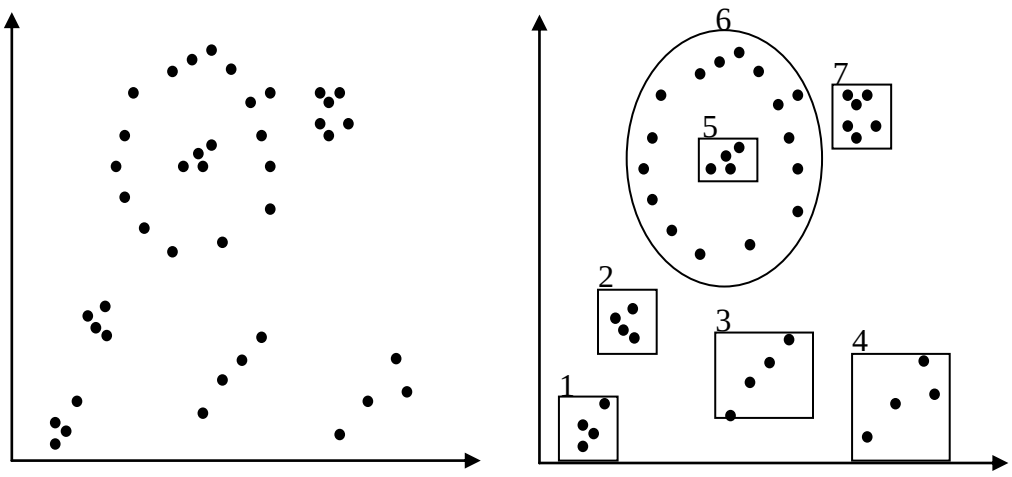
Kümeleme analizinin özellikleri aşağıda kısaca özetlenmiştir:

- Denetimsiz öğrenmedir.

- Önceden tanımlanan sınıf ve sınıf-etiketli öğrenme örnekleriyle çalışmamaktadır.
- Kümeleme veri dağılımını anlamada fayda sağlar.
- Bir veri madenciliği fonksiyonudur.

Basit bir kümeleme örneği

Şekil 3.3' de gösterilmiştir. Bu örnekte soldaki grafikte yer alan veriler giriş olarak verilmiş ve kümeleme işlemi sonucunda ortaya çıkan 7 adet küme sağda gösterilmiştir.



Şekil 3.3. Veri kümeleme örneği

3.2.1. Kümeleme analizinde kullanılan başlıca metotlar

Veri kümeleme güçlü bir gelişme göstermektedir. Veri tabanlarında toplanan veri miktarının artmasıyla orantılı olarak, kümeleme analizi son zamanlarda veri madenciliği araştırmalarında aktif bir konu haline gelmiştir.

Literatürde pek çok kümeleme algoritması bulunmaktadır. Kullanılacak olan kümeleme algoritmasının seçimi, veri tipine ve amaca bağlıdır. Genel olarak başlıca kümeleme yöntemleri şu şekilde sınıflandırılabilir:

1- Bölümleme yöntemleri
2- Hiyerarşik yöntemler
3- Model tabanlı yöntemler.

Bölümleme Yöntemleri

Bölümleme yönteminde ilk önce örneklem kümesi içinden rastgele k tane merkez seçilir. Daha sonra her bir noktanın küme merkezlerine olan uzaklıkları hesaplanır ve bu uzaklığı minimum yapan yeni küme merkezleri bulunarak güncellenir. Küme merkezlerinde hiçbir değişim olmayıncaya kadar, noktaların küme merkezlerine olan uzaklığının hesaplanması ve bu uzaklığı minimum yapan küme merkezlerinin bulunarak güncellenmesi işlemi tekrarlanır. K-ortalamalar (k-means) ve k-medoids birer bölümleme kümeleme algoritmasıdır. Aşağıda tez kapsamında kullanılan k-ortalama algoritması anlatılmıştır.

K-ortalamalar

Veri madenciliğinde kümelemede kullanılan ve en çok bilinen uygulamalardan biri K-ortalama ‘dır. Öncelikle, K sayıda gözlem N gözlem içinden küme sayısına göre rastgele seçilir ve ilk kümelerin merkezi olur. İkinci olarak, kalan her bir N-K gözlem için öklid uzaklık cinsinden en yakın küme bulunur. Her gözlem en yakın kümelere atandıktan sonra, kümenin merkezi yeniden hesaplanır. Son olarak, tüm gözlemler dağıtıldıktan sonra, gözlemler ile kümenin merkezi arasındaki öklid uzaklık hesaplanarak en yakın kümeye atanıp atanmadığı tespit edilir. Kümelemede uygulanan K-ortalama algoritması pek çok araştırmada kullanılmıştır [Liao ve Wen, 2007].

K-ortalamalar algoritması bölümleme yöntemleri olarak adlandırılan algoritmalardan biridir. Bölümleme kümeleme problemi şöyle ifade edilmiştir: d boyutlu metrik uzayda verilen n nesnesinin, aynı kümelerdeki nesneler diğer kümelerdekine kıyasla daha benzer olacak şekilde k kümeye yerleştirerek bölümlenmesinin yapılmasıdır. K

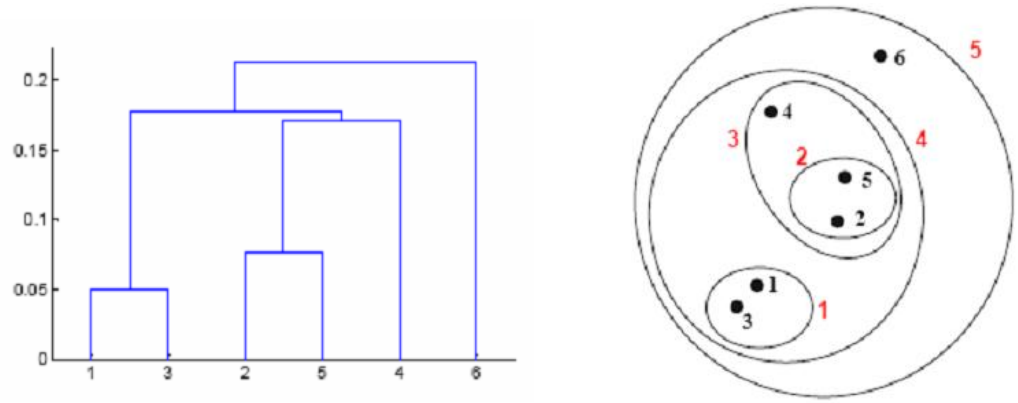
değeri probleme göre belirlenebilir veya belirlenmez. Hata kareler ölçütü gibi bir kümeleme ölçütünün olması gerekir.

Bu sorunun çözümü şöyledir: Bir kümeleme kriteri seçilir, sonra her bir veri nesnesi için bu kriterleri optimize edecek küme seçimi yapılır. K-ortalamalar algoritması k kümelerini, her bir kümeyi temsil edecek bir nesnenin keyfi seçimiyle başlatır. Kalan her nesne bir kümeye atanır ve kümeleme kriteri küme ortalamasını hesaplayabilmek için kullanılır. Bu ortalamalar yeni küme noktaları olarak kullanılır ve her bir nesne kendisine en benzer olan kümeye yeniden atanır. Bu kümeler yeniden hesaplanır ve kümelerde hiç bir değişim gözlenilmediği duruma ve değişim istenen hata düzeyinin altına düşürülünceye kadar bu döngü devam ettirilir.

Hiyerarşik Metotlar

Hiyerarşik kümeleme nesnelerin yakınlık ilişkisine göre oluşturulan kümelerden bir ağaç inşa eder. Hiyerarşik kümeleme aşağıdaki özelliklere sahiptir:

- Bir veri tabanını bir kaç kümeye ayrıştırır.
- Bu ayrıştırma dendogram adı verilen bir ağaç sayesinde yapılır (Bkz. Şekil 3.4)
- Bu ağaç, yapraklardan gövdeye doğru veya gövdeden yapraklara doğru kurulabilir. Dendogram istenen seviyede kesilerek kümeler elde edilir.



Şekil 3.4. Dendogram yapısına bir örnek

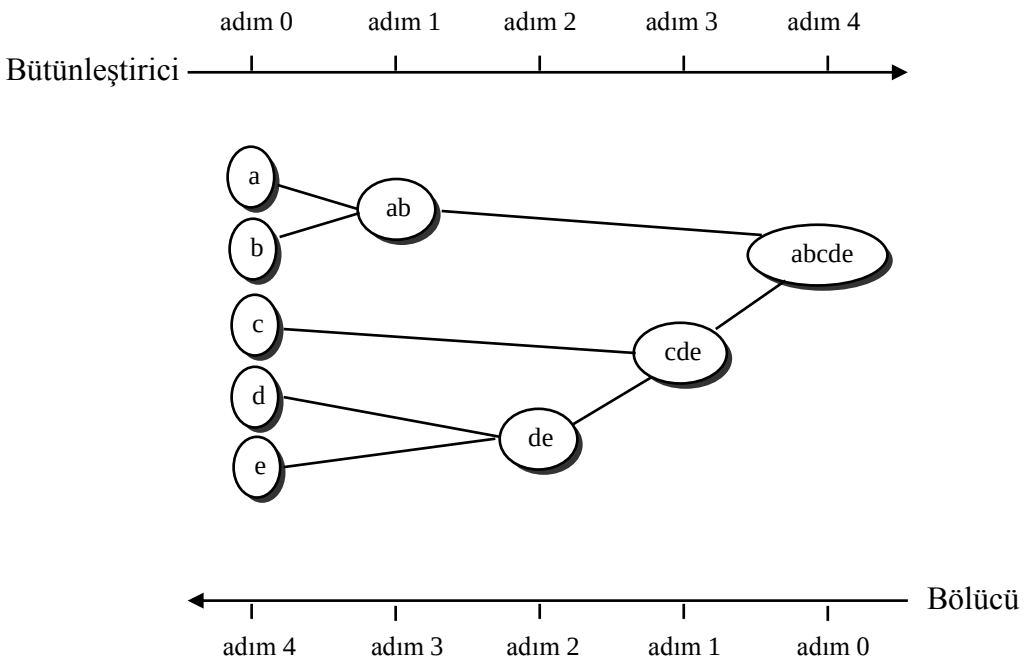
Bir hiyerarşik kümeleme metodu veri nesnelerini bir küme ağacına gruplayarak çalışır. Hiyerarşik kümeleme yöntemleri, hiyerarşik ayrışmanın yukarıdan-aşağıya veya aşağıdan-yukarıya oluşturulmasına bağlı olarak bütünleştirici ve bölücü hiyerarşik kümeleme olarak sınıflandırılabilir (Bkz. Şekil 3.5.). Saf hiyerarşik kümeleme yöntemlerinin kalitesi, bir kez birleştirme veya ayırma kararı işletildiğinde uyum gösterememesinden etkilenmektedir.

- Aşağıdan yukarıya ya da bir diğer ifadeyle bütünleştirici yaklaşıma göre hiyerarşik kümeleme şu şekildedir:

    - Her bir nesne için farklı bir grup oluşturarak başla,
    - Bazı kurallara göre grupları birleştir. Örneğin, merkezler arasındaki uzaklık,ortalama vb.,
    - Bir sonlandırma durumuna ulaşılıncaya kadar devam et. Yani, bütün nesneler tek bir küme içinde kalana kadar ya da istenen sayıda küme elde edene kadar birleştirme işlemi devam eder.

- Yukarıdan aşağıya ya da bir diğer ifadeyle bölücü yaklaşıma göre hiyerarşik kümeleme şu şekildedir:

    - Aynı kümedeki bütün nesnelerle başla,
    - Bir kümeyi daha küçük kümelere böl,
    - Bir sonlandırma durumuna ulaşılıncaya kadar devam et. Yani, her nesne ayrı bir küme oluşturana ya da istenilen küme sayısı elde edilene kadar ayrılma işlemi devam eder



Şekil 3.5. Bütünleştirici ve bölücü hiyerarşik kümelemenin {a,b,c,d,e} veri nesneleri üzerinde gösterimi

Model Bazlı Kümeleme Metotları

Model tabanlı kümeleme metotları, verilen veri ile bazı matematiksel modellerin arasındaki uygunluğu optimize etmeye çalışır. Bu metotlar verinin olasılık dağılımlarının bir karışımından elde edildiğini varsayar. Model tabanlı kümeleme metotları iki yaklaşımdan oluşur:

- İstatistiksel Yaklaşım
- Sinir Ağları Yaklaşımı

İstatistiksel Yaklaşım

Kavramsal kümeleme iki adımlı bir işlemdir: ilk olarak kümeleme yapılır sonrasında bunu tanımlama takip eder.

Kavramsal kümelemedeki bir çok metot, kavram veya kümelerin belirlenmesinde olasılık ölçümlerini kullanan istatistiksel yaklaşımı benimser.

COBWEB, artımlı kavramsal kümelemede popüler ve basit bir metottur. Bu metotta girdi nesneleri kategorik alan değer çiftleri olarak tanımlanır. COBWEB sınıflama ağacı formunda bir hiyerarşik kümeleme oluşturur. Sınıflama ağacındaki her bir nokta bir kavrama karşılık gelmektedir ve bu nokta altında sınıflama nesneleri özetleyen kavramın olasılıklı tanımlamasını içerir

Sinir Ağları Yaklaşımı

Sinir ağları yaklaşımı ile kümeleme her bir kümeyi bir "temsilci" olarak sunma eğilimindedir. Bir temsilci, kümenin bir prototipi olarak rol oynar ve belli bir veri örneğine veya nesneye karşılık gelmek zorunda değildir. Yeni nesneler, bazı uzaklık ölçütlerine bağlı olarak temsilcisi en benzer olan kümeye dağıtılabilirler. Bir kümeye atanan bir nesnenin alanları, kümenin temsilcisinin alanlarından tahmin edilebilir.

3.3. Birliktelik Kuralları

Birliktelik kuralları, büyük veri kümeleri arasında birliktelik ilişkileri bulurlar [Han ve Fu, 1999]. Toplanan ve depolanan verinin her geçen gün gittikçe büyümesi yüzünden, şirketler veritabanlarındaki birliktelik kurallarını ortaya çıkarmak istemektedirler. Birliktelik kurallarının amacı, kullanıcı tarafından belirlenen en küçük destek ve güven değerlerini sağlayan kuralların bulunmasıdır.

Birliktelik kurallarının kullanıldığı en tipik örnek market sepeti uygulamasıdır. Bu işlem, müşterilerin yaptıkları alışverişlerdeki ürünler arasındaki birliktelikleri bularak müşterilerin satın alma alışkanlıklarını analiz eder. Bu tip birlikteliklerin keşfedilmesi, müşterilerin hangi ürünleri bir arada aldıkları bilgisini ortaya çıkarır ve market yöneticileri de bu bilgi ışığında daha etki satış stratejileri geliştirebilirler[Özekes, 2003].

Örneğin, bir alışveriş merkezinde, ekmek alan müşterilerin 80%'i süt de almaktadır. İlişki kuralları algoritmasını uygulamanın asıl amacı rastsal verilerin analizi ile eş

zamanlı ilişkileri ortaya çıkarmak ve karar verirken referans olarak kullanmaktır [Hsu, 2008].

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 [Adriaans ve Zantinge, 1996].

Örneğin bir A ürününü satın alan müşteriler aynı zamanda B ürününü da satın alıyorlarsa, bu durum birliktelik buralı ile gösterilir [Zaki, 1999]:

A => B [destek = %2, güven = %60]

Buradaki destek ve güven ifadeleri, kuralın ilginçlik ölçüleridir. Sırasıyla, keşfedilen kuralın kullanışlığını ve doğruluğunu gösterirler. Birliktelik kuralı için %2 oranındaki bir destek değeri, analiz edilen tüm alışverişlerden %2'sinde A ile B ürünlerinin birlikte satıldığını belirtir. %60 oranındaki güven değeri ise A ürününü satın alan müşterilerinin %60'ının aynı alışverişte B ürününü de satın aldığını ortaya koyar [Zaki, 1999].

3.3.1. Apriori algoritması

Birliktelik kuralı için en çok bilinen strateji Apriori'dir [Liao ve Wen, 2005]. Apriori' de kullanıcı en küçük destek eşiğini verir ve algoritma bu eşik değerinden büyük tüm veri kümesini arar. İkinci adımda, ilk adımda bulunan veri kümelerinden kurallar oluşturulmaktadır. Algoritma her kural için güven değerini hesaplar ve kullanıcı tarafından tanımlanan güven eşik değerini aşan kuralları saklar.

Uygulamada görülen en önemli problemlerden biri ise destek ve güven eşiklerinin belirlenmesi olmuştur [Plasse ve ark., 2007].

Apriori algoritması veri tabanındaki verileri tekrarlayarak kaydeder ve her kayıttan sonra geniş veri kümelerini oluşturur. İşlemleri indirgemek için aday veri kümeleri için sadece destek seviyeleri hesaplanır [Liao ve Wen, 2005].

Apriori algoritmasında k öğeli sık geçen öğe küme adayları, (k-1) öğeli sık geçen öğe kümelerinden faydalanılarak bulunur. Ancak bu algoritma veri tabanının pek çok kere taranmasını gerektirmektedir. Takipteki taramalarda bir önceki taramada bulunan sık geçen öğe kümeleri aday kümeleri adı verilen yeni potansiyel sık geçen öğe kümelerini üretmek için kullanılır. Aday kümelerin destek değerleri tarama sırasında hesaplanır ve aday kümelerinden minimum destek metriğini sağlayan kümeler o geçişte üretilen sık geçen öğe kümeleri olur. Sık geçen öğe kümeleri bir sonraki geçiş için aday küme olurlar. Bu süreç yeni bir sık geçen öğe kümesine rastlanıncaya kadar devam eder.
Bu algoritmadaki temel yaklaşım eğer *k*-öğe kümesi minimum destek metriğini sağlıyorsa bu kümenin alt kümelerinin de minimum destek metriğini sağladığıdır [Han ve Kamber, 2001].

Veri madenciliğindeki yöntemler bahsedildikten sonra, çalışmanın bir sonraki bölümünde bankacılık sektöründe personel seçimi ve performans değerlendirilmesine yönelik bir VM uygulamasına yer verilmiştir. Weka yazılımı kullanılarak gerçekleştirilen madencilik sürecinde, bu bölümde bahsedilen sınıflandırma ve kümeleme algoritmalarından yararlanılacaktır. Sınıflandırma tekniklerinden karar ağaçları, Bayes sınıflayıcısı ve yapay sinir ağı algoritmaları uygulanarak sonuçları karşılaştırılacak; kümelemede ise k-ortalama algoritmasından yararlanılacaktır.

4. BANKACILIK SEKTÖRÜ ÇALIŞANLARINI DEĞERLENDİRMEYE YÖNELİK BİR UYGULAMA

Bu bölümde, banka şubelerinde çalışan satış personellerinin değerlendirilmesine yönelik bir veri madenciliği uygulamasına yer verilmiştir. Çalışmanın bu bölümünde, öncelikle Weka yazılımı hakkında bilgi verilecek, ardından da bir önceki bölümde bahsedilmiş olan sınıflandırma ve kümeleme algoritmalarından yararlanılarak bir uygulama gerçekleştirilecektir.

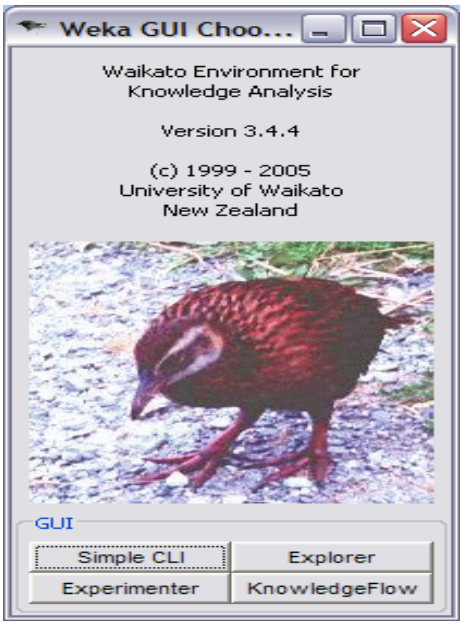
4.1. WEKA Yazılımı

Weka, Yeni Zelanda'daki Waikato Üniversitesi tarafından geliştirilmiş olup "Waikato Environment for Knowledge Analysis" kelimelerinin baş harflernin kısaltmasıdır. [Witten ve Frank, 2005]. Weka başta Yeni Zelanda'da tarımsal verinin işlenmesi amacıyla geliştirilmiştir. Bununla birlikte sahip olduğu makine öğrenme metotları ve veri mühendisliği kabiliyeti öyle hızlı ve köklü bir şekilde gelişmiştir ki, şimdi veri madenciliği uygulamalarının tüm formlarında yaygın olarak kullanılmaktadır [Frank ve ark., 2004].

Weka, bir öğrenen makinalar algoritmaları kolleksiyonu olduğu gibi yeni algoritmaların geliştirilmesi için de çok uygundur. GNU (General Public License) altında yayınlanmış, Java dilinde kodlanmış, açık kaynaklı bir yazılımdır [Kirkby ve Frank, 2005]. Ayrıca WEKA, Windows, Linux ve Masintosh gibi farklı işletim sistemleri üzerinde çalışabilen bir programdır [Witten ve Frank, 2005]. Weka Grafiksel Kullanıcı Arayüzü (Bkz. Resim 4.1.), WEKA'nın grafiksel çevresine erişim için kullanılmaktadır.

Weka pencerinin alt kısmında ise dört adet seçenek bulunmaktadır:

1. Simple CLI: WEKA komutlarının direkt olarak işlenmesine olanak sağlayan basit bir komut satırı arayüzü sağlar.



Resim 4.1. WEKA grafiksel kullanıcı arayüzü seçim penceresi

2. Explorer: Verinin WEKA ile keşfi için bir arayüzdür. Bu arayüzde VM ile sınıflandırma, kümeleme ve birliktelik kuralı uygulamaları kolaylıkla gerçekleştirilmektedir.

Weka Explorer ile, Bayes sınıflayıcısı, karar ağaçları, karar kuralları, regresyon, yapay sinir ağları gibi sınıflandırma algoritmaları; K-ortalama, Cobweb gibi kümeleme algoritmaları; Apriori gibi birliktelik kuralları kolaylıkla uygulanabilmektedir.

Weka Explorer' da önişleme, sınıflama, kümeleme, birliktelik kuralları, özellik seçme ve görselleştirme panelleri bulunmaktadır.

Önişleme : Veri dosyalarının yüklendiği, veri tabanının seçildiği ve verinin çeşitli yollarla değiştirildiği keşif sürecinin ilk adımıdır.

Sınıflama: Sınıflandırma ve regresyon algoritmalarının uygulanıp değerlendirildiği paneldir. Sınıflandırma fonksiyonları, kuralları, karar ağaçları, Bayes ağları, sinir ağları gibi sınıflandırma algoritmaları bu panelde yer almaktadır.

Kümeleme: K-ortalama, cobweb gibi kümeleme algoritmalarının yer aldığı paneldir.

Birliktelik kuralları: Verilerden birliktelik kurallarının çıkarıldığı paneldir.

Özellik şeçme: Veri kümesindeki ilişkili verilerin seçildiği paneldir.

Görselleştirme: Özellikler arasındaki ilişkiler iki boyutlu grafiklerle izlenebildiği paneldir.

3. Experimenter: Deneylerin gerçekleştirilmesi ve öğrenme planları arasındaki istatistiksel testleri yürüten bir arayüzdür. Bir veri setine farklı teknikleri uygulayarak yada aynı tekniği farklı parametrelerle tekrarlayarak, tek seferde birden fazla deneyin gerçekleştirilmesine izin veren bir araçtır.

4. Knowledge Flow: Weka veri madenciliği paketi ile sağlanan fonksiyonerliğin alternatif bir arayüzüdür. Bu arayüz temel olarak Explorer ile aynı işlevleri sürükle-bırak arayüzü ile yerine getirmektedir. Experimenter tarafından desteklenmeyen ek özellikleri ve experimenter de bulunan bazı eksik özellikleri ile gelişmekte olan bir bölümdür.

4.2. Bankacılık Sektörü Çalışanlarını Değerlendirmeye ve Personel Seçimine Yönelik Veri Madenciliği Uygulaması

Günümüzde firmaların kendilerine rekabet avantajı sağlaması açısından gün geçtikçe artan rekabet koşulları içerisinde personel kalitesi giderek daha da önemli bir hal almaktadır. Etkili bir personel seçimi mekanizması ile doğru insanı doğru yetenekler ile doğru yerde bulunmasının sağlanması organizasyonlar için kritik bir süreç olmaktadır. Türkiye Bankalar Birliği verilerine göre 2007 yıl sonu itibari ile sektörde 46 banka, 7618 şube ve 158534 çalışan yer almaktadır [Türkiye Bankalar Birliği, 2008]. Bu kadar çok çalışanın olduğu bir sektörde rekabet avantajı elde etmek adına insan kaynağı şüphesiz çok önemlidir.

Bu çalışma kapsamında, Türkiye' de faaliyet gösteren bir bankanın insan kaynakları veri tabanı üzerindeki verilerden yararlanılmıştır. Banka şubelerinde satışa yönelik hizmet gösteren çalışanların değerlendirildiği bu çalışmada, personelin çalıştığı il, çalıştığı şubenin bankanın diğer şubeleriyle kıyaslandığındaki sınıfı, şubelerdeki yürüttükleri ticari veya bireysel rolü, belirli bir periyot içerisindeki TPY (Ticari Portföy Yöneticisi) veya BPY (Bireysel Portföy Yöneticisi) rolündeki performans düzeyi, TPY veya BPY rolünü sürdürdüğü dönem sayısı, bankadaki unvanı, hizmet süresi, emeklilik durumu, yıllık tezkiye puanı gibi görev yerine, hizmetine ve pozisyonuna ilişkin bilgileri; yaşı, medeni hali, cinsiyeti gibi demografik bilgileri ile öğrenim durumu, yabancı dili, Sermaye Piyasası Kurumu (SPK) tarafından lisanslama belgesine sahip olup olmadığı gibi eğitim durumuna ve sertifikalarına ilişkin bilgileri kullanılmıştır. Çalışmada kullanılan veriler çoğunlukla kategorik özellikler içermiştir. Çalışmada personelin yaşı, hizmet süresi, TPY-BPY olarak görevde bulunduğu dönem sayısı, performans ortalaması gibi özellikler uzman görüşleri de dikkate alınarak kategorik hale getirilmiştir.

Çalışma kapsamında, Bankanın insan kaynakları ve performans veritabanlarından gerekli bilgiler elde edilmiştir. Veri tabanında tutulan bilgilerin kodlanmış halde tutuluyor olması çalışma sırasında veri temizleme sürecinin oldukça kısalmasını sağlamıştır. Veri tabanındaki onlarca farklı tablo üzerinde tutulan veriler PL/SQL sorgulama dili kullanılarak birleştirilmiştir. Elde edilen veri daha sonra Microsoft Excel üzerine aktarılmış ve ön işlemler yapılmıştır.

Bu çalışmada, çalışanlar öncelikle performanslarına göre gruplara ayrılmış, bu aşamada veri madenciliğinde kümelemeden yararlanılmıştır. Elde edilen performans sınıfları daha sonra sınıflandırma ile personel seçimi ve atamasında karar kuralları oluşturmada çıktı olarak kullanılmıştır. Sonuçlar çerçevesinde, TPY ve BPY 'lerin performanslarının değerlendirilmesi ile personel atamalarına ilişkin karar kuralları oluşturulmuştur.

4.2.1. Problemin tanımlanması ve amacın belirlenmesi

Gün geçtikçe artan rekabet koşulları içerisinde personel kalitesi günümüzde firmaların kendilerine rekabet avantajı sağlaması açısından oldukça önemli olmaktadır. Geleneksel insan kaynakları yönetimi yaklaşımına ek olarak, etkili bir personel seçimi mekanizması ile organizasyon için gerekli yetenekleri bulmak acil bir ihtiyaç olmaktadır [Chien ve Chen, 2008].

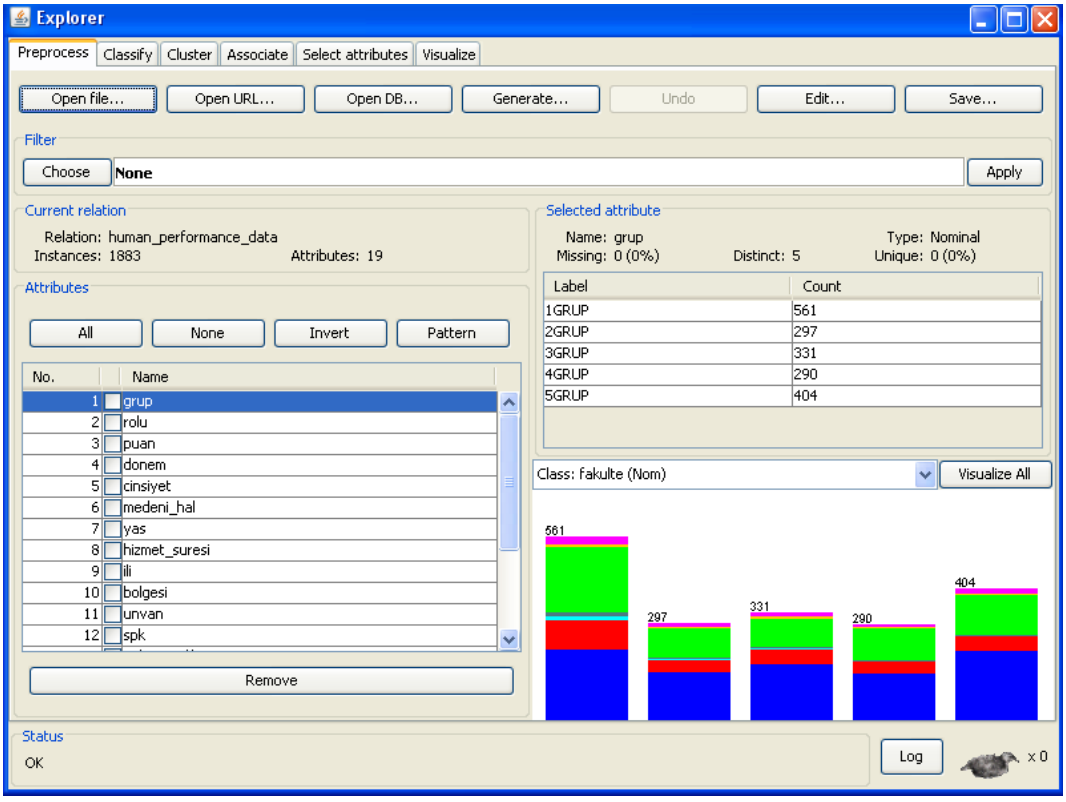
Tez kapsamında yapılan veri madenciliği çalışması için ülkemizde faaliyet gösteren bir bankanın şubelerinde çalışan Ticari ve Bireysel Portföy Yöneticileri'ne ilişkin personel ve performans verileri alınmıştır. Elde edilen veriler çerçevesinde, TPY ve BPY olarak görev alacak personelin seçiminde kriterler oluşturulması ve performansının belirlenmesi amaçlanmıştır. Yapılan bu çalışma ile personel atamalarındaki boşluğun doldurulması, atanan personellerin performans düzeylerinin öngörülebilir hale gelmesi, doğru personelin doğru özelliklerle doğru yerde görevlendirilmesinin sağlanması ile personel seçimi sürecinde fayda sağlanması amaçlanmıştır.

4.2.2. Veri toplama ve hazırlama

Veri madenciliği sürecinin en zaman alıcı adımlarından birisi veri temizleme ve ön işleme sürecidir. Tez kapsamında kullanılacak verilere karar verildikten sonra elde edilen veri üzerinde temizleme ve hazırlama süreci üzerinde durulmuştur. Veri temizleme sürecinde bilindiği gibi eksik, hatalı ya da boş veriler sıkıntı yaratmaktadır. Ancak, bu çalışmada verilerin veri tabanında oldukça düzgün tutuluyor olması veri temizleme sürecini önemli oranda azaltmıştır.

Bu çalışmada, şubelerde görev yapmakta olan personellerin çalıştığı ili, şubesinin olduğu bölgesi, şubesinin banka içindeki sınıfı, yürüttüğü rol (TPY-BPY), TPY-BPY olarak görev yaptığı süre, fiilen yürüttüğü unvanı, bankadaki hizmet süresi, emeklilik durumu, performans puanı, yöneticisinin son 2 yılda çalışanı hakkındaki öznel değerlendirmesi (tezkiye) gibi iş yaşamına ilişkin bilgileri ile cinsiyeti, medeni hali,

yaşı gibi demografik bilgileri ve öğrenim durumu, üniversitesi ve fakültesi, yabancı dili, yabancı dil seviyesi SPK lisanslama belgesine sahip olup olmaması gibi eğitimine ilişkin bilgileri içeren 19 farklı özellikten yararlanılmıştır. Çalışmanın toplam 1883 kayıttan oluşması ile 19 x 1883 lük bir matris elde edilmiştir. Şekil 4.1.'de çalışmada kullanılan veriler ışığında oluşan Weka Explorer ekran görüntüsü yer almaktadır.



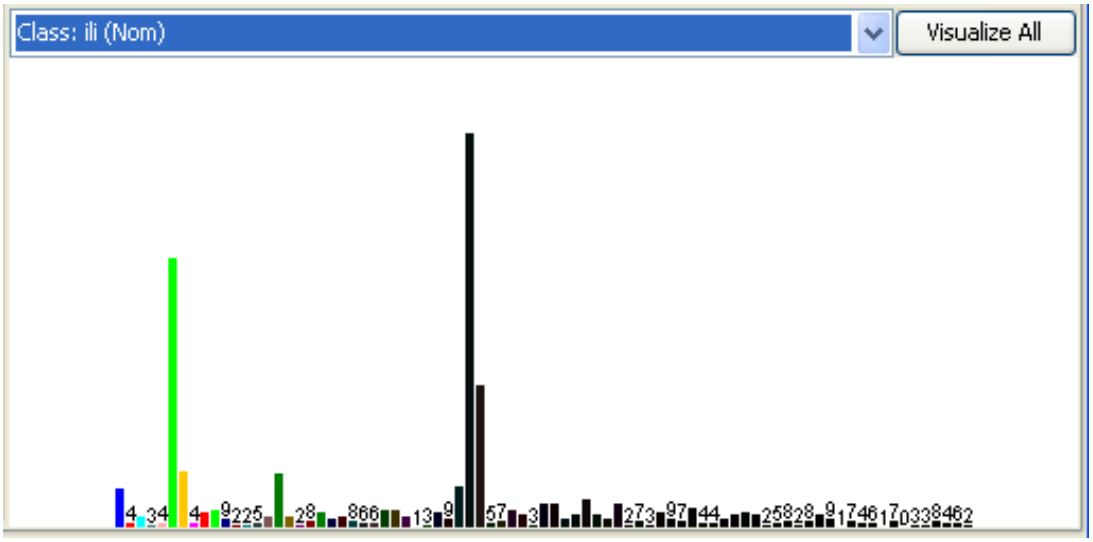
Şekil 4.1. Weka Explorer ekran görüntüsü

Özellikler

Bu çalışmada, personelin görev yerine ilişkin bilgiler, iş yaşamına ilişkin bilgiler, eğitim durumuna ilişkin bilgiler ile yaş, cinsiyet, medeni hal gibi demografik özelliklerini içeren 19 özellik kullanılmıştır.

Görev yerine ilişkin özellikler:

İli: Personelin çalışmakta olduğu ili temsil etmektedir. WEKA' da il trafik kodları ile gösterilmiştir (Bkz. Şekil 4.2).

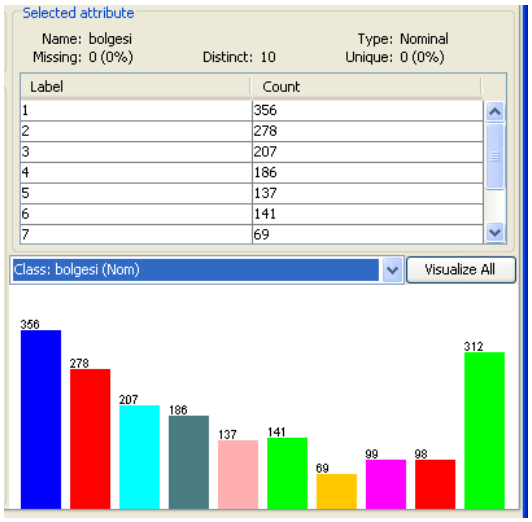


Şekil 4.2. Personelin çalıştığı illere göre dağılımı

Bölgesi: Personelin çalıştığı şubenin hangi bölge müdürlüğüne bağlı olduğunu göstermektedir. Aynı ildeki şubelerin farklı bölgelere bağlı olabilmesinden dolayı bu bilgiye ihtiyaç duyulmuştur. Banka organizasyonu içinde yer alan bölgeler Şekil 4.3 ' deki gibi sınıflandırılmıştır.

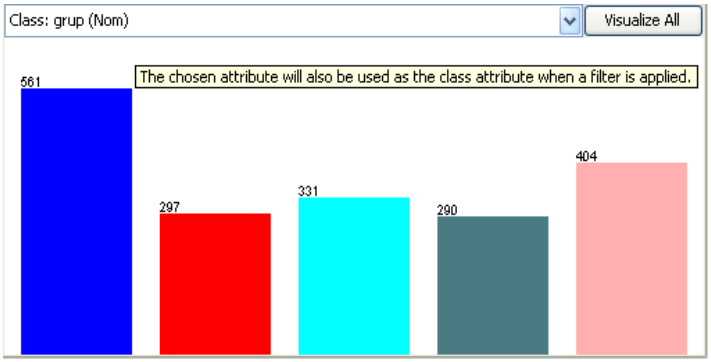
Grup: Personelin çalışmış olduğu şubenin banka içindeki sınıfını temsil etmektedir. Bankada şubeler 5 farklı sınıfta değerlendirildiği için 5 farklı grup yer almaktadır. 1. gruptaki şubeler performansı en iyi olan A sınıfı ya da 1. sınıf şubeler iken 5. grupta yer alan şubeler ise performans seviyesi en alt seviyede olan E sınıfı veya 5. sınıf şubeleri temsil etmektedir (Bkz. Şekil 4.4).

BÖLGE ADI	TANIMLAMA
ANKARA	0
İSTANBUL AVRUPA	1
EGE	2
ÇUKUROVA	3
İSTANBUL ANADOLU	4
MARMARA	5
KARADENİZ	6
DOĞU ANADOLU	7
İÇ ANADOLU	8
AKDENİZ	9



Şekil 4.3. Çalışılanların bağlı olduğu bölgelere yönelik tanımlamalar

ŞUBE SINIFI	TANIMLAMA
A	1GRUP
B	2GRUP
C	3GRUP
D	4GRUP
E	5GRUP

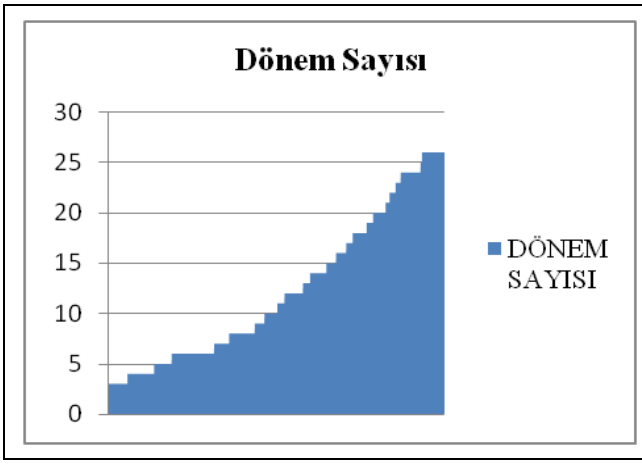


Şekil 4.4. Şube sınıflarına yönelik tanımlamalar

İş yaşamına ilişkin özellikler:

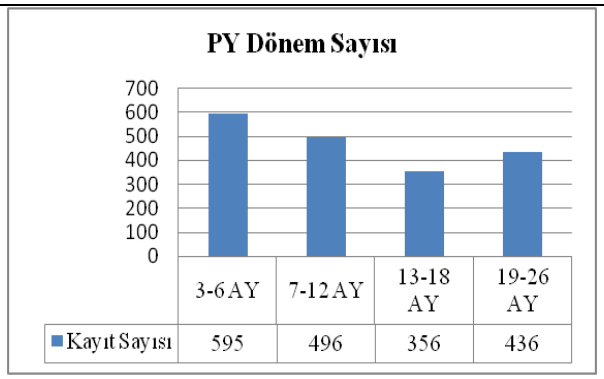
Rol: Şube satış personelini ticari veya bireysel müşterilere hizmet sunmasına göre Ticari Portföy Yöneticisi (TPY) ve Bireysel Portföy Yöneticisi (BPY) olmak üzere 2 gruptan oluşmaktadır. Bu çalışmada 1138 TPY, 745 BPY değerlendirilmiştir. TPY' ler ve BPY' ler sırası ile {T, B} ile tanımlanmıştır.

Dönem Sayısı: 26 aylık dönem içinde çalışanın kaç ay TPY-BPY olarak çalıştığını göstermektedir. Bu periyot içinde 3 aydan daha az portföy yöneticiliği yapan personel dikkate alınmamıştır. Sayısal olarak tutulan bu değer uzman görüşleri alınarak kategorize edilmiştir (Bkz. Şekil 4.5. ve Şekil 4.6).



Min.	3
Maks.	26
Ort.	12,213
Std. Sapma	7,271

Şekil 4.5. Kategorize öncesi dönem sayısı



DÖNEM SAYISI	KAYIT SAYISI
3-6 AY	595
7-12 AY	496
13-18 AY	356
19-26 AY	436

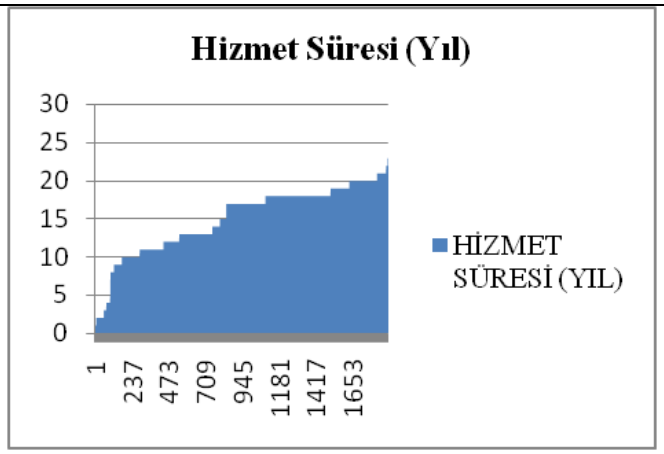
Şekil 4.6. Kategorize sonrası dönem sayısı

Unvan Grubu: Çalışanların ticari veya bireysel rolleri dışında unvan gruplarına göre {YÖNETİCİ, UZMAN, YETKİLİ, MEMUR} olarak 4 farklı grupta gösterilmiştir (Bkz. Çizelge 4.1). Memur grubu bankada memur unvanı ile çalışanları, uzman grubu yönetici adayı, uzman yardımcısı veya uzman olarak çalışanları, yetkili grubu memur grubundan yükselip yetki alan personeli, yönetici grubu ise yönetmen veya müdür yardımcısı seviyesini temsil etmektedir.

Çizelge 4.1. Unvan gruplarına yönelik tanımlamalar

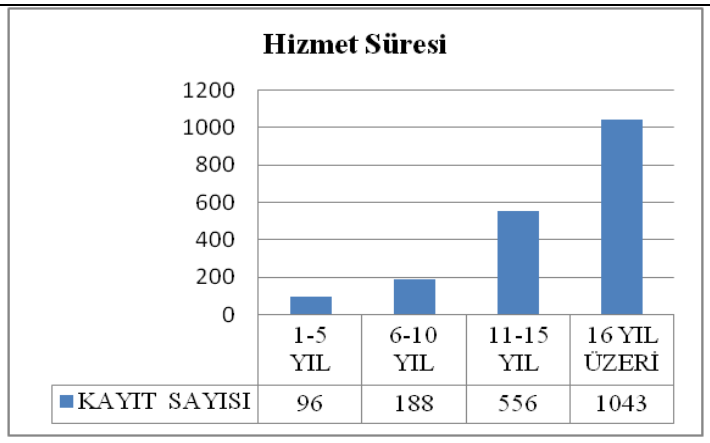
UNVAN GRUBU	YÖNETİCİ	UZMAN	YETKİLİ	MEMUR
KAYIT SAYISI	689	15	109	1070

Hizmet Süresi (Yıl): Personelin bankada geçen fiili hizmet süresini göstermektedir. Uzman görüşleri dikkate alınarak hizmet süresi kategorik hale getirilmiştir (Bkz. Şekil 4.7. ve Şekil 4.8).



Min.	1
Maks.	24
Ort.	14,994
Std. Sapma.	4,546

Şekil 4.7. Kategorize öncesi hizmet süresi dağılımı



HİZMET SÜRESİ	KAYIT SAYISI
1-5 YIL	96
6-10 YIL	188
11-15 YIL	556
16 YIL ÜZERİ	1043

Şekil 4.8. Kategorize sonrası hizmet süresi dağılımı

Emeklilik Durumu: Bankanın stratejik planları doğrultusunda çalışanların emekliliği hak ettiği tarihlere göre Çizelge 4.2.' deki gibi 4 kategoride sınıflandırılmıştır.

Çizelge 4.2. Emeklilik durumuna göre tanımlamalar

EMEKLİLİK HAKEDİŞ TARİHİ	TANIMLAMA	KAYIT SAYISI
2008 VE ÖNCESİ	1	70
2009	2	63
2010	3	81
2011 VE SONRASI	4	1669

Tezkiye (Yönetici Değerlendirmesi): Şube müdürlerinin son 2 yıldaki çalışan personeli hakkındaki kanaatini göstermektedir. Burada yöneticiler çalışanları için 1 ile 4 aralığında notlar vermektedirler. Çizelge 4.3.' de bu tanımlamalar yer almaktadır.

Çizelge 4.3. Tezkiyelere göre tanımlamalar

YÖNETİCİ DEĞERLENDİRMESİ	TANIMLAMA	KAYIT SAYISI
YOK	YOK	25
YETERSİZ	1	0
ORTA	2	37
BAŞARILI	3	862
ÇOK BAŞARILI	4	959

Çalışana ilişkin özel bilgiler:

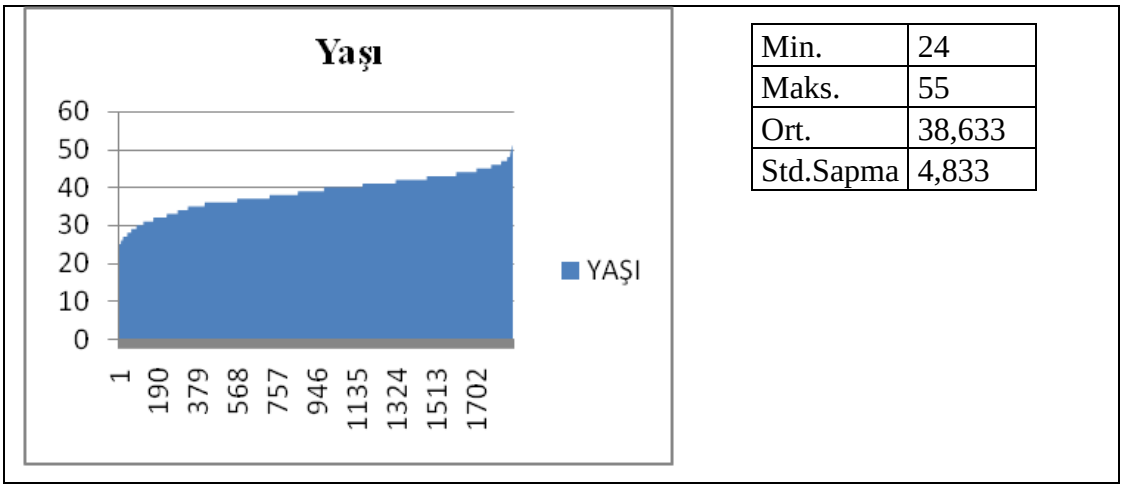
Cinsiyet: Kadın ve erkek sırasıyla {K, E} olarak tanımlanmıştır. Çalışanlardan 1096' sı kadın, 787' si ise erkektir.

Medeni Hal: 4 kategoride ele alınmış ve Çizelge 4.4.' de bu kategoriler gösterilmiştir.

Çizelge 4.4. Medeni hale ilişkin tanımlamalar

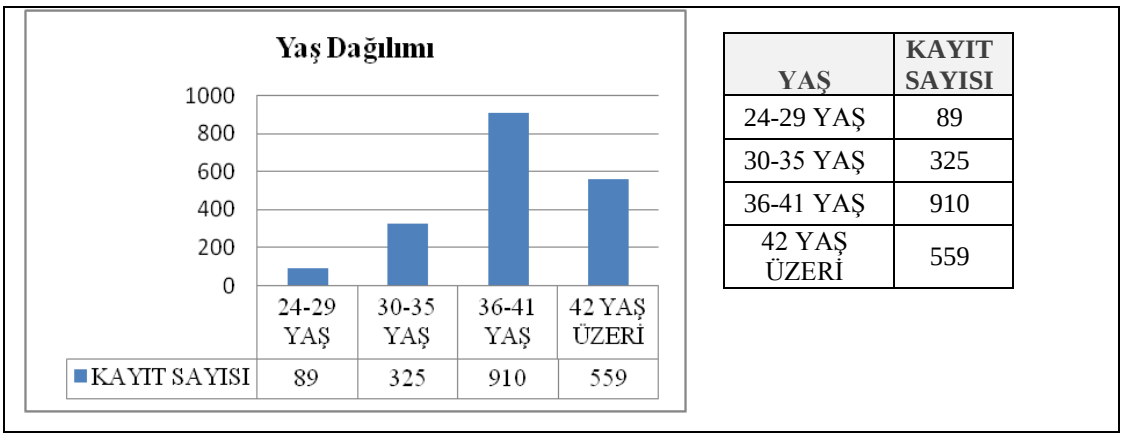
MEDENİ HAL	TANIMLAMA	KAYIT SAYISI
BEKAR	1	299
EVLİ VE ÇOCUKLU	2	1388
EVLİ	3	110
BOŞANMIŞ VE ÇOCUKLU	4	86

Yaş: Personelin doğum tarihlerine göre yaşları hesaplanmış, daha sonra ise bu veri anlamlı gruplarda kategorize edilmiştir (Bkz. Şekil 4.9. ve Şekil 4.10).



Min.	24
Maks.	55
Ort.	38,633
Std.Sapma	4,833

Şekil 4.9. Kategorize öncesi yaş dağılımı



YAŞ	KAYIT SAYISI
24-29 YAŞ	89
30-35 YAŞ	325
36-41 YAŞ	910
42 YAŞ ÜZERİ	559

Şekil 4.10. Kategorize sonrası yaş dağılımı

Eğitim durumuna ilişkin değişkenler:

Öğrenim Durumu: 4 kategoride tanımlanmıştır (Bkz. Çizelge 4.5).

Çizelge 4.5. Öğrenim durumuna yönelik tanımlamalar

ÖĞRENİM DURUMU	TANIMLAMA	KAYIT SAYISI
LİSE VE ALTI	1	704
2 YILILK YÜKSEK OKUL	2	182
LİSANS	3	984
YÜKSEK LİSANS	4	13

Üniversite Kategorisi: İşe alım politikası doğrultusunda çalışanlar, mezun olduğu üniversitelere göre Çizelge 4.6.' daki gibi sınıflandırılmıştır.

Çizelge 4.6. Mezun olunan üniversiteye yönelik tanımlamalar

ÜNİVERSİTE	TANIMLAMA	KAYIT SAYISI
ÜNV. MEZUNU DEĞİL	0	886
İŞE ALIMDA BANKA İÇİN ÖNCELİKLİ ÜNİVERSİTELER	1	10
ANKARA'DAKİ DİĞER ÜNİVERSİTELER	2	200
İSTANBUL'DAKİ DİĞER ÜNİVERSİTELER	3	94
İZMİR 'DEKİ ÜNİVERSİTELER	4	127
DİĞER YURTİÇİ ÜNİVERSİTELERİ	5	566
YURTDIŞI ÜNİVERSİTELERİ	6	0

Fakülte Kategorisi: Çalışanlar mezun oldukları bölümlere göre aşağıdaki gibi kategorize edilmiştir (Bkz. Çizelge 4.7).

Çizelge 4.7. Mezun olunan fakülteye yönelik tanımlamalar

FAKÜLTE TÜRÜ	TANIMLAMA	KAYIT SAYISI
YOK (ÜNV. MEZUNU DEĞİL)	0	886
AÇIKÖĞRETİM	1	252
İKTİSADİ-İDARİ BİLİMLER / BANKACILIK	2	591
DİĞER FAKÜLTELER	3	154

SPK Belgesi: Sermaye Piyasası Kurumu tarafından verilen SPK lisanslama belgesi olup olmadığına göre {VAR, YOK} şeklinde sınıflandırılmıştır. SPK belgesi olan personel sayısı 210 iken, belgesi olmayanların sayısı 1673' tür.

Yabancı Dil: İngilizce, Almanca, Fransızca bilen ya da yabancı dil bilmeyen olmak üzere 4 sınıfta gösterilmiştir (Bkz. Çizelge 4.8).

Çizelge 4.8. Yabancı dil bilgisine yönelik tanımlamalar

YABACI DİL	KAYIT SAYISI
İNGİLİZCE	15
ALMANCA	9
FRANSIZCA	1
YOK	1858

Yabancı Dil Seviyesi: Yabancı dil bilen personel yabancı dil seviyelerine göre sınıflanmıştır (Bkz. Çizelge 4.9). Yabancı dil bilmeyen personelin seviyesi ile "YOK" olarak alınmıştır.

Çizelge 4.9. Yabancı dil seviyesine yönelik tanımlamalar

YABACI DİL SEVİYESİ	KAYIT SAYISI
İYİ	9
ORTA	16
YOK	1858

Performansa ilişkin değişkenler

Puan (Performans Başarı Düzeyi): Bu çalışma kapsamında sınıflandırma sonucunda tahmin edilecek özelliktir. 26 aylık bir periyottaki TPY ve BPY' lerin bankadaki performans birimi tarafından hesaplanan performans puanları alınmıştır. Alınan bu puanlar her TPY-BPY için ait olduğu şube grubunun ortalama performans puanı ile oranlanmıştır.

Örneğin, bit TPY 'nin başarı puanı hesaplanırken;

Yeni aktif müşteri sayısı,
Çapraz satış oranı artışı,
Vadesiz mevduat artış miktarı,
Vadeli mevduat + yatırım artış miktarı,
Nakdi kredi artış miktarı,
Gayri nakdi kredi artış miktarı,
Takibe düşen kredi miktarı…
gibi kriterler dikkate alınarak portföy yöneticisinin (PY) toplam puanı hesaplanmaktadır.

Bu çalışmada, öncelikle PY 'lerin toplam puanı ticari ya da bireysel olmasına göre içinde bulunduğu şube sınıfının ortalama grup puanına oranlanarak her PY için bir başarı oranı belirlenmiştir.

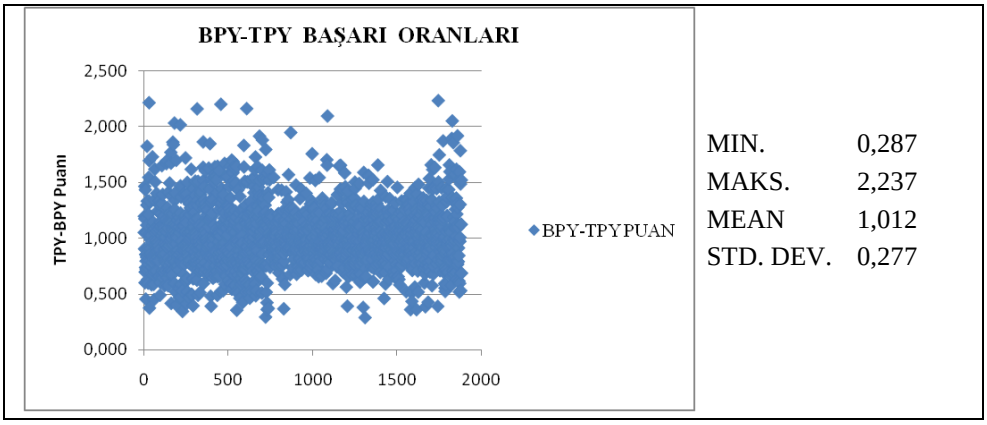
PY başarı oranı = PY toplam puanı / Grup ortalama puanı

Örneğin, bir PY 'nin toplam puanı belirtilen kriterlere göre 56 olarak hesaplanmış olsun. Çalışanın kendi şube grubundaki PY' lere ilişkin grup ortalama puanı 43 ise;

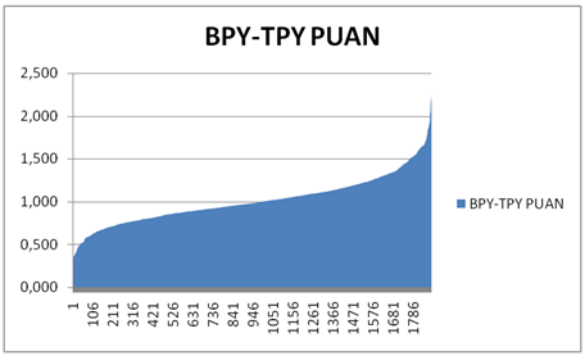
PY başarı oranı = 56 / 43 = 1,302 olarak hesaplanmıştır.

Bu başarı oranı, ilgili PY 'nin kendi şube sınıf ortalamasının üzerinde performans gösterdiğini belirtmektedir. Bu başarı oranı 1'in ne kadar altında ise PY, grup ortalamasının o kadar altında; 1'in ne kadar üzerinde ise PY grup ortalamasının o derece üzerinde performans göstermiştir.

PY başarı oranları dikkate alındığında ise Şekil 4.11.' de ve Şekil 4.12' de gösterilen dağılımlar ortaya çıkmıştır.



Şekil 4.11. Portföy yöneticilerine ilişkin başarı dağılımı



Şekil 4.12. Kümeleme öncesi portföy yöneticilerine ilişkin puan dağılımı

Çalışma kapsamında elde edilen bu sayısal değerler uzman görüşleri de dikkate alınarak kategorik hale getirilmiştir. Bu süreçte, Banka uzmanlarının isteği dikkate alınarak PY 'ler PY başarı oranına göre çok başarılıdan başarısıza kadar gruplara ayrılmıştır. Banka uzmanlarının bu gruplamadaki istediği, grup sayısının 5'i geçmemesi olmuştur. Bu aşamada ise kümelemede en yaygın olarak kullanılan k-ortalama algoritması kullanılmıştır.

K-ortalama algoritmasının PY başarı oranlarına uygulanması:

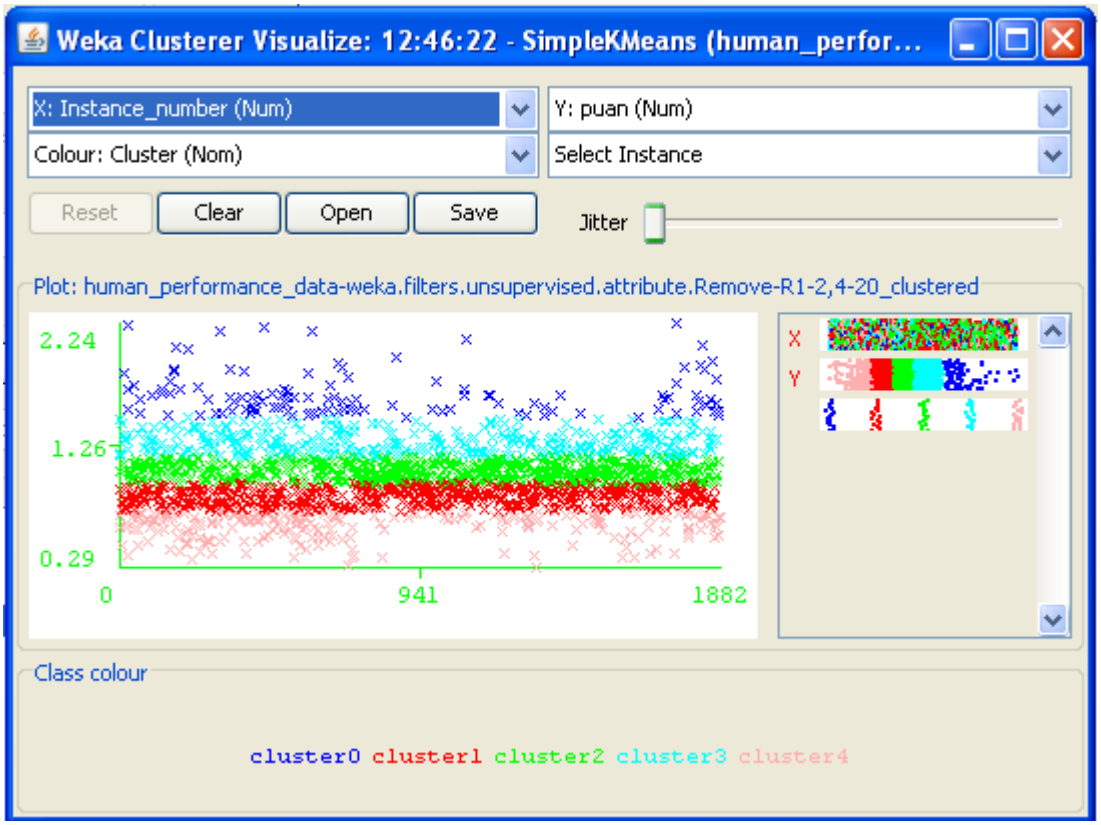
K-ortalama algoritması küme sayısı 2'den 5'e kadar WEKA 'da uygulanmış ve Çizelge 4.10' daki hata kareleri elde edilmiştir. Hata kareleri toplamı en az olan küme sayısı performans başarı düzeyi belirlemede kullanılmıştır.

Çizelge 4.10. K-ortalama algoritmasına göre küme sayısı ve hata kareleri toplamları

küme sayısı	**2**	**3**	**4**	**5**
hata kareleri toplamı	15,450	8,232	4,972	3,457

Burada belirtilmesi gereken nokta küme sayısı beşten fazla olduğunda hata karelerinin bir süre daha azalmasına karşın yöneticilerin performans değerlendirmede en fazla beş sınıf istemesidir. Hata kareleri toplamını en küçük olan küme sayısı 5' e ilişkin K-ortalama algoritması sonuç özeti EK-1' de gösterilmiştir.

WEKA' da küme sayısı=5 için oluşan kümeler aşağıdaki gibidir (Bkz. Şekil 4.13).



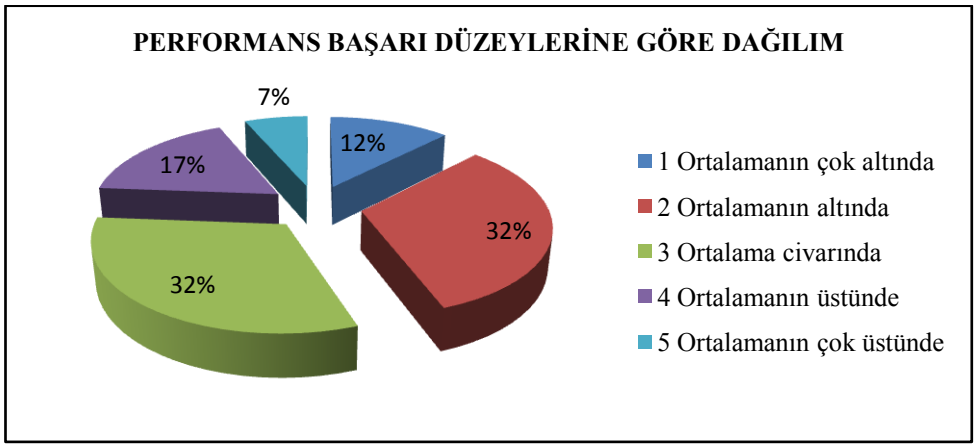
Şekil 4.13. k=5 için K-ortalama algoritması sonuçlarına göre oluşan kümeler

Bu sonuçlara göre 5 kümeye karşılık gelen ve Çizelge 4.11.' de gösterilen şu başarı düzeyleri oluşmuştur:

Çizelge 4.11. K-ortalama algoritması sonucu oluşan performans düzeyleri

BAŞARI DÜZEYİ	AÇIKLAMA	MİN.	MAKS.	ORT.	KAYIT SAYISI
1	Ortalamanın çok altında	0,287	0,728	0,604	233
2	Ortalamanın altında	0,730	0,950	0,852	605
3	Ortalama civarında	0,951	1,167	1,049	597
4	Ortalamanın üstünde	1,168	1,471	1,286	326
5	Ortalamanın çok üstünde	1,479	2,237	1,664	122

Elde edilen başarı düzeylerine göre 1, beklentilerin çok altında kalanları gösterirken 5, beklentilerin çok üstünde başarılı olan personeli tanımlamaktadır. Şekil 4.14. ‘ de kümeleme sonrası başarı düzeylerinin dağılımı gösterilmiştir.



Şekil 4.14. Kümeleme sonrası başarı düzeyleri

Kümeleme ile performans düzeyleri de belirlendikten sonra veri ön işleme adımı sona ermiş ve düzenlenen veriye ilişkin veri örneği Çizelge 4.12.’ de gösterilmiştir. Çalışmada kullanılan verilere ilişkin özellikler ve tanımlamalar ise EK-2’ de özetlenmiştir.

Çizelge 4.12. Düzenlenmiş veri örneği

Grup	Rol	Puan Başarı Düzeyi	Dönem Sayısı	Cins.	Med. Hal	Yaş	Hizmet süresi	Çalış. İl	Bölge	Unvan	SPK	Yabancı Dil	Yabancı Dil Sev.	Yönetici Değerl.	Emeklilik	Öğr. Durumu	Üniv.	Fakülte
2GRUP	T	3	19-26AY	E	2	30-35YAS	6-10YIL	06	0	YETKILI	YOK	INGILIZCE	ORTA	0	4	3	1	3
2GRUP	T	2	7-12AY	K	1	36-41YAS	11-15YIL	68	8	YETKILI	VAR	YOK	YOK	2	4	3	1	2
1GRUP	B	3	19-26AY	E	2	42YASUSTU	16YILUSTU	06	0	YONETICI	VAR	YOK	YOK	4	1	3	2	2
2GRUP	B	4	13-18AY	E	2	36-41YAS	11-15YIL	33	3	YONETICI	VAR	YOK	YOK	4	4	3	2	2
3GRUP	B	4	13-18AY	K	2	36-41YAS	11-15YIL	10	5	YONETICI	VAR	YOK	YOK	4	4	3	2	2
3GRUP	B	4	3-6AY	K	2	36-41YAS	11-15YIL	34	4	YETKILI	VAR	YOK	YOK	4	4	3	2	2
3GRUP	B	5	7-12AY	K	2	30-35YAS	11-15YIL	28	6	YETKILI	VAR	YOK	YOK	4	4	3	2	2
1GRUP	T	3	7-12AY	E	2	36-41YAS	11-15YIL	06	0	YONETICI	YOK	INGILIZCE	ORTA	4	4	3	2	2
1GRUP	T	3	19-26AY	K	1	30-35YAS	11-15YIL	06	0	YONETICI	YOK	INGILIZCE	IYI	4	4	3	2	2
1GRUP	T	3	7-12AY	E	1	36-41YAS	11-15YIL	06	0	YONETICI	YOK	INGILIZCE	ORTA	4	4	3	2	2
5GRUP	T	2	19-26AY	E	2	30-35YAS	11-15YIL	13	7	YETKILI	VAR	YOK	YOK	4	4	2	0	0
4GRUP	B	5	3-6AY	K	2	36-41YAS	11-15YIL	34	4	YETKILI	VAR	YOK	YOK	4	4	3	2	2
3GRUP	T	4	19-26AY	E	2	36-41YAS	11-15YIL	09	2	YONETICI	YOK	INGILIZCE	ORTA	4	4	3	2	2
1GRUP	T	2	19-26AY	E	2	36-41YAS	11-15YIL	34	1	YONETICI	YOK	INGILIZCE	ORTA	3	4	3	3	2
4GRUP	B	5	7-12AY	K	2	30-35YAS	11-15YIL	28	6	YETKILI	VAR	YOK	YOK	4	4	3	2	2
5GRUP	B	3	3-6AY	K	4	36-41YAS	16YILUSTU	33	3	YONETICI	VAR	YOK	YOK	4	4	3	2	2
5GRUP	B	5	7-12AY	K	2	30-35YAS	11-15YIL	28	6	YETKILI	VAR	YOK	YOK	4	4	3	2	2
1GRUP	T	2	13-18AY	K	2	36-41YAS	16YILUSTU	34	1	YONETICI	VAR	YOK	YOK	4	4	3	2	2
1GRUP	T	1	13-18AY	E	2	30-35YAS	6-10YIL	34	1	YONETICI	YOK	INGILIZCE	IYI	4	4	3	3	2
1GRUP	T	1	19-26AY	E	2	42YASUSTU	16YILUSTU	35	2	YETKILI	YOK	INGILIZCE	ORTA	3	4	3	4	3
1GRUP	T	5	3-6AY	K	1	24-29YAS	1-5YIL	34	1	UZMAN	YOK	INGILIZCE	ORTA	4	4	3	5	2
1GRUP	B	4	7-12AY	K	4	42YASUSTU	16YILUSTU	35	2	YONETICI	YOK	ALMANCA	IYI	3	3	1	0	0
…																		
…																		
…																		

4.2.3. WEKA'da programın çalıştırılması:

Bu aşamada, veri önişleme sürecinde Çizelge 4.12.'de de gösterildiği gibi düzenlenmiş veri ARFF formatına getirilerek WEKA' da çalıştırılmıştır. Çalışmada kullanılan ARFF uzantılı veri dosyası örneği Çizelge 4.13.'de gösterilmiştir.

Çizelge 4.13. ARFF uzantılı veri dosyası örneği

```
@relation human_performance_data

@attribute grup{1GRUP,2GRUP,3GRUP,4GRUP,5GRUP}
@attribute rolu{T,B}
@attribute puan{1,2,3,4,5}
@attribute donem{3-6AY,7-12AY,13-18AY,19-26AY}
@attribute cinsiyet{K,E}
@attribute medeni_hal{1,2,3,4}
@attribute yas{24-29YAS,30-35YAS,36-41YAS,42YASUSTU}
@attribute hizmet_suresi{1-5YIL,6-10YIL,11-15YIL,16YILUSTU}
@attribute
ili{01,02,03,04,05,06,07,08,09,10,11,12,13,14,15,16,17,18,19,20,21,2
2,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,
45,46,47,48,49,50,51,52,53,54,55,56,57,58,59,60,61,62,63,64,65,66,67
,68,69,70,71,72,73,74,75,76,77,78,79,80,81}
@attribute bolgesi{1,2,3,4,5,6,7,8,9,0}
@attribute unvan{YONETICI,MEMUR,UZMAN,YETKILI}
@attribute spk{YOK,VAR}
@attribute yabanci_dil{YOK,INGILIZCE,ALMANCA,FRANSIZCA}
@attribute yabanci_dil_seviye{YOK,IYI,ORTA}
@attribute tezkiye_ortalamasi{YOK,1,2,3,4}
@attribute emeklilik{1,2,3,4}
@attribute ogrenim_durumu{1,2,3,4}
@attribute universite{0,1,2,3,4,5,6}
@attribute fakulte{0,1,2,3}

@data

1GRUP,B,3,7-12AY,E,2,42YASUSTU,16YILUSTU,07,9,YONETICI,YOK,
INGILIZCE,ORTA,3,1,1,0,0

1GRUP,B,4,13-18AY,K,2,42YASUSTU,16YILUSTU,06,0,YONETICI,YOK,
```

Çizelge 4.13. (Devam) ARFF uzantılı veri dosyası örneği

```
YOK,YOK,3,2,2,0,0

1GRUP,B,2,3-6AY,E,2,42YASUSTU,16YILUSTU,06,0,YONETICI,YOK,
YOK,YOK,4,4,1,0,0

1GRUP,B,4,13-18AY,E,2,42YASUSTU,16YILUSTU,26,5,YONETICI,YOK,
YOK,YOK,3,4,3,5,3

1GRUP,B,4,19-26AY,K,4,42YASUSTU,16YILUSTU,06,0,YONETICI,YOK,
YOK,YOK,3,1,3,2,3

1GRUP,B,1,19-26AY,E,2,42YASUSTU,16YILUSTU,58,7,YONETICI,YOK,
YOK,YOK,3,3,1,0,0

1GRUP,B,1,3-6AY,K,2,36-41YAS,16YILUSTU,35,2,YETKILI,YOK,
YOK,YOK,3,3,1,0,0

1GRUP,B,4,19-26AY,E,2,42YASUSTU,16YILUSTU,06,0,YONETICI,YOK,
YOK,YOK,4,1,1,0,0

1GRUP,B,2,19-26AY,K,2,42YASUSTU,16YILUSTU,34,1,YONETICI,YOK,
INGILIZCE,IYI,4,1,3,3,3

1GRUP,B,2,19-26AY,E,2,42YASUSTU,16YILUSTU,06,0,YONETICI,YOK,
YOK,YOK,4,4,1,0,0
.
.
.
.
.
.
```

4.2.4. Sınıflandırma algoritmalarının uygulanması ve algoritma sonuçları

ARFF uzantılı veri dosyası WEKA' da çalıştırılmış,sınıflandırma algoritmalarından ID3, J4.8, PART, Saf Bayes, OneR ve MultilayerPerceptron algoritmaları uygulanmış ve sonuç özetleri ise sırasıyla değerlendirilmiştir.

WEKA, sınıflandırma algoritmalarının sonuçlarını değerlendirirken şu çıktıları bize sunmaktadır:

Düzensizlik matrisi: Yakınsaklık matrisi olarak da adlandırılır. Doğru olarak sınıflandırılan örneklerin sayısı bu matrisin köşegeni üzerindeki elemanlarının toplamına eşittir. Doğru olarak sınıflandırılan kayıt yüzdesi bize madencilik algoritmalarını karşılaştırma imkanı sunmaktadır.

True Positive (TP): Sınıflandırma algoritması tarafından herhangi bir sınıfa atanan kayıtlardan gerçekte o sınıfa ait olanların oranını yüzdesel olarak gösterir.

False Positive (FP): Sınıflandırma algoritması tarafından herhangi bir sınıfa atandığı halde gerçekte o sınıfa ait olmayan kayıtların oranını gösterir.

Kesinlik: Gerçekte herhangi bir sınıfa ait olan kayıtların hangi oranda sınıflandırma algoritması tarafından o sınıfa atandığı gösterir.

Kappa istatistiği: Tahmin doğruluğunun ölçüsüdür

ID3 algoritması için sonuç özeti**:**

Bölüm 3.1.1' de bahsedildiği gibi ağaç bölünmesinde bilgi kazancı kriterini kullanır. Kesikli veri üzerinde çalışır. EK-3' de gösterilen algoritma sonuç özeti incelendiğinde, düzensizlik matrisinin köşegeni üzerindeki kayıt sayısının toplam kayıt sayısına oranı olan doğru sınıflandırılan kayıt oranının %98,40 olduğu görülmektedir.

J4.8 algoritması için sonuç özeti:

C4.5 karar ağacının WEKA tarafından javada kodlanan 8. versiyonudur. Kesikli ve sürekli veri üzerinde karar ağacı oluşturur. Sayısal özellikler, kayıp değerler, gürültülü veri ile başa çıkabilmekte ve ağaçtan kurallar oluşturmaktadır. Algoritma

sonuç özetine EK-4' de yer verilmiştir. Buna göre doğru sınıflandırılan kayıt oranının %65,90 olduğu görülmektedir.

PART algoritması sonuç özeti:

J4.8'deki gibi kullanıcı tarafından tanımlanan parametreleri kullanarak kısmi karar ağacından kurallar oluşturur. EK-5' de yer verilen algoritma sonuç özetine göre oluşan düzensizlik matrisinden de anlaşılacağı gibi doğru sınıflandırılan kayıt oranının %67,34 olduğu, hatalı sınıflandırılan kayıt sayısının ise %32,66 olduğu gözlemlenmektedir.

Saf Bayes algoritması sonuç özeti:

Bilindiği gibi saf Bayes algoritması Bayes teoremine dayanan standart olasılıklı bir sınıflandırma yöntemidir. Weka' da elde edilen algoritma sonuç özeti EK-6' da gösterilmiş olup, bu algoritma sonucunda doğru sınıflandırılan kayıt oranı %42,75' de kalmıştır.

OneR algoritması sonuçları:

Basit sınıflandırma kuralları bulmamızı sağlayan basit ve ucuz bir sınıflama kuralıdır. Şaşırtıcı derecede yüksek kesinlikte kurallar oluşturur. Tek bir özellik üzerinde tek seviyeli karar ağacı oluşturur. Özelliklerden bir tanesi seçilir ve o özelliğe göre dallar oluşturulur. Her bir dal o özelliğin farklı bir değerini temsil eder. Her dalda en iyi kuralı veren bellidir ve ardından hata oranları hesaplanır. Her özellik için ayrı kural kümesi oluşturur. OneR algoritması sonuçlarına göre, doğru sınıflandırılan kayıt oranı %39,57 ile oldukça düşük seviyede kalmıştır. Algoritma sonuç özetine EK-7' de yer verilmiştir.

MultilayerPerceptron algoritması sonuç özeti:

Geriye yayılımı kullanan bir sinir ağıdır ve üç katmandan oluşmaktadır: girdi katmanı, saklı katman ve sonuç katmanı. WEKA'da yapay sinir ağlarına yönelik özel bir kullanıcı ara yüzü yer almaktadır. EK-8' de yer verilen algoritma sonuçlarına göre doğru sınıflandırılan kayıt oranı %46,57 olmuştur.

4.2.5. Sonuçların karşılaştırılması ve yorumlanması

Weka sınıflama panelinde yer alan karar ağacı algoritmaları (ID3, J4.8, PART), Bayes algoritması (saf Bayes), OneR sınıflandırma kuralı ile yapay sinir ağı algoritması (MultilayerPerceptron) uygulandıktan sonra, bu algoritmaların sonuçları karşılaştırılmış ve Çizelge 4.14.' de gösterilmiştir.

Çizelge 4.14. Sınıflandırma algoritma sonuçlarının karşılaştırılması

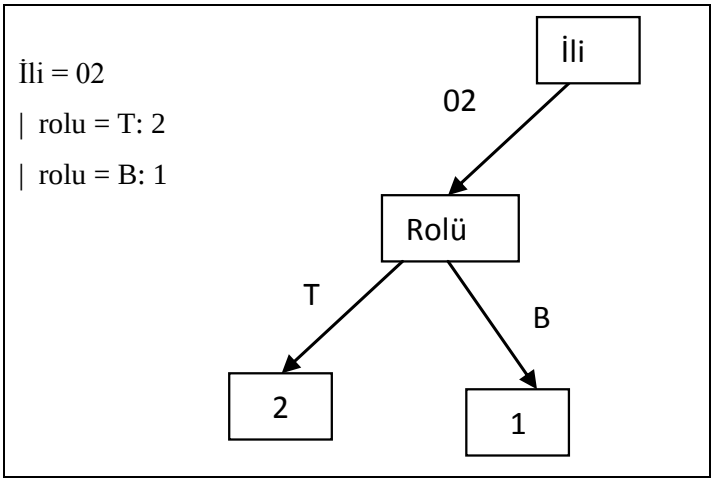
	Doğru sınıflandırılan kayıt (%)	Hatalı sınıflandırılan kayıt (%)	Kappa istatistiği	Ort. mutlak hata	Ort. hata karekök	Göreli mutlak hata (%)	Göreli hata karekök (%)
ID3	98,354	1,646	0,978	0,007	0,058	2,275	15,085
J4.8	65,906	34,094	0,534	0,174	0,295	58,218	76,308
PART	67,339	32,661	0,557	0,178	0,296	58,465	76,469
Saf Bayes	42,751	57,249	0,223	0,264	0,372	88,373	96,228
OneR	39,565	60,435	0,132	0,242	0,492	80,914	127,224
Multilayer Perceptron	46,575	53,425	0,247	0,217	0,440	72,672	113,795

Madencilikte kullanılan algoritma sonuçları incelendiğinde doğru olarak sınıflandırılan kayıt sayısının/oranının en yüksek olduğu algoritma %98,35 ile ID3 algoritması olmuştur. Madencilik sonuçlarından bahsedilirken ID3 algoritmasının çıktılarından söz edilmiştir.

Sonuçların Yorumlanması:

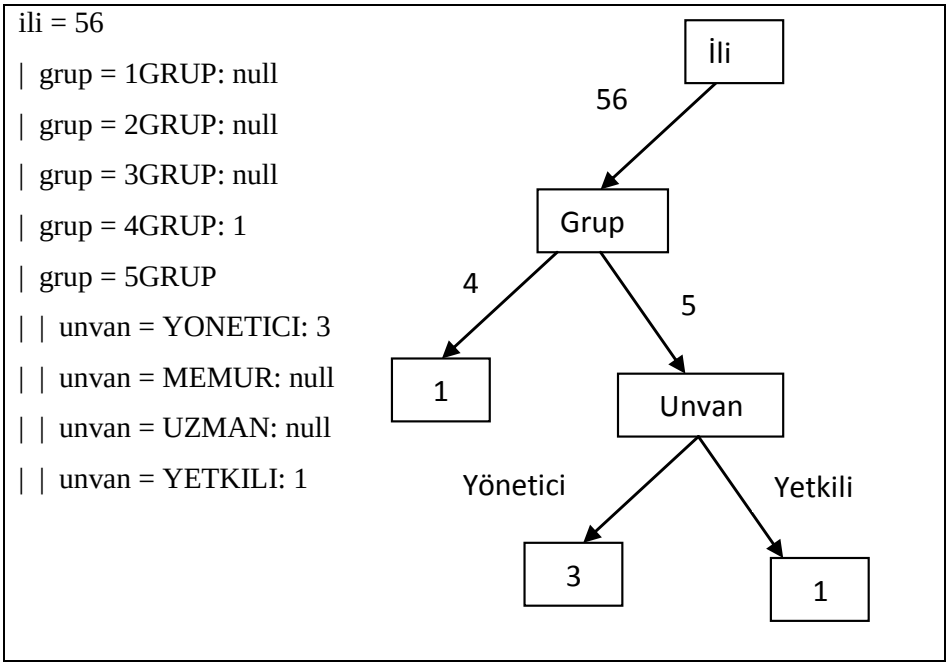
WEKA' da sınıflandırma algoritmalarının karşılaştırılmasında sonra doğru sınıflandırılan kayıt sayısı oranının en yüksek olduğu ID3 algoritmasının sonuçları ele alınmış ve yorumlanmıştır.

ID3 algoritması sonuçlarını değerlendirirken karar ağacındaki ilk dallanmanın illere göre olduğu görünmektedir. Bu da bizlere personel seçiminde ilk olarak personel ihtiyacı olan ile bakmamız gerektiğini söylemektedir. İlden sonraki dallanmalar ise her il için farklılık göstermiştir. Aşağıda elde edilen sonuçların bir bölümü özetlenmiştir.



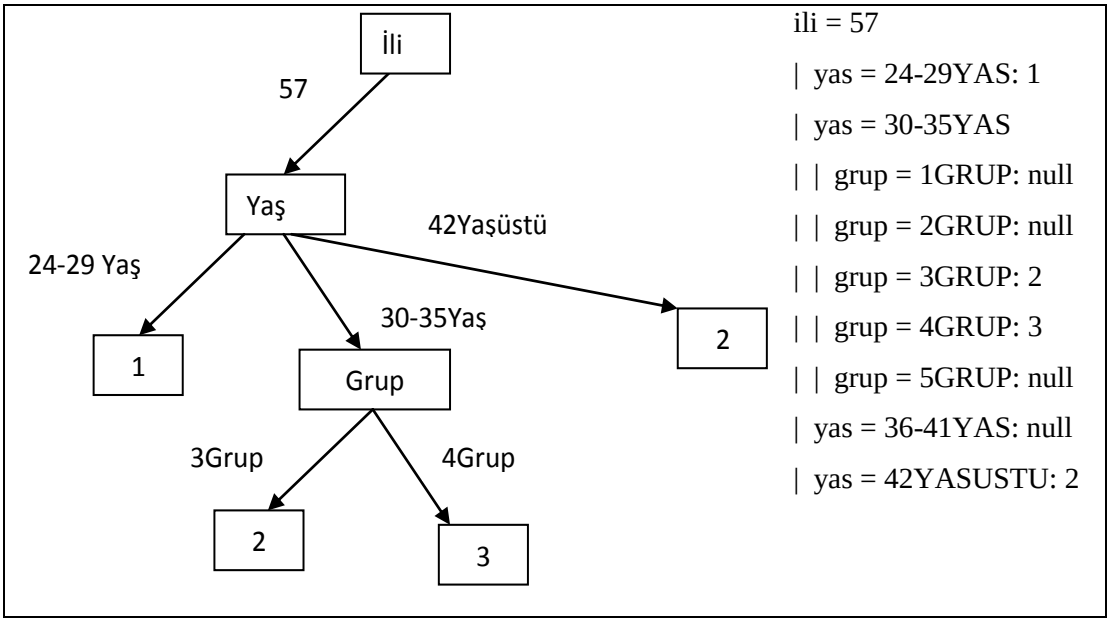
Şekil 4.15.'02' iline ilişkin karar ağacı

Şekil 4.15.' de olduğu gibi personelin çalıştığı il kodu =02 (Adıyaman) ve rolü TPY ise başarı düzeyi 2; rolü BPY ise başarı düzeyi 1 olarak en alt seviyede kalmaktadır. Bu sonuç bu ilde çalışan satış personelinin yetersiz olduğunu göstermektedir. Bu ilde çalışan satış personellerinin eğitimine ağırlık verilmesi gerekmektedir.



Şekil 4.16.‘56’ iline ilişkin karar ağacı

Eğer personelin çalıştığı il kodu=56 (Siirt) ise Şekil 4.16. ’ da gösterildiği gibi, oluşan karar ağacına göre 5. sınıf şubelerde yönetici unvanında çalışan personelin performans düzeyi 5. sınıftaki şube satış personeli başarı ortalamasının üzerinde yer alarak performans puanı 3 olmuştur. Ancak, aynı ilde görev yapan yetkili personeller ise ortalamanın oldukça altında performans göstermişlerdir. Bu sonuç bizlere portföy yöneticilerinin “yönetici” unvanındaki personel arasından seçilmesinin daha doğru olacağı sonucunu vermektedir.



Şekil 4.17.‘57’ iline ilişkin karar ağacı

Şekil 4.17‘ de gösterildiği gibi personelin çalıştığı il kodu=57 (Sinop) ise karar ağacındaki dallanma öncelikle personelin yaşından başlamıştır. Yani bu ilde, 30-35 yaş aralığında 4. sınıf şubelerde çalışan personel diğer personele göre daha yüksek performans göstermiştir. 24-29 yaş aralığındaki genç personellerin ise oldukça düşük performans gösterdiği görülmektedir.

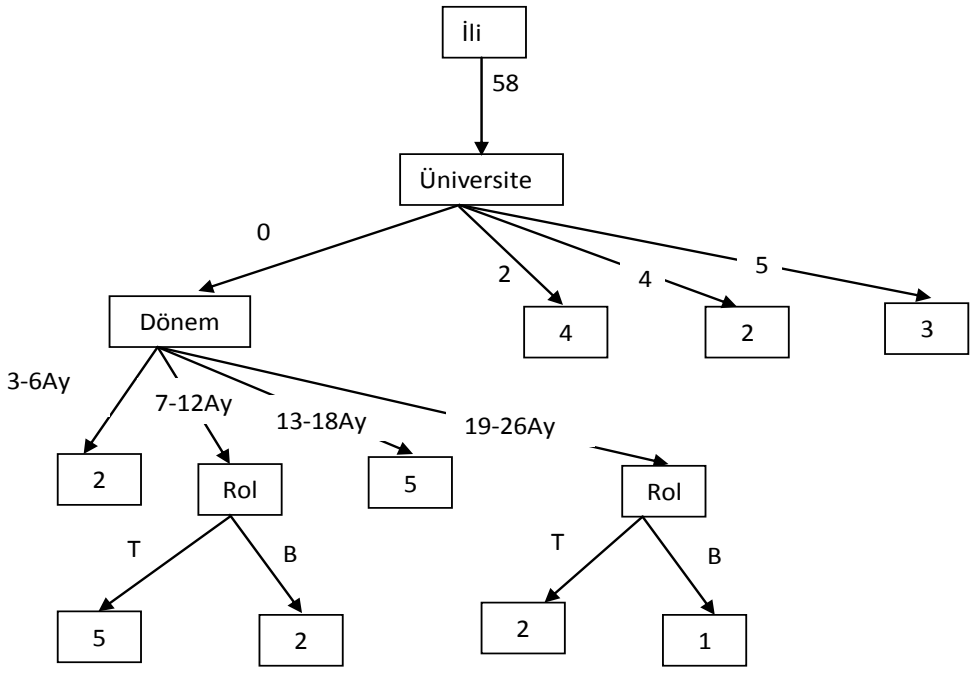
Çizelge 4.15. ‘58’ iline ilişkin oluşan karar kuralı

```
ili = 58
| universite = 0
| | donem = 3-6AY: 2
| | donem = 7-12AY
| | | rolu = T: 5
| | | rolu = B: 2
| | donem = 13-18AY: 5
| | donem = 19-26AY
| | | rolu = T: 2
| | | rolu = B: 1
```

Çizelge 4.15. (Devam) ‘58’ iline ilişkin oluşan karar kuralı

| universite = 1: null
| universite = 2: 4
| universite = 3: null
| universite = 4: 2
| universite = 5: 3
| universite = 6: null

Eğer personelin çalıştığı il kodu=58 (Sivas) ise karar ağacındaki dallanma üniversiteden devam etmektedir (Bkz. Şekil 4.18. ve Çizelge 4.15). Personelin üniversitesi=0 ise yani üniversite mezunu değilse ve 7-12 ay arasında bir TPY’ lik süresine sahipse performans düzeyi 5 ile en üst seviyede gerçekleşmiş, BPY ise aynı başarıyı gösterememiş ve performans puanı 2 ile ortalamanın altında kalmıştır. 13-18 ay arasında satış personeli olarak çalışanların performans düzeyi 5 ile en üst seviyede gerçekleşmiş ancak, 18 ay üzerindeki TPY/BPY’ lik süresinde başarı düzeyinde ciddi bir düşüş gözlemlenmiştir. Oluşan bu ağaç yapısı 58 kodlu ilde çalışan ve üniversite mezunu olmayan satış personelleri için ideal sürenin 13-18 ay arasında olduğunu göstermektedir.



Şekil 4.18.‘58’ iline ilişkin karar ağacı

Yine aynı ilde (Sivas), Ankara'daki üniversitelerden mezun olanlar (üniversite=2) için performans düzeyi 4 ile ortalamanın oldukça üzerinde iken, İzmir'deki üniversitelerden mezun olanların (üniversite=4) performans düzeyi 2 ile ortalamanın altında kalmış, yurtiçindeki diğer üniversitelerden mezun olanlar için (üniversite=5) performans düzeyi 3 ile ortalama civarında seyretmiştir. Diyebiliriz ki, Ankara' da ki üniversitelerden mezun olanlar bu ilde daha başarılı olmaktadır.

ID3 algoritmasının çıktılarına göre bazı iller için oluşan karar kuralları ise EK-9' da gösterilmiştir. Bu karar kurallarından ise sırasıyla aşağıda bahsedilmiştir.

Eğer personelin çalıştığı il kodu=60 (Tokat) ise 2. sınıf şubelerde çalışan 1-5 yıl arası toplam hizmet süresi olan personeller performans puanı 5 ile çok başarılı iken yine 2. sınıf şubelerde çalışan ve hizmet süresi 6 ile 10 yıl arasında değişen kadınların performans puanı 5, erkeklerinki ise 3' tür. Yani bankada 6 ile 10 yıl arasında hizmeti süresi olan kadınlar 2. sınıf şubelerde erkeklere göre daha başarılıdır.

Eğer personelin çalıştığı il kodu=67 (Zonguldak) ise; 3.sınıf şubeler için yöneticilerin çalışanlarına yönelik kişisel kanaatlerine göre dallanma gerçekleşmiştir. Yöneticilerin orta veya başarılı bulduğu personeller performans puanlarına göre ortalama başarının üzerinde yer almış ve performans puanları 3 olarak hesaplanmıştır. Yöneticisinin çok başarılı bulduğu personellerin performans puanı 4 olarak hesaplanmış ve gerçekten de ortalamanın oldukça üzerinde başarı göstermişlerdir. Yani bu ildeki 3. sınıf şubelerdeki yöneticilerin personellerini tarafsız olarak değerlendirebildiğini görmekteyiz.

Yine aynı ildeki 3. sınıf şubeler değerlendirildiğinde (grup=3GRUP) 42 yaş üzerindeki çalışanların çok başarısız olduğu ve performans düzeyinin 1 ile en alt seviyede kaldığı ancak 30-35 yaş arasındaki BPY' lerin performans seviyesinin 5, TPY' lerin ise 4 olduğu, 36-41 yaş aralığında ise performanslarının 5 ile en üst seviyede olduğu gözlemlenmiştir. Bu da bize 30 ile 41 yaş aralığında çalışanların bu ildeki 3. sınıf şubelerde çok başarılı olduğunu göstermektedir.

Eğer personelin çalıştığı il kodu=78 (Karabük) ise kadınlar performans puanı=3 ile ortalamanın üzerinde başarı gösterirken erkekler aynı oranda başarılı olamamıştır.

Eğer personelin çalıştığı il kodu=79 (Kilis) ise,7-18 ay aralığında satış personeli olarak çalışanların performans düzeyi 1 ile en alt seviyede olmuşken 19 ay ve üzerinde bu görevi yürütenlerin performans düzeyi 2 olmuştur. Yani bu ilde genel olarak performans düzeyi düşük olmasına karşın satış personelinin sürekliliği önemlidir.

Eğer personelin çalıştığı il kodu=80 (Osmaniye) ise; 24-29 yaş aralığındaki çalışanların performans düzeyi 4 iken, 36-41 yaş aralığında SPK lisansı olanların performans düzeyi 4, SPK lisansı olmayanların ise performans düzeyi 3 olmuştur. Yani bu ilde 36-41 yaş aralığında SPK belgesi olanlar başarılı olmuştur.

Benzer sonuçlar 81 il için de elde edilmiş ve sonuçları yöneticilere iletilmiştir. Elde edilen sonuçlar neticesinde, her il için öne çıkan özellikler belirlenmiştir. Sınıflandırma sonucunda iller bazında öne çıkan özelliklerin tespiti ve performans seviyelerinin belirlenmesi ile yöneticilere personel seçimi sürecinde karar desteği sağlanmıştır. Elde edilen bilgiler çerçevesinde, yöneticiler istenilen bir ilde performansı başarılı olan personellerin öne çıkan özelliklerine bakarak o ile ataması düşünülen personelin belirlenmesinde veri madenciliği sonuçlarından yararlanabilecektir.

5. SONUÇ VE DEĞERLENDİRME

Büyük miktarlardaki veriye ulaşmanın kolaylaştığı günümüz bilgi endüstrisinde "bilgi çıkarımı" giderek önem kazanmış, verilerin yararlı bilgilere dönüştürülebilmesi ihtiyacı ile birlikte veri madenciliği giderek dikkat çekmeye başlamıştır. Başta bankacılık, finans ve pazarlama olmak üzere sağlık, insan kaynakları, telekomünikasyon, sigortacılık gibi pek çok alanda veri madenciliği uygulanmaktadır.

Firmaların kendilerine rekabet avantajı sağlamak için en önemli kaynağı ise şüphesiz insandır. Bu çalışma ise personel seçiminde karar kuralları oluşturmak ve etkili bir personel değerlendirme süreci ile firmaların en önemli kaynağı olan insan kaynağını etkin ve verimli şekilde kullanmak üzerine odaklanmıştır. Veri madenciliği yöntemlerinden kümeleme ve sınıflandırma ile etkili bir personel seçim mekanizması geliştirilerek özellikle personel seçimi sürecinde fayda sağlanması amaçlanmıştır.

Literatürde insan kaynakları yönetimine ilişkin çok az sayıda veri madenciliği uygulaması yer almaktadır. Özellikle bankacılık sektörü gibi çok sayıda personel çalıştıran bir sektörde, şube personellerinin seçimine yönelik veri madenciliği uygulamasına literatürde rastlanılmamıştır. Bu çalışma ile, bankacılık sektöründe personel seçimine ilişkin karar kuralları oluşturularak bu alandaki boşluğun doldurulması sağlanmıştır.

Bu çalışmada, bankacılık sektöründe çalışan satış personellerinin performansları değerlendirilmiş, kümeleme yöntemlerinden k-ortalama ile personellerin performans başarı düzeylerine göre sınıflandırılması sağlanmıştır. Elde edilen performans düzeyleri daha sonra sınıflandırma ile karar kuralları oluşturmada çıktı olarak kullanılmıştır. Çalışanların yaş, medeni hal, cinsiyet gibi demografik bilgileri, öğrenim durumu, yabancı dili, SPK belgesi gibi eğitim durumlarına ilişkin bilgileri, çalıştığı şubesine ve iş yaşamındaki pozisyonuna ilişkin bilgileri dikkate alınarak veri madenciliğinde sınıflandırma algoritmaları kullanılmıştır. WEKA' da gerçekleştirilen

madencilik uygulamasında sınıflandırma algoritmalarından ID3, J4.8, PART, Saf Bayes, OneR ve MultilayerPerceptron algoritmaları karşılaştırılmıştır. WEKA çıktılarına göre ID3 algoritması hatalı sınıflandırılan kayıt oranı ve ortalama mutlak hata açısından en iyi sonucu sağlamış ve ID3 algoritmasının sonuçları üzerinde durulmuştur.

ID3 algoritmasının çıktıları incelendiğinde ise karar ağacında dallanmanın personelin çalıştığı ilden başladığı gözlemlenmiştir. Yani karar kuralları oluşturmada ilk dikkat edilecek nokta olarak "çalışılan il" ön plana çıkmıştır. Daha sonraki dallanmaların ise illere göre değişiklik gösterdiği ve her ilde farklı özelliklerin ön plana çıkabildiği gözlemlenmiştir. Oluşan karar kuralları ile her ildeki personelin performans başarı düzeyleri belirlenmiş, böylece yöneticilerin personel değerlendirme ve personel seçimi sürecinde karar kurallarına sahip olması sağlanarak personel seçimi ve performans değerlendirme sürecinde fayda sağlanmıştır.

Veri madenciliği uygulaması neticesinde çalışanların performanslarına göre değerlendirilmesi yapılmış, hangi özelliklerdeki personelin hangi şubede ne oranda başarılı olduğuna yönelik kurallar oluşturulmuştur. Bu kurallar dikkate alınarak, bir personelin özelliklerine göre hangi şubelere atanabileceği ya da ataması düşünülen şubede hangi düzeyde performans gösterebileceği öngörülebilmektedir. Bu sayede, çalışanlara daha gerçekçi performans hedefleri de verilebilecektir.

Sonuç olarak, doğru insanın doğru özellikler ile doğru yerde kullanılması ve doğru hedeflere odaklanması ile ciddi anlamda bir fayda sağlanacak, gerek çalışanlar gerekse de işletme için verimlilik ve etkinlik artacaktır. Tez kapsamında gerçekleştirilen çalışma sonuçları, veri madenciliğinin insan kaynakları yönetimindeki uygulanabilirliğini göstermiştir. Benzer uygulamalar, bankacılık sektörünün yanı sıra finans, sigortacılık, pazarlama gibi farklı sektörlerde de gerçekleştirilebilinir.

KAYNAKLAR

Abascal E., Lautre I.G., Mallor F., "Data mining in bicriteria clustring problem", ***European Journal of Operational Researh,*** 173: 705-716 (2006).

Adriaans, P., Zantinge, D., "Data Mining", ***Addison Wesley Longman***, Harlow, 159 (1996).

Aktürk H., Korukoğlu S., " Veri Madenciliği Teknolojisini Kullanarak Fiyat Değişimlerinde Paralellik Gösteren Hisse Senetlerinin Bulunması Ve Risk Azaltılması"***, Akademik Bilişim,*** Çanakkale, 2-3 (2008).

Akbulut S., "Veri madenciliği teknikleri ile bir kozmetik markanın ayrılan müşteri analizi ve müşteri segmentasyonu", Yüksek Lisans Tezi, ***Gazi Üniversitesi Fen Bilimleri Enstitüsü***, Ankara, 21-22 (2005).

Akpınar H., "Veritabanlarında bilgi keşfi ve veri madenciliği", ***İstanbul Üniversitesi İşletme Fakültesi Dergisi***, 29: 1-22 (2000).

İnternet: Alpaydın, E., "Zeki veri madenciliği: Ham veriden altın bilgiye ulaşma yöntemleri", www.cmpe.boun.edu.tr/~ethem/files/papers/veri-maden_2k-notlar.doc (1999).

Altıntaş T., "Veri madenciliği metotlarından olan kümeleme algoritmalarının uygulamalı etkinlik analizi", Yüksek Lisans Tezi , ***Sakarya Üniversitesi Fen Bilimleri Enstitüsü*** , Sakarya, 14-16 (2006).

Aydoğan F., "E-ticarette veri madenciliği yaklaşımlarıyla müşteriye hizmet sunan akıllı modüllerin tasarımı ve gerçekleştirimi", Yüksek Lisans Tezi, ***Hacettepe Üniversitesi Fen Bilimleri Enstitüsü,*** Ankara, 10-18, 66-74, 88-95 (2003).

Baykasoğlu A., Özbakir L., "MEPAR-miner:Multi-expression programming for classification rule mining", ***Europan Journal of Operational Research*** , 183: 767-784 (2007).

Ben-David, A., Sterling L., "Generating rules from examples of human multiattribute decision making should bu simple", ***Expert Systems with Application,*** 31: 390-396 (2006).

Berson, A., Smith, S. and Thearling, K., " Buildind data mining applications for CRM", ***McGraw Hill***, USA, 510 (1999).

Berry, M., Linoff, G., "Data Mining Techniques for Marketing Sales and Customer Support", ***John Wiley & Sons,*** 2-12 (1997).

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(1): 280-290 (2008).

Chien C.-F., Chen L.-F., "Using Rough Set Theory To Recruit And Retain High Potential Talents For Semiconductor Manufacturing ", ***IEEE Transactions On Semiconductor Manufacturing***, 20 (4) : 528-541 (2007).

Cho V., Ngai E. W. T., "Data mining for selection of insurance sales agents" , ***Expert Systems***, 20(3): 123-132 (2003).

Chu, W., Lin T.Y., "Foundations and Advances in Data Mining 1st ed." ,***Springer Publisherss***, USA, 25, 100 (2005).

Çetinyokuş T., "Veri Küplerinin Bütünleşik Kullanımına Yönelik Yeni Bir OLAP Mimarisi", Doktora Tezi, ***Gazi Üniversitesi Fen Bilimleri Enstitüsü***, Ankara, 7-10 (2008).

Dolgun M. Ö., Zor İ., "Bir Alışveriş Merkezinden Yapılan Satışlar İçin Sepet Analizi" , ***SPSS Türkiye***, 1-4 (2006).

Fayyad, U., Piatetsky-Shapiro G., Smyth P., "From Data Mining to Knowledge Discovery in Databases," ***American Association for Artificial Intelligence***, 3(17): 37-54 (1996).

Frank, E., Hall, M., Holmes, G., Kirkby, R., Pfahringer, B., Witten, I., H., "WEKA: A Machine Learning Workbench for Data Mining", ***University of Waikato***, New Zealand**,** 7-10 (2004).

Fu S.-Y. K., Anderson D., Courtney M., Hu W., "The relationship between culture, attitude, social networks and quality of life in midlife Austrilian and Taiwanese citizens", ***Maturitas 10.1016:*** (2007).

Giudici, P., "Applied Data Mining: Statistical Methods for Business and Industry 1st ed.", ***John Wiley & Sons***, England, 1-15, 85-110 (2003).

Guha S., Rastogi R., Shim K., "ROCK: A Robust Clustring Algorithm For Categorial Attributes", ***Information Systems,*** 25(5): 345-366 (2000).

Han, J. ve Kamber, M., "Data Mining: Concepts and Techniques 1st ed.", ***Morgan Kaufmann***, USA, 3-16, 279-326 (2001).

Han J. - Fu Y., "Mining Multiple-Level Association Rules in Large Databases",***IEEE Transactions on Knowledge and Data Engineering***, 11 (5): 798-805 (1999).

Hand D.J., "Data mining: statistics and more ?", ***The American Statistician***, 52: 112-118 (1998).

Holsheimer M. and Siebes A., "Data mining: The search for knowledge in databases.", ***Technical Report*** , CWI, Netherlands, 12 (1994).

Hsia T.-C., Shie A.-J.,Chen L.-C., "Course planning of extension education to meet market demand by using data mining techniques – an example of Chinkuo technology university in Taiwan" ***,Expert Systems with Applications***, 34: 596–602 (2008).

Hsu C.-C., Chen Y.-C., "Mining of mixed data with appliation to catalog marketting", ***Expert Systems with Applications,*** 32: 12-23 (2007).

Hsu M.-H., "A personalized English learning recommender system for ESL students", ***Expert Systems with Applications,*** 34: 683–688 (2008).

Jacobs P., "Data Mining: What general managers need to know", ***Harvard Management Update***, 4 (10): 8-9 (1999).

Kalıkov A., "Veri madenciliği ve bir e-ticaret uygulaması", Yüksek Lisans Tezi, ***Gazi Üniversitesi Fen Bilimleri Enstitüsü***, Ankara, 22-38 (2006).

Kantardzic M., "Data Mining: Concepts, Models, Methods, and Algorithms", ***IEEE Press & John Wiley,*** USA, 1-18, 154-155 (2002).

KDnuggets, "In what industries/sectors were your data mining clients in 2007-2008?", http://www.kdnuggets.com/polls/2008/industry-data-mining-clients.htm (2008).

Kirkos E., Spathis C., Manolopoulos Y., "Data Mining techniques for detection of fraudulent financial statements", ***Expert Systems with Applications*** 32: 995-1003 (2007).

Kovalerchuk, B., "Data Mining in Finance: Advances in Relational and Hybrid Methods", ***Kluwer Academic,*** New York, 1-19 (2000).

Liao S.-H., Wen C.-H., "Artificial neural networks classification and clustring of methologies and applications - literature analysis from 1995 to 2005", ***Expert Systems With Applications,*** 32: 1-11 (2007).

Özekes S., "Veri Madenciliği Modelleri ve Uygulama Alanları", ***İstanbul Ticaret Üniversitesi Dergisi, 2003:*** 65-82 (2001).

Piramuthu S., "Evaluating feature selection methods for learning in data mining applications", ***Thirty-First Annual Hawai International Conference on System Sciences***, 5: 294 (1998).

Plasse M., Niang N., Saporta G., Villeminot A., Leblond L., "Combined use of association rules mining and clustring methods to find relevant links between binary rare attributes in a large data set" , ***Computationel Statistics & Data Analysis,*** 52: 596-613 (2007).

Springer, "The Knowledge Discovery Process", http://www.springer.com/cda/content/document/cda_downloaddocument/9780387333335-c2.pdf?SGWID=0-0-45-424299-p173660317 (2007).

Questier F., Put R., Coomans D., Walczak B., Heyden Y.V., "The use of CART and multivariate regression trees for supervised and unsupervised feature selection",***Chemometrics And Intellegent Labaratory Systems*** , 76: 45-54 (2005).

Seow H.-V., Thomas L.C. , "To ask or not to ask, that is the question", ***European Journal of Operational Reserah***, 183: 1513-1520 (2007).

Türkiye Bankalar Birliği, "50. Yılında Türkiye Bankalar Birliği ve Türkiye'de Bankacılık Sistemi 1958-2007", ***Türkiye Bankalar Birliği, İstanbul,*** 98-99 (2008).

Two Crows Corporation; "Introduction to Data Mining and Knowledge Discovery," http://www.twocrows.com/intro-dm.pdf (2005).

Witten, I., H., Frank, E., "Data Mining: Practical Machine Learning Tools and Techniques 2nd ed.", ***Morgan Kaufmann***, USA, 365-415 (2005).

Yılmaz L., "A Decision Support System Using Data Mining", Yüksek Lisans Tezi, ***Yeditepe Üniversitesi***, İstanbul, 16-22 (2002).

Zaki, M. J., "Parallel and Distributed Association Mining: A Survey", ***IEEE Concurrency Special issue on Parallel Mechanisms for Data Mining***, 7 (5), 14-25 (1999).

EKLER

EK-1 K-ortalama algoritması k=5 için sonuç özeti

Çizelge 1.1. K-ortalama algoritması küme sayısı=5 için sonuç özeti

```
=== Run information ===


Scheme:       weka.clusterers.SimpleKMeans -N 5 -S 10
Relation:                                       human_performance_data-
weka.filters.unsupervised.attribute.Remove-R1-2,4-20
Instances:    1883
Attributes:   1
              puan
Test mode:    evaluate on training data

=== Model and evaluation on training set ===
kMeans
======

Number of iterations: 20
Within cluster sum of squared errors: 3.457392603809491

Cluster centroids:

Cluster 0
        Mean/Mode:  1.664
        Std Devs:   0.1755
Cluster 1
        Mean/Mode:  0.8522
        Std Devs:   0.0629
Cluster 2
        Mean/Mode:  1.0494
        Std Devs:   0.0613
Cluster 3
        Mean/Mode:  1.2859
        Std Devs:   0.0834
Cluster 4
        Mean/Mode:  0.6049
        Std Devs:   0.1044
```

EK-1 (Devam) K-ortalama algoritması k=5 için sonuç özeti

Çizelge 1.1. (Devam) K-ortalama algoritması küme sayısı=5 için sonuç özeti

```
Clustered Instances

0       122 (  6%)
1       605 ( 32%)
2       597 ( 32%)
3       326 ( 17%)
4       233 ( 12%)
```

EK-2 Çalışmada kullanılan özellikler

Çizelge 2.1. Modellemede kullanılan verilere ilişkin özellikler ve tanımlamaları

Özellik	**Tanımlama**
İli	01: Adana 02: Adıyaman … 81: Düzce
Bölgesi	0: Ankara 1: İstanbul Avrupa yakası 2: Ege 3: Çukurova 4: İstanbul Avrupa yakası 5: Marmara 6: Karadeniz 7: Doğu Anadolu 8: İç Anadolu 9: Akdeniz
Grup (Şube Sınıfı)	1GRUP: A sınıfı (1. sınıf) şubeler 2GRUP: B sınıfı (2. sınıf) şubeler 3GRUP: C sınıfı (3. sınıf) şubeler 4GRUP: D sınıfı (4. sınıf) şubeler 5GRUP: E sınıfı (5. sınıf) şubeler
Rol	T: Ticari portföy yöneticisi B: Bireysel portföy yöneticisi
Dönem Sayısı	3-6AY 7-12AY 13-18AY 19-26AY
Unvan	YÖNETİCİ MEMUR UZMAN YETKİLİ

EK-2 (Devam) Çalışmada kullanılan özellikler

Çizelge 2.1. Modellemede kullanılan verilere ilişkin özellikler ve tanımlamaları

Özellik	**Tanımlama**
Hizmet Süresi	1-5YIL 6-10YIL 11-15YIL 16YILÜSTÜ
Emeklilik	1: 2008 ve öncesi 2: 2009 yılı 3: 2010 yılı 4: 2011 ve sonrası yıllarda
Tezkiye (yönetici değerlendirmesi)	YOK: Yok 1: Yetersiz 2: Orta 3: Başarılı 4: Çok başarılı
Cinsiyet	K: Kadın E: Erkek
Medeni Hal	1: Bekar 2: Evli ve çocuklu 3: Evli 4:Boşanmış ve çocuklu
Yaş:	24-29YAŞ 30-35YAŞ 36-41YAŞ 42YAŞÜSTÜ
Öğrenim durumu	1: Lise veya altı 2: 2 yıllık yüksek okul 3: Üniversite 4: Yüksek lisans

EK-2 (Devam) Çalışmada kullanılan özellikler

Çizelge 2.1. Modellemede kullanılan verilere ilişkin özellikler ve tanımlamaları

Özellik	**Tanımlama**
Üniversite Kategorisi	0: Üniversite mezunu değil 1: İşe alımda Banka için öncelikli üniversitelerden mezun 2: Ankara' daki diğer üniversitelerden mezun 3: İstanbul' daki diğer üniversitelerden mezun 4: İzmir' deki üniversitelerden mezun 5: Diğer yurtiçi üniversitelerden mezun 6: Yurtdışı üniversitelerinden mezun
Fakülte Kategorisi	0: Yok (Üniversite mezunu değil) 1: Açıköğretim fakültesi 2: İktisadi ve idari bilimler / bankacılık fakülteleri 3: Diğer fakülteler
SPK belgesi	VAR YOK
Yabancı dil	İNGİLİZCE ALMANCA FRANSIZCA YOK
Yabancı dil seviyesi	İYİ ORTA YOK
Puan (performans başarı düzeyi)	1: Ortalamanın çok altında 2: Ortalamanın altında 3: Ortalama civarında 4: Ortalamanın üstünde 5: Ortalamanın çok üstünde

EK-3 ID3 algoritması için sonuç özeti

Çizelge 3.1. ID3 algoritması için sonuç özet tablosu

```
=== Summary ===

Correctly Classified Instances        1852               98.3537 %
Incorrectly Classified Instances        31                1.6463 %
Kappa statistic                          0.9779
Mean absolute error                      0.0068
Root mean squared error                  0.0583
Relative absolute error                  2.2753 %
Root relative squared error             15.0854 %
Total Number of Instances             1883

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
  0.996     0.003      0.979     0.996     0.987      1         1
  0.997     0.014      0.971     0.997     0.984      1         2
  0.988     0.005      0.988     0.988     0.988      1         3
  0.954     0.001      0.997     0.954     0.975      1         4
  0.951     0          1         0.951     0.975      1         5

=== Confusion Matrix ===

   a   b   c   d   e   <-- classified as
 232   0   1   0   0 |   a = 1
   2 603   0   0   0 |   b = 2
   0   7 590   0   0 |   c = 3
   3   7   5 311   0 |   d = 4
   0   4   1   1 116 |   e = 5
```

EK-4 J4.8 algoritması için sonuç özeti

Çizelge 4.1. J4.8 algoritması için sonuç özet tablosu

```
=== Summary ===

Correctly Classified Instances        1241               65.9055 %
Incorrectly Classified Instances       642               34.0945 %
Kappa statistic                          0.5342
Mean absolute error                      0.1739
Root mean squared error                  0.2949
Relative absolute error                 58.2177 %
Root relative squared error             76.3077 %
Total Number of Instances             1883

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
  0.592     0.047      0.639     0.592     0.615      0.933    1
  0.774     0.196      0.651     0.774     0.707      0.888    2
  0.714     0.161      0.673     0.714     0.693      0.886    3
  0.479     0.053      0.655     0.479     0.553      0.9      4
  0.434     0.014      0.688     0.434     0.533      0.957    5

=== Confusion Matrix ===

   a   b   c   d   e   <-- classified as
 138  49  31  11   4 |   a = 1
  34 468  70  25   8 |   b = 2
  25 115 426  26   5 |   c = 3
  17  63  83 156   7 |   d = 4
   2  24  23  20  53 |   e = 5
```

EK-5 PART algoritması sonuç özeti

Çizelge 5.1. PART algoritması sonuç özet tablosu

```
=== Summary ===

Correctly Classified Instances        1268               67.3394 %
Incorrectly Classified Instances       615               32.6606 %
Kappa statistic                          0.5565
Mean absolute error                      0.1747
Root mean squared error                  0.2955
Relative absolute error                 58.4646 %
Root relative squared error             76.4693 %
Total Number of Instances             1883

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
  0.639     0.051      0.639     0.639     0.639      0.942    1
  0.75      0.167      0.681     0.75      0.714      0.886    2
  0.73      0.141      0.707     0.73      0.718      0.892    3
  0.589     0.076      0.617     0.589     0.603      0.91     4
  0.303     0.01       0.673     0.303     0.418      0.951    5

=== Confusion Matrix ===

  a   b   c   d   e   <-- classified as
 149  44  23  14   3 |  a = 1
  31 454  74  41   5 |  b = 2
  29  92 436  38   2 |  c = 3
  12  54  60 192   8 |  d = 4
  12  23  24  26  37 |  e = 5
```

EK-6 Saf Bayes algoritması sonuç özeti

Çizelge 6.1. Saf Bayes algoritması sonuç özet tablosu

```
=== Summary ===

Correctly Classified Instances         805               42.7509 %
Incorrectly Classified Instances      1078               57.2491 %
Kappa statistic                          0.2231
Mean absolute error                      0.264
Root mean squared error                  0.3719
Relative absolute error                 88.3731 %
Root relative squared error             96.2276 %
Total Number of Instances             1883

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
  0.288     0.058      0.411     0.288     0.338      0.77     1
  0.481     0.278      0.45      0.481     0.465      0.672    2
  0.514     0.298      0.445     0.514     0.477      0.669    3
  0.282     0.083      0.414     0.282     0.336      0.706    4
  0.393     0.065      0.296     0.393     0.338      0.822    5

=== Confusion Matrix ===

   a   b   c   d   e   <-- classified as
  67  78  50  18  20 |   a = 1
  48 291 194  37  35 |   b = 2
  27 180 307  54  29 |   c = 3
  15  72 117  92  30 |   d = 4
   6  25  22  21  48 |   e = 5
```

EK-7 OneR algoritması sonuçları

Çizelge 7.1. OneR algoritması sonuç tablosu

```
=== Summary ===

Correctly Classified Instances         745               39.5645 %
Incorrectly Classified Instances      1138               60.4355 %
Kappa statistic                          0.1324
Mean absolute error                      0.2417
Root mean squared error                  0.4917
Relative absolute error                 80.9141 %
Root relative squared error            127.2237 %
Total Number of Instances             1883

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
  0.137     0.018      0.516     0.137     0.217      0.56     1
  0.512     0.334      0.421     0.512     0.462      0.589    2
  0.606     0.481      0.369     0.606     0.459      0.563    3
  0.11      0.033      0.414     0.11      0.174      0.539    4
  0.041     0.007      0.294     0.041     0.072      0.517    5

=== Confusion Matrix ===

   a   b   c   d   e   <-- classified as
  32 103  91   6   1 |   a = 1
  18 310 259  14   4 |   b = 2
   6 208 362  18   3 |   c = 3
   4  82 200  36   4 |   d = 4
   2  34  68  13   5 |   e = 5
```

EK-8 MultilayerPerceptron algoritması sonuç özeti

Çizelge 8.1. MultilayerPerceptron algoritması sonuç özet tablosu

```
=== Summary ===

Correctly Classified Instances         877               46.5746 %
Incorrectly Classified Instances      1006               53.4254 %
Kappa statistic                          0.2471
Mean absolute error                      0.2171
Root mean squared error                  0.4398
Relative absolute error                 72.6722 %
Root relative squared error            113.795  %
Total Number of Instances              1883

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
  0.193     0.033      0.455     0.193     0.271      0.708    1
  0.798     0.54       0.412     0.798     0.543      0.698    2
  0.442     0.135      0.603     0.442     0.51       0.741    3
  0.16      0.012      0.743     0.16      0.263      0.682    4
  0.27      0.04       0.32      0.27      0.293      0.797    5

=== Confusion Matrix ===

   a   b   c   d   e   <-- classified as
  45 153  16   5  14 |   a = 1
  14 483  89   3  16 |   b = 2
  19 288 264   6  20 |   c = 3
  12 181  61  52  20 |   d = 4
   9  68   8   4  33 |   e = 5
```

EK-9 ID3 algoritması sonucunda bazı iller için oluşan karar kuralları

Çizelge 9.1. ID3 algoritması sonucunda bazı iller için oluşan karar kuralları

```
ili = 60
| grup = 1GRUP: null
| grup = 2GRUP
| | hizmet_suresi = 1-5YIL: 5
| | hizmet_suresi = 6-10YIL
| | | cinsiyet = K: 5
| | | cinsiyet = E: 3
| | hizmet_suresi = 11-15YIL
| | | yas = 24-29YAS: null
| | | yas = 30-35YAS: 2
| | | yas = 36-41YAS: 3
| | | yas = 42YASUSTU: null
| | hizmet_suresi = 16YILUSTU: 4
| grup = 3GRUP
| | yas = 24-29YAS: null
| | yas = 30-35YAS
| | | rolu = T: 4
| | | rolu = B: 5
| | yas = 36-41YAS: 5
| | yas = 42YASUSTU: 1
| grup = 4GRUP
| | rolu = T: 4
| | rolu = B: 2
| grup = 5GRUP: 2
ili = 67
| grup = 1GRUP: null
| grup = 2GRUP
| | donem = 3-6AY: 4
| | donem = 7-12AY
| | | yas = 24-29YAS: null
```

EK-9 (Devam) ID3 algoritması sonucunda bazı iller için oluşan karar kuralları

Çizelge 9.1. ID3 algoritması sonucunda bazı iller için oluşan karar kuralları

| | | yas = 30-35YAS: 2
| | | yas = 36-41YAS
| | | | rolu = T: 3
| | | | rolu = B: 5
| | | yas = 42YASUSTU: 3
| | donem = 13-18AY: null
| | donem = 19-26AY: 4
| grup = 3GRUP
| | tezkiye_ortalamasi = YOK: null
| | tezkiye_ortalamasi = 1: null
| | tezkiye_ortalamasi = 2: 3
| | tezkiye_ortalamasi = 3: 3
| | tezkiye_ortalamasi = 4: 4
| grup = 4GRUP
| | rolu = T: 2
| | rolu = B: 1
| grup = 5GRUP: 3
ili = 78
| cinsiyet = K: 3
| cinsiyet = E
| | rolu = T: 2
| | rolu = B: 1
ili = 79
| donem = 3-6AY: null
| donem = 7-12AY: 1
| donem = 13-18AY: 1
| donem = 19-26AY: 2
ili = 80
| yas = 24-29YAS: 4

EK-9 (Devam) ID3 algoritması sonucunda bazı iller için oluşan karar kuralları

Çizelge 9.1. ID3 algoritması sonucunda bazı iller için oluşan karar kuralları

| yas = 30-35YAS: 1
| yas = 36-41YAS
| | spk = YOK: 3
| | spk = VAR: 4
| yas = 42YASUSTU: 2

ÖZGEÇMİŞ

Kişisel Bilgiler

Soyadı, adı : BİLEN, Hamdi

Uyruğu : T.C.

Doğum tarihi ve yeri : 02.12.1983 Ankara

Medeni hali : Bekar

Telefon : 0 (536) 430 45 35

e-mail : hamdibilen@yahoo.com

Eğitim

Derece	Eğitim Birimi	Mezuniyet tarihi
Lisans	Gazi Üniversitesi/ Endüstri Müh.,Ankara	2006
Lise	Başkent Lisesi (YDA), Ankara	2001

İş Deneyimi

Yıl	Yer	Görev
2006 -	Özel bir banka	Uzman Yrd.

Yabancı Dil

İngilizce

Hobiler

Futbol, Bilardo, Yüzme