T.C.
MALTEPE ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

KÜMELEME ANALİZİ KULLANILARAK BENZİN İSTASYONLARININ OPERASYONEL DEĞERLENDİRİLMESİ

Halil Çağdaş DARAKÇI

Yüksek Lisans Tezi

Tez Danışmanı
Yrd. Doç. Dr. Şenol Zafer ERDOĞAN

İSTANBUL – 2011

T.C.
MALTEPE ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

KÜMELEME ANALİZİ KULLANILARAK BENZİN İSTASYONLARININ OPERASYONEL DEĞERLENDİRİLMESİ

YÜKSEK LİSANS TEZİ

Halil Çağdaş DARAKÇI

Tez Danışmanı
Yrd. Doç. Dr. Şenol Zafer ERDOĞAN

İSTANBUL – 2011

Bu tez çalışması, Maltepe Üniversitesi Fen Bilimleri Enstitüsü Yönetim Kurulu'nun ..... / ..... / ..... tarih ve ..... / ..... sayılı kararıyla oluşturulan jüri tarafından ***Bilgisayar Mühendisliği Yüksek Lisansı Tezi*** olarak kabul edilmiştir.

JÜRİ

Yrd. Doç. Dr. Şenol Zafer ERDOĞAN
Danışman

Yrd. Doç. Dr. Turgay Tugay BİLGİN
Üye

Yrd. Doç. Dr. Birim Balcı Demirci
Üye

ÖZET

Yüksek Lisans Tezi, Kümeleme Analizi Kullanılarak Benzin İstasyonlarının Operasyonel Değerlendirilmesi, T.C. Maltepe Üniversitesi, Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Anabilim Dalı.

Son yıllarda bilgi teknolojilerinin hızla büyümesi ile birlikte saklanan veri miktarları sürekli bir artış göstermektedir. Bir anlam ifade etmeyen çok fazla sayıda veri yığınları oluşmaya başlamıştır. Şirketler rekabet koşullarını arttırabilmek için büyük boyutlardaki veriler arasından faydalı bilgiler çıkarabilme yoluna yönelmişlerdir. Veri madenciliği ile büyük veri kümelerinden yararlı bilgiler çıkarılabilir hale gelmiştir.

Veri madenciliği sayesinde müşterilerin ilgi alanları, ürün pazarlama stratejileri oluşturma, performans ve verimlilik analizi yapabilme gibi hususlarda kurumlara yardımcı olunur.

Bu çalışmada veri madenciliği kavramının nasıl oluştuğu, bir veri madenciliği uygulamasında hangi süreçler olduğu, veri madenciliğinde kullanılan metodlar ve teknikler, k-ortalamalar kümeleme yöntemi ve bir akarkayıt firmasına ait satış verilerinden benzin istasyonlarının operasyonel değerlendirmesi anlatılmaktadır.

Bu tez 2011 yılında yapılmıştır ve 85 sayfadan oluşmaktadır.

Anahtar Kelimeler: Veri Madenciliği, Kümeleme Analizi, K-ortalamalar algoritması

ABSTRACT

Master Thesis, Operational Evaluations of Gas Stations Using Cluster Analysis. T.C. Maltepe University, Graduate School of Natural and Applied Sciences, Department of Computer Engineering.

In recent years, along with the rapid growth of information technology shows a steady increase in the amount of data is stored. Large amount of data that does not make sense to the masses began to form. Companies, in order to increase the competitive conditions in order to obtain useful information from huge data started efforts. Useful information can be obtained by data mining from large data sets have become.

Data mining is help assistance to corporations about subject which are interest of customers, marketing product strategies, to make the performance and efficiency analysis.

In this study of how the concept of data mining, which is a data mining application that processes occur, methods and techniques used in data mining, k-means clustering method and the operational assessment of petrol stations from a fuel oil company's sales data.

This thesis has been completed in 2011 and consists of 85 pages.

Keywords: Data Mining, Clustering Analysis, K-means Algorithm

TEŞEKKÜR

Tezime başlamamdan bitirmeme kadar bana devamlı yol gösteren ve destek veren tez danışmanım Sayın Yrd. Doç. Dr. Şenol Zafer Erdoğan'a, Yrd. Doç. Dr. T. Turgay Bilgin'e, Sayın Canan Beşel'e, Sayın İlknur Dönmez'e ve bana her zaman destek olan aileme teşekkürü bir borç bilirim.

İÇİNDEKİLER

ÖZET …… IV
ABSTRACT …… V
TEŞEKKÜR …… VI
İÇİNDEKİLER ……. VII
ŞEKİLLER …… IX
TABLOLAR …… X
DENKLEMLER……. XI
1. VERİ, BİLGİ VE ÖĞRENME…….. 1
1.1. Veri …… 1
1.2. Bilgi …… 1
1.2.1. Yüzeysel Bilgi…… 2
1.2.2. Çok Boyutlu Bilgi…….. 2
1.2.3. Saklı Bilgi…….. 2
1.2.4. Derin Bilgi……. 3
1.3. Öğrenme…… 3
2. VERİ TABANI, VERİ AMBARI VE SORGU ARAÇLARI…….. 5
2.1. Veri Tabanı…….. 5
2.1.1. Hiyerarşik Veri Tabanları…….. 7
2.1.2. İlişkisel Veri Tabanları…….. 8
2.1.3. Nesne Yönelimli Veri Tabanları…… 9
2.2. Veri Ambarı……. 9
2.2.1. Veri Ambarının Karakteristik Özellikleri…….. 10
2.2.2. Veri AmbarınınYapısı…… 11
2.2.3. Veri Ambarı Projelerinde Kaçınılması Gereken Hatalar……... 12
2.2.4. Veri Ambarı İhtiyacı…….. 13
2.3. Veri Ambarı ile Veri Tabanı Arasındaki Farklar…… 14
2.4. Sorgu Araçları …… 14
3. VERİ MADENCİLİĞİ KAVRAMI…….. 15
3.1. Veri Madenciliği Tanımı…….. 15
3.2. Veri Madenciliği Tarihçesi……... 17
4. VERİ TABANLARINDA BİLGİ KEŞFİ SÜRECİ…….. 19
4.1. Problemin Tanımlanması…….. 20
4.2. Verilerin Hazırlanması…….. 21
4.2.1. Toplama…….. 21
4.2.2. Değer Biçme……... 21
4.2.3. Birleştirme ve Temizleme…….. 22
4.2.4. Veri Seçimi ve Örnekleme……..... 23
4.2.5. Dönüştürme…… 24
4.3. Modelin Kurulması ve Değerlendirilmesi…… 24
4.4. Modelin Kullanılması…… 27
4.5. Modelin İzlenmesi…… 27
5. VERİ MADENCİLİĞİ METODLARI VE TEKNİKLERİ…… 28
5.1. Tahmin Edici Modeller…… 28
5.1.1. Sınıflama ve Regresyon Teknikleri…… 29
5.1.1.1. Karar Ağaçları…….. 29
5.1.1.2. Yapay Sinir Ağları…… 30

5.1.1.3. Genetik Algoritmalar…….. 30
5.1.1.4. K-En Yakın Komşu…… 31
5.2. Tanımlayıcı Modeller……. 31
5.2.1. Birliktelik Kuralları……... 32
5.2.2. Kümeleme Modelleri…… 33
5.2.2.1. Yoğunluk Tabanlı Yöntemler…… 35
5.2.2.2. Hiyerarşik Yöntemler……. 35
5.2.2.3. Model Tabanlı Yöntemler…….. 37
5.2.2.4. Bölme Yöntemleri…….. 38
5.2.2.4.1. K-Ortalamalar Kümeleme Yöntemi …… 40
5.2.2.4.1.1. K-Ortalamalar Algoritmasının Adımları…… 42
5.2.2.4.1.2. Geometrik Hesaplama…… 45
5.2.2.4.1.3. Aritmetik Hesaplama……. 47
6. K-ORTALAMALAR YÖNTEMİ İLE BENZİN İSTASYONLARININ OPERASYONEL DEĞELENDRİLMESİ…… 54
6.1. Problem Tanımı…… 55
6.2. Verilerin Hazırlanması…….. 55
6.2.1. Bir Akaryakıt Firmasına Ait Benzin İstasyonlarının Akaryakıt Satış Bilgilerinin Toplanması…….. 55
6.2.2. Değer Biçme…… 58
6.2.3. Verileri Birleştirme ve Temizleme…….. 58
6.2.4. Kullanılacak Verilerin Seçimi ve Dönüştürülmesi…….. 60
6.3. K-Ortalamalar Kümeleme Yönteminin Verilere Uygulanması ve Modelin Oluşturulması…….. 63
6.3.1. KNIME……... 63
6.3.2. Modelin Oluşturulması……..……. 66
6.4. Modelin Değerlendirilmesi…… 75
7. SONUÇ……... 77
KAYNAKLAR…… 79
ÖZGEÇMİŞ…… 85

ŞEKİLLER

Şekil 3.1 Veri Madenciliğine Katkıda Bulunan Disiplinler..... 18
Şekil 4.1 Veri Madenciliği Süreci..... 19
Şekil 4.2 Veri Tabanlarında Bilgi Keşfi Sürecinde Yer Alan Adımlar..... 20
Şekil 5.1 Veri Madenciliği Modelleri..... 28
Şekil 5.2 Veri Kümele Örneği..... 34
Şekil 5.3 DBSCAN Algoritması ile Kümeleme..... 35
Şekil 5.4 Dendogram Yapısına Bir Örnek..... 36
Şekil 5.5 Bütünleştirici ve Bölücü Hiyerarşik Kümelemenin {a,b,c,d,e}
Veri Nesneleri Üzerinde Gösterimi..... 37
Şekil 5.6 K-ortalamalar Kümele Algoritması..... 42
Şekil 5.7 K-ortalamalar Algoritmasının Adımları..... 44
Şekil 5.8 Geometrik Hesaplama Yöntemiyle İlk Kümelerin Belirlenmesi..... 46
Şekil 5.9 Noktaların Kümelere Dahil Edilmesi Sonrasında Yeni Küme
Merkezleri..... 46
Şekil 5.10 Her Döngü Sonrasında Küme Sınırları Değişmektedir..... 47
Şekil 5.11 Aritmetik Hesaplamaya Uygulanacak Veriler..... 48
Şekil 5.12 Aritmetik Hesaplamada Seçilen İlk Küme Merkezleri..... 49
Şekil 5.13 Aritmetik Hesaplamada İkinci Döngüde Oluşan Küme Merkezler..... 50
Şekil 5.14 Aritmetik Hesaplamada Üçüncü Döngüde Oluşan Küme Merkezleri... 52
Şekil 6.1 KNIME Programı Ekran Görüntüsü..... 64
Şekil.6.2 KNIME Programında Veri Kaynağı Ekleme ve Çıkış Alabilme
Nesneleri..... 64
Şekil.6.3 KNIME Programında Veri Madenciliği Algoritmaları Seçim Paneli..... 65
Şekil.6.4 KNIME Programında Görselleştirme Araçları..... 66
Şekil 6.5 KNIME Programında K-ortalamalar Kümeleme Yöntemi İçin Oluşturulan
Veri Akışı..... 67
Şekil 6.6 K-ortalamalar Nesnesinin Özellikleri.....
67
Şekil 6.7 KNIME ile Küme Merkez Değerlerinin Hesaplanması..... 74
Şekil 6.8 KNIME Interactive Table..... 74

TABLOLAR

Tablo 5.1 Kümeleme İçin Kullanılacak Veriler..... 50
Tablo 5.2 Kümeleme Sonucunda Oluşan Gruplama..... 53
Tablo 6.1 Satış Kayıtları Tablosu..... 56
Tablo 6.2 İstasyon Tablosu..... 56
Tablo 6.3 Müşteri Tablosu..... 57
Tablo 6.4 Ürün Tablosu..... 57
Tablo 6.5 Ürün Tablosu Verileri..... 59
Tablo 6.6 Ürün Tablosu Verilerinin Birleştirimesi..... 59
Tablo 6.7 Satış Kayıtları Tablosunundan Günlük Satış Tablosu Oluşturulması..... 60
Tablo 6.8 Günlük Satış Tablosu Verileri..... 62
Tablo 6.9 K=3,4,5 İçin Küme Merkez Değerleri..... 68
Tablo 6.10 K=3,4,5 İçin Kümelerde Yer Alan İstasyon Sayısı..... 69
Tablo 6.11 K=10,15,49,86 İçin Kümelerde Yer Alan İstasyon Sayısı..... 70
Tablo 6.12 Gürültülü Verilerin Silinmesi Sonrası Kümeleme..... 72
Tablo 6.13 Gürültülü Verilerin Silinmesi Sonrası Küme Merkez Değerleri..... 73
Tablo 6.14 Kapatılan İstasyonlar..... 76

DENKLEMLER

5.1 Öklit uzaklık ölçüsü.....39
5.2 Öklit uzaklık ölçüsü.....39
5.3 Pearson uzaklık ölçüsü.....40
5.4 Manhattan uzaklık ölçüsü.....40
5.5 Minkowski uzaklığı.....40

1. VERİ, BİLGİ VE ÖĞRENME

1.1. Veri

Veri; tek başına değersiz, bir amaca yönelik bilginin hammaddesidir. Veri sadece sayılar veya harfler değil; sayılar, harfler, ölçüm değerleri, grafikler, resimler ve onların anlamıdır. Bilgi ise bir amaca yönelik işlenmiş veridir. Başka bir ifade ile bilgi, bir soruya yanıt vermek için veriden çıkardığımız sonuç olarak tanımlanabilir [1].

Veri, oldukça esnek bir yapıdadır. Temel olarak varlığı bilinen, işlenmemiş, ham haldeki kayıtlar olarak adlandırılırlar. Bu kayıtlar ilişkilendirilmemiş, düzenlenmemiş yani anlamlandırılmamışlardır. İşlenerek farklı bir boyut kazanan bir veri, daha sonra bu haliyle kullanılmak üzere kayıt altına alındığında, farklı bir amaç için veri halini koruyacaktır [2].

1.2. Bilgi

Bilgi, potansiyel olarak ilginç ve faydalı olan veriler arasındaki ilişkidir. Keşif ise, gizlenmiş veya daha önceden bilinmeyen nesnelerin bulunması anlamına gelmektedir [3].

Veri madenciliğinde bilgi, veriden çıkarılışına göre 4 farklı tipe ayrılır.

1.2.1. Yüzeysel Bilgi:

Veri tabanlarından kolaylıkla yapılandırılmış sorgu dili olan SQL, (Structured Query Language), ile ulaşılabilen bilgi türüdür.

1.2.2. Çok Boyutlu Bilgi:

Bu tür bilgi çevrimiçi analitik işleme (OLAP, Online Analytical Processing) yöntemi araçları ile çözümlenebilen bilgidir. OLAP araçları ile veriden istenilen bilgileri keşfetmek ve farklı sıralamalar yapma olanağı bulunmaktadır. Fakat OLAP araçları ile öğrenilecek birçok bilginin SQL kullanılarak da öğrenilebileceğini anlamak önemlidir. OLAP araçlarının avantajı, bu çeşit araştırmaların ve çalışmaların çözümlemelerini en iyi (optimal) şekilde yapmasıdır. Yinede OLAP; veri madenciliği kadar güçlü değildir, en iyi çözümleri arayamaz [4].

1.2.3. Saklı Bilgi:

Veri tabanları üzerinde SQL sorguları ile yapılan aramaların çok uzun süreler alması ile ulaşılan bilgilere saklı bilgi denir. Bu tür bilgilere ulaşmakta SQL sorgu ifadeleri kullanmak yerine daha hızlı sonuç verebilen makine öğrenimi (machine-learning) ve örüntü tanıma algoritmaları kullanılmaktadır. SQL sorguları ile yapılacak sorgularda sonuçların elde edilebilmesi aylarca sürebildiği durumlarda örüntü tanıma algoritmaları ile sonuçlara birkaç saatte ulaşılabilmektedir.

1.2.4. Derin Bilgi:

Veri tabanı içerisinde gizlenen bilgidir. Bu tür bilgileri ortaya çıkarmak için veri tabanlarında nereler üzerinde arama yapılacağını gösteren ipuçlarına ihtiyaç duyulmaktadır.

1.3. Öğrenme

Öğrenme, kuramsal düşüncelerden uygulama ve tecrübelerden elde edilen bilgilerle insan inançlarını, değerlerini, tutum ve davranışlarını değiştirme sürecidir. Arzulanan sonuçlara erişebilmek amacıyla davranış ve eylemlerin değiştirilmesine götüren yeni bilgi ve sezgisel kavrayış kazanma sürecidir [5].

En basit şekilde ifade etmek gerekirse öğrenme, bir şeye uyum sağlama şeklidir. Diğer bir yandan insanlar, öğrenme süreçlerini daha kolaylaştırmak ve hızlandırmak için öğrenebilir bilgisayar sistemleri oluşturmaya başladılar. Uzmanlar son yıllardaki gelişmelere dayanarak bilgisayarlara kendi kendine öğrenme yetkinliğinde programların uygulanabileceğini düşünmektedirler. Bu düşüncelerle birlikte yapay zekadan yararlanarak bilgisayar programları geliştirilmiştir.

1970'li yılların sonlarında araştırmacılar, jeolojik örneklerin sınıflandırılması ya da hastalıkların teşhisi gibi sınırlı alanlarda, basit kurallarla uzman sistemler oluşturmuşlardır. Uzman sistemler binlerce kuraldan oluşur. Basit sistemler için bile çok fazla doğru kural bulmak gerekir. Akıllı uzman sistemler yaratmak için ilgilenilen alanla ilgili ihtiyaç duyulan her veri uzman sistemin temeli olarak kaydedilir ve uzmanlar tarafından doğru kurallar bulmak için kullanılır [6].

1980'li yılların başlarında ise araştırmacılar makine öğrenimi ile birlikte yeni keşifler yapmışlardır. Bu keşifler ile; Minsky ve Papert tarafından yeniden değerlendirilen algılayıcılardan daha güçlü sinir ağları için oluşturulan yeni yapılar ve genetik algoritmalar ile nesnelerin karar ağaçlarıyla rasgele olarak sınıflandırılması için yapılan basit algoritmalar gibi yaklaşımları modelleme mümkün olmuştur. Aynı zamanda bilgisayar ve bilgi teknolojilerinin de ilerlemesinden dolayı daha güçlü hale gelen bilgisayarlarla, yeni algoritmaları gerçek problemlere uygulama olanağı bulunmuştur [6].

Makine öğrenimi üzerine yapılan çalışmalar devam ederken çok daha ilgiç bir gelişme ortaya çıkmıştır. Bu gelişme, çağdaş toplumda ortaya çıkan veri patlamasıdır. Bu veri patlamasının sonucunda ihtiyaç duyulan faydalı bilgiye ulaşmak gün geçtikçe daha zorlaşmaktadır.

2. VERİ TABANI, VERİ AMBARI VE SORGU ARAÇLARI

Günümüzde bilgi teknolojilerinde yaşanan büyümenin sonucu olarak saklanan verilerin büyüklüğü çok fazla artış göstermektedir. Depolanacak verilerin büyüklükleri her geçen gün katlanarak artış göstermektedir. Günlük kullanımdaki bilgisayarların veri depolama kapasiteleri bile artık gigabyte seviyelerine gelmiş, internet ortamında tutulan verilerin ise boyutları terabyte hatta petabyte seviyelerine geldiği görülmektedir. Bu bilgi kapasitesinin büyümesi ile birlikte bu verileri işleyecek yada depolayacak bilgisayarlarında gelişerek, erişilebilir hale geldiği görülmektedir. Bunun sonucu olarak da detaylı ve doğru bilgiye ulaşabilmek mümkün hale gelmiştir. Örneğin eskiden bir akaryakıt istasyonunda yapılan satışlar toplanarak günlük ne kadar satış yapıldığı bilgisine ulaşılabilirdi. Günümüzde artık yapılan satışların anlık kayıtlarının tutulabilmesi sonucu hangi zaman dilimlerinde ne kadar satış yapılabildiği ya da müşterilere verilen bağlılık kartları sayesinde müşteri kayıtları tutularak müşteri davranışları bile gözlenebilir duruma gelmiştir. Bu sayede detaylı olarak kayıtlar tutulabilmekte ve ihtiyaç duyulduğu takdirde belirli amaçlar doğrultusunda bu kayıtlardan doğru sonuçlar üretilebilmektedir.

2.1. Veri Tabanı

Veri tabanı bir veya daha fazla uygulamaya hizmet vermek için bir araya toplanmış birbirleriyle ilişkili veriler toplamıdır. Veri tabanı sadece verinin alınması değil aynı zamanda o veri üzerinde değişiklik yapılmasına da imkan vermektedir. Veri tabanı bilgisayarda veri depolamak ve işlemek amacıyla kullanılmaktadır. Veri tabanı, çeşitli tiplerdeki varlıklara, bu varlıkların özniteliklerine ve bunlar arasındaki ilişkilere ev sahipliği yapan bir yapıdır. Bir veri tabanında soyutlama katmanları kullanılarak gerçek dünyanın kavramları bilgisayar ortamına adapte edilebilmektedir [7].

Fiziksel veri tabanı, disk üzerinde bulunan dosya ve indeks koleksiyonu ve bunlara ulaşmak için kullanılan depolama yapılarıdır. Kavramsal veri tabanı, gerçek hayatın bir soyutlamasıdır. Bu soyutlamayı gerçekleştirmek için veri tabanı yönetim sistemi, bir veri tabanı tanımlama dili kullanır. Veri tanımlama dili kavramsal veri tabanını veri modeli olarak tanımlayabilmemizi saglar. Kavramsal veri tabanı, organizasyon tarafından kullanılan verinin bütünü temsil eder [8].

Görünüm katmanı veya alt şema, kavramsal veri tabanının bir parçasıdır. Bu alt şemalar veri düzenleme dili ile oluşturulabilir. Görünümler, bir veri tabanında güvenliğin sağlanması için de önemlidir. Çünkü bir görünüm, kullanıcının veri tabanının sadece bir bölümünü görmesine izin vermektedir [13].

Geleneksel dosya yönetim sisteminin dezavantajları şu şekilde sıralanabilir:

- Veri tekrarına yol açması
- Aynı verinin farklı kullanıcılar tarafından aynı anda güncellenmeye çalışılması. Veri bütünlüğü bu durumda tehlikeye girer.
- Veri depolama alanının gereksiz yere kullanılması.
- Veriye ulaşmak için kullanılan dilin kullanılan programa özgü olması ve kullanışlı olmaması [13].

Bir veri tabanı sistemi ise;

- Veri tekrarını önler.
- Veri bütünlüğünü sağlar.
- Veri depolama alanı ihtiyacını azaltır.
- SQL gibi bir standart sorgulama dili kullanılarak veriye ulaşabilmeyi sağlar [9].

Bununla birlikte veri tabanı sisteminin dezavantajları da bulunmaktadır. Bunlar şu şekilde sıralanabilir:

- Kurulum ve bakımının zor ve pahalı olması
- Bütünleşik sistemdeki bir bölüm veriye ulaşamamak tüm sistemin çalışmamasına sebep olur.

İşletmeleri veri tabanı yaklaşımına götüren pek çok problem mevcuttur. Bunlardan bazıları şunlardır:

- Basit ihtiyaçlara çabuk yanıtlar alınamaması.
- Düşük veri kalitesi ve doğruluğu
- Değişime hızlı ayak uyduramama
- Yüksek gelişim maliyetleri
- Gerçek dünya için geçersiz veri modeli kullanımı [10].

Veri tabanı sistemlerinin başlıca üç özelliği vardır:

- Özerklik: Bir veri tabanı diğer veri tabanlarıyla etkileşimde olmak için kendi kontrol politikasını oluşturabilir.
- Heterojenlik: Veri modelleri, sorgulama dilleri, veri tabanından veri tabanına farklılık gösterebilir.
- Dağıtım: Fiziksel olarak farklı ortamlarda yerleşmiş bulunan veri tabanları [13].

Veri tabanlarını;

- Hiyerarşik veri tabanları,
- İlişkisel veri tabanları,
- Nesne yönelimli veri tabanları, olarak başlıca üç kısımda incelemek mümkündür [11].

2.1.1. Hiyerarşik veri tabanları

IBM'in yönetim bilişim sistemleri ile popüler olan hiyerarşik veri tabanı yönetim sistemleri, kullanıcılara verileri ağaç yapısında modellemesine olanak tanımaktadır. Bu ağaç yapısı bir köke bağlı bir veya daha fazla daldan oluşmaktadır. Hiyerarşik veri tabanları içindekileri bir ağaca benzeyen hiyerarşik model içinde organize eder. Hiyerarşik ağaç sadece bir veri tabanı içindeki veri elemanlarını tanımlamakla kalmaz,

ayın zamanda bu veri elemanları arasındaki ilişkiyi de tanımlar. Çeşitli hiyerarşik veri tabanı modelleri mevcuttur. Bunlar; bire-bir, bire çok ve çoklu şeklindedir [7].

2.1.2. İlişkisel veri tabanları

İlişkisel bir veritabanı, her birine eşsiz bir isim atanmış tablolar setidir. Her tablo, özelliklerin bir setinden ve büyük bir kayıtlar setinden meydana gelir. ilişkisel bir tablodaki her kayıt, eşsiz bir anahtar tarafından özdeşleştirilen ve bir özellikler seti tarafından tanımlanan bir nesneyi yansıtır [56].

İlişkisel veri tabanlarının kullanılması hiyerarşik veri tabanlarına nispetle bir takım kolaylıkları da beraberinde getirmiştir. Bu kolaylıklar şu şekilde sıralanabilir:

- Yeni veri tabanı kayıtları veri tabanına sokulabilir ve bir öğenin herhangi bir parçası değiştirilebilir veya silinebilir.
- Öğenin kayıt numarası ve alan ismine bakmak suretiyle veri tabanı içerisindeki verinin yerinin bulunmasının kolaylaşması.
- Veri kayıtlarını sıralama ve yeniden düzenleyebilme ve veri elemanlarını birleştirebilme [12].

İlişkisel veri tabanlarının barındırdığı bazı kısıtlarda şu şekilde sıralanabilir:

- Veri türlerinin tanımlanmasındaki yetersizlikler nedeni ile uygunsuzluk. Çoğu veri tabanı sistemi tamsayı, karakter dizisi, tarih gibi standart türdeki veri tiplerini içerir. Kullanıcının kendisi yeni veri tipleri ve bu veri tipleri arasındaki yeni işlem tiplerini tanımlayamaz.
- Karmaşık bir veri yapısını, çok yapısal olan hiyerarşik veya ilişkisel olarak gösterememe.
- Örneğin, C gibi programlama dillerine göre verilerin idaresini yapan yordamların zayıflığı.

- Bir yapıyı modellemedeki zayıflık: İlişkisel sistem klasik satır sütun yapısındadır. Bu yaklaşımda karmaşık sorgulama ve işlemlerin yapılamayacağı açıktır [13].

İlişkisel model ile veri tabanını tasarlayan kişilerin ve son kullanıcıların işlerini kolaylaştırmak amaçlanmıştır. Veriler basit bir tablo yapısında tutularak anlaşılması rahat, kolay bir yapı oluşturulmaktadır. İlişkisel veri tabanı ise tabloların oluşturduğu bütün bir yapı olarak göze çarpmaktadır.

2.1.3. Nesne yönelimli veri tabanları

Nesne yönelimli veri tabanı, bir veri tabanının kavramsal modelinin karmaşık nesnelerini oluşturan ve kullanabilen bir veri tabanıdır [14].

Nesneye dayalı veri tabanı seçiminin ve kullanımının sebebi, veri tabanının yeterliliğinin arttırılmış olmasıdır. Bunlar; yüksek yeterliliğe sahip yüksek düzeyde sorgu dili, giriş ve düzeltme işlemleri, aynı zamanda olabilirlik kontrolü ve geri alma, karmaşık nesnelerin saklanması, indeksler ve hızlı erişim imkanlarıdır [13].

Nesneye dayalı bir veri tabanında, yapısı gereği arama işlemleri çok hızlı yapılabilir. Özellikle büyük tablolarla uğraşırken ilişkisel veri tabanlarından çok daha hızlı sonuca ulaşırlar. Ancak çalışma mantığı tümüyle değişir [63].

2.2. Veri Ambarı

Veri ambarcılığı çeşitli şekillerde tanımlanmıştır. Veri ambarcılığının babası sayılan Bill Inmon veri ambarını 1992'de şu şekilde tanımlamıştır: Veri ambarı, yönetimin karar sürecini desteklemede kullanılan, konuya yönelik, entegre, zamana bağlı, kalıcı veri topluluğudur [15].

Başka bir tanıma göre ise veri ambarı, basitleştirilmiş biçimde hareket sistemlerinden özetlenen ve kümelenen verinin saklandığı yerdir. Son kullanıcılar için yapılan veri erişimi ve raporlama araçları kullanıcıların karar desteği için veriyi elde etmelerini sağlamaktadır [16].

Veri ambarlarının üç çeşidi vardır:

- Tüm kuruma hizmet eden kurumsal (geleneksel) veri ambarı,
- İşletmedeki belirli bir iş birimini veya bölümü desteklemek üzere tasarlanmış minyatür bir veri ambarı olan veri pazarı (data mart),
- Veri ambarı tekniklerinin hareket sistemlerine uyarlandığı operasyonel veri deposu [17].

2.2.1. Veri ambarının karakteristik özellikleri

Veri ambarının karakteristik özellikleri şu şekildedir:

- Konuya Yönelik Olma: Operasyonel veri ihtiyacı, uygulamanın anlık ihtiyaçları ile ilgilidir ve o anda geçerli iş kurallarına dayanır. Veri ambarı dünyası ise müşteri, mal veren, ürün ve etkinlik gibi temel konular etrafında organize olur. Veri ambarındaki veri karar vermeye yöneliktir ve zaman derinliği çok daha fazla olduğundan daha karmaşık ilişkilere olanak tanır.

- Veri Entegrasyonu: Sistemlerden veri ambarına veri aktarılırken veri entegre edilir ve hepsi aynı formata getirilir. Böylece değişik kaynaklardan gelen veri, veri ambarında tek ve genel olarak üzerinde anlaşmaya varılmış bir şekilde yer alır. Veri ambarındaki veri, temiz, geçerliliği onaylanmış ve uygun biçimde kümelenmiş olmalıdır. Benzer biçimde verinin doğru olması da gerekmektedir.
- Kalıcı Ortam: Veri ambarında veri her zaman eklenir, üzerine yazılmaz. Veri tabanı sürekli olarak yeni veri alır ve eskisiyle entegre eder. Veri ambarlarında veri yükleme belirli zamanlarda yapılır ve kullanıcılar ancak bundan sonra veriye erişebilirler.
- Zamana Bağlı Olma: Veri ambarındaki veri referans alınan zaman birimi ile birlikte kaydedilir ve veri bir kez doğru biçimde kaydedildikten sonra kullanıcılar tarafından güncellenemez. Veri ambarındaki veri tipik olarak 3-10 yıllık bir zaman dilimini kapsar [18].

2.2.2. Veri ambarının yapısı

Veri ambarında dört tip veri bulunur. Bunlar;

- Şu Anda Geçerli Detay Veri: Şu anda geçerli detay veri en son olan olayları içerir. Bu veri en alt düzey tanelikte saklanırsa çok yer tutabilir. Bu yüzden bu veri genellikle şu anda geçerli hareket verisinin temizlenip veri ambarına yüklenmiş biçimidir.
- Eski Detay Veri: Çoğu veri ambarında detay verinin belirli bir zaman sonra diskten bir büyük veri saklama ortamına aktarılmasını öngören kurallar vardır. Detay veriye hala erişilebilir, ancak daha yavaş bir ortamda olduğu için erişim hızı yavaştır.
- Özetlenmis Veri: Standart rakamları önceden tahmin edip veriyi buna göre özetlemek veri ambarının sorgulamalara daha hızlı cevap vermesini ve daha çok kullanılmasını sağlar. Ayrıca özetlemeler sayesinde hesaplamalar tekrar tekrar yapılmak zorunda kalmaz, ancak daha fazla veri saklama kapasitesi gerekir.

- Meta Data: Veri ambarının en önemli bileşenlerinden biri meta datadır. Veri ambarında verilerin tanımlandıgı kısımdır. Meta data "veri hakkında veri" anlamındadır. Meta data her veri elementinin anlamını, hangi elementlerin hangileriyle nasıl ilişkili olduğunu ve kaynak verisi ile erişilecek veri gibi bilgileri içermektedir [19].

Yukarıda belirtilen veri tiplerinin hepsinin aynı kayıt ortamında saklanma zorunluluğu bulunmamaktadır. Ancak kullanılacak olan yazılımın hepsine erişebilmesi gerekmektedir.

2.2.3. Veri ambarı projelerinde kaçınılması gereken hatalar

Veri ambarları, işletmelerin birden fazla departmanını ile ilişkili maliyeti yüksek yatırımlardır. Bir veri ambarı projesinin başarısı için veri ambarı projelerinde yöneticilik yapmış tecrübeli yöneticilerin uyarılarını dikkate almak faydalı olabilecektir.

Veri ambarı enstitüsü bir yayınında, veri ambarı projelerinde kaçınılması gereken pek çok hatadan birkaçını şu şekilde sıralamaktadır:

- Yanlış sponsorluk zinciriyle ise başlamak,
- Ulaşılamayacak hedefler koymak ve gerçekler ortaya çıktığında üst düzey yöneticileri hüsrana uğratmak,
- Yalnızca elde oldugu için veri ambarına enformasyon yüklemek,
- Veri ambarı veri tabanı tasarımını, hareket veri tabanı tasarımıyla bir tutmak,
- Kullanıcıya değil, teknolojiye önem veren bir veri ambarı yöneticisi seçmek,
- Geleneksel olarak işletme içinde üretilen veriye odaklanıp, veri, metin, resim ve hatta ses ve video görüntülerinden oluşan harici verinin potansiyel değerini ihmal etmek,
- Üst üste çakışan veya kafa karıştırıcı şekilde tanımlanmış veri sağlamak, performans, kapasite ve ölçeklenebilirlik sözlerine inanmak,

- Veri ambarı bir kez kurulduktan sonra tüm sorunlarımızın hallolduğuna inanmak [20].

2.2.4. Veri ambarı ihtiyacı

Bir işletmenin büyüklüğü veri ambarı ihtiyacının bir ölçüsü değildir. İşletmenin bir veri ambarına ihtiyacı olup olmadığına karar verirken işe bazı anahtar göstergelere bakarak başlanabilir [21].

Bu göstergelerden bazıları şunlardır:

- İşletme değişen, rekabetin çok yoğun oldugu bir pazarda faaliyet gösteriyorsa,
- Müşteriler hakkında sağlıklı enformasyon elde etme ihtiyacı varsa,
- Kazanç sağlayacak ve/veya verimliliği arttıracak enformasyona dayalı ürünler veya hizmetler oluşturma fırsatları varsa,
- Sık kullanılan ve birbiriyle ilişkili kurumsal veri birçok değişik yerde ve farklı sistemlerde bulunuyorsa,
- "Aynı veri ama farklı sonuç" şeklindeki sorun işletmede sürekli bir rahatsızlık haline gelmişse,
- Gerçek karar destek sistemlerine ihtiyaç varsa,
- Kullanıcılar daha etkili ve anlık sorgulama ve raporlama yapmak istiyorlarsa,
- Bir enformasyon dağıtımı alt yapısına ihtiyaç varsa [13].

Veri ambarı kurmaya karar veren firmaları şu özellikleri göze çarpmaktadır:

- Yönetimde enformasyona dayalı bir yaklaşım,
- Rekabetin yoğun olduğu, hızla değişen pazarlarda bulunmak,
- Çok sayıda ve farklı özelliklerde müşterilere sahip olmak,
- Verinin farklı sistemlerde bulunuyor olması,
- Aynı verinin farklı sistemlerde değişik biçimlerde gösteriliyor olması,
- Verinin çok teknik ve deşifresi zor biçimde kayıtlı olması [21].

2.3. Veri Ambarı ile Veritabanı Arasındaki Farklar

Veritabanı içerisindeki bilgiler genelde anlık bilgilerdir. Yani belirli bir süre sonunda güncelliğini kaybedecek olan bilgilerdir. Ancak veri ambarı içerisinde veriler genelde yığılarak birikirler ve verilerin geçerliliği çok daha uzun sürmektedir. Veri ambarı içerisinde ne kadar çok kayıt olursa, yapılan incelemelerin sonucu da o kadar doğru olacaktır. Oysa veritabanı içerisindeki kayıt sayısının çok fazla olması durumunda, veritabanını kullanan canlı sistemlerin performansları düşecek ve bu da verilere erişimi çok yavaşlatacaktır. Bu durum, canlı sistemlerde en istenmeyen durumdur [22].

2.4. Sorgu Araçları

Bir veri tabanı içerisinde istenilen bilgiye ulaşmak için sorgu dilleri kullanılmaktadır. Yapılandırılmış sorgu dili (SQL), verileri yönetmek ve tasarlamak için kullanılmaktadır. Bir veri madenciliği aracı bir sorgu aracının yerini alamaz fakat veri madenciliği aracı kullanıcıya ilave imkanlar sağlar. Veri madenciliği ve SQL ile ilgili; tam olarak ne arandığı biliniyorsa SQL, ne arandığı belli belirsiz biliniyorsa veri madenciliği kullanılır. Genellikle ne arandığı belli belirsiz olduğu durumlar, ne arandığı bilindiği durumlardan çok daha fazladır. Bu durum da son zamanlarda veri madenciliğine olan ilginin sebebini açıklamaktadır [4].

3. VERİ MADENCİLİĞİ KAVRAMI

Veri madenciliği kavramını anlamak için veri madenciliği tanımını ve tarihçesini anlamak gerekir.

3.1. Veri Madenciliği Tanımı

Aşağıda bir çok kaynaktan alınmış veri madenciliğini tanımlayan ifadeler bulunmaktadır.

Veri madenciliği, veri tabanı bilgi keşfi sürecinde bir adımdır ve verideki örüntüleri ortaya çıkarmak için kullanılan algoritmaları kapsar. Ortaya çıkarılan bilgi daha sonra bir öngörü (prediction) veya sınıflandırma (classification) modeli kurmak, eğilimleri ve birliktelikleri belirlemek, mevcut bir modeli yenilemek veya üzerinde madencilik çalışması yapılmış bir veri tabanının özetini çıkarmak için kullanılabilir [23].

Veri madenciliği, önceleri bilinmeyen, geçerli ve etkin bilginin büyük veri tabanlarından çekilmesi ve daha sonra bu bilginin son iş kararlarını almak için kullanılmasını kapsayan bir süreçtir [24].

Veri madenciliği, aksi halde keşfedilemeyebilecek olan eğilimleri ve örüntüleri bulmak için, çok miktardaki verinin otomatikleştirilmiş analizidir [25].

Veri madenciliğini amacı, mevcut veri içindeki geçerli, alışılmamış, kullanışlı ve anlaşılır korelasyonları ve örüntüleri saptamaktır [26].

Veri madenciliği, muazzam boyuttaki veriden şirketlerin daha iyi kararlar almalarına yardımcı olup, pazarda rekabetçi olarak kalmalarını sağlayabilecek ilginç bilgileri keşfetme sürecidir [27].

Veri madenciliği, anlamlı örüntüler ve kurallar keşfetmek için büyük miktardaki veriyi, otomatik veya yarı otomatik yöntemlerle araştırma ve analiz etme sürecidir [28].

Veri madenciliği, eldeki verilerden üstü kapalı, çok net olmayan, önceden bilinmeyen ancak potansiyel olarak kullanışlı bilginin çıkarılmasıdır. Bu da; kümeleme, veri özetleme, değişikliklerin analizi, sapmaların tespiti gibi belirli sayıda teknik yaklaşımları içerir [29].

Veri Madenciliği genel anlamda; büyük miktarda veri içerisinden, gizli kalmış, değerli, kullanılabilir bilgilerin açığa çıkarılması biçiminde ifade edilmektedir. Veri madenciliğinin temelinde klasik istatistiksel teknikler vardır, fakat bu istatistiksel teknikleri doğrudan veri madenciliği olarak tanımlamamak gereklidir. Veri madenciliği, istatistiksel yöntemlerin gelişmiş halinini yanı sıra diğer yöntemlerden özgün farkları bulunmaktadır [30].

Veri Madenciliği, makine öğrenimi, istatistik, veri tabanı yönetim sistemleri, veri ambarlama, görselleştirmeler gibi farklı disiplinlerde kullanılan yaklaşımları birleştiren veri analiz yaklaşımıdır [31].

Veri madenciliği büyük miktardaki verinin analiz edilerek anlamlı bilgiler ve kurallar keşfetmek amacıyla geliştirilen bir yaklaşımdır [32].

Veri madenciliği, büyük veri yığınlarının içersinden önceden bilinmeyen, geçerli kurallar ve ilişkiler bulmaya yarayan veri analiz aracıdır. Bu yaklaşım, istatistiksel modeller, matematiksel algoritmalar ve makine öğrenme metodlarını içerir (denemelerle otomotik performaslarını geliştiren algoritmalar; yapay sinir ağları ve karar ağaçları vb.) Sonuç olarak veri madenciliği verinin toplanması ve yönetilmesinin yanında analizi ve tahminini de içerir [32,33].

3.2. Veri Madenciliğinin Tarihçesi

İstatistik sözcüğü ''statista'' (İtalyancada devlet işleriyle uğraşan kişi) ile ''stato'' devlet ve durum anlamına gelen sözcüklerden türetilmiştir [64].

İstatistik, 16. yüzyılda önce İtalya'da, sonraları Fransa, Hollanda ve Almanya'ya yayılarak kullanılmıştır. 17. ve 18. yüzyıllarda üniversitelerde öğretilmeye başlanarak, 19. yüzyılın başlarına kadar devletlerin coğrafya, ekonomi ve nüfus gibi önemli belirleyici niteliklerinin analizinde kullanılmıştır [34]. Önceleri, teknik bir disiplin olarak ele alınırken günümüzde bir bilim dalı olarak kendini kabul ettirmiştir. Ulusal ve uluslararası verilerin analizinde, devletlerin sonuçlarının birbiri ile karşılaştırılmasında, sonuçların izlenmesinde güncellik, güvenilirlik istatistik bilim dalı ile sağlanmaktadır [64].

İstatistik, verileri değerlendirmek ve analiz etmek amaçlı yöntemlerden oluşmaktadır. Bilgisayarların verileri analiz etmek için kullanılmaya başlamasının ardından istatiksel çalışmalar artış göstermiştir. Bilgisayarlar sayesinde, yapılması daha önce mümkün olmayan istatistiksel araştırmalar yapılabilir hale gelmiştir. Veri madenciliği ile istatistik birlikte kullanılmaya başlanarak veri yığınlarından istenilen bilginin çıkarılması ve analizinin yapılması daha kolaylaşmıştır.

Veri madenciliğinin kökeni ilk sayısal bilgisayar olan ENIAC (Electrical Numerical Integrator And Calculator)'a kadar dayanmaktadır. [35] Bilgisayarların efektif kullanımı verilerin depolanması ile başlamaktadır. İlk haliyle karmaşık hesaplamaları yapmaya yönelik geliştirilen bilgisayarlar, kullanıcı ihtiyaçları doğrultusunda veri depolama işlemleri için de kullanılmaya başlanarak veri tabanları ortaya çıkmıştır. Veri tabanlarının genişlemesiyle birlikte donanımsal olarak bu verilerin tutulacakları ortamların da genişlemesi gerekeceği için veri ambarı kavramının ortaya çıkışı bu dönemde olmuştur. Saklanmak istenen veriler, bir ambar gibi fiziksel sürücülerde tekrar kullanılmak üzere saklanarak gittikçe büyüyen veri tabanlarının organizasyonu, düzenlenmesi ve yönetimi de buna paralel olarak güç bir hal almaya başlamıştır [64].

Veri modelleme kavramı ilk olarak basit veri modelleri olan "Hiyerarşik ve Şebeke" veri modelleri ortaya çıkmıştır [35]. Ancak bu kavram kullanıcıların ihtiyaçlarını bütünüyle karşılayamamıştır ve geliştirilmiş veri modelleri ortaya çıkmıştır.

Veri madenciliği, kavramsal olarak 1960'lı yıllarda, bilgisayaların veri analiz problemlerini çözmek için kullanılmaya başlamasıyla ortaya çıkmıştır. Bilgisayar yardımıyla, kısa zamanda istenilen verilere ulaşmak yeterli hale gelmemeye başlamıştır. Bu yönteme önceleri veri taraması (data dredging), veri yakalanması (data fishing) gibi isimler verilse de 1990'lı yıllara gelindiğinde bilgisayar mühendisleri tarafından veri madenciliği ismi konulmuştur [64].

Şekil 3.1: Veri Madenciliğine Katkıda Bulunan Disiplinler [70]

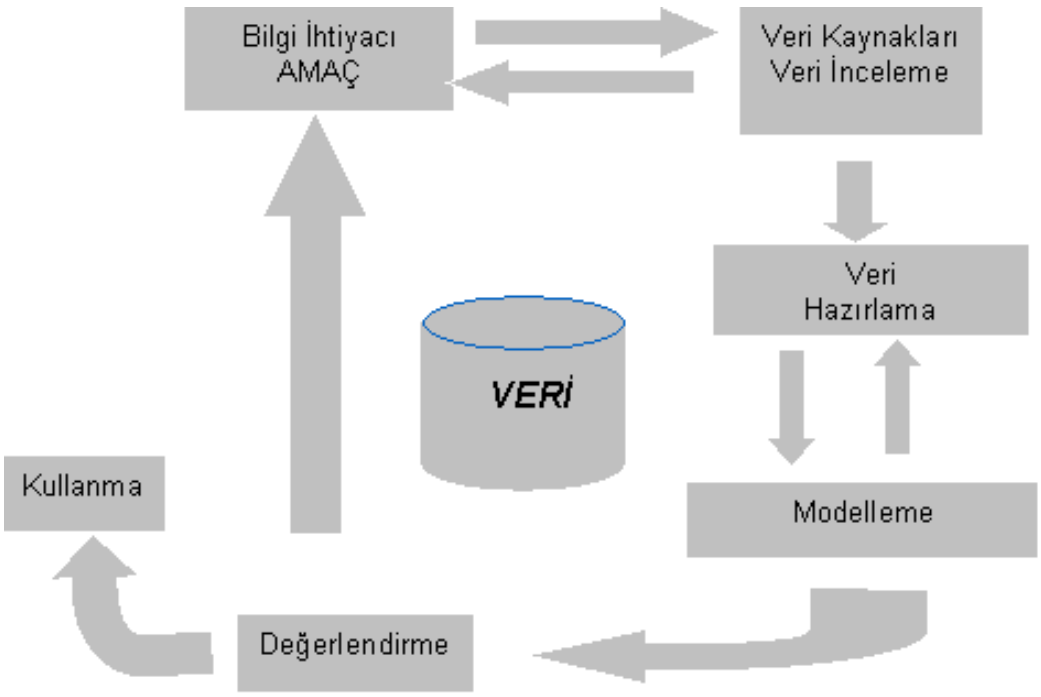
Veri madenciliği Şekil 3.1'de görülen istatistik, makine öğrenimi (machine learning), veritabanları, otomasyon, pazarlama, araştırma gibi disiplinler ve kavramlardan esinlenilerek çeşitli yaklaşımlar geliştirilmiştir [36].

4. VERİ TABANLARINDA BİLGİ KEŞFİ SÜRECİ

Ne kadar etkin olursa olsun hiç bir veri madenciliği algoritması, üzerinde inceleme yapılan işin ve verilerin özelliklerinin bilinmemesi durumunda fayda sağlaması mümkün değildir. Bu nedenle yukarıda tanımlanan tüm aşamalardan önce, iş ve veri özelliklerinin öğrenilmesi / anlaşılması başarının ilk şartı olacaktır. Başarılı bir veri madenciliği projelerinde izlenmesi gereken adımlar aşağıdadır [37].

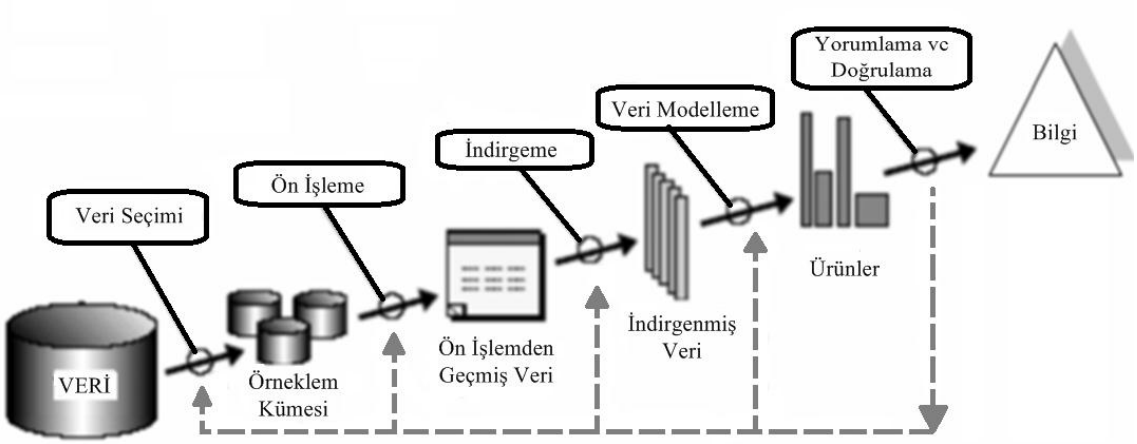
1. Problemin tanımlanması,
2. Verilerin hazırlanması,
3. Modelin kurulması ve değerlendirilmesi,
4. Modelin kullanılması,
5. Modelin izlenmesi.

Veri madenciliği süreci Şekil 4.1'de gösterilmiştir.



Şekil 4.1 Veri Madenciliği Süreci [38]

Veri tabanlarında bilgi keşfi sürecinde yer alan temel adımlar Şekil 4.2'de gösterilmiştir.



Şekil 4.2 Veri Tabanlarında Bilgi Keşfi Sürecinde Yer Alan Adımlar [23]

4.1. Problemin Tanımlanması

Veri madenciliği çalışmalarının birinci ve en önemli adımı problemin tanımlanmasıdır. Projenin hangi amaca yönelik yapılacağının ve elde edilecek sonuçların başarı düzeylerinin nasıl ölçüleceğinin belirlenmesi veri madenciliği çalışmalarının başarılı olmasının en önemli şartıdır. Ayrıca doğru tahminlerin getireceği kazanımlar ve yanlış tahminlerin getireceği maliyetlere ilişkin tahminlere de bu aşamada yer verilmelidir. Bu aşamada mevcut iş probleminin nasıl bir sonuç üretilmesi durumunda çözüleceğinin, üretilecek olan sonucun fayda-maliyet analizinin başka bir ifadeyle üretilen bilginin işletme için değerinin doğru analiz edilmesi gerekmektedir [38,39].

4.2. Verilerin Hazırlanması

Veri madenciliğinin en önemli aşamalarından biri olan verinin hazırlanması aşaması, analistin toplam zaman ve enerjisinin %50 - %85'ini harcamasına neden olmaktadır. Bu aşamada firmanın mevcut bilgi sistemleri üzerinde ürettiği sayısal bilginin iyi analiz edilmesi, veriler ile mevcut iş problemi arasında ilişki olması gerektiği unutulmamalıdır. Proje kapsamında kullanılacak sayısal verilerin, hangi iş süreçleri ile oluşturulduğu da bu veriler kullanılmadan analiz edilmelidir. Bu sayede analist veri kalitesi hakkında fikir sahibi olabilir [38,39].

Verilerin hazırlanması verilerin toplanması, değer biçilmesi, birleştirmesi ve temizlenmesi, seçilmesi ve dönüştürme adımlarından oluşmaktadır.

4.2.1. Toplama

Tanımlanan problem için gerekli olduğu düşünülen verilerin ve bu verilerin toplanacağı veri kaynaklarının belirlenmesi adımıdır. Verilerin toplanmasında kuruluşun kendi veri kaynaklarının dışında, nüfus sayımı, hava durumu, merkez bankası kara listesi gibi veri tabanlarından veya veri pazarlayan kuruluşların veri tabanlarından faydalanılabilir [38,39].

4.2.2. Değer biçme

Veri madenciliğinde kullanılacak verilerin farklı kaynaklardan toplanması, doğal olarak veri uyumsuzluklarına neden olacaktır. Bu uyumsuzlukların başlıca olanları farklı

zamanlara ait olmaları, kodlama farklılıkları (örneğin bir veri tabanında cinsiyet özelliğinin e/k, diğer bir veri tabanında 0/1 olarak kodlanması), farklı ölçü birimleridir. Ayrıca verilerin nasıl, nerede ve hangi koşullar altında toplandığı da önem taşımaktadır. Bu nedenlerle, iyi sonuç alınacak modeller ancak iyi verilerin üzerine kurulabileceği için, toplanan verilerin ne ölçüde uyumlu oldukları bu adımda incelenerek değerlendirilmelidir [40].

4.2.3. Birleştirme ve temizleme

Bu adımda farklı kaynaklardan toplanan verilerde bulunan ve bir önceki adımda belirlenen sorunlar mümkün olduğu ölçüde giderilerek veriler tek bir veri tabanında toplanır. Ancak basit yöntemlerle ve baştan savma olarak yapılacak sorun giderme işlemlerinin, ileriki aşamalarda daha büyük sorunların kaynağı olacağı unutulmamalıdır [39,40]. Hatalı veya analizin yanlış yönlenmesine sebep olabilecek verilerin temizlenmesine çalışılır. Genellikle yanlış veri girişinden veya bir kereye özgü bir olayın gerçekleşmesinden kaynaklanan verilerin, önemli bir uyarıcı enformasyon içerip içermediği kontrol edildikten sonra veri kümesinden atılması tercih edilir [38,40].

Veri tabanında yer alan tutarsız ve hatalı veriler gürültü olarak değerlendirilmektedir. Bu gibi durumlarda verinin söz konusu sorunlardan temizlenmesi gerekecektir [41]. Ayrıca yanlış ve aşırı uçta bulunan verilerin ortadan kaldırılması da verilerin temizlenmesi konusuna girer [42].

Bu kayıp verilerin yaratacağı sorunları ortadan kaldırmak için kullanılan teknikler şu şekilde özetlenebilir:

- *Eksik değer içeren kayıtlar veri kümesinden atılabilir*: Eğer kayıp verilerin kayıt sayısı, toplam kayıt sayısına oranlandığında sonuçların hassasiyetini etkilemeyecek kadar küçükse bu yöntem kullanılabilir.

- *Kayıp değerlerin yerine genel bir sabit kullanılabilir*: Örneğin eğitim seviyesi eksik olan bilgiler yerine "X" gibi bir sabit atanabilir. Bu çalışma sonucunda eğitim durumunun X olması anlamlı bir sonuçmuş gibi çıkabilir yada eğitim durumunun X olması belirleyici ve ayırt edici bir özellikmiş gibi çıkabilir. Yani bu bilgiler veri madenciliğini yanıltabilir. Bununla beraber bu yöntemin kullanılması, bazı durumlarda veri madenciliğinin gerçek amacına hizmet ederek bilinmeyeni ortaya çıkarabilir.
- *Değişkenin tüm verilerinin ortalaması veya sadece bir sınıfa alınarak eksik değerin yerine bu değer kullanılabilir*: Örneğin elimizde sigorta hasar ihbar verileri olsun. Teminat tutarı eksik olan kısma teminat tutarlarının ortalaması alınıp atanabilir.
- *Verilere uygun bir tahmin yapılarak eksik değerin yerine konabilir*: Eldeki eksik olmayan veriler kullanılarak bir regresyon denklemi ve regresyon katsayıları elde edilerek kayıp veriler tahmin edilebilir. Regresyon dışında zaman serileri analizi, Bayesyen sınıflandırma, karar ağaçları, maksimum beklenti gibi veri madenciliğinde kullanılan diğer yöntem ve teknikler de kayıp verilerin tahmin edilmesinde kullanılabilir [41,42].

4.2.4. Veri Seçimi ve Örnekleme

Bu adımda kurulacak modele bağlı olarak veri seçimi yapılır. Örneğin tahmin edici bir model için, bu adım bağımlı ve bağımsız değişkenlerin ve modelin eğitiminde kullanılacak veri kümesinin seçilmesi anlamını taşımaktadır. Sıra numarası, kimlik numarası gibi anlamlı olmayan ve diğer değişkenlerin modeldeki ağırlığının azalmasına da neden olabilecek değişkenlerin modele girmemesi gerekmektedir. Bazı veri madenciliği algoritmaları konu ile ilgisi olmayan bu tip değişkenleri otomatik olarak elese de, pratikte bu işlemin kullanılan yazılıma bırakılmaması daha akılcı olacaktır. Modelde kullanılan veri tabanının çok büyük olması durumunda tesadüfîliği bozmayacak şekilde örnekleme yapılması uygun olabilir. Günümüzde hesaplama

olanakları ne kadar gelişmiş olursa olsun, çok büyük veri tabanları üzerinde çok sayıda modelin denenmesi zaman kısıtı nedeni ile mümkün olamamaktadır. Bu nedenle tüm veri tabanını kullanarak bir kaç model denemek yerine, tesadüfî olarak örneklenmiş bir veri tabanı parçası üzerinde birçok modelin denenmesi ve bunlar arasından en güvenilir ve güçlü modelin seçilmesi daha uygun olacaktır [40].

4.2.5. Dönüştürme

Verilere aldıkları değerlere göre dönüştürme işlemi yapılır. Örneğin yapılan harcamalar verisi, harcama>1000 ve harcama<=1000 olarak gruplanarak model içerisinde kullanılabilir veya uç değerler varsa bunların modele aynen katılması uygun olmayabilir. Bu durumda veriler üzerinde standartlaştırma veya normalizasyon işlemleri uygulamak uygun olacaktır [41,43].

Ayrıca modelde kullanılan algoritma, verilerin gösteriminde önemli rol oynayacaktır. Örneğin bir uygulamada bir yapay sinir ağı algoritmasının kullanılması durumunda kategorik değişken değerlerinin evet/hayır olması; bir karar ağacı algoritmasının kullanılması durumunda ise örneğin gelir değişken değerlerinin yüksek/orta/düşük olarak gruplanmış olması modelin etkinliğini artıracaktır [44].

4.3. Modelin Kurulması ve Değerlendirilmesi

Tanımlanan problem için en uygun modelin bulunabilmesi, olabildiğince çok sayıda modelin kurularak denenmesi ile mümkündür. Bu nedenle veri hazırlama ve model kurma aşamaları, en iyi olduğu düşünülen modele varılıncaya kadar yinelenen bir süreçtir. Model kuruluş süreci denetimli (supervised) ve denetimsiz (unsupervised) öğrenimin kullanıldığı modellere göre farklılık göstermektedir [45].

Örnekten öğrenme olarak da isimlendirilen denetimli öğrenimde, bir denetçi tarafından ilgili sınıflar önceden belirlenen bir kritere göre ayrılarak, her sınıf için çeşitli örnekler verilir. Sistemin amacı verilen örneklerden hareket ederek her bir sınıfa ilişkin özelliklerin bulunmasıdır [38].

Öğrenme süreci biter ve daha sonrasında verilmiş olan yeni örnekler üzerine tanımlanan kurallar uygulanır. Bunun sonucunda oluşturulan model yardımı ile yeni örneklerin hangi sınıfa dahil olduğu bilgisi elde edilir.

Denetimli öğrenmede hedef, girdi ölçülerinin sayısını temel alarak çıktı ölçüsünün değerini tahmin etmektir [65]. Örnekten öğrenme olarak da isimlendirilen denetimli öğrenimde, bir denetçi tarafından ilgili sınıflar önceden belirlenen bir kritere göre ayrılarak, her sınıf için çeşitli örnekler verilir. Sistemin amacı verilen örneklerden hareket ederek her bir sınıfa ilişkin özelliklerin bulunması ve bu özelliklerin kural cümleleri ile ifade edilmesidir. Öğrenme süreci tamamlandığında, tanımlanan kural cümleleri verilen yeni örneklere uygulanır ve yeni örneklerin hangi sınıfa ait olduğu kurulan model tarafından belirlenir [66].

Denetimsiz öğrenmede ise çıktı ölçüsü yoktur ve hedef girdi ölçüleri kümesi arasındaki birliktelik ve örüntüleri betimlemektir [65]. Denetimsiz öğrenmede, kümeleme analizinde olduğu gibi ilgili örneklerin gözlenmesi ve bu örneklerin özellikleri arasındaki benzerliklerden hareket ederek sınıfların tanımlanması amaçlanmaktadır [66].

Bir modelin doğruluğunun test edilmesinde kullanılan en basit yöntem basit geçerlilik (simple validation) testidir. Bu yöntemde tipik olarak verilerin %5 ile %33 arasındaki bir kısmı test verileri olarak ayrılır ve kalan kısım üzerinde modelin öğrenimi gerçekleştirildikten sonra, bu veriler üzerinde test işlemi yapılır. Bir sınıflama modelinde yanlış olarak sınıflanan olay sayısının, tüm olay sayısına bölünmesi ile hata oranı, doğru olarak sınıflanan olay sayısının tüm olay sayısına bölünmesi ile ise doğruluk oranı hesaplanır. (Doğruluk Oranı=1-Hata Oranı) [39,46].

Az sayıda veriye sahip olunduğu durumda, kullanılabilecek başka bir yöntem ise çapraz geçerlilik (cross validation) testidir. Bu yöntemde ilk olarak veri kümesi iki eşit parçaya ayılır. Sonrasında ilk adım olarak A parçası üzerinde model eğitimi ve B parçası üzerinde test işlemi gerçekleştirilir. İkinci adımda aynı işlemler tam tersi yönde yani B parçası üzerinde model eğitimi ve A parçası üzerinde test işlemi gerçekleştirilir. Bu işlemlerden elde edilen hata oranlarının ortalaması kullanılır.

Bir kaç bin veya daha az satırdan meydana gelen küçük veri tabanlarında, verilerin n gruba ayrıldığı n katlı çapraz geçerlilik (n-fold cross validation) testi tercih edilebilir. Verilerin örneğin 10 gruba ayrıldığı bu yöntemde, ilk aşamada birinci grup test, diğer gruplar öğrenim için kullanılır. Bu süreç her defasında bir grubun test, diğer grupların öğrenim amaçlı kullanılması ile sürdürülür. Sonuçta elde edilen on hata oranının ortalaması, kurulan modelin tahmini hata oranı olacaktır [67].

Bootstrapping küçük veri kümeleri için modelin hata düzeyinin tahmininde kullanılan bir başka tekniktir. Çapraz geçerlilikte olduğu gibi model bütün veri kümesi üzerine kurulur. Daha sonra en az 200, bazen binin üzerinde olmak üzere çok fazla sayıda öğrenim kümesi tekrarlı örneklemelerle veri kümesinden oluşturularak hata oranı hesaplanır [38].

Bir model oluşturulmaya çalışılırken, aynı tip teknikle farklı değişkenlerin kullanıldığı veya farklı algoritma ve araçların denendiği birbirinden farklı modeller kurulabilir. Model oluşturulurken hangi tekniğin kullanacağına karar vermek oldukça güçtür. Bu nedenle model oluşturulmaya çalışılırken farklı teknikler ile denemeler yapılarak en uygun modelin bulunmasında fayda bulunmaktadır.

Önemli diğer bir değerlendirme kriteri modelin anlaşılabilirliğidir. Bazı uygulamalarda doğruluk oranlarındaki küçük artışlar çok önemli olsa da, birçok işletme uygulamasında ilgili kararın niçin verildiğinin yorumlanabilmesi çok daha büyük önem taşıyabilir. Çok ender olarak yorumlanamayacak kadar karmaşıklaşsalar da, genel olarak karar ağacı ve kural temelli sistemler model tahmininin altında yatan nedenleri çok iyi ortaya koyabilmektedir [38].

Kurulan modelin doğruluk derecesi ne denli yüksek olursa olsun, gerçek dünyayı tam anlamı ile modellediğini garanti edebilmek mümkün değildir. Yapılan testler sonucunda geçerli bir modelin doğru olmamasındaki başlıca nedenler, model kuruluşunda kabul edilen varsayımlar ve modelde kullanılan verilerin doğru olmamasıdır [38].

4.4. Modelin Kullanılması

Kurulan ve geçerliliği kabul edilen model doğrudan bir uygulama olabileceği gibi, bir başka uygulamanın alt parçası olarak kullanılabilir. Kurulan modeller risk analizi, kredi değerlendirme, dolandırıcılık tespiti gibi işletme uygulamalarında doğrudan kullanılabileceği gibi, promosyon planlaması simülasyonuna entegre edilebilir veya tahmini envanter düzeyleri yeniden sipariş noktasının altına düştüğünde, otomatik olarak sipariş verilmesini sağlayacak bir uygulamanın içine gömülebilir [37].

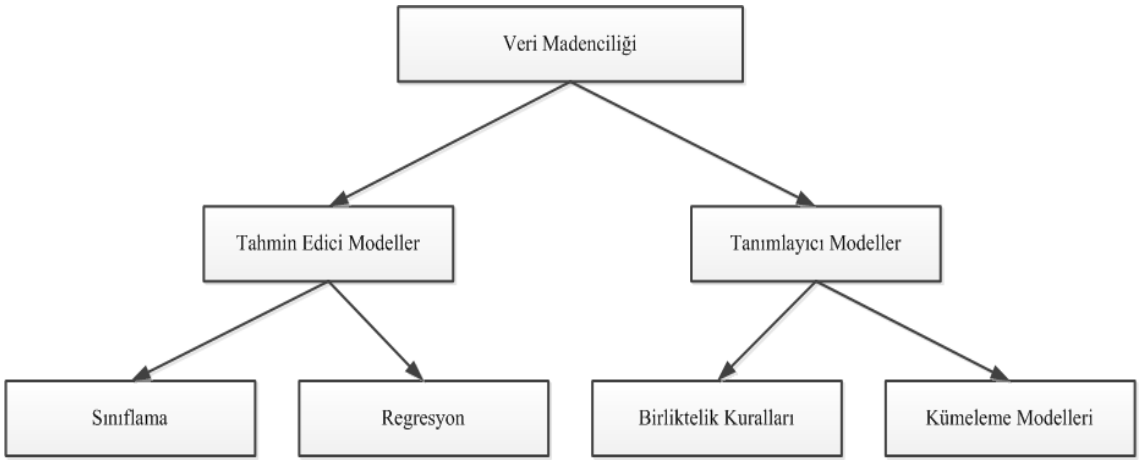
4.5. Modelin İzlenmesi

Zaman içerisinde bütün sistemlerin özelliklerinde ve dolayısıyla ürettikleri verilerde ortaya çıkan değişiklikler, kurulan modellerin sürekli olarak izlenmesini ve gerekiyorsa yeniden düzenlenmesini gerektirecektir. Tahmin edilen ve gözlenen değişkenler arasındaki farklılığı gösteren grafikler model sonuçlarının izlenmesinde kullanılan yararlı bir yöntemdir [37].

5. VERİ MADENCİLİĞİ METODLARI 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 [23].

Veri madencilği metodları, tahmin edici ve tanımlayıcı olmak üzere iki ana başlık altında toplanmaktadır.



Şekil 5.1 Veri Madenciliği Modelleri [68]

5.1. Tahmin Edici Modeller

Tahmin edici modellerde, sonuçları bilinen verilerden hareket edilerek bir model geliştirilmesi ve kurulan bu modelden yararlanılarak sonuçları bilinmeyen veri kümeleri için sonuç değerlerin tahmin edilmesi amaçlanmaktadır. Örneğin bir banka önceki dönemlerde vermiş olduğu kredilere ilişkin gerekli tüm verilere sahip olabilir. Bu verilerde bağımsız değişkenler kredi alan müşterinin özellikleri, bağımlı değişken ise kredinin geri ödenip ödenmediğidir. Bu verilere uygun olarak kurulan model, daha

sonraki kredi taleplerinde müşteri özelliklerine göre verilecek olan kredinin geri ödenip ödenmeyeceğinin tahmininde kullanılmaktadır [47].

Sınıflama ve Regresyon modelleri tahmin edici modellerdir.

5.1.1. Sınıflama ve Regresyon Teknikleri

Sınıflandırma ve regresyon, önemli veri sınıflarını ortaya koyan veya gelecek veri eğilimlerini tahmin eden modelleri kurabilen veri analiz modelleridir. Mevcut verilerden hareket ederek geleceğin tahmin edilmesinde faydalanılan ve veri madenciliği teknikleri içerisinde en yaygın kullanıma sahip olan sınıflama ve regresyon modelleri arasındaki temel fark, tahmin edilen bağımlı değişkenin kategorik (kesikli) veya süreklilik gösteren bir değere sahip olmasıdır. Kesikli değerler sınıflandırma, sürekli değerler ise regresyon yöntemleriyle kullanılırlar. Sınıflandırma iki adımdan oluşur; eğitim veri setinden model kurulması ve bu modelin sınıfları bilinmeyen diğer veri setinde uygulanmasıdır. Sınıflandırma ve regresyon modellerinde kullanılabilen başlıca teknikler aşağıda açıklanmıştır [47].

5.1.1.1. Karar ağaçları

Karar ağacı, karar düğümü, dal ve yaprak yapılarından oluşan ağaca benzer bir yapıdır. Kategorik değerli hedef fonksiyonlar için görsel olarak bazı kurallar, şartlar tanımlayarak, kararın verilmesine imkan sağlamaktadır. Karar ağaçları birçok test uygular ve hedefin tahmini için en uygun sıraya ulaşmayı dener. Her test, yaprak düğümüne ulaşana kadar daha fazla test yapmayı sağlayan dalları oluşturur. Kökten hedef yaprağa doğru giden yol, kuralı ifade eder ve bu hedefi sınıflandırır [47].

Kuruluşlarının ucuz, yorumlanmalarının kolay olması, veri tabanları sistemleri ile kolayca bütünleştirilmeleri ve güvenirliklerinin iyi olması nedenleriyle yaygın

kullanılmaktadırlar. Karar ağacı, karar vericiye hangi faktörlerin önemli olduğunu ve bu faktörlerin kararın farklı çıktılarıyla nasıl ilişkilendirileceğini anlamasında yardımcı olur [48].

5.1.1.2. Yapay sinir ağları

Yapay sinir ağları sınıflandırma ve regresyon algoritmalarının en karmaşık olanıdır. Genellikle kara kutu olarak düşünülür. Çünkü ilişkilerin bulunmasını açıklayamaz. Sinir ağı eğitimi için çok sayıda veri gerektirir, bu da zaman kaybına neden olur. Ancak bir kere eğitildikten sonra yeni olaylar için tahminleri çok hızlı, neredeyse gerçek zamanlı gerçekleştirir. Sinir ağları doğrudan sayılarla işlem yaptığı için bağımlı ya da bağımsız tüm nümerik olmayan veriler sayılara dönüştürülmelidir. Bu sistemin sinir ağları olarak nitelendirilmesinin nedeni, çalışma mekanizmasının insan beynine benzemesidir [47].

5.1.1.3. Genetik algoritmalar

Genetik algoritmalar evrimsel hesaplama (evolutionary computing) metotlarının örnekleridir ve optimizasyon tipli algoritmalardır [49]. Genetik algoritmalar da yapay sinir ağları gibi biyolojik işlemlerden kaynağını almıştır. Yüzyıllar boyu süren adaptasyonlar ve doğal seleksiyon sonucunda çevre koşullarına en fazla uyum sağlayanlar hayatta kalmışlardır. Genetik algoritmalarında benzer bir çalışma biçimi vardır.

Algoritma ilk olarak popülasyon adı verilen bir çözüm kümesi (öğrenme veri kümesi) ile başlatılır. Bir popülasyondan alınan sonuçlar bir öncekinden daha iyi olacağı beklenen yeni bir popülasyon oluşturmak için kullanılır. Evrim süreci (yeni

popülasyonlar yaratma iterasyonu) tamamlandığında bağımlılık kuralları veya sınıf modelleri ortaya konmuş olur [50].

5.1.1.4. K-En yakın komşu

Bir nesnenin niteliklerini inceleyerek bir sınıf veya gruba yerleştirerek ona en yakın olan nitelikleri aynı gruba dâhil etme prensibine göre çalışır. K, olay ya da gruptaki benzer unsurları ifade etmekte olup, rakamsal verilerdeki ilişkileri ve dizileri tespit etmek için kullanılır. Diğer tahmin modellerinden farklı olarak eğitim seti modelin kendisidir. K-NN, uzaklık kavramına bağlıdır ve uzaklıkların belirlenmesi için ölçü gerektirir. Uygun ölçünün seçimi en hassas noktadır, çünkü eğitim verisinde kullanılan farklı ölçüler tamamen farklı tahminlere neden olabilir [47].

K-NN algoritmasında iki temel parametre vardır:

- Kullanılan en yakın olayın sayısı
- Benzerliği ölçecek bir ölçü

Bu teknolojilerin çoğu, uzun zamandır kullanımda olmakla beraber küçük veri hacimleri için uygulanmışlardır. Ancak, veri ambarları ve OLAP uygulamaları ile analiz edilen veri hacimleri arttırılmıştır [51].

5.2. Tanımlayıcı Modeller

Tanımlayıcı modellerde, karar vermeye rehberlik etmede kullanılabilecek mevcut verilerdeki örüntülerin tanımlanması sağlanmaktadır. Veri setindeki bağıntı ve profilleri

belirleyen ve yorumu kullanıcıya bırakan modellerdir. Dolayısıyla karar vermeye rehberlik etmede kullanılabilecek mevcut verilerdeki bağıntıların tanımlanması sağlanmaktadır. X/Y aralığında geliri, evi ve arabası olan, ayrıca çocukları okul çağında olan aileler ile, çocuğu olmayan ve geliri X/Y aralığından düşük olan ailelerin satın alma örüntülerinin birbirlerine benzerlik gösterdiğinin belirlenmesi tanımlayıcı modellere bir örnektir [46].

5.2.1. Birliktelik Kuralları

Birliktelik kuralları, büyük veri kümeleri arasında birliktelik ilişkileri bulurlar [52]. Depolanan verilerin gün geçtikçe büyümesinden ötürü, şirketler veritabanlarındaki birliktelik kurallarını ortaya çıkarmak istemektedirler. Büyük miktardaki kayıtlardan faydalı birliktelik ilişkileri keşfetmek, şirketlerin karar alma işlemlerini kolaylaştırarak daha verimli hale getirmektedir.

Birliktelik kurallarının kullanıldığı en tipik örnek market sepeti uygulamasıdır. Bu işlem, müşterilerin yaptıkları alısverişlerdeki 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 etkili satış stratejileri geliştirebilirler. Örnegin bir müşteri süt satın alıyorsa, aynı alışverişte sütün yanında ekmek alma olasılıgı nedir? Bu tip bir bilgi ışığında rafları düzenleyen market yöneticileri ürünlerindeki satış oranını arttırabilirler. Örnegin bir marketin müşterilerinin süt ile birlikte ekmek satın alan oranı yüksekse, market yöneticileri süt ile ekmek raflarını yan yana koyarak ekmek satışlarını arttırabilirler [53].

5.2.2. Kümeleme Modelleri

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 [54].

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 [1].

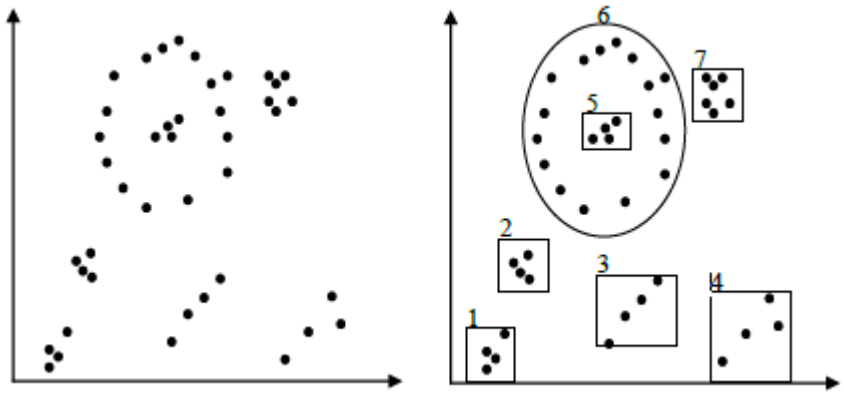
Kümeleme analizi, nesnelerin birbiri ile benzer olan altdizinlere gruplanmasını yapan işleme denir. Böylece nesneler, örneklenen grubun özelliklerini iyi yansıtan etkili bir temsil gücüne sahip olmuş olur. Kümeleme, bir denetimsiz öğrenme yöntemidir.

Kümeleme analizi için başka bir tanım da şu biçimde yapılmaktadır; kümeleme analizi, temel amacı nesneleri (birimleri) sahip oldukları karakteristik özellikleri baz alarak gruplamak olan çok değişkenli teknikler grubudur. Kümeleme analizi, nesneleri küme içerisinde çok benzer biçimde, kümeler arasında farklı olacak biçimde kümeler. Kümeleme işlemi başarılı olursa, bir geometrik çizim yapıldığında nesneler küme içerisinde birbirine çok yakın, kümeler ise birbirinden uzak olacaktır [55].

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 [56].

Aşağıdaki Şekil 5.2'de basit bir kümeleme örneği gösterilmiştir. Bu örnekte soldaki grafikte yer alan veriler giriş olarak verilmiş ve kümeleme işlemi sonucunda ortaya çıkan 7 adet küme sağda gösterilmiştir.



Şekil 5.2 Veri Kümele Örneği [56]

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-Yoğunluk Tabanlı Yöntemler

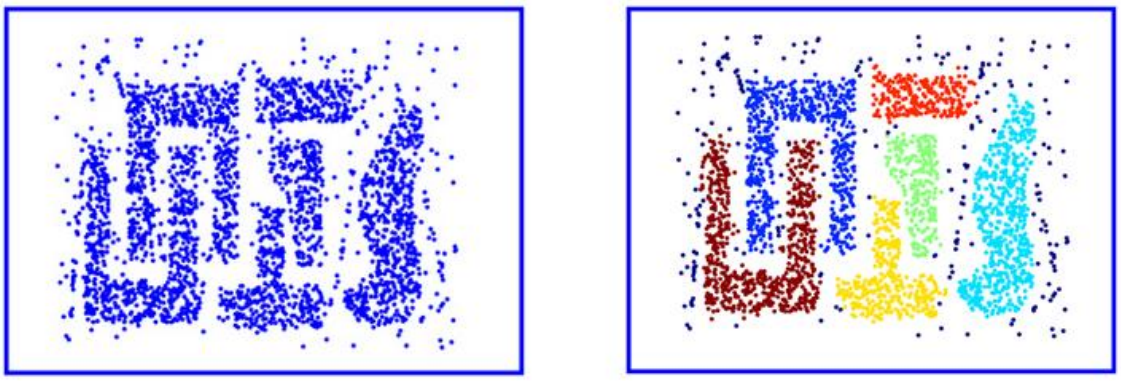
2-Hiyerarşik Yöntemler

3-Model Tabanlı Yöntemler

4-Bölümleme Yöntemleri

5.2.2.1. Yoğunluk Tabanlı Yöntemler

Yoğunluk tabanlı kümelemede alanlar veri yoğunluğunun fazla ve az olmasına göre belirlenir. Üretilen demetler birbirinden farklı şekillerde olabilir. Yoğunluk tabanlı kümeleme güçlü bir yöntem olmasından ötürü istisna verilerden ve gürültülü verilerden etkilenmez. Yoğunluk tabanlı kümeleme işlemi belirlenen veri demetine son veri eklenene kadar devam eder bu aşamalar sürecinde yoğunluk parametresinin verilmesi gerekmektedir [57]. Birçok kümeleme yöntemi nesnelerin birbirleri arasındaki farklılıklarına göre kümeleme yaparken, bu yöntem nesnelerin yoğunluğuna göre gruplama yapar. Yoğunluktan kasıt, analiz edilen nesnelerin sayısıdır. Yoğunluk tabanlı yöntemlere örnek olarak DBscan verilebilir. Aşağıda Şekil 5.3'de bir veri setine DBSCAN algoritması uygulandıktan sonrası oluşturulan kümeler gösterilmektedir.



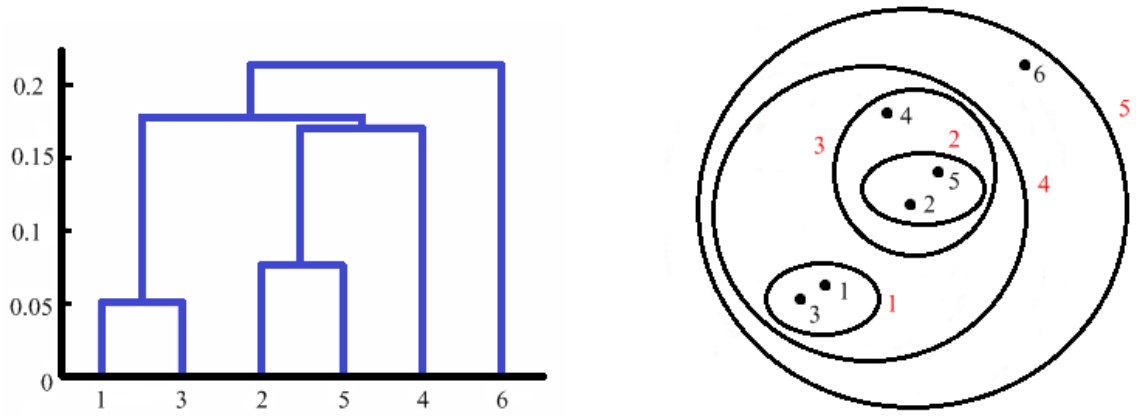
Şekil 5.3 DBSCAN Algoritması ile Kümeleme [57]

5.2.2.2. Hiyerarşik Yöntemler

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. Şekil 5.4'de örnek bir dendogram gösterilmektedir.
- Bu ağaç, yapraklardan gövdeye doğru veya gövdeden yapraklara doğru kurulabilir. Dendogram istenen seviyede kesilerek kümeler elde edilir [56].



Şekil 5.4 Dendogram Yapısına Bir Örnek [56]

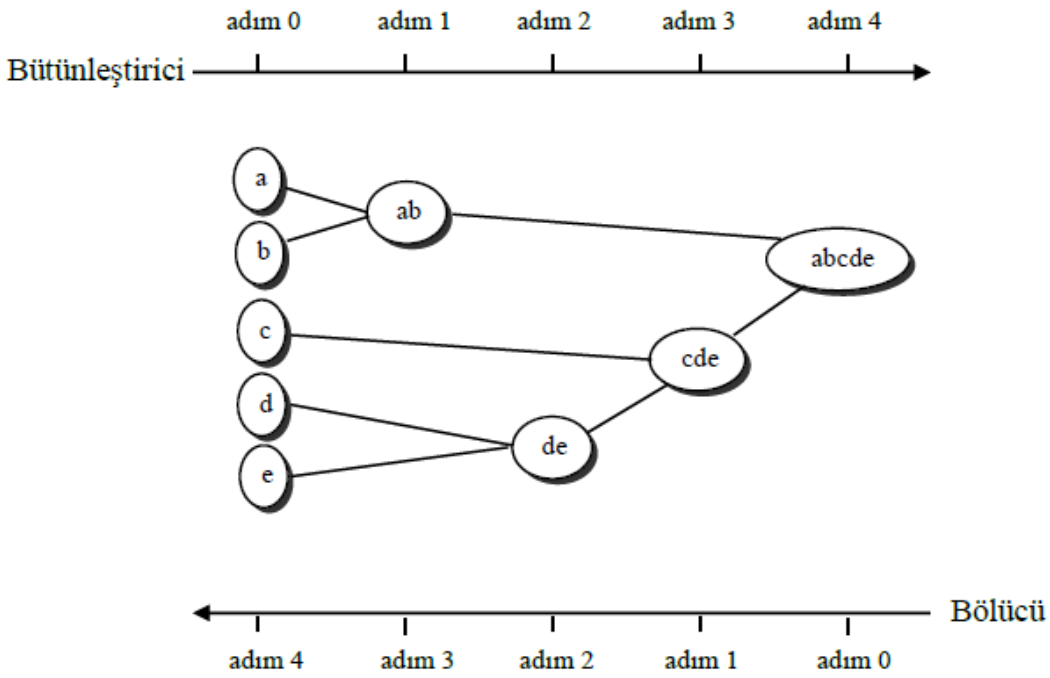
Bir hiyerarşik kümeleme metodu veri nesnelerini bir küme ağacına gruplayarak çalışır. Hiyerarşik kümeleme yöntemleri, hiyerarşik ayrışmanın yukarıdan-aşağıya veya aşağıdan-yukarıya oluşturulmasına bağlı olarak bütünleştirici ve bölücü hiyerarşik kümeleme olarak sınıflandırılabilir (Bkz. Şekil 5.5.). 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 [56].



Şekil 5.5 Bütünleştirici ve Bölücü Hiyerarşik Kümelemenin {a,b,c,d,e} Veri Nesneleri Üzerinde Gösterimi [56]

5.2.2.3. Model Tabanlı Yöntemler

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 [56].

5.2.2.4. Bölme Yöntemleri

Bölme yöntemlerinde, n veri tabanındaki nesne sayısı ve k oluşturulacak küme sayısı olarak kabul edilir. Bölme algoritması n adet nesneyi, k adet kümeye böler ($k < n$). Kümeler tarafsız bölme kriteri olarak nitelendirilen bir kritere uygun oluşturuldugu için

aynı kümedeki nesneler birbirlerine benzerken, farklı kümedeki nesnelerden farklıdırlar [1].

Bölmeli yöntemlerde işlemler şu sıra ile yapılmaktadır:

- İlk olarak başlangıçta kullanılacak küme merkezleri gelişi güzel olarak seçilir.
- Nesnelerin, belirlenen kümelerin merkezlerine olan uzaklıklarına göre yeni küme merkezleri oluşturulur.
- Bu işlemler birbirinden farklı, kendi içerinde homojen, birbirleri arasında benzerlik bulunmayan k adet küme oluşturuluncaya kadar sürdürülür.

En iyi bilinen ve en yaygın kullanılan bölümleme yöntemleri k-ortalamalar (k-means) yöntemi, k-medoids yöntemi ve bunların varyasyonlarıdır. 1967 yılında Mac Queen tarafından bulunan k-ortalamalar algoritması, kümeleme problemini çözen en basit gözetimsiz ögrenme algoritmalarından biridir. Bölümleyici kümeleme tekniklerinden birisi olan k-ortalamalar kümeleme yöntemi, endüstriyel ve bilimsel uygulamalarda en yaygın olarak kullanılan kümeleme algoritmaları arasında yer alır. Bu tezde k-ortalamalar kümeleme yöntemi kullanılmıştır.

Bölmeli kümeleme yöntemlerinde, benzerlik ölçümüdünde en çok kullanılan uzaklık ölçüleri aşağıda gösterilmiştir.

1-Öklit uzaklık ölçüsü: Öklit uzaklık ölçüsü kullanılarak iki birim arasındaki uzaklık formülü aşağıdaki gibidir. En çok kullanılan ölçü birimidir.

$$d(i,j)=\sqrt{\left(x_{i1}-x_{j1}\right)^2+\left(x_{i2}-x_{j2}\right)^2+\ldots+\left(x_{ip}-x_{jp}\right)^2} \tag{5.1}$$

Eğer kümeleme analizinde kullanılacak olan değişkenler belirli bir önem derecesinde ağırlıklandırılmışsa, öklit uzaklık ölçüsü formülü aşağıdaki gibi olur.

$$d(i,j)=\sqrt{w_1\left(x_{i1}-x_{j1}\right)^2+w_2\left(x_{i2}-x_{j2}\right)^2+\ldots+w_p\left(x_{ip}-x_{jp}\right)^2} \tag{5.2}$$

2-Pearson uzaklık ölçüsü: Pearson uzaklık ölçüsü kullanılarak iki birim arasındaki uzaklık formülü aşağıdaki gibidir.

$$d(i,j)=\sqrt{\left(x_{i1}-x_{j1}\right)^2/S_1^2+\left(x_{i2}-x_{j2}\right)/S_2^2+\cdots+\left(x_{ip}-x_{jp}\right)^2/S_p^2} \quad (5.3)$$

Bu formülde kullanılan Sp, uzaklığın hesaplandığı değişkene ait varyanstır. Bununla birlikte farklı gruplar hakkında önceden bilgi sahibi olunmadığı için, uzaklık hesaplanmasında S değerinin kullanılması doğru olmaz. Bu nedenle pearson uzaklık ölçüsü yerine genellikle öklit uzaklık ölçüsü tercih edilir.

3-Manhattan uzaklık ölçüsü: Manhattan uzaklık ölçüsü kullanılarak iki birim arasındaki uzaklık formülü aşağıdaki gibidir. Bu ölçü de birimler arasındaki mutlak uzaklık kullanılır.

$$d(i,j)=\left(\left|x_{i1}-x_{j1}\right|+\left|x_{i2}-x_{j2}\right|+\cdots+\left|x_{ip}-x_{jp}\right|\right) \quad (5.4)$$

4-Minkowski uzaklığı: Minkowski uzaklık ölçüsü genel bir formüldür. Formülde yer alan m değerinin alacağı farklı değerlere göre yeni formüller türetir. Minkowski uzaklık ölçüsü kullanılarak iki birim arasındaki uzaklık fomülü aşağıdaki gibidir.

$$d(i,j)=\left[\left|x_{i1}-x_{j1}\right|^m+\left|x_{i2}-x_{j2}\right|^m+\cdots+\left|x_{ip}-x_{jp}\right|^m\right]^{1/m} \quad (5.5)$$

5.2.2.4.1. K-Ortalamalar Kümeleme Yöntemi

En iyi bilinen ve yaygın kullanılan algoritmalardan biri olan k-ortalamalar kümeleme yöntemi, verileri sınıflandıran bir kümeleme algoritmasıdır. Verilen nesneleri nitelik

veya özelliklerine göre k adet sınıfa ayırmak amacıyla kullanılır. Sınıflandırma, verilerin en yakın veya benzer oldukları küme merkezleri (centroid) etrafına yerleştirilmesi ile gerçekleştirilir [58].

Bu çalışmada kullanılan k-ortalamalar algoritması, J. MacQueen [59] tarafından 1967 yılında tanıtılmıştır. Çalışma yönteminde, Öklit uzaklığı temel alınarak kümeleme yapılmaktadır. Bu yöntem yıllardır endüstriyel ve bilimsel uygulamalarda en yogun kullanılan kümeleme algoritması haline gelmiştir.

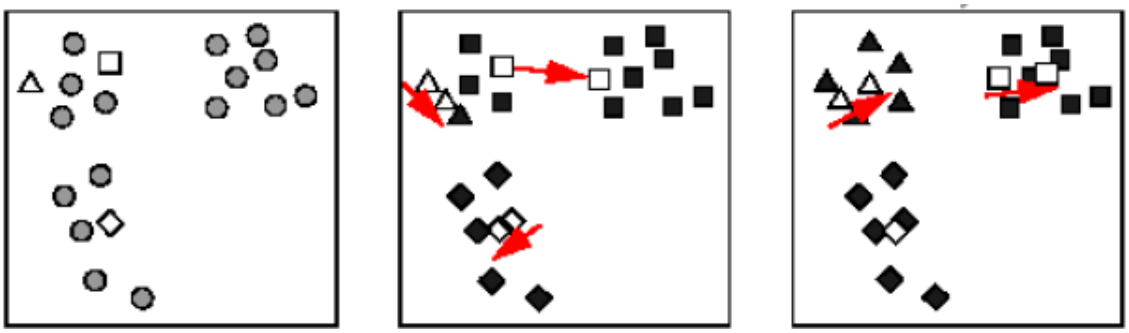
Algoritmaya k-ortalamalar adı verilmesinin nedeni, algoritmanın çalışmasından önce sabit bir küme sayısına ihtiyaç duyulmasıdır. Küme sayısı k ile gösterilir ve elemanlarının birbirlerine olan yakınlıklarına göre oluşacak grup sayısını ifade eder. Buna göre k önceden bilinen ve kümeleme işlemi bitene kadar değeri değişmeyen sabit bir pozitif tamsayıdır [58].

Çok yaygın kullanımı olan bu algoritmanın aşagıda belirtildiği gibi bir takım zayıf yanları da bulunmaktadır:

- Algoritmanın başında giriş parametresi olarak bir k sayısına ihtiyacı vardır. Elde edilecek olan sonuçlar k sayısına göre değişkenlik gösterebilir. Eger küme sayısı belirli degil ise deneme yoluyla en uygun sayı bulunur.
- Aşırı gürültü ve istisna veriler algoritmayla hesaplanan ortalamayı değiştirdiği için k-ortalamalar algoritması gürültü ve istisnaya karsı çok duyarlıdır. Algoritma
  uygulanmadan önce veriler gürültü veya istisnadan temizlenebilir.
- Çakışan kümelerde iyi sonuç vermez.
- Her eleman aynı anda verilen bir kümenin içindedir veya dışındadır.
- K-ortalamalar algoritması sadece sayısal veriler ile kullanılabilmektedir. Kategorik verilerin kümelenmesi için k-ortalamalar algoritması bir çözüm sunmaz [60].

Aşağıda Şekil 5.6'da K-ortalamalar algoritmasının çalışma şekline bir örnek verilmiştir. Aşağıdaki örnekte k değeri 3 olarak seçilmiştir. Beyaz renkle belirtilen Δ,□,◊

simgeleride Şekil 5.6 (a)'da başlangıçta rastgele seçilen küme merkezlerini temsil etmektedir. Şekil 5.6 (b)'de noktaların hangi küme merkezine yakınsa onun simgesini aldığı görülmektedir. Ayrıca ilk oluşan bu üç kümenin ortalamaları alınarak yeni merkez noktalarının belirlenmesi gösterilmektedir. Değişen küme merkezleri Şekil 5.6 (b)'de oklarla gösterilmektedir. Şekil 5.6 (c)'de aynı işlemin tekrar edildiğinde küme merkezlerinin değişimi görülmektedir. Bu şekilde baslangıç durumunda rastgele seçilen küme merkezleri, sürekli yinelemeler ile gerçek kümelenme alanlarının ortasına doğru yaklaşır. Bu işleme merkeze yakınsama (convergence) denir. Merkeze yakınsama minimum seviyeye geldiğinde veya durduğunda kümeleme işlemi sona erer [58].



(a) İlk küme merkezleri (b) İlk ortalama hesabı (c) Merkeze yakınsama

Şekil 5.6 K-ortalamalar Kümele Algoritması [69]

5.2.2.4.1.1. K-Ortalamalar Algoritmasının Adımları

Algoritmanın ilk adımında öncelikle küme merkezlerini veya diğer bir deyişle küme ortalamasını temsil edecek k adet eleman belirlenir. MacQueen algoritmasında küme merkezleri ilk k adet elemandan seçilir. Ancak elemanların değerleri birbirine çok yakın ise, seçim rastgele yapılır veya birbirinden uzak elemanlar seçilir. Bu noktaların her biri prototip olarak adlandırılır. Belirlenen bu elemanlar tek elemanlı başlangıç kümeleridir ve ilk küme merkezlerini oluştururlar. Kümenin ağırlıklı ortalama değerine sahip olan ya da bu değere en yakın olan elemanı küme merkezi olarak adlandırılır [58].

İkinci adımda, okunan elemanlar kendilerine en yakın k adet küme merkezinden birine dahil edilir. Elemanların küme merkezine olan yakınlık derecesini bulmak amacıyla çeşitli geometrik yöntemler kullanılır. Bunlardan bir tanesi, kümeler arasındaki sınırları belirleyerek, elemanların hangi küme merkezine daha yakın olduklarını tespit eder. Bunun için önce iki küme merkezi bir doğruyla birleştirilir. Bu doğrunun orta noktasından geçen ve doğruyu dik kesen baska bir doğru daha geçirildiğinde, bu dogru iki kümenin sınırı olarak kabul edilir. Bulunan sınır çizgisi dikkate alınarak, elemanların hangi kümeye dahil edilecegi ortaya çıkar [58].

Noktalar arası uzaklığın hesaplanmasında en çok kullanılan yöntem Öklit bağıntısıdır. İki boyutlu bilgilerde, iki küme merkezinin birleştirilmesinde doğru kullanılırken, boyut sayısı arttığında doğru yerine düzlem kullanılır. Çok boyutlu bilgilerde çok boyutlu düzlemler kullanılır. Algoritmanın geometrik gösteriminde küme sınırları yukarıda anlatıldığı şekilde belirlenmektedir. Bilgisayar programları ile geliştirilen k-ortalamalar algoritmalarında ise, düzlemler yerine noktalar arasındaki uzaklıklar hesaplanarak, noktaların merkeze yakınlığı dikkate alınmaktadır [28].

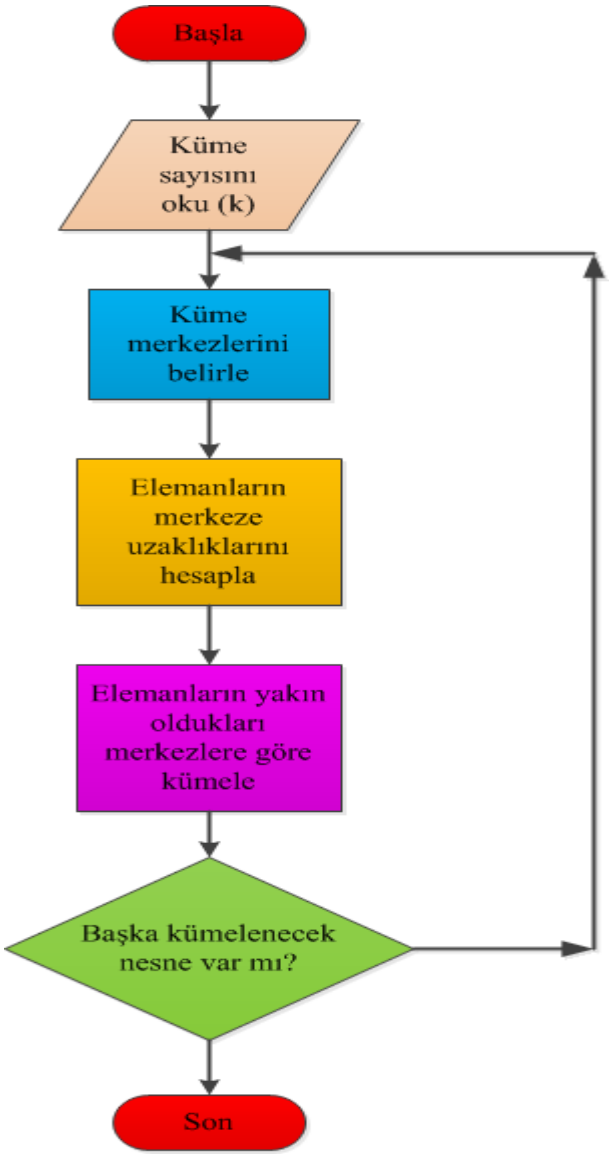
Üçüncü adımda her bir kümeye eklenen yeni eleman ile, küme elemanlarının ağırlıklı ortalaması tekrar hesaplanarak yeni bir küme merkezi bulunur. Ağırlıklı ortalama kümenin her bir boyutundaki bütün elemanların ortalama değerlerinin alınması ile hesaplanır. Algoritmanın basında seçilen elemanlar küme merkezini oluştururken, ikinci döngü sonucunda bulunan yeni küme merkezleri artık bir küme elemanı değil, sadece bir ortalama değerdir. Bundan sonraki seçim işlemlerinde küme merkezini bu yeni eleman temsil eder. Her bir döngüde elemanlar farklı bir kümeye dahil edilebilirler [58].

Kümeleme işlemi, tüm elemanların tekrar aynı veya farklı bir kümeye dahil edilmesiyle devam eder. Elemanların bir kümeye dahil edilmesi ve küme merkezlerinin tekrar hesaplanması işlemlerine ait döngü, küme sınırlarının değişimi bitene kadar devam eder. K-ortalamalar algoritması ile uygulamalarda genellikle birkaç düzine döngü sonrası kararlı bir küme grubu ortaya çıkar [58].

K-ortalamalar algoritması bilgisayar programına uygulanırken asagıdaki adımlar izlenir:

1. Küme sayısı (K) okunur. Bu değer algoritmaya dışardan verilir.
2. K adet rastgele küme merkezi belirlenir.İlk K eleman merkez olabilir.
3. Elemanların merkezlere yakınlıkları hesaplanır.
4. Elemanlar, yakın oldukları merkezlere göre kümelenir.
5. Kümelerin ortalamaları hesaplanarak yeni küme merkezleri belirlenir.
6. Kümelenecek baska eleman var mı? Hayır ise (3)'e git.Evet ise dur [58].

Algoritmanın akış diyagramı Şekil 5.7'de görülmektedir.



Şekil 5.7 K-ortalamalar Algoritmasının Adımları [58]

Bazı uygulamalarda k-ortalamalar algoritmasının çalıştırılması sonucunda, verilerin büyük çoğunluğunun aynı kümeye dahil edildikleri görülür. Büyük kümenin çevresinde bir kaç küçük küme de yer alır. Bunun nedeni verilerin büyük çoğunluğunun birbirine yakın özellikler taşıması ve az sayıda verinin farklı özellikte olmasıdır. Bu tipteki uygulamalara örnek olarak, sahtekarlık tespiti ve üretim hataları verilebilir. Her iki uygulamada da önemli sayıda veri, istenen özellikleri taşımaktadır. Gürültü/istisna adı verilen az sayıda veri, istenilen özelliklerin dışında kaldıkları için büyük kümenin dışında konumlanmışlardır [58].

K-ortalamalar algoritması ile verileri kümelere ayırmak için geometrik veya aritmetik hesaplama yöntemleri kullanılmaktadır. Bu yöntemler asagıda örneklerle tanıtılmıştır.

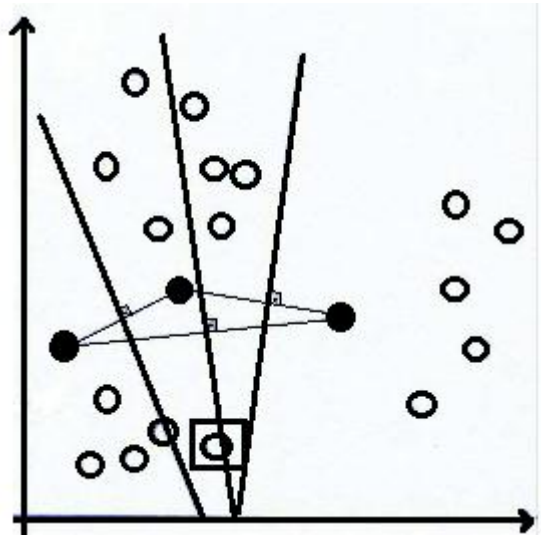
5.2.2.4.1.2. Geometrik Hesaplama

Bu yöntemde, kümelenecek veriler koordinat sisteminde birer nokta olarak ele alınır. Noktalar arası uzaklığın hesaplanmasında en çok kullanılan yöntem Öklit bağıntısıdır. Küme sınırlarının belirlenmesi için iki boyutlu sistemde doğru, boyut sayısı arttığında ise doğru yerine düzlem kullanılır. Çok boyutlu bilgilerde çok boyutlu düzlemler kullanılır. Geometrik hesaplama yöntemini bir örnek üzerinden açıklamak için, 20 elemanın üç kümeye bölünmesi aşağıda şekillerle anlatılmıştır [58].

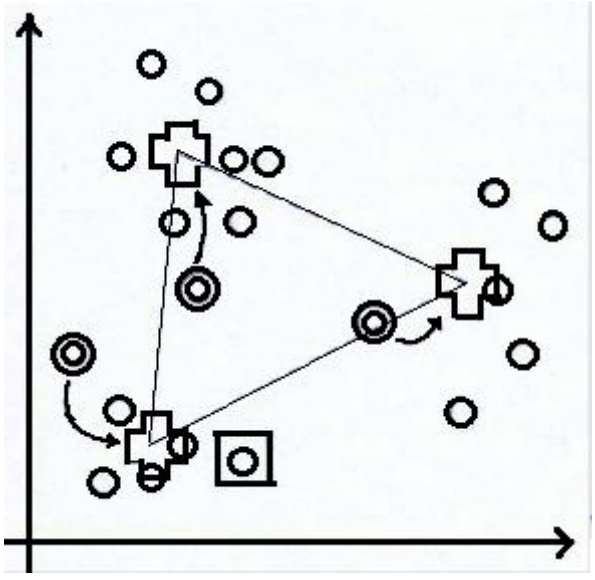
İlk adımda başlangıç küme merkezlerini temsil edecek 3 eleman seçilir. Şekil 5.8'de içi dolu çember şekliyle gösterilen noktalar ilk küme merkezleridir. İkinci adımda, diğer bütün elemanları yakın oldukları bir küme merkezlerine dahil etmek için her bir kümenin sınırı belirlenir. Küme merkezlerinden eşit uzaklıkta bulunan noktalar küme sınırını oluşturur. Bu noktaları belirlemek için küme merkezleri birer doğruyla birleştirilir. Doğruların orta noktalarından geçen ve doğruları dik kesen başka doğrular çizildiginde, kümelerin sınırları ortaya çıkarılır. Şekil 5.8'de ilk küme merkezleri noktalı çizgiler ile birleştirilmiştir. Bunları dik kesen kalın çizgiler küme sınırlarıdır.

Kümelerin sınır çizgilerine bakılarak diğer elemanların hangi kümeye dahil edileceği görülür [58].

Üçüncü adımda, her bir kümedeki elemanların ortalama değerleri alınarak küme merkezleri yeniden hesaplanır. Şekil 5.9'da yeni küme merkezleri artı sembolünün içinde yer almaktadır. Önceki küme merkezlerinin çevresine çember çizilerek, merkezlerin nereden nereye değiştiği gösterilmiştir. Şekil 5.8'de kare içine alınarak işaretlenen eleman, ilk döngü sonrasında ikinci kümeye dahil olmustur. İkinci döngü sonrasında ise Şekil 5.9'da görüldüğü gibi birinci kümeye dahil edilmiştir [58].

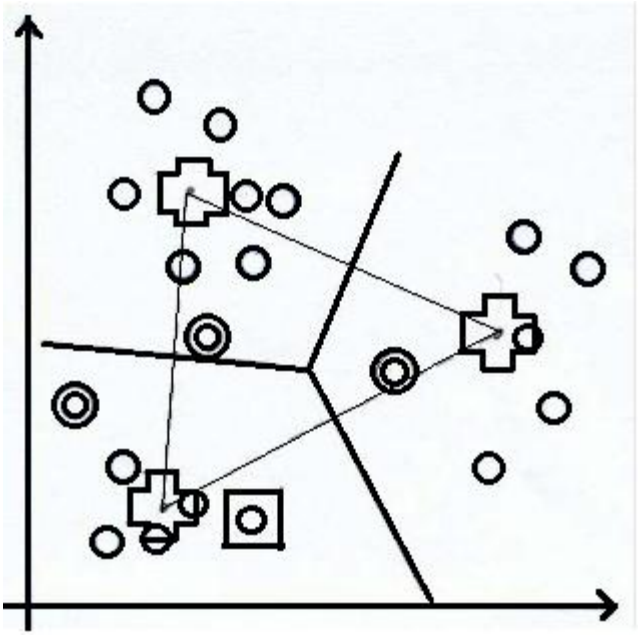


Şekil 5.8 Geometrik Hesaplama Yöntemiyle İlk Kümelerin Belirlenmesi [58]



Şekil 5.9 Noktaların Kümelere Dahil Edilmesi Sonrasında Yeni Küme Merkezleri [58]

Döngü içinde ikinci adım tekrarlandıktan sonra, bütün elemanlar en yakın oldukları küme merkezlerine yeniden dahil edilirler. Her döngüde küme merkezlerinin değişimiyle küme sınırları da değişmektedir. Şekil 5.10'da yeni küme sınırları gösterilmektedir. İşlem döngüsü, küme sınırlarının değişimi durana kadar sürer [58].



Şekil 5.10 Her Döngü Sonrasında Küme Sınırları Değişmektedir [58]

5.2.2.4.1.3. Aritmetik Hesaplama

Bilgisayar programları ile geliştirilen k-ortalamalar algoritmalarında aritmetik hesaplama yöntemlere başvurulmaktadır. Aritmetik hesaplamada, dogrular veya düzlemler yerine eleman değerleri arasındaki uzaklıklar hesaplanarak, elemanların merkezlere yakınlığı bulunmaktadır [28].

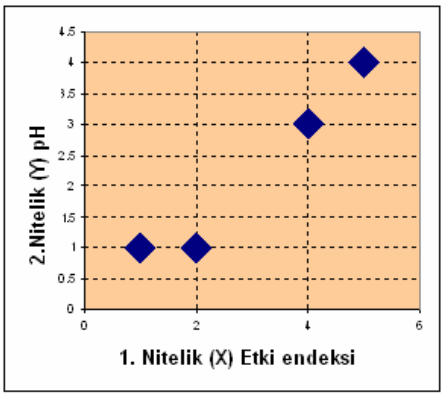
Aritmetik hesaplama yöntemini bir örnek üzerinden açıklamak için, tablo 5.1'de görülen veriler kullanılmıştır. Tabloda 4 çeşit ilaca ait veriler yer almaktadır. Her bir ilaç nesne

olarak, ilaçların etki endeksi ve ph degeri nitelik değerleri olarak ele alınmıştır. Bu örnekte veriler iki kümeye (k=2) ayrılacaktır [58].

Aritmetik hesaplamada, tablodaki her bir ilaca (nesneye) ait nitelikler, nitelik uzayında (x,y) ile bir koordinat noktasını göstermektedir. Her bir nitelik bir koordinat boyutuna karşılık gelmektedir. Buna göre, A ilacı (1,1), B ilacı (2,1), C ilacı (4,3) ve D ilacı (5,4) noktası ile temsil edilmektedir. İlaçların iki boyutlu koordinat sistemindeki görüntüsü Şekil 5.11'de yer almaktadır [58].

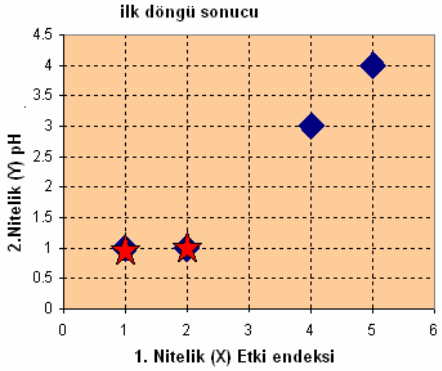
Nesne	1. Nitelik (x) etki endeksi	2. Nitelik (y) pH
İlaç A	1	1
İlaç B	2	1
İlaç C	4	3
İlaç D	5	4

Tablo 5.1 Kümeleme İçin Kullanılacak Veriler [61]



Şekil 5.11 Aritmetik Hesaplamaya Uygulanacak Veriler [61]

Birinci adımda, ilk küme merkezleri olarak, (0,0) noktasına yakınlıkları nedeniyle A ve B ilaçları seçilmiştir. Küme merkezlerinin koordinatları (1,1) ve (2,1)'dir. İlk küme merkezleri Şekil 5.12'de gösterilmiştir [58].



Şekil 5.12 Aritmetik Hesaplamada Seçilen İlk Küme Merkezleri [61]

İkinci adımda, seçilen ilk küme merkezleri ile diğer elemanlar arasındaki uzaklık hesaplanır. Öklit bağıntısı kullanılarak hesaplanan uzaklık matrisindeki her sütun bir ilacı (nesneyi) temsil etmektedir. Uzaklık matrisinin ilk satırında nesnelerin ilk küme merkezine olan uzaklıkları, ikinci satırında aynı şekilde, nesnelerin ikinci küme merkezine olan uzaklıkları yer alır. Örneğin ilk satırda A ilacının koordinat noktası ile, B, C ve D ilaçlarının koordinat noktaları arasında hesaplanan uzaklık değerleri yer almaktadır. İkinci satırda B ilacının koordinat noktası ile, A, C ve D ilaçlarının ilaçlarının koordinat noktaları arasında hesaplanan degerler yer almaktadır. Hesaplanan ilk uzaklık matrisi D aşagıdaki gibidir [58];

$$\mathbf{D}^0 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 1 & 0 & 2.83 & 4.24 \end{bmatrix} \quad \begin{matrix} \mathbf{c}_1 = (1,1) \\ \mathbf{c}_2 = (2,1) \end{matrix}$$

$$\begin{matrix} A & B & C & D \end{matrix}$$

$$\begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} \quad \begin{matrix} X \\ Y \end{matrix}$$

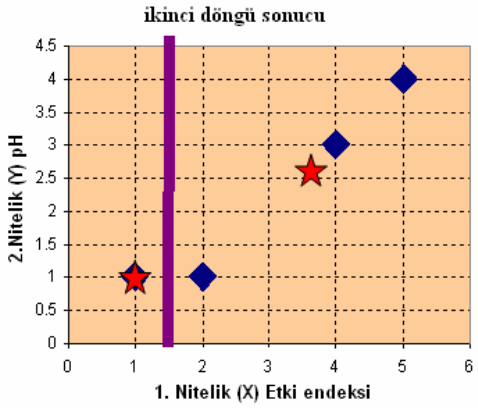
Üçüncü adımda, uzaklık matrisinde bulunan değerlere göre nesneler bir kümeye dahil edilir. Her nesne için matrisin satırları arasındaki değerlerin en küçüğü bulunur. Örneğin D nesnesi için birinci ve ikinci satır değerleri 5 ve 4.24'dür. En küçük değer 4.24 olduğu için D nesnesi ikinci kümeye dahil edilir. Aynı şekilde, C nesnesinin satır değerleri 3.62 ve 2.83'dür. En küçük deger 2.83 oldugu için C nesnesi ikinci kümeye dahil edilir. A ve B nesneleri, bu aşamada merkezleri oldukları kümelerin içinde kalırlar. Nesnelerin dahil edildikleri kümeler, ayrı bir küme matrisinde gösterilir. Asagıdaki G isimli küme matrisinde, ilk satır birinci kümeyi, ikinci satır da ikinci

kümeyi göstermektedir. Nesneler kümeye dahil ise 1, degil ise 0 değerini alırlar. Matristen de görüldüğü gibi, A nesnesi birinci kümenin hem merkezi hem de tek elemanıdır. B nesnesi ise ikinci kümenin merkezidir ve bu kümede C ve D nesneleri diğer küme elemanlarıdır [58].

$$\mathbf{G}^0 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix} \quad \begin{matrix} A & B & C & D \end{matrix}$$

Küme matrisinin oluşturulmasıyla ilk döngü sona erer. İkinci döngü yeni küme merkezlerinin belirlenmesi işlemi ile başlar. Birinci kümede tek eleman olduğu için küme merkezinin değeri (1,1) değişmez. İkinci kümenin yeni küme merkezini hesaplamak için her bir nesnenin x ve y degerlerinin aritmetik ortalaması alınır. Elde edilen ortalama x ve y değeri, yeni küme merkezinin koordinat değeridir. İkinci kümenin üç elemanı ile ortalama x ve y değerleri aşağıdaki gibi hesaplanır [58];

$$c_2 = (\frac{2+4+5}{3}, \frac{1+3+4}{3}) = (\frac{11}{3}, \frac{8}{3})$$



Şekil 5.13 Aritmetik Hesaplamada İkinci Döngüde Oluşan Küme Merkezleri [61]

İkinci döngüde hesaplanan yeni küme merkezleri ile ortaya çıkan kümeler Şekil 5.13'de gösterilmektedir. Şekilde de görüldüğü gibi ikinci kümenin merkezi bir nesne değil bir noktadır. Yeni küme merkezleri oluşturulduktan sonra, diğer nesnelerin yeni merkezlere

olan uzaklıkları tekrar hesaplanır. Hesaplama birinci döngünün ikinci adımında olduğu gibi Öklit bağıntısını kullanarak yapılır. İkinci döngüde oluşan uzaklık matrisi aşağıdaki gibidir [58];

$$\mathbf{D}^1 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 3.14 & 2.36 & 0.47 & 1.89 \end{bmatrix} \quad \begin{matrix} \mathbf{c}_1 = (1,1) \\ \mathbf{c}_2 = (\frac{11}{3}, \frac{8}{3}) \end{matrix}$$

$$\begin{matrix} A & B & C & D \end{matrix}$$

$$\begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} \quad \begin{matrix} X \\ Y \end{matrix}$$

Uzaklık matrisinde bulunan en küçük değerlere göre nesneler bir kümeye dahil edilir. Bunun için ilk döngüde olduğu gibi her nesne için matrisin birinci ve ikinci satırlarındaki değerlerin en küçük olanı bulunur. En küçük uzaklığa göre her bir nesne en yakın olduğu kümeye dahil edilir. Küme matrisinin oluşturulmasıyla, yeni kümelenme ortaya çıkar. Bu sefer A ve B nesneleri birinci kümenin elemanları haline gelirler, C ve D nesneleri ise ikinci kümeyi oluştururlar. G küme matrisi şöyledir [58];

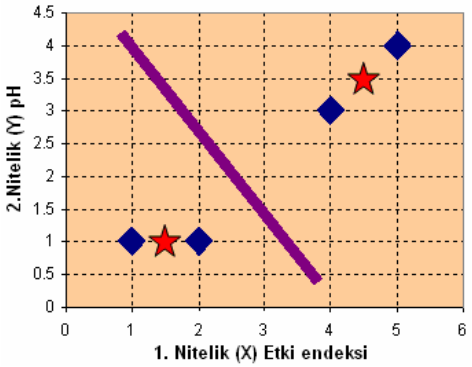
$$\mathbf{G}^1 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}$$

$$\begin{matrix} A & B & C & D \end{matrix}$$

Üçüncü döngü yeni küme merkezlerinin belirlenmesi işlemi ile başlar. Yeni merkezlerin koordinatlarını hesaplamak için, her bir kümeyi oluşturan nesnelerin x ve y değerlerinin aritmetik ortalaması alınır. Birinci ve ikinci kümelerin yeni küme merkezleri şöyledir [58];

$$c_1 = (\frac{1+2}{2}, \frac{1+1}{2}) = (1\frac{1}{2}, 1) \qquad c_2 = (\frac{4+5}{2}, \frac{3+4}{2}) = (4\frac{1}{2}, 3\frac{1}{2})$$

Bu döngüde her iki küme merkezi de nesne değil, nokta olmuştur. Şekil 5.14 üçüncü döngüde ortaya çıkan kümelenmeyi göstermektedir [58].



Şekil 5.14 Aritmetik Hesaplamada Üçüncü Döngüde Oluşan Küme Merkezleri [61]

Küme nesnelerinin yeni küme merkezlerine olan uzaklıkları tekrar hesaplanır. Oluşan uzaklık matrisi şöyledir;

$$\mathbf{D}^2 = \begin{bmatrix} 0.5 & 0.5 & 3.20 & 4.61 \\ 4.30 & 3.54 & 0.71 & 0.71 \end{bmatrix} \quad \begin{matrix} \mathbf{c}_1 = (1\frac{1}{2}, 1) \\ \mathbf{c}_2 = (4\frac{1}{2}, 3\frac{1}{2}) \end{matrix}$$

$$\begin{matrix} A & B & C & D \end{matrix}$$

$$\begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} \begin{matrix} X \\ Y \end{matrix}$$

Uzaklık matrisinden, en küçük uzaklığa göre her bir nesne en yakın olduğu kümeye dahil edilir. Böylece ortaya çıkan küme matrisi şöyledir [58];

$$\mathbf{G}^2 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}$$

$$\begin{matrix} A & B & C & D \end{matrix}$$

Küme matrisinden görüldüğü gibi A ve B nesneleri birinci, C ve D nesneleri ise ikinci kümenin elemanları olarak kalmışlardır. İkinci ve üçüncü döngü sonucunda elde edilen küme matrisleri birbirinin aynı olmuş ve üçüncü döngü sonucunda kümeleme değişmemiştir. Bu sonuçla, nesnelerin artık kümeler arasında hareket etmeyeceği anlaşılmaktadır. Böylece k-ortalamalar ile kümeleme işlemi kararlı bir noktaya gelmiştir. Yeni bir döngüye ihtiyaç bulunmamaktadır. İlaçların kümeleme sonucu tablo 5.2'de görülmektedir [58].

Nesne	1. Nitelik (x) etki endeksi	2. Nitelik (y) pH	Sonuç kümeleme
İlaç A	1	1	1
İlaç B	2	1	1
İlaç C	4	3	2
İlaç D	5	4	2

Tablo 5.2 Kümeleme Sonucunda Oluşan Gruplama [61]

6. K-ORTALAMALAR YÖNTEMİ İLE BENZİN İSTASYONLARININ OPERASYONEL DEĞERLENDİRİLMESİ

Günümüzde bilgi teknolojilerinin gelişimi ile birlikte saklanan veri miktarı sürekli büyümektedir. Şirketlerin ileriye dönük planlama yapabilmesi ve karar alma süreçlerinde bu büyük miktardaki veri içerisinden fayda sağlayabilecekleri anlamlı bilgiler çıkarmaları önemli yer tutmaktadır.

Rekabetin yoğun yaşandığı akaryakıt şirketleri içinde yapacağı yatırımları planlayabilmek ve gözlemleyebilmek için veri madenciliği önemli bir yer tutmaktadır.

Bir akaryakıt şirketi için yatırım gerektiren en önemli alanlardan birisi son kullanıcıya hizmet satabildikleri benzin istasyonlarıdır. Ülkemizdeki benzin istasyonlarının büyük çoğunluğu şahıs şirketleri olup akaryakıt şirketleri ile bayilik anlaşmaları yapmaktadırlar. Bu anlaşmanın sonucunda akaryakıt şirketleri, benzin istasyonunun kendi marka değerini taşıyacağı için bazı yatırımlar yaparlar. Bu yatırımlara örnek olarak otomasyon altyapısı, iletişim altyapısı, şirket standardında çevre düzenlemesi vb. yatırımlar verilebilir. Bu yatırımların yanırısıra, altyapılarında belirli aralıklarla bakımları ve onarımı gerekmektedir. Binlerce benzin istasyonu olduğu düşünüldüğünde, akaryakıt şirketine istasyonların yüksek maliyet getirdiği ortaya çıkmaktadır.

Gerek bu maliyetlerden ötürü gerekse ileriye dönük stratejik planlar yapabilmek için akaryakıt şirketlerinin, benzin istasyonları hakkında operasyonel değerlendirme yapabilme ihtiyacı doğmuştur.

6.1. Problem Tanımı

Tez kapsamında yapılan veri madenciliği çalışması için ülkemizde faaliyet gösteren bir akaryakıt şirketine ait benzin istasyonlarından elde edilen akaryakıt satış verileri alınmıştır. Elde edilen veriler neticesinde, akaryakıt şirketine bağlı olan benzin istasyonlarından operasyonel olarak verimi düşük olan istasyonların tesbiti amaçlanmıştır. Yapılan bu çalışma ile benzin istasyonlarının operasyonel verimliliğinin tesbiti, ileriye dönük bayilik anlaşmalarında karar alma süreçlerine fayda sağlanması amaçlanmıştır.

6.2. Verilerin Hazırlanması

Veri madenciliği adımlarında en fazla zamanı veri hazırlama aşaması tutmaktadır. Modelin kurulması aşamasında yaşanacak problemlerde tekrar uzun zaman alan bu aşamaya dönülmesi gerekebilir. Problem tanımında belirtilen amaca yönelik verilerin hazırlanmasında veri toplama, birleştirme, temizleme, seçim ve dönüştürme aşamaları izlenmiştir.

6.2.1. Bir Akaryakıt Firmasına Ait Benzin İstasyonlarının Akaryakıt Satış Bilgilerinin Toplanması

Bu çalışmada, ülkemizde önde gelen bir akaryakıt şirketine ait kayıtlar kullanılmıştır. 2011 yılı itibariyle akaryakıt şirketine bağlı anlaşmalı istasyon sayısı yaklaşık olarak 2500 civarındadır. 2008 yılından bu yana istasyonlardan yapılan tüm satışların kayıtları,

uydu iletişimi yardımı ile anlık olarak merkezdeki bir veritabanında kayıt altına alınmaktadır.

Tez kapsamında 2008 ve 2009 yıllarındaki istasyonlara ait satış kayıtları, istasyonlara ait kayıtlar, satılan ürün türlerine ait kayıtlar ve müşteri kayıtları kullanılmıştır. Tüm kayıtlar tek bir veri tabanı üzerinden alınmıştır ve çalışma için yeni oluşturulan veri tabanına aktarılmıştır. Toplanılan bu tablolarda ne tür kayıtlar tutulduğu bilgisi aşağıda gösterilmiştir.

Satış Kayıtları Tablosu
Satış ID
Müşteri ID
İşlem tarihi
İstasyon ID
Ürün ID
Miktar
Tutar
Plaka

Tablo 6.1 Satış Kayıtları Tablosu

Satış kayıtları tablosu istasyonlardan gelen satışların anlık olarak kayıt altına alındığı tablodur. Tablo 6.1'de satış kayıtları tablosunda ne tür bilgiler tutulduğu gösterilmektedir.

İstasyon Tablosu
İstasyon ID
İstasyon Adı
İl
İlçe
Adres
Telefon
Email
Plaka
Kontak Kişi
Kontak Kişi Tel
Aktif-Pasif
Kayıt zamanı

Tablo 6.2 İstasyon Tablosu

Benzin istasyonlarına ait bilgiler istasyon tablosunda yer almaktadır. Tablo 6.2'de istasyon tablosunda ne tür bilgiler tutulduğu gösterilmektedir.

Müşteri Tablosu
Müşteri ID
Ad
Soyad
Cinsiyet
Doğum günü
Doğum yeri
İşlem zamanı
Ev tel
GSM tel
Email
Adres
İlçe
İl

Tablo 6.3 Müşteri Tablosu

Benzin satışının yapıldığı müşterilere ait kayıtlar müşteri tablosunda yer almaktadır. Tablo 6.3'de müşteri tablosunda ne tür bilgiler tutulduğu gösterilmektedir.

Ürün Tablosu
Ürün ID
Ürün Adı
Ürün Grup ID

Tablo 6.4 Ürün Tablosu

İstasyonlarda satışı yapılan ürün bilgileri ürün tablosunda yer almaktadır. Tablo 6.4'de ürün tablosunda ne tür bilgiler tutulduğu gösterilmektedir.

6.2.2. Değer Biçme

Veriler, tek bir veri kaynağından temin edildiğinden verilerin birbiri ile uyumunda bir sıkıntı yaşanmamıştır.

6.2.3. Verileri Birleştirme ve Temizleme

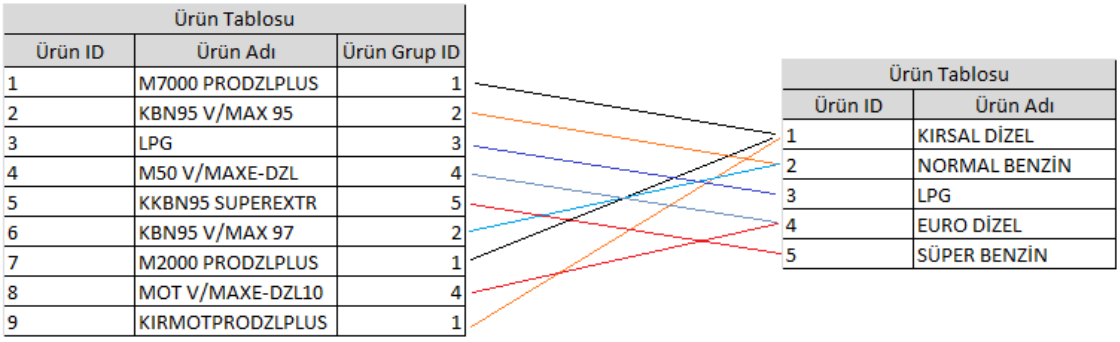
Satış kayıtları tablosu incelendiğinde kayıtlar arasındaki bazı kayıtların yanlış yada test amaçlı girilmiş olabileceği düşünülerek, bu kayıtların çalışmayı etkilememesi için tablodan silinmesine karar verildi. Satın alınan ürüne ait miktar bilgisi alanına 0 olarak değer girilmiş tüm kayıtlar tablodan silindi.

Tablo 6.5'de gösterilen ürün tablosuna ait veriler incelendiğinde 9 adet ürünün olduğu görülmektedir. Bu ürünler birbirlerinden farklı olmasına karşın aslında 5 adet ürün grubunda yer almaktadır. Aynı grup altında yer alıp farklı olmasının sebebi ürün çeşitliliği oluşturmaktır. Tablo incelendiğinde, "Ürün Grup ID"'si 1 olan ürünlerin kırsal dizel tipinde ürünler olduğu görülmektedir. "Ürün Grup ID"'si 2 olan ürünlerin normal benzin tipinde ürünler olduğu görülmektedir. "Ürün Grup ID"'si 3 olan tek bir ürün bulunmaktadır, bu üründe lpg'dir. "Ürün Grup ID"'si 4 olan ürünlerin euro dizel tipinde ürünler olduğu görülmektedir ve son olarak "Ürün Grup ID"'si 5 olan bir tek süper benzin tipinde ürün yer almaktadır.

Ürün Tablosu		
Ürün ID	Ürün Adı	Ürün Grup ID
1	M7000 PRODZLPLUS	1
2	KBN95 V/MAX 95	2
3	LPG	3
4	M50 V/MAXE-DZL	4
5	KKBN95 SUPEREXTR	5
6	KBN95 V/MAX 97	2
7	M2000 PRODZLPLUS	1
8	MOT V/MAXE-DZL10	4
9	KIRMOTPRODZLPLUS	1

Tablo 6.5 Ürün Tablosu Verileri

Aynı grupta yer alan ürünlerin aynı tip ürünler olduğu görülerek daha az veri tipi ile uğraşmak için birleştirilmesine karar verilmiştir. Ürün tablosunun son hali tablo 6.6'da gösterilmektedir. Satış kayıtları tablosundaki "Ürün ID" sütunundaki verilerde yeni oluşturulan ürün tablosuna göre güncellenmiştir.

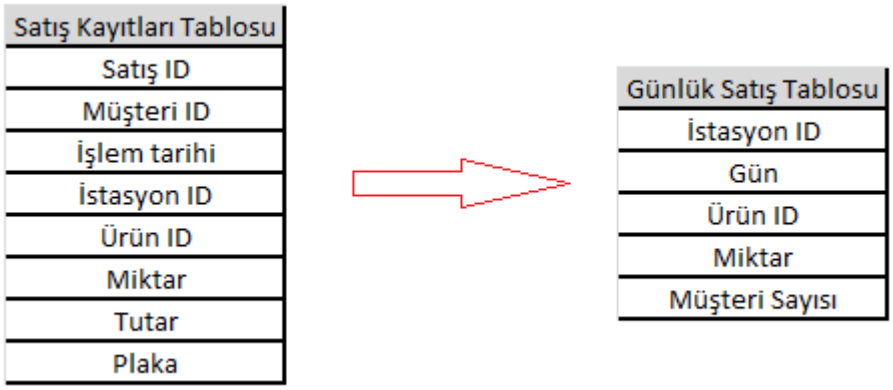


Tablo 6.6 Ürün Tablosu Verilerinin Birleştirimesi

6.2.4. Kullanılacak Verilerin Seçimi ve Dönüştürülmesi

2008 ve 2009 yıllarına ait benzin istasyonlarının satış bilgilerinin tutulduğu satış kayıtları tablosu incelendiğinde veri miktarının çok büyük olduğu görüldü. Aylık kayıtlar olarak bakıldığında ortalama her ay 2 milyon kayıt olduğu görüldü. Tüm kayıtların iki seneyi kapsadığı düşünüldüğünde toplam satış kaydı sayısının yaklaşık olarak 48 milyon civarında olduğu görüldü. Uygulanacak veri madenciliği tekniklerinden k-ortalamalar kümeleme yöntemi ile 48 milyon veri ile kümeleme yapmanın, çalışma kapsamında kullanılan bilgisayarlar düşünüldüğünde mümkün olamayacağı görüldü.

Satış kayıtları tablosundaki veriler istasyon tablosundaki verilerle birlikte incelendiğinde satış kayıtlarında 2171 istasyona ait kayıt olduğu görülmüştür. Gerek satış tablosundaki verilerin sayısını düşürmek, gerekse kümeleme analizinde kullanılacak olan veri kümesinde yer alan verilerin istasyon başına aynı veya yakın değerlerde olmasını sağlamak amacıyla satış kayıtları tablosundan aşağıdaki gibi günlük satış tablosuna indirgeme yapılmıştır.



Tablo 6.7 Satış Kayıtları Tablosunundan Günlük Satış Tablosu Oluşturulması

Günlük satış tablosunda hangi veri tiplerinin olması gerektiği düşünüldüğünde, öncelikle hangi istasyona ait satış bilgisi geldiğini anlayabilmek için "İstasyon ID" sütunu seçilmiştir. İstasyonun hangi üründen ne kadar satış yaptığını görebilmek için

“Ürün ID” ve “Miktar” verileride yeni tabloya dahil edilmiştir. İstasyonun gün içerisinde bir üründen ne kadar sattığı bilgisini elde edebilmek için o gün içerisindeki belirlenmiş ürüne ait satış miktarları toplanarak günlük satış tablosundaki “Miktar” alanına girilmiştir, aynı zamanda kaydın hangi güne ait olduğu bilgisi için “Gün” alanına gün bilgisi girilmiştir. Son olarak k-ortalamalar kümeleme yönteminde kullanabilmek amacıyla istasyona gelen günlük müşteri sayılarını hesaplayabilmek için müşteri tablosundan da yararlanılarak “Müşteri Sayısı” alanı eklenmiştir.

Satış kayıtları tablosundaki “Tutar” bilgilerinin iki senelik süreçte zamlardan etkilenerek değişkenlik göstermesinden ötürü yeni tabloda yer almasına gerek görülmemiştir. Ayrıca satışın hangi plakalı araca satıldığı bilgisi istasyon ikmal personeli tarafından elle giriş yapıldığından yanlış veri gelme olasılığı bulunmaktadır. Bu yüzden “Plaka” alanına da yeni tabloda yer verilmemiştir.

İstasyonların operasyonel verimlilik analizinde kullanılacak olan k-ortalamalar kümeleme yönteminde seçilecek verilerin sayısal veriler olması gerekmektedir. İstasyonların operasyonel değerlendirmesinde kullanılmak üzere niteliklerine göre veriler istasyona ait günlük “Normal Benzin”, “Süper Benzin”, “Kırsal Dizel”, “Euro Dizel”, “LPG” miktarları ve “Müşteri Sayısı” olarak seçilmiştir.

Günlük satış tablosunun istasyon başına günlük verilerin bir kısmını gösteren veri tabanı görüntüsü Tablo 6.8’de gösterilmektedir.

STATION_ID	NORMAL_BENZIN	SUPER_BENZIN	KIRSAL_DIZEL	EURO_DIZEL	LPG	GUN	MUSTERI_SAYISI
000006	0.00	0.00	0.00	42.39	0.00	20080113	1
000016	0.00	9.67	0.00	0.00	0.00	20080113	1
000244	0.00	24.11	0.00	0.00	0.00	20080113	1
000305	0.00	15.82	0.00	18.79	0.00	20080113	2
000317	0.00	53.00	0.00	0.00	0.00	20080113	1
000334	0.00	24.11	0.00	0.00	0.00	20080113	1
000079	0.00	24.11	0.00	0.00	0.00	20080114	1
000162	0.00	67.22	0.00	0.00	0.00	20080114	2
000218	0.00	0.00	0.00	37.59	0.00	20080114	1
001283	0.00	0.00	0.00	37.73	0.00	20080114	1
000015	0.00	0.00	0.00	54.45	0.00	20080115	1
000154	0.00	16.34	0.00	0.00	0.00	20080115	1
000104	0.00	0.00	0.00	75.75	0.00	20080117	1
000001	0.00	950.31	490.17	680.86	0.00	20080118	86
000002	26.10	223.18	0.00	76.33	0.00	20080118	13
000004	0.00	26.23	0.00	0.00	29.18	20080118	2
000006	0.00	273.61	0.00	173.03	0.00	20080118	12
000009	0.00	128.92	38.14	38.17	31.87	20080118	9
000010	0.00	32.68	0.00	0.00	0.00	20080118	1
000014	0.00	0.00	58.15	48.04	0.00	20080118	2
000015	0.00	122.52	30.12	181.16	59.84	20080118	12
000016	86.68	65.49	0.00	28.30	35.67	20080118	6
000017	0.00	4.91	0.00	45.89	31.03	20080118	4
000022	36.71	0.00	0.00	0.00	0.00	20080118	1
000026	0.00	305.80	830.81	261.38	596...	20080118	86
000028	0.00	1.65	0.00	0.00	0.00	20080118	1
000037	32.59	0.00	0.00	0.00	0.00	20080118	1
000038	40.01	6.53	0.00	57.65	0.00	20080118	3
000039	0.00	35.94	79.98	0.00	14.45	20080118	6
000045	0.00	16.28	0.00	0.00	0.00	20080118	1
000047	0.00	0.00	19.92	19.01	0.00	20080118	2

Tablo 6.8 Günlük Satış Tablosu Verileri

Günlük satış tablosunun son halindeki kayıt sayısı incelendiğinde tablo içerisinde 1220444 kaydın olduğu görülmüştür. K-ortalamalar kümeleme yönteminde kullanılmak için bu kayıt sayısınında yüksek olduğu görülmüştür. Kayıt sayısını düşürebilmek için her istasyondan gelişigüzel 50 kayıt olacak şekilde örnekleme yapılmıştır. Örnekleme sonucunda k-ortalamalar kümeleme yönteminde kullanılacak olan veri setinin kayıt sayısının 105374 olduğu görülmüştür.

6.3. K-Ortalamalar Kümeleme Yönteminin Verilere Uygulanması ve Modelin Oluşturulması

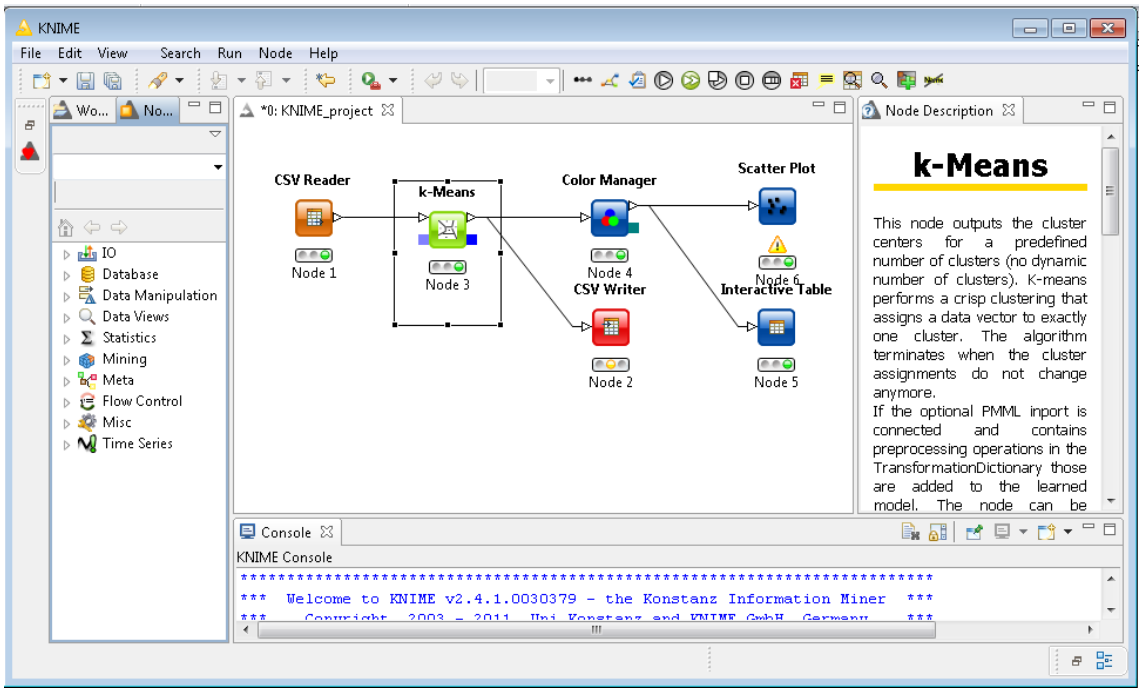
K-ortalamalar kümeleme yönteminin küçük k değerleri seçildiği durumlarda çok büyük veri kümelerine uygulanabilirliği ve hızı bu çalışmada seçilmesine etken olmuştur.

Çalışmada kullanılan veri kümesi SQL Server veri tabanında saklanmaktadır. K-ortalamalar kümeleme yöntemi ile küme merkezlerinin bulunmasında KNIME programından yararlanılmıştır.

6.3.1. KNIME

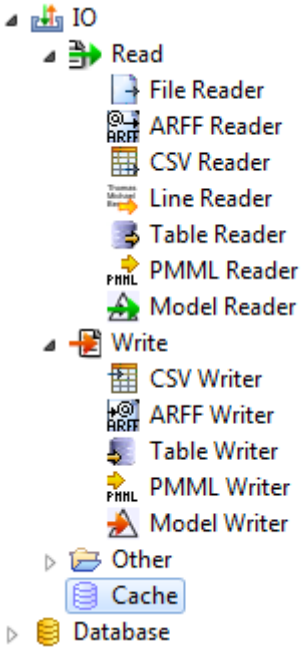
Konstanz Information Miner (KNIME) adlı araç Konstanz Üniversitesi görsel veri madenciliği araştırma grubu tarafından Eclipse Rich Client Platform üzerinde geliştirilmiştir [62].

KNIME, Eclipse tabanlı olmasının sağladığı avantaj sayesinde modüler ve görsel veri akış sistemi geliştirme ortamı sunmasının yanı sıra eğitim ve araştırma amaçlı ortak çalışma ortamı da sunmaktadır [62].



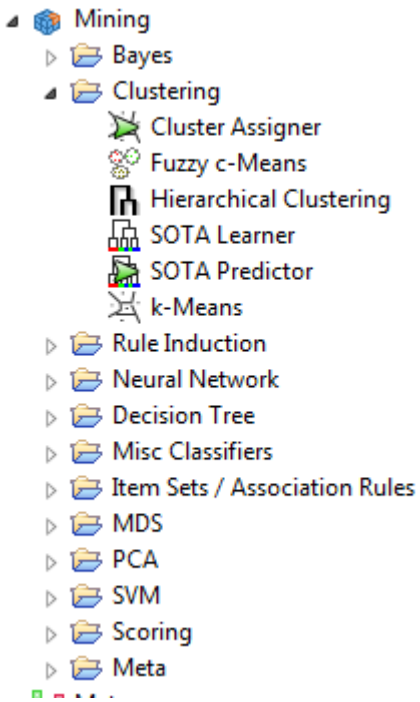
Şekil 6.1 KNIME Programı Ekran Görüntüsü

Bu program Eclipse platformunun desteklediği tüm işletim sistemlerinde kurulum gerektirmeden çalışabilmektedir. Kendi içinde bir JRE barındırmaktadır. Bu sayede işletim sisteminde Java kurulu olmasına gerek yoktur [62].



Şekil 6.2 KNIME Programında Veri Kaynağı Ekleme ve Çıkış Alabilme Nesneleri

KNIME programında Şekil 6.2'de de görüldüğü gibi farklı formatlarda veri girişi ve veri çıkışı yapabilen nesneler bulunmaktadır. Alınan verilerin ne kadarını bellekte tutabileceğini, ne kadarını diskte tutabileceğini ayarlamaya da olanak sağlamaktadır. Bu özellik büyük veritabanlarında çalışırken bellek bitmesi sorunu ile karşılaşma ihtimalini düşürmektedir [62].

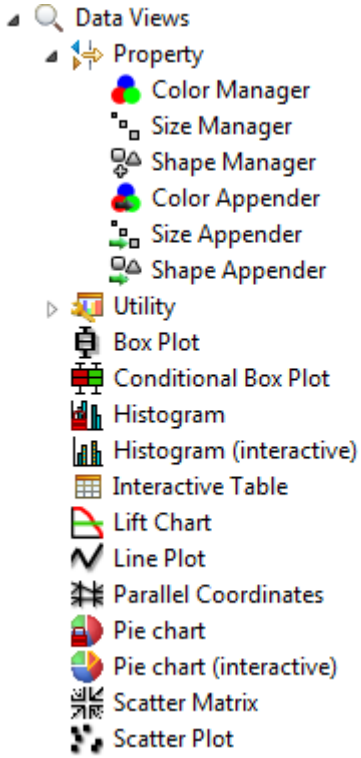


Şekil 6.3 KNIME Programında Veri Madenciliği Algoritmaları Seçim Paneli

Şekil 6.3'de de gösterildiği üzere KNIME içerisinde veri önişleme için kullanılabilecek bir çok veri madenciliği algoritması bulunmaktadır.

"Node Repository" sekmesinde altında yer alan nesneler sürükle bırak yöntemi ile orta alanda yer alan proje alanına bırakılır. Sonrasında nesnelerin giriş ve çıkış yönündeki okların fare aracılığı ile birleştirerek bir veri akış tasarımı yapılabilmektedir. Veri akış diyagramının çalışma yapısı tüm nesnelerin tek-tek çalıştırılmasına dayanır [62].

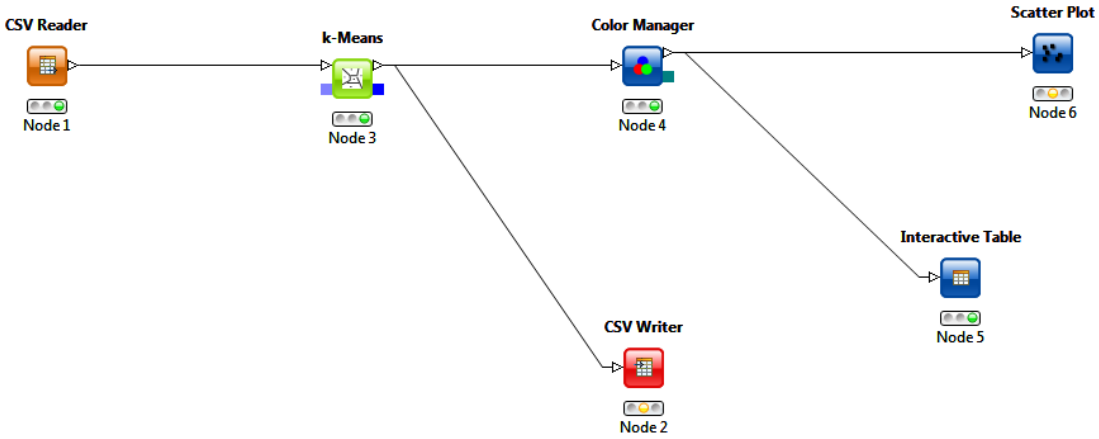
Şekil 6.4'de KNIME programı içerisinde görselleştirmede kullanılan nesneler gösterilmiştir. Uygulanan veri madenciliği algoritmaları bu nesneler sayesinde görselleştirilerek daha anlaşılır sonuçların çıkarımı yapılabilir.



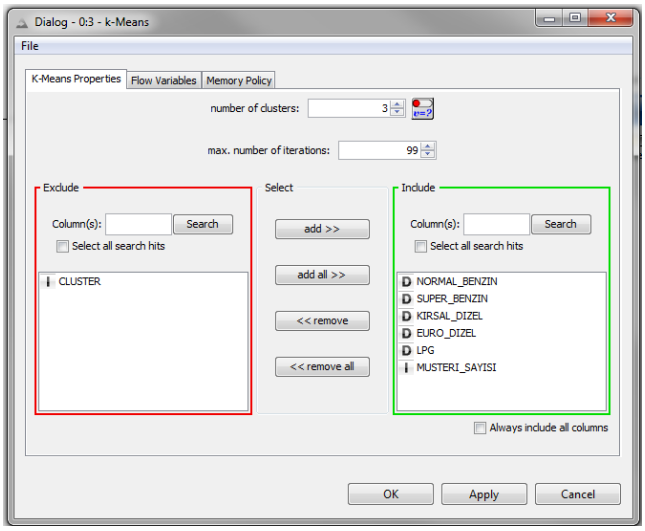
Şekil 6.4 KNIME Programında Görselleştirme Araçları

6.3.2. Modelin Oluşturulması

Daha önceki veri madenciliği adımlarında uygulamada kullanılacak olan veri setine ait nesneler belirlenmişti. Tekrar etmek gerekirse kısaca bu nesneler; normal benzin, süper benzin, kırsal dizel, euro dizel, lpg kullanım miktarları ve müşteri sayısı olarak belirlenmişti.



Şekil 6.5 KNIME Programında K-ortalamalar Kümeleme Yöntemi İçin Oluşturulan Veri Akışı



Şekil 6.6 K-ortalamalar Nesnesinin Özellikleri

Her bir istasyonun ortalama değeri KNIME sayesinde oluşturulmuş olan küme merkezlerinin değerlerine olan uzaklıkları hesaplanarak, bir istasyonun hangi küme merkezine yakın olduğu bilgisine ulaşılmaktadır. Uzaklık hesabında öklit uzaklık ölçüsü kullanılmıştır. Tüm istasyonlar en yakın oldukları küme merkezinin bulunduğu kümeye dahil olurlar.

KNIME ile küme merkez değerlerini hesaplamak için ilk başta k değerine sırasıyla 3,4,5 değerleri veriler üç farklı kümeleme yapıldı. Sonrasında istasyon ortalama değerleri ile oluşan küme merkez değerleri arasındaki uzaklıklar hesaplanarak istasyonlara en yakın kümeler bulunarak istasyonların hangi kümede olduğu bulundu.

NORMAL_BENZIN	SUPER_BENZIN	KIRSAL_DIZEL	EURO_DIZEL	LPG	MUSTERI_SAYISI	K=3
191.54	256.78	958.58	322.73	197.56	47	0
3768.26	341.5	44.87	477.41	386.42	107	1
31.21	317.54	5070.9	547.48	256.36	93	2

NORMAL_BENZIN	SUPER_BENZIN	KIRSAL_DIZEL	EURO_DIZEL	LPG	MUSTERI_SAYISI	K=4
277.05	239.83	578.12	294.05	185.73	41	0
4013.65	349.1	46.99	497.09	395.52	111	1
30.26	317.53	2902.13	467.64	256.77	77	2
29.33	335.13	7716.25	589.77	238.51	108	3

NORMAL_BENZIN	SUPER_BENZIN	KIRSAL_DIZEL	EURO_DIZEL	LPG	MUSTERI_SAYISI	K=5
2203.97	308.27	47.57	357.42	376.38	83	0
6214.8	453.21	45.98	710.77	466.64	152	1
95.71	233.47	711.22	294.42	160.92	37	2
28.06	335.73	7985.7	590.26	240.21	110	3
31.37	317.16	3077.62	479.68	257.38	79	4

Tablo 6.9 K=3,4,5 İçin Küme Merkez Değerleri

Tablo 6.9'da k=3,4,5 için elde edilen küme merkez değerleri görülmektedir. Turuncu renkle işaretlenen kümelerin, bulunduğu kümeler arasında verimliliği en az olan küme olduğu görülmektedir. Tüm istasyon ortalama değerlerinin en yakın olduğu küme merkezleri hesaplanarak ait oldukları kümeler tesbit edildi. Tablo 6.10'da hangi kümede kaç istasyon olduğu görülebilmektedir. Tabloda verimliliği en az kümeler turuncu renkle belirtilmiştir. Satış bakımından en kötü kümeler K=3 için 1. küme, K=4 için 1. küme, K=5 için 3. kümedir. Satış bakımından en kötü kümelerin istasyon sayıları incelendiğinde istasyon sayılarının yüksek olduğu görülmektedir. Toplamda 2171 istasyon olduğu düşünüldüğünde K=3 için 1929/2171=%88, K=4 için 1503/2171=%69, K=5 için 1432/2171=%65 gibi değerler elde edilmektedir.

K=3	K=4	K=5	
1929	1503	214	1.KÜME
8	122	30	2.KÜME
234	482	1432	3.KÜME
	64	59	4.KÜME
		436	5.KÜME

Tablo 6.10 K=3,4,5 İçin Kümelerde Yer Alan İstasyon Sayısı

Tüm istasyonların %88'i, %69'u veya %65'i satış olarak düşük kümede yer aldıkları için bu istasyonlara operasyonel olarak verimi düşük istasyonlar denilmesi doğru olmayacaktır. Daha anlamlı değerler elde etmek için K değerinin büyütülmesine karar verildi. Aynı işlemlerin K=10,15,49,86 değerleri ile yapılması sonucunda kümelerin aldığı istasyon sayıları tablo 6.11'de gösterilmektedir.

Tablo 6.11 de K=10,15,49,86 değerleri için satış verimliliği en düşük olan kümeler turuncu renkle gösterildi. En kötü küme istasyon sayısı ile toplam istasyon sayısı oranı sırası ile %43, %34, %17, %10 olduğu görülmektedir. En düşük verimli istasyonların olduğu düşünülen kümelerin sayıları karar alma süreçleri için istenilen seviyelere gelmiş olmasına rağmen bazı olumsuzluklar göze çarpmaktadır. Öncelikle k-ortalamalar algoritmasında kullanılan K değerinin yüksek değerlere çıkmış olması algoritmanın çalışma hızını düşürmektedir. Ayrıca K=86 ile elde edilen kümelerdeki istasyon sayıları incelendiğinde bazı kümelerin istasyon sayılarının çok düşük olması bazılarında istasyon dahi olmaması bu kümelerin gürültülü veriler içerdiğini göstermektedir. Gürültülü verilerin veri setinden temizlenmesi ile birlikte yapılacak k-ortalamalar kümeleme algoritmasının daha doğru sonuçlar vereceği düşünülmüştür.

K=10	K=15	K=49	K=86	
941		37	10	0.KÜME
10	190	3	3	1.KÜME
566	0	95	51	2.KÜME
15	744	3	1	3.KÜME
50	5	2	6	4.KÜME
64	119	12	9	5.KÜME
232	9	31	23	6.KÜME
69	92	7	6	7.KÜME
25	36	22	18	8.KÜME
199	254	18	14	9.KÜME
	97	1	12	10.KÜME
	34	371	23	11.KÜME
	538	32	26	12.KÜME
	27	1	1	13.KÜME
	18	14	7	14.KÜME
	8	0	0	15.KÜME
		0	0	16.KÜME
		91	54	17.KÜME
		95	126	18.KÜME
		5	12	19.KÜME
		117	16	20.KÜME
		18	19	21.KÜME
		3	5	22.KÜME
		7	6	23.KÜME
		8	26	24.KÜME
		4	5	25.KÜME
		13	9	26.KÜME
		28	8	27.KÜME
		3	6	28.KÜME
		15	1	29.KÜME
		60	10	30.KÜME
		4	3	31.KÜME
		11	2	32.KÜME
		13	1	33.KÜME
		15	10	34.KÜME
		7	10	35.KÜME
		6	6	36.KÜME
		35	13	37.KÜME
		135	238	38.KÜME
		31	8	39.KÜME
		310	69	40.KÜME
		70	69	41.KÜME
		81	59	42.KÜME
		9	12	43.KÜME
		23	17	44.KÜME
		214	7	45.KÜME
		32	16	46.KÜME

		4	3	47.KÜME
		55	48	48.KÜME
			7	49.KÜME
			3	50.KÜME
			8	51.KÜME
			16	52.KÜME
			3	53.KÜME
			3	54.KÜME
			2	55.KÜME
			113	56.KÜME
			46	57.KÜME
			7	58.KÜME
			65	59.KÜME
			11	60.KÜME
			8	61.KÜME
			0	62.KÜME
			195	63.KÜME
			21	64.KÜME
			7	65.KÜME
			1	66.KÜME
			1	67.KÜME
			26	68.KÜME
			5	69.KÜME
			0	70.KÜME
			5	71.KÜME
			44	72.KÜME
			14	73.KÜME
			155	74.KÜME
			11	75.KÜME
			90	76.KÜME
			4	77.KÜME
			37	78.KÜME
			78	79.KÜME
			21	80.KÜME
			12	81.KÜME
			30	82.KÜME
			7	83.KÜME
			2	84.KÜME
			9	85.KÜME

Tablo 6.11 K=10,15,49,86 İçin Kümelerde Yer Alan İstasyon Sayısı

İlk olarak kümelerdeki istasyon sayısı 10 ve altında olan kümelere ait kayıtlar veri setinden silindi ve K=5 ve K=10 için k-ortlamalar kümeleme algoritması uygulandı. Sonrasında kümelerdeki istasyon sayısı 20 ve altında olan kümelere ait kayıtlar veri setinden silinerek K=5 ve K=10 için k-ortalamalar kümeleme algoritması uygulandı ve son olarak kümelerdeki istasyon sayusu 30 ve altında olan kümelere ait kayıtlar veri setinden silinerek K=5 ve K=10 için k-ortalamalar kümeleme algoritması uygulandı.

Tablo 6.12'de gürültülü olduğu düşünülen verilerin silinmesi sonrası oluşturulan kümeler görülmektedir. Tabloda turuncu renkle belirtilen kümeler bulunan en kötü kümelerdir. Son olarak yapılan istasyon sayısı 30'un altındaki kümelere ait istasyonlara ait verilerin silinmesi neticesinde çıkan kümelede verimi en düşük olan küme 4. kümenin istasyon sayısının 422 olduğu görülmektedir. Aynı zamanda diğer kümelerdeki istasyon sayılarınında daha orantılı dağıldığı görülmektedir. İstasyonun operasyonel veriminin düşük olduğu düşünülen kümenin istasyon sayısının tüm istasyonların sayısına oranı 422/2171 = %19 gibi bir değer çıkması şirket için ihtiyaç duyulan planlama ve karar verme aşamalarına katkı sağlayabileceğini göstermektedir.

<10 K=5	<10 K=10	<15 K=5	<15 K=10	<20 K=5	<20 K=10	<30 K=5	<30 K=10	
302	132	308	121	399	186	328	113	0.KÜME
125	78	123	78	119	67	245	171	1.KÜME
599	133	584	108	607	143	268	120	2.KÜME
838	133	840	138	746	382	810	188	3.KÜME
307	593	316	582	300	511	520	422	4.KÜME
	448		426		231		304	5.KÜME
	166		174		262		223	6.KÜME
	176		190		104		275	7.KÜME
	220		136		95		124	8.KÜME
	92		218		190		231	9.KÜME

Tablo 6.12 Gürültülü Verilerin Silinmesi Sonrası Kümeleme

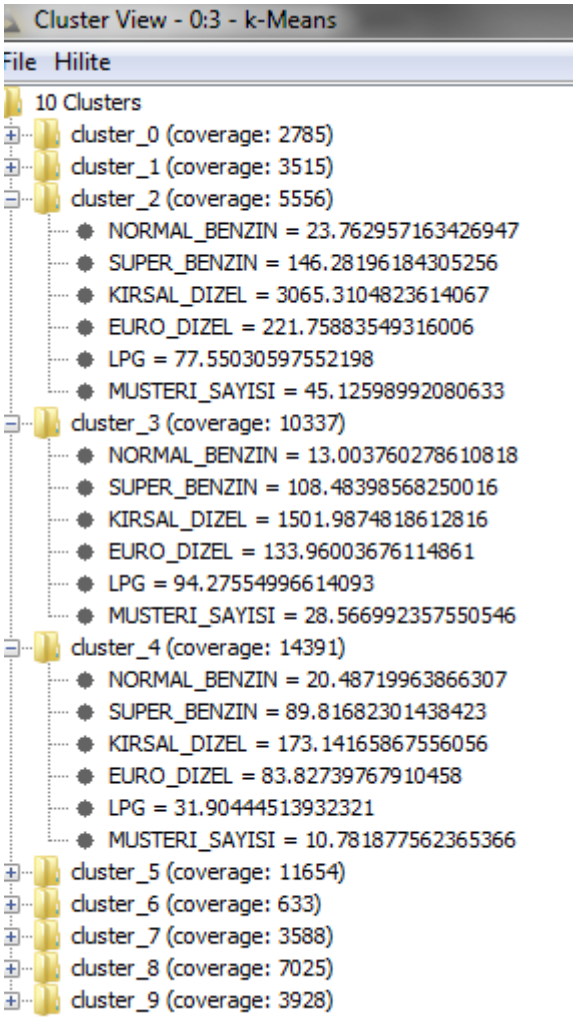
NORMAL_BENZIN	SUPER_BENZIN	KIRSAL_DIZEL	EURO_DIZEL	LPG	MUSTERI_SAYISI	<30 İÇİN K=10
8.82	317.56	2284.21	723.76	102.76	77	0
24.44	152.01	4051.06	229.47	82.61	50	1
23.76	146.36	3065.63	221.86	77.55	45	2
13	108.52	1502.39	133.93	94.17	28	3
20.43	89.83	173.24	83.83	31.93	10	4
14.97	78.05	810.4	91.04	46.05	17	5
1392.64	137.88	1595.46	151.6	125.8	35	6
607.68	72.42	49.92	65.79	54.62	16	7
15.27	103.05	2232.23	122.09	89.63	32	8
10.27	253	1078.02	586.34	158.05	54	9

Tablo 6.13 Gürültülü Verilerin Silinmesi Sonrası Küme Merkez Değerleri

Tablo 6.13'de son olarak hesaplanan kümeleye ait küme merkez değerleri gösterilmiştir. Turuncu renkle aranılan satış değerleri düşük olan en kötü küme işaretlenmiştir.

Şekil 6.7'de KNIME programı ile k-ortalamalar algoritması sonrası oluşan kümelere ait merkez noktaların değerleri gösterilmiştir.

Şekil 6.8'de KNIME programında "Color Manager" ve "Interactive Table" nesneleri kullanılarak kümeler renklere ayrılarak tabloda gösterilmiştir.



Şekil 6.7 KNIME ile Küme Merkez Değerlerinin Hesaplanması

Table View - 0:5 - Interactive Table

File Hilite Navigation View Output

Row ID	D NORMA...	D SUPER...	D KIRSAL...	D EURO_...	D LPG	I MUSTE...	S Cluster
33256	0	0	3,041.67	0	0	11	cluster_2
33257	0	0	2,994.39	0	183.17	17	cluster_2
33258	0	14.58	3,662.49	18.19	0	18	cluster_1
33259	0	72.36	3,053.91	200	538.33	97	cluster_2
33260	0	0	3,300	0	0	7	cluster_2
33261	0	142.49	3,252.72	641	210.06	97	cluster_2
33262	0	51.48	3,409.06	214.83	108.1	46	cluster_2
33263	565	5	712.87	0	0	15	cluster_5
33264	0	9.8	436.41	22.91	0	12	cluster_4
33265	0	35.74	399.19	19.17	0	9	cluster_4
33266	0	42.78	637.32	87.59	34.26	22	cluster_5
33267	0	50.12	382.88	166.34	0	15	cluster_4
33268	0	163.47	628.2	349.25	0	35	cluster_5
33269	0	230.87	679.6	176.1	0	31	cluster_5

Şekil 6.8 KNIME Interactive Table

6.4. Modelin Değerlendirilmesi

K-ortalamalar kümeleme yöntemi ile yapılan çalışmada oluşturulan model ile hangi istasyonların satış açısından verimlerinin düşük olduğunu tesbit edemilmek mümkün hale gelmiştir. Akaryakıt firmasının, tesbit edilen kümedeki istasyonlarla yapılacak muhtemel yeni anlaşmalarda bu veriler göz önüne alınacaktır. Belkide bir kısmı ile anlaşma yenilemiyerek, istasyonlar kapatılacaktır.

Bu bölümde tesbit edilen kümenin ne derece güvenilir olduğunu sayısal değerler ile incelenecektir.

Veri tabanında yer alan istasyon tablosunda istasyonların şirkete dahil oldukları tarih ile anlaşmalarının bittiği tarih bilgileri bulunmaktadır. Tabloda 2009 yılına ait kapatılan istasyonlar tesbit edildi. Daha sonrasında bu istasyonlara ait ortalama kullanım miktarları bulunarak son yapılan k-ortlamalar kümeleme algoritması sonucu istasyonların hangi kümeye ait olduğu bilgilerine ulaşılmıştır.

Kapatılan istasyonların toplam sayısı 40'dır. Bu istasyonlardan 62 istasyon ID'li istasyonun 9. kümede, 87 istasyon ID'li istasyonun 5. kümede, 2513 istasyon ID'li istasyonunda 0. kümede yer aldığı görülmüştür. Bunlar haricinde kalan 37 istasyonun 4. kümede olduğu görülmüştür.

K-ortalamalar kümeleme yöntemi ile oluşturulan modelde de en kötü kümenin 4. küme olması oluşturulan modelin 37/40 yani %92.5 doğru sonuç ürettiği gözlenmiştir.

Kapatılan istasyon bilgileri ve küme değerleri tablo 6.14'de gösterilmiştir. Yeşil olarak belirtilen istasyonlar 4. küme içerisinde yer almayan istasyonlardır. Yani bu istasyonlar en kötü küme içerisinde yer almayan istasyonlardır.

STATION_	NORMAL_BENZIN	SUPER_BENZIN	KIRSAL_DIZEL	EURO_DIZEL	LPG	MUSTERI_SAYISI	<30 için K=10
62	0	1822.56	725.84	1180.67	0	146	9
87	21.94	155.86	866.59	345.7	0	59	5
154	0	275.75	152.5	291.85	0	30	4
326	4.17	29.14	68.16	27.68	0	4	4
466	5.31	80.91	91.87	92.42	116.72	18	4
557	0.35	7.35	3.93	0.84	0	1	4
582	0	32.13	241.3	60.56	0	17	4
764	4.75	14.09	13.02	13.81	96.46	8	4
1191	101.45	25.31	105.89	86.65	0	11	4
1407	5.62	52.44	120.57	75.25	0	12	4
1450	159.05	103.04	81.49	112.47	1.67	15	4
1499	0	8.3	6.96	2.21	0	1	4
1541	93.91	5.18	148.62	15.13	7.9	8	4
1560	0	20.23	51.45	11.16	0	3	4
1587	0.84	33.33	45.99	5.43	0	5	4
1672	11	3.9	24.35	1.89	0	2	4
1713	3.95	12.94	26.19	23.21	14.11	11	4
1747	0.5	1	86.37	1.6	0	2	4
1822	1.33	2.99	3.19	2.27	17.75	1	4
1828	0.35	4.06	25.84	19.91	5.79	1	4
1829	11.64	12.87	23.3	10.22	0	2	4
1857	0	5.18	2.5	8.37	0	1	4
1870	1.92	6.41	6.88	1.06	0	1	4
1882	13.53	11.13	26.78	2.62	0.28	2	4
1913	33.63	8.06	9.39	27.49	7.09	1	4
2058	5.12	52.41	32.58	13.27	0.72	5	4
2208	137.85	73.37	58.72	7.64	0	10	4
2223	1.08	35	180.54	45.8	0	6	4
2254	5.32	7.59	105.68	7.13	31.69	4	4
2318	40.81	7.34	110.13	0.9	0	3	4
2325	0	3.78	15.16	3.14	0	1	4
2333	5.7	9.76	10.1	5.75	8.5	1	4
2358	27.05	20.17	8.47	4.8	0	1	4
2377	0.33	39.38	91.29	68.22	0	5	4
2399	1.63	36.84	93.25	2.68	0	1	4
2424	3.83	6.37	10.86	4.72	14.08	1	4
2436	0.62	4.23	23.76	2.73	0	1	4
2508	0	15.2	72.18	14.55	0	3	4
2513	0	42.73	2077.03	608.32	0	87	0
2523	0	21.99	67.56	0	0	1	4

Tablo 6.14 Kapatılan İstasyonlar

7. SONUÇ

Gelişen ve değişen çevre koşulları, sınırların kalkması ile küreselleşen dünya, farklı pazarlama ve ar-ge (araştırma geliştirme) yöntemleri "veri"nin değil "bilgi"nin önemini her geçen gün daha da artacak şekilde ortaya koymaktadır. Günümüzde yararlı bilgiye ulaşmak giderek önem kazanmış, verilerin yararlı bilgiye dönüştürülmesi ihtiyacı ile birlikte veri madenciliği giderek önem kazanmaya başlamıştır. Birbirinden farklı bankacılık, sağlık, sigortacılık gibi bir çok alanda veri madenciliği uygulanmaktadır.

Rekabet avantajını sağlamak için firmalar, daha iyi pazarlama yapabilme, daha iyi ürün elde edebilme, daha iyi yatırımlar yapabilme, daha iyi planlama yapabilme gibi alanlarda yararlı bilgiye ihtiyaç duymaktadırlar. Bu çalışmada veri madenciliği yöntemlerinden k-ortalamalar kümeleme yöntemi ile bir akaryakıt firmasına ait istasyon satış kayıtları kullanılarak, yatırım yapılan istasyonların verimliliklerinin analizi yapılarak verimi düşük istasyonlar belirlenerek, yapılan yatırımlar hakkında karar alma süreçlerine faydalı bilgi sağlamak amaçlanmıştır.

Veri seti olarak ülkemizde önde gelen bir akaryakıt şirketine ait istasyon satış kayıtları kullanılmıştır. Veri seti üzerinde kümeleme yönteminde kullanılabilecek sayısal özellikler seçildi. Sonrasında gereksiz veriler temizlendi ve kümeleme için uygun olabilecek formata dönüştürüldü. Veri setinin büyüklüğünden ötürü örnekleme yapılarak kümelemede kullanılacak veri seti hazırlandı. Farklı sayıda K değerleri için k-ortalamalar yöntemi uygulandı. K-ortalamalar kümeleme yöntemlerinde KNIME programından yararlanıldı. Kümeleme yöntemi sonucunda bir adet çok büyük küme ve yanında ufak kümelerin oluştuğu görüldü. Ufak kümelerin veri seti üzerindeki gürültülü veriler olduğuna karar verildi ve kümeleme sonuçlarını olumsuz etkilediği düşünülerek ufak sayıda istasyon içeren kümelerdeki istasyon verileri küme merkezi hesaplamasından çıkarıldı. K-ortalamalar kümeleme yöntemi tekrar uygulanarak kümeler oluşturuldu. Küme merkez değerleri özellikleri incelendiğinde en az satış değerlerine sahip kümedeki istasyonların, verimi en düşük istasyonlar olduğuna karar

verildi. Son olarak gerçekten kapanan istasyonlar ile seçilen kümedeki istasyonların büyük bir yüzde ile örtüştüğü görüldü.

Sonuç olarak k-ortalamalar kümeleme yöntemi ile oluşturulan model ile bir akaryakıt firmasına ait istasyonların operasyonel değerlendirmesi yapılarak çıkarılan analizlerin, ileriye dönük yatırımlara ve istasyonlar hakkındaki karar alma süreçlerinde yol haritası oluşturulabileceği görülmüştür.

KAYNAKLAR

1. Han, J., Kamber, M., Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, San Francisco, 2001
2. Sertaç Öğüt, Veri Madenciliği Kavramı ve Gelişim Süreci, Yeditepe Üniversitesi, Istanbul, 2009
3. Yurtsever, U., Veri Madenciliği ve Uygulaması, Yüksek Lisans Tezi, Sakarya Üniversitesi Sosyal Bilimler Enstitüsü, Sakarya. 2002
4. Adriaans, P., Zantinge, D., Data Mining, Addison Wesley Longman Limited, Harlow, 159p., 1996
5. Marquardt, M. , Reynolds, A. Global Learning Organizations, Irwin, New York USA, 1994
6. Yalçıntas, G., Veri Madenciliği, Yüksek Lisans Tezi, Gazi Üniversitesi Fen Bilimleri Enstitüsü, Ankara., 2003
7. Corey, M., Abbey, M., Abramsom, I., Oracle & Data Warehousing-A practical Guide to Successful Data Warehouse Analysis, Oracle Press, 1998
8. Allsopp, D. J., Harrison, A., Sheppard, C., A Database Architecture for Reusable CommonKADS Agent Specification Components, Knowledge-Based Systems, pp: 275–283, 2002
9. Anand, S. S., Bell, D. A., Hughes, J. G., EDM: A General Framework for Data Mining Based on Evidence Theory, Data and Knowledge Engineering, pp. 189–223, 1996
10. Westphal, C., Blaxton, T., Data Mining Solutions, Methods and Tools for Solving Real-World Problems, New York, pp: 21-30, 1998
11. Heijst, G., Spek, R., Kruizinga, E., Corporate Memories as a Tool for Knowledge Management, Expert Systems With Applications, pp: 41– 54., 1997
12. Hendriks, P. H. J., Vriens, D. J., Knowledge-Based Systems and Knowledge Management: Friends or Foes?, Information and Management, pp: 113–125, 1999
13. Çoban A., İmalat Sanayinde Veri Madenciliği Destekli Tedarikçi Seçimi Uygulaması, Sakarya Üniversitesi Fen Bilimleri Entstitüsü, 2006

14. Knight, B., MA, J., Temporal Management Using Relative Time in Knowledge-Based Process Control, Engineering Applications Artificial Intelligence, pp: 11-19, 1997

15. Wang, X., Abraham, A., Smith, KA., Web traffic mining using a oncurrent neuro-fuzzy approach In Proceedings of the 2nd International Conference on Hybrid Intelligent Systems, Computing Systems: Design, Management and Applications, Santiago, Chile, pp: 853–862, 2002

16. Laudon, K.C., Laudon, J. P., Essential of Management Information Systems, Prentice Hall, New Jersey, 2002

17. Inmon, W. H., Building the Data Warehouse, John Wiley and Sons, New York, pp:102-125, 1993

18. Bonfiati, A., Casati, F., Dayal, U., Shan M.-C., Warehousing Workflow Data: Challenges and Opportunities, Proceedings of VLDB, Rome, Italy, pp: 649–652, 2001

19. Cheung, D., Kao, B., Lee, J., Discovering User Access Patterns on the World Wide Web, In Proceedings of the 1st Pacific-Asia Conference on Knowledge Discovery and Data Mining, Vol. 10, pp: 463–70, 1997

20. Chang, G., Healey, M., J., Mchugh, J.A.M., Wang, J.T.L., Web Mining. In Mining the World Wide Web-An Information Search Approach, Dordetch: Kluwer; pp:43-47, 2001

21. Han, J., Pei, J., Yin, Y., Mining Frequent Patterns Without Candidate Generation, In Proceedings of the 2000 ACMSIGMOD International Conference on Management of Data, pp:1–12, Dallas, TX, USA 2000

22. Tantuğ, A.C., Veri Madenciliği ve Denetleme, Yüksek Lisans Tezi, İstanbul eknik Üniversitesi Fen Bilimleri Enstiüsü, İstanbul, 2002

23. Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P. and Uthurusamy, R., Advances in Knowledge Discovery and Data Mining, MIT Press, Cambridge, 1996

24. Cabena, P., Hadjinian, P., Stadler, R., Verhees, J. and Kamber, M., Discovering Data Mining: From Concept to Implementation, Prentice Hall, New Jersey, 1998

25. Fabris, P., Advanced Navigation, CIO, May 15 1998

26. Chung, H., Gray, M., Special Section: Data Mining, Journal of Management Information Systems, 1999

27. Hui, S.C. and Jha, G., Data Mining for Customer Service Support, Information & Management, 1-13, 2000

28. Berry, M.J.A. and Linoff, G.S., Mastering Data Mining, John Wiley & Sons, New York, 2000

29. Grossman, R.L., Kamath, C., Kegelmeyer P., Kumar, V., Namburu R. R., Data Mining For Scientific and Engineering Applications, Kluwer Academic Publishers, Netherlands, 2001

30. Koyuncugil A. S., Veri Madenciliği veya Bir Başka Deyişle Akıllı Algoritmalarla İstatistik Kullanımı, Emniyet Genel Müdürlüğü Polis Dergisi Bilişim Özel Sayısı, 38-40, 2004

31. Deogun, J.S., Raghavan, V.V., Sarkar, A. ve Sever, H., Data mining: Trends in research and development. Rough sets and data mining: Analysis for imprecise data, 9-45. New York: Kluwer Academic Publishers, 1997

32. www.data-miners.com

33. Adriaans, P., Zantinge, D., Data Mining, New York: Addison Wesley, 1996

34. http://www.varyans.net/dogusu_gelisimi.htm

35. Öğüt, S., Veri madenciliği kavramı ve gelişim süreci, Erişimi http://www..sertacogut.com/papers/Sertac_Ogut_Veri_Madenciligi_Kavrami_ve_Gelisim_Sureci.pdf, 2008

36. Akpınar, H., Business Intelligence & Data Mining, Dönence Basın ve Yayın Hizmetleri, İstanbul, 2004

37. Shearer C., The Crisp-DM model: The new blueprint for data mining, Journal of Data Warehousing, 13-23, 2000

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

39. Tiryaki, S., Lojistik Alanında Bir Veri Madenciliği Uygulaması, Yüksek Lisans Tezi, Fen Bilimleri Enstitüsü, İstanbul Teknik Üniversitesi, İstanbul, 2006

40. Arslan, H., Sakarya Üniversitesi Web Erişim Kayıtlarının Web Madenciliği İle Analizi, Yüksek Lisans Tezi, Elektronik-Bilgisayar Eğitimi ABD, Fen Bilimleri Enstitüsü, Sakarya Üniversitesi, Sakarya, 2008

41. Özkan, Y., Veri Madenciliği Yöntemleri - Papatya Yayıncılık Eğitim A.Ş., İstanbul, Türkiye. 2008

42. Silahtaroğlu, G.,Kavram ve Algoritmalarıyla Temel Veri Madenciliği Papatya Yayıncılık Eğitim A.Ş., İstanbul, Türkiye, 2008

43. Shmueli, G., Patel, N.R., Bruce, P.C.,. Data Mining for Business Intelligence - John Wiley and Sons, Ltd., Hoboken, NJ, USA, 2007

44. http://yunus.hacettepe.edu.tr/~hcingi/ist376a/6Bolum.doc

45. Witten I., Frank E., Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann Publishers, San Fransisco, 120-129, 267-277, 2000

46. Akpınar H., Veritabanlarında bilgi keşfi ve veri madenciliği, İstanbul Üniversitesi İşletme Fakültesi Dergisi, 1-22, 2000

47. Baysal A.C., Bayi değerlendirmesi için veri madenciliği uygulaması, İstanbul Üniversitesi Fen Bilimleri Enstitüsü, 2008

48. Bounsaythip C. ve Runsala E. R., Research Report Overview of data mining for customer behaviour modeling, VTT Information Technology, 2001

49. Dunham, M.H., Data Mining Introductory and Advanced Topics, Pearson Education Inc., New Jersey, 2003

50. Shah S., Kursak A., Data mining and genetic algorithms based gene / SNP selection, Artificial Intelligence in Medicine, 183 -196, 2004

51. Berson A., Smith S. ve Thearling K.,. Data mining applications for CRM, www.thearling.com, 2004

52. Mining Multiple-Level Association Rules in Large Databases, Han J. - Fu Y., IEEE Transactions on Knowledge and Data Engineering, vol 11 no.5, 1999

53. Özekeş S., Veri Madencilği Modelleri ve Uygulama Alanları, İstanbul Ticaret Üniversitesi Dergisi, http://www.iticu.edu.tr/yayin/dergi/d3/M00041.pdf

54. Hsu M.-H., A personalized English learning recommender system for ESL students, Expert Systems with Applications, 683–688, 2008

55. Hair J. F., Anderson R., Black W. C., Multivariate Data Analysis with Readings, New Jersey: Prentice Hall Inc., 1995

56. Bilen H., Bankacılık Sektöründe Personel Seçimi ve Performans Değerlendirmesine İlişkin Veri Madenciliği Uygulaması, Gazi Üniversitesi Fen Bilimleri Enstitüsü, Ankara, 2009

57. Öznohutçu A., Density Based Approach, Beykent Üniversitesi Mühendislik Mimarlık Fakultesi Bilgisayar Mühendisliği Bölümü

58. Dinçer E., Veri Madenciliğinde K-Means Algoritması ve Tıp Alanında Uygulanması, Kocaeli Üniversitesi Fen Bilimleri Enstitüsü, 2006

59. MacQueen, J., Some methods for classification and analysis of multi-variate observations. In: Proc. of the Fifth Berkeley Symp. on Math., Statistics and Probability, LeCam, L.M., and Neyman, J., Berkeley: U. California Press, 281., 1967

60. Berkhin, Pavel.: Survey of Clustering Data Mining Techniques, Accrue Software Inc., San Jose, California, USA, 2002

61. http://people.revoledu.com/kardi/tutorial/kMean/

62. Bilgin T.T, Veri Akışı Diyagramları Tabanlı Veri Madenciliği Araçları ve Yazılım Geliştirme Ortamları, Maltepe Üniversitesi, Bilgisayar Mühendisliği Bölümü, 2009

63. Burma Z. A., Veri Tabanı Yönetim Sistemleri-I, Seçkin Yayıncılık, 2005

64. Dönmez Z. S., Bayi Performans Değerlendirmesinde Bir Veri Madenciliği Uygulaması, İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü, 2008

65. Koyuncugil A.S., Özgülbaş N., Veri Madenciliği: Tıp ve Sağlık Hizmetlerinde Kullanımı ve Uygulamaları, Başkent Üniversitesi Sağlık Bilimleri Fakültesi, Sağlık Kurumları İşletmeciliği Bölümü, Ankara, 2009

66. Ayık Y.Z., Özdemir A., Yavuz O., Lise Türü ve Lise Mezuniyet Başarısının, Kazanılan Fakulte İle İlişkisinin Veri Madenciliği Tekniği İle Analizi, Atatürk Üniversitesi

67. Terzi Ö., Küçüksille E. U., Ergin G., İlker A., Veri Madenciliği Süreci Kullanılarak Güneş Işınımı Tahmini, Süleyman Demirel Üniversitesi Teknik Eğitim Fakultesi, 2011

68. Çelik M., Veri Madenciliğinde Kullanılan Sınıflandırma Yöntemleri ve Bir Uygulama, İstanbul Üniversitesi Sosyal Bilimler Entitüsü Ekonometri Anabilim Dalı, 2009

69. Böhm, C., Powerful Database Support For High Performance Data Mining, Ludwig-Maximilian Üniversitesi Enformatik ve Matematik Bölümü,.Münih, Almanya, 2001

70. Dolgun M. Ö., Büyük Alışveriş Merkezleri İçin Veri Madenciliği Uygulamaları, Hacettepe Üniversitesi İstatistik Anabilim Dalı, 2006

ÖZGEÇMİŞ

1982 yılında Ankara'da doğdu. İlköğretimini Uşak Mehmetçik İlköğretim okulunda, ortaöğretimini Uşak Orhan Dengiz Anadolu Lisesi'nde tamamladı.

2000 yılında Anadolu Üniversitesi Elektrik-Elektronik Mühendisliği Bölümünü kazandı ve 2005 yılında mezun oldu. 2005-2006 yılları arasında askerlik görevini yedek subay olarak tamamladı. 2006-2009 yılları arasında NEC Australia'da sistem desktek uzmanı olarak görev yaptı. 2010-2011 İŞNET'te network danışmanı olarak görev yaptı. Şubat 2011'den bu yana Vodafone'da network güvenlik uzmanı olarak görev yapmaktadır.

2009 yılında, Maltepe Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Yüksek Lisans programında yüksek lisans öğrenimi yapmaya hak kazandı.