T.C.
BAHÇEŞEHİR ÜNİVERSİTESİ

DİJİTAL YAYINCILIKTA İÇERİK İZLEME ORANLARINA GÖRE MÜŞTERİ KÜMELENMESİ

Yüksek Lisans Tezi

ÖZGÜR TEKİNGÖZ

İSTANBUL, 2016

T.C.
BAHÇEŞEHİR ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

BİLGİ TEKNOLOJİLERİ

DİJİTAL YAYINCILIKTA İÇERİK İZLEME ORANLARINA GÖRE MÜŞTERİ KÜMELENMESİ

Yüksek Lisans Tezi

ÖZGÜR TEKİNGÖZ

Tez Danışmanı: Doç. Dr. Mehmet Alper TUNGA

İSTANBUL, 2016

BAHÇEŞEHİR ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİ TEKNOLOJİLERİ

Tezin Adı: Dijital Yayıncılıkta İçerik İzleme Oranlarına göre Müşteri Kümelenmesi
Öğrencinin Adı Soyadı: Özgür Tekingöz
Tez Savunma Tarihi:

Bu tezin Yüksek Lisans tezi olarak gerekli şartları yerine getirmiş olduğu Fen Bilimleri Enstitüsü tarafından onaylanmıştır.

Doç. Dr. Nafiz Arıca
Enstitü Müdürü
_____

Bu tezin Yüksek Lisans tezi olarak gerekli şartları yerine getirmiş olduğunu onaylarım.

Unvan
Program
Koordinatörü
_____

Bu Tez tarafımızca okunmuş, nitelik ve içerik açısından bir Yüksek Lisans Tezi olarak yeterli görülmüş ve kabul edilmiştir.

Jüri Üyeleri	İmzalar
Tez Danışmanı Doç. Dr. Mehmet Alper TUNGA	_____
Üye Doç. Dr. Ahmet KIRIŞ	_____
Üye Yücel Batu SALMAN	_____

ÖZET

DİJİTAL YAYINCILIKTA İÇERİK İZLEME ORANLARINA GÖRE MÜŞTERİ KÜMELENMESİ

Özgür Tekingöz

Bilgi Teknolojileri

Tez Danışmanı: Doç. Dr. Mehmet Alper TUNGA

Ocak 2015, 56 Sayfa

Veri madenciliği, veriler içerisinde gizlenen bilgileri ortaya çıkarma sürecidir. Müşterilerin sınıflandırılması, profil çıkarılması, segmentasyonun çıkarılması ve kümeleme analizi şirketler için değerli müşterileri belirlemede kullanılan uygulamalardır. Müşteri segmentasyonun yapılması, grup bazlı pazarlama strateji geliştirilmesini sağlar.

Bu tez kapsamında dijital yayıncılık sektöründe hizmet veren bir firmanın kendi kanallarının yayın içeriklerinden yola çıkarak(ulusal kanalların yayın içeriklerini içermiyor), müşterilerin izledikleri içerik türlerinin oranlarına göre kümelenmesi hedeflenmiştir.

Çalışma sırasında veri madenciliği kavramı incelenip, veri madenciliğin yöntemlerinden kümeleme analizi kapsamlı şekilde anlatılmış, kümeleme analizi yapılırken kullanılan SPSS ve WEKA programlarından bahsedilmiş ve bir telekomünikasyon firmasına ait örnek data seti alınarak kümeleme analizi yapılmaya çalışılmış ve sonuçlar anlatılmaya çalışılmıştır.

Anahtar Kelimeler: Veri Madenciliği, Kümeleme Analizi, SPSS, WEKA

ABSTRACT

CUSTOMER CLUSTERING IN DIGITAL BROADCASTING ACCORDING TO CONTENT RATINGS

Özgür Tekingöz

Information Technologies

Thesis Supervisor: Associate Professor Mehmet Alper TUNGA

January 2015, 56 Pages

Datamining is a process of finding hidden information from large data. Classification, profiling, segmentation and clustering analysis are datamining applications to define valuable customers for companies. Customer segmentation provides companies to develop marketing programs.

The purpose of this thesis is to determine clustering of customers in broadcasting according to content ratings. First of all data mining techniques and clustering analysis are defined.After that, we tell about SPSS and WEKA. These programs are used for clustering analysis. Finally, results of the thesis are discussed and interpreted.

Keywords: Data Mining, Clustering, SPSS, WEKA

İÇİNDEKİLER

TABLOLAR

ŞEKİLLER

KISALTMALAR

Ci	:	i. küme merkezi
d_{ij}	:	i ile j. veri noktaları arasındaki öklid uzaklığı
D	:	Uzaklık matrisi
J	:	n tane datanın küme merkezlerine toplam uzaklığı
MIN	:	Minimum
MAX	:	Maksimum
N	:	Veri matrisinde yer alan bir birim
P	:	Bir veri tabanında nesne
Sim	:	Benzerlik matrisi
Sim_{ij}	:	Benzerlik matrisi elemanları
SPSS	:	Statistical Package for the Social Sciences
YSA	:	Yapay Sinir Ağları
X	:	Veri Matrisi
WEKA	:	Waikato Environment for Knowledge Analysis

SEMBOLLER

Ortalama değer	:	μ
Standart sapma	:	σ
Yüzde	:	yüzde

1. GİRİŞ

Günümüzde yaşanan yazılım teknolojilerindeki gelişmeler, artan veri sayısının çok hızlı bir şekilde depolanıp saklanabilmesine, verinin işlenmesine ve kullanılabilir bir bilgiye çok hızlı şekilde dönüştürülmesine imkân sağlamaktadır. Artan veri sayısı, bu verilerden nasıl faydalanılacağı ve verilerin anlamlı ve kullanılabilir hale getirilmesi gerektiği problemini ortaya çıkarmıştır. Bu problem doğrultusunda, günümüzde problemin çözümüne yönelik veri madenciliği kavramı ortaya çıkmıştır.

Veri madenciliği, çok fazla miktarda verinin içinden daha önceden bilinmeyen fakat önemli olan bilgiyi elde edebilme işlemidir. Burada önceden bilinmeyenden kasıt, erişilecek sonucun kestirilmemesi anlamını taşımaktadır.

Veri madenciliği günümüzde çok yaygın bir şekilde kullanılmaktadır. Örneğin pazarlama sektöründe, telekomünikasyon sektöründe, bankacılık ve sigortacılık sektörlerinde çok yaygın bir şekilde kullanılmaktadır. Bu sektörler içerisinde en yaygın kullanıldığı uygulama alanlarından biri, müşteri segmentasyonudur.

Müşteri segmentasyonu, mevcut veya hedef olarak belirlenen müşteri bilgilerinden yola çıkarak, analiz yöntemlerinin belirlenerek, müşterilerin ortak özelliklerine göre kümelenmesidir. Veri madenciliği, elimizdeki mevcut veriyi kullanarak müşteriyi kümeleme, tanımlama veya müşteri hareketleri hakkında tahminde bulunma gibi çalışmalarda yardımcı olur. Bu kümeleme, CRM projeleriyle birlikte bir segmente yönelik kampanyalar veya müşterinin hareketi tahmin edilerek buna yönelik bir öneri getirmek için kullanılabilir.

Bu çalışmada, dijital yayıncılık sektöründe hizmet veren bir firmanın kendi kanallarının yayın içeriklerinden yola çıkarak(ulusal kanalların yayın içeriklerini içermiyor), müşterilerin izledikleri içerik türlerinin oranlarına göre kümelenmesi hedeflenmiştir.

Tezin ilk kısmında veri madenciliği kavramı ele alınmış, veri madenciliği süreci, veri madenciliğinin yaygın olarak kullanıldığı alanlar, veri analizinin yapılması için gereken tekniklerden bahsedilmiştir.

Tezin ikinci kısmında sonuç alınabilmesi için kullanılan tekniklere temel oluşturan Kümeleme Analizinden ayrıntılı bir şekilde bahsedilmiştir.

Tezin üçüncü kısmında elimizdeki verinin kullanılabilir hale getirilmesi için yazılan uygulamadan, veri madenciliğinde kullanılan WEKA, SPSS programları ve bu programların içinden kullanılan fonksiyonlardan bilgiler verilmiştir.

Tezin son kısmında yapılan kümeleme analizi sonucunda ortaya çıkan sonuçlar yorumlanmıştır.

2. LİTERATÜR TARAMASI

Günümüzde artan veri sayısından dolayı veri madenciliği kavramı çok fazla önem kazanmıştır. Literatür gözden geçirildiğinde veri madenciliği kavramının çok fazla kullanıldığı anlaşılacaktır. Veri madenciliği kavramının kullanıldığı önemli alanlardan biri müşteri bazlı elde edilen bütün verilerden yola çıkılarak müşteri segmentasyonu gerçekleştirilmesidir. Dünyada ve Türkiye'de birçok farklı konuda müşteri segmentasyonu yapılmış araştırmalar vardır.

Akbulut (2006), Türkiye'deki bir kozmetik şirketinde ayrılma potansiyeli olan müşteri kitlelerini belirleyerek, bu müşterilere göre pazarlama stratejisi belirlenmesini amaçlamıştır. Kozmetik firması, alışveriş sıklığı 18 ayın üzerindeki müşterilerin profilinin belirlenmesi talebinde bulunmuştur. Bu talep doğrultusunda alışveriş sepet tutarı 100 ile 250 TL arasında olan ve alışveriş sepet tutarı 400 TL'den fazla olan müşterilere yılbaşı ve doğum günü gibi özel günlerde karakterlerine uygun olarak promosyon paketleri ve ürünler hazırlanarak hediye edilmesi, alışveriş harcaması 250 ile 400 TL arasında olan üyeler içinse reklam ve tanıtımları lokasyon bazında özelleştirilmesi sonucuna ulaşmıştır.

Turan(2010), Telekomünikasyon firmasında bulunan müşterilerin bilgilerinin alınarak, bunların incelenmesi, müşterilerin bireysel ve toplu olarak göstermiş oldukları davranışsal hareketlerden yola çıkarak profil modellemesi çalışması yapmıştır. Bu çalışma kapsamında yüksek kazancı olup faturasını zamanında ödemeyen müşteriler veya yüksek öğrenim görmüş fakat evinde internet olmayan müşteriler kimlerdir tespitleri yapılmıştır. Bu tespitler üzerinden düzenli ödeme yapılmayan bölgelerdeki müşteriler aranarak onların düzenli ödeme yapabilmeleri için öneriler sunulabilir çıkarımında bulunmuştur.

Güçdemir (2013), Uluslar arası bir TV üreticisi firmanın müşteri tabanının benzer özellikler gösteren müşteri gruplarına bölünmesi ve aynı zamanda bu grupların göreli önemlerinin bulunmasını amaçlamıştır. Bu çalışma kapsamında müşteriler 5 gruba

ayrılmıştır. Birinci grup en zengin olan müşterileri içinde bulunduran grup. Bu grup şirketin gelir kaynaklarının artmasındaki en önemli grup. İkinci grup şirket ile uzun yıllardır ilişkisi olan müşterilerin grubu. Üçüncü grup şirketin sipariş miktarının artmasını sağlayan grup. Dördüncü gruptaki müşterilerin çoğunluğu yeni müşterilerden oluşuyor. Son grup ise yakın zamanda sipariş vermemiş ve şirket ile ilişkileri kopmak üzere olan gruplar.

3. VERİ MADENCİLİĞİ

Günlük hayatımıza bilgisayarların daha çok girmesi ile birlikte, hayatta yaptığımız her işlem kayıt altına alınmaya başlandı. Örneğin marketlerde yaptığımız alışverişler, hizmet aldığımız telekomünikasyon firmalarındaki işlem geçmişimiz, bankalarda yaptığımız işlemler, hastaneye muayene olmaya gittiğimizdeki kayıt altına alınan kişisel bilgilerimiz, devlet dairelerinde yapılan işlemlerimiz, hatta telefon görüşmelerimiz ve görüntülerimiz vb. veritabanlarında tutulmaya başlandı. Etrafımızda bu kadar veri varken, bu verilerin hepsi bilgiye dönüşmeyi beklemektedir.

1990'lı yıllarla beraber büyük verilerin anlamlı hale getirilebilmesini sağlayan teknikler geliştirilmeye başlanmıştır. Bu teknikler verilerin işlenmesine ve bilgiye dönüştürülmesine olanak sağlamıştır. Veri madenciliği kavramının 2000'li yıllarda dünyada oldukça önemli bir yer aldığı görülmektedir.

3.1 TANIM

Veri madenciliği kavramına yönelik çok fazla tanım bulmak mümkündür. Veri madenciliği büyük veriler içerisinden herkesin kendi ihtiyaçlarına göre gerekli bilgiyi bulup çıkarmasıdır.

Peter Cabena ve diğ.(1998, s. 12), "Veri madenciliği büyük veri içeren veritabanlarından bilinmeyen fakat uygulanabilir ve geçerli olan bilginin elde edilmesi ve bu elde edilen bilginin şirket kararları kapsamında kullanılmasıdır." şeklinde açıklamıştır.

Piatetsky-Shapiro veri madenciliğini, bilinmeyen fakat büyük ihtimalle yararlı bilgilerin tekdüze olmayan bir zamanda ulaşılabilmesi olarak açıklamıştır. Piatetsky ve Shapiro (1991)

J. Han göre veri madenciliği; veri ambarları veya veritabanlarındaki büyük miktardaki veriler içerisinden sonucunda ilgi uyandırılabilecek bilgileri bulabilme aşamasıdır (Hand 1998).

Gartner Group veri madenciliğini, istatistiksel ve matematiksel tekniklerle birlikte örüntü tanıma teknolojilerini kullanarak, veri depolarında saklı kalan verilerin çıkarılması ile anlamlı yeni örüntülerin ve eğilim hareketlerinin keşfedilmesi sürecidir (Akpınar 2000, s.1-22).

David veri madenciliğinin matematiksel algoritmaları kullanarak büyük miktardaki verilerin düzenli bir biçimde hareket ettiğini bulmaya çalıştığını açıklamıştır. Veri madenciliği varsayımlarda bulunur ve sonuçların birleştirilmesinde insan yeteneğini kullanır. Veri madenciliği sadece bilim değil, aynı zamanda bir sanattır (Davis 1999).

Hand'e göre veri madenciliği, veritabanı teknolojisi, istatistik, makine öğrenme, örüntü tanıma ile etkileşim içerisinde olan yeni bir düzen ve büyük veritabanlarında eskiden bilinemeyen etkileşimlerin ikincil analizidir (Hand 1998).

Kitler ve Wang'e göre veri madenciliği, tahmin edilebilen önemli değişkenlerin büyük miktardaki potansiyel parametreden ayrılabilmesini sağlama yeteneğidir. Kitler ve Wang (1998)

Bransten, veri madenciliğini insanların hiçbir zaman bulmayı düşenemediği bireysel veya toplumsal eğilimlerin bulunabilmesini sağladığını belirtmiştir (Bransten 1999).

3.2 UYGULAMA ALANLARI

Veri madenciliğinin şekil 3.1' de görüldüğü gibi günümüzde çok farklı kullanım yerleri bulunmaktadır. Örneğin e-ticaret, pazarlama, bankacılık, tıbbi araştırmalar, sigortacılık ve telekomünikasyon gibi alanlarda yaygın şekilde kullanılmaktadır. Kullanım yerlerine göre aşağıdaki gibi sınıflandırılmıştır (Akpınar 2000).

3.2.1 Pazarlama

a) Müşteri satın alma davranışı belirlenmesi
b) Müşteri demografik özellikleri arasındaki bağlantıların ortaya konulması
c) Yeni müşterilerin kazanılması, var olan müşterilerin kaybedilmemesi
d) Müşteri ilişkileri yönetimi
e) Müşteri önemi derecelendirme
f) Satış varsayımlarında bulunmak

3.2.2 Bankacılık

a) Mali belirtiler arasındaki bağlantıların ortaya çıkarılması
b) Kredi kartı dolandırıcılıklarının ve sahtekârlıklarının belirlenmesi
c) Kredi kartı kullanımına göre müşterilerin gruplara ayrılması
d) Müşteri kredi isteklerinin değerlendirilmesi

3.2.3 Elektronik Ticaret

a) Saldırıların çözümlenmesi
b) E-CRM uygulamalarının yönetimi
c) Kullanıcıların davranışlarına göre web sitelerin düzenlenmesi
d) Web sayfalarına yapılan ziyaretlerinin çözümlenmesi

3.2.4 Sigortacılık

a) Poliçe talebinde bulunabilecek müşterilerin tahmin edilmesi
b) Risk içeren müşterilerin belirlenmesi
c) Sigorta dolandırıcılıklarının tespiti

3.2.5 Telekomünikasyon

a) İletişim ağlarında sorunlu bölgelerin tespiti
b) Müşteri davranışlarına göre yeni hizmetlerin
c) Kullanıcı davranışlarının belirlenmesi
d) Kaçak hat kullanımlarının belirlenmesi

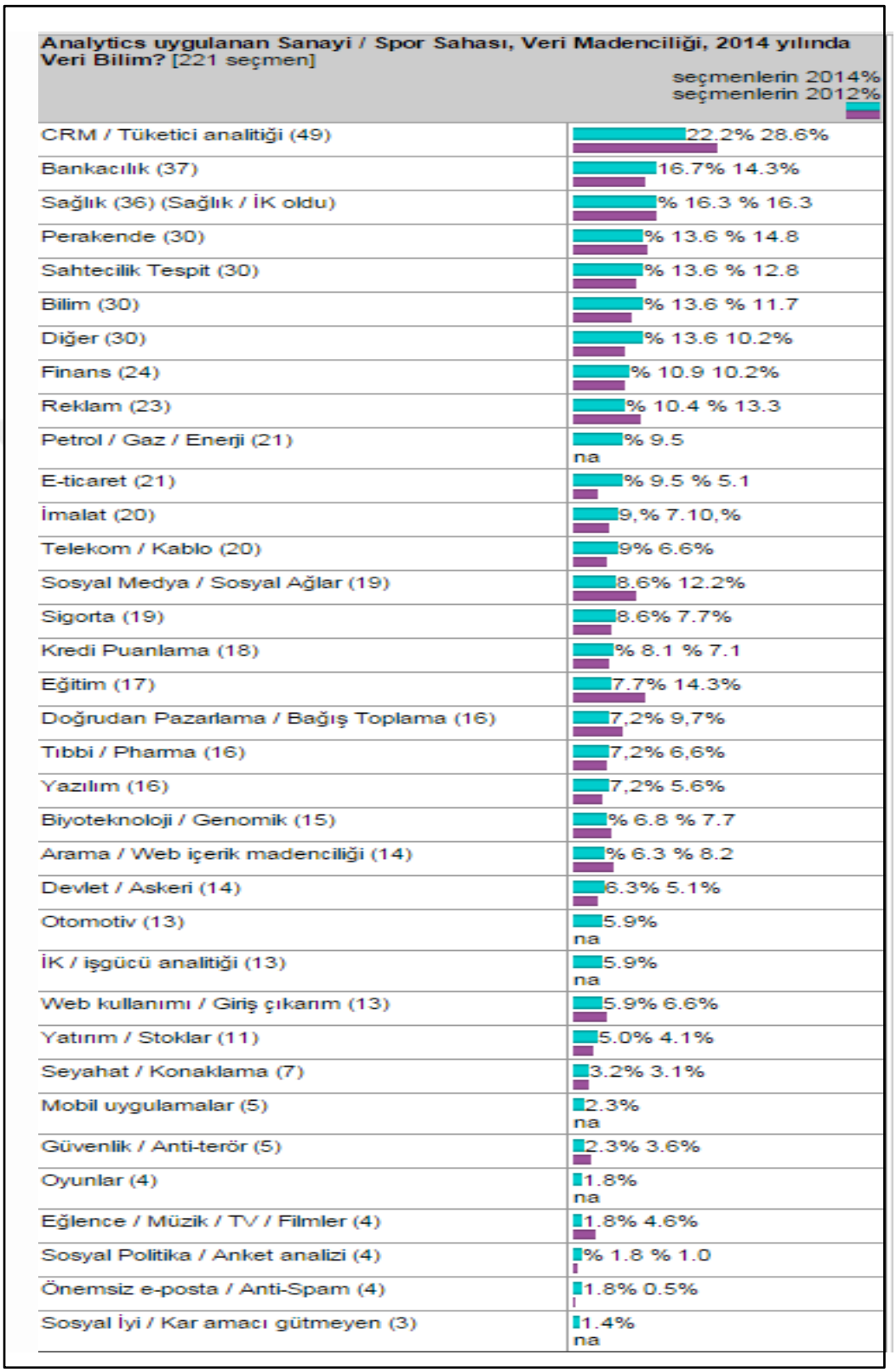
3.2.6 Tıbbi Araştırmalarda

a) DNA içerisindeki genlerin sıralarının belirlenmesi
b) Protein analizlerinin yapılması
c) Hastalık tanıları
d) Hastalık haritalarının belirlenmesi
e) Sağlık politikalarına yön verilmesi

Bunların dışında da veri madenciliğinin kullanılabileceği ve faydalı olabileceği alanlardan bazıları şunlardır:

a) Taşımacılık ve ulaşım
b) Turizm ve otelcilik
c) Devlet kurumları
d) Eğitim
e) Bilim ve mühendislik

Şekil 3.1: Aralık 2014' deki çalışmaya göre veri madenciliği kullanım yerleri



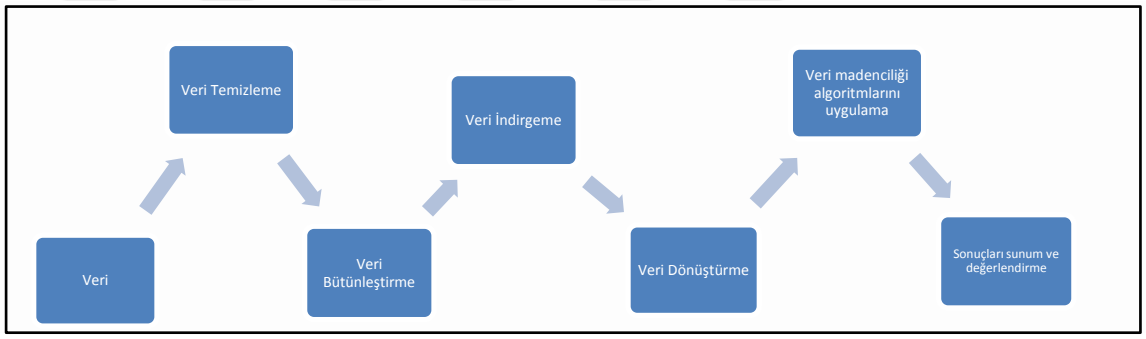
Kaynak: http://www.kdnuggets.com/polls/2014/industries-applied-analytics-data-mining-data-science.html

3.3 SÜREÇLERİ

Veri madenciliğini bir süreç olarak değerlendirmek gerekiyor. Söz konusu süreç şekil 3.2 ’de belirtilen adımları içermektedir.(Han 2001)

a) Verinin temizlenmesi
b) Verinin bütünleştirilmesi
c) Verinin indirgenmesi
d) Verinin dönüştürülmesi
e) Veri madenciliği algoritmasının uygulanması
f) Algoritma sonuçlarının sunum ve değerlendirilmesi

Şekil 3.2: Veri madenciliği süreçleri



3.3.1 Verinin Temizlenmesi

Analizi yapılacak datanın uygun olmadığı durumlarda, eksik verilerin yerine yenilerinin konulması ve tutarsız verilerin düzeltilmesi gibi durumlarda verinin temizlenmesi gerekmektedir. Eksik verilerin yenilerinin konulması için aşağıda belirtilen yöntemlerden biri kullanılabilir (Han 2000).

i. Eksik olan değerler data içerisinden atılabilir.
ii. Kayıp verinin yerine sabit bir değer verilebilir. Örneğin “geçersiz” değeri eksik veriler için kullanılabilir. Ancak bütün kayıp değerler yerine aynı sabit değerin kullanımı sorun yaratacaktır.
iii. Kayıp verinin yerine değişken bazında ortalama hesaplanarak eksik değerler yerine bu değer kullanılabilir.

iv. Kayıp verinin yerine veri kümesinin tüm değerlerinin ortalaması hesaplanarak bu değer kullanılabilir.
v. Verilere uygun tahmin yapılarak, örneğin regresyon ya da karar ağacı modeli kurularak eksik değer tahmin edilebilir ve eksik değer yerine kullanılabilir.

3.3.2 Verinin Bütünleştirilmesi

Farklı kaynakların bir araya gelmesi ile oluşan verilerin analizinin yapılabilmesi için farklı biçimdeki verilerin tek çeşit olması yani bütünleştirilmesi gerekmektedir. Veri bütünleştirilmesi yapılabilmesi için veri ambarı oluşturulmuş olması gerekmektedir. Ancak bu yapı oluşturulmamış ise veri bütünleştirme işlemi direkt olarak veriler üzerinden uygulanacaktır.

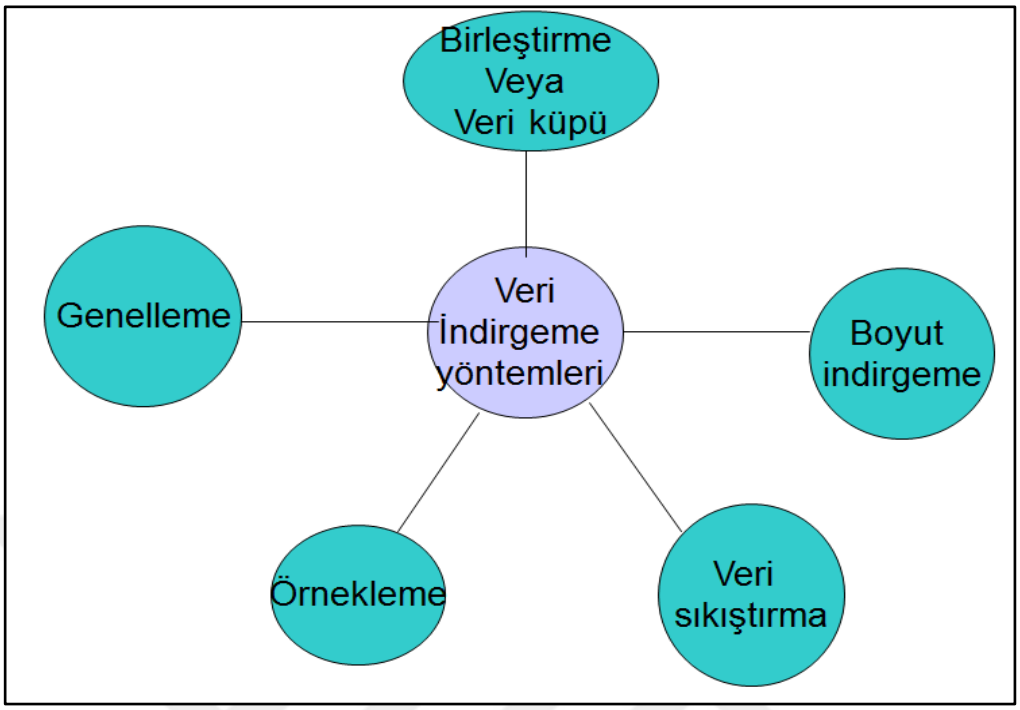
3.3.3 Verinin İndirgenmesi

Veri Madenciliği uygulamalarında bazı durumlarda verinin çözümleme aşamaları uzun sürebilir. Eğer çözümleme sonuçlarının değişmeyeceği düşünülüyorsa değişkenlerin veya verinin sayısı azaltılabilir. Verinin indirgenmesi şekil 3.3.'de bahsedildiği gibi çeşitli biçimlerde yapılabilir.(Han 2000)

i. Veri birleştirilmesi ve veri küpleri
ii. Veri boyutu indirgeme
iii. Veriyi sıkıştırma
iv. Veri örnekleme
v. Veri genelleme

Verinin indirgenmesi sırasında, veriler çok boyutlu veri küpleri haline dönüştürülerek yapılan çözümlemeler belirlenen veri küplerine göre yapılır. Verilerin arasında seçme yapılarak, fazla ve gereksiz veriler veri kümesinden silinir ve boyut indirgeme yapılabilir. Verileri sıkıştırılması sırasında, büyük veri kümelerinin sıkıştırılarak daha az yer kaplamaları sağlanır. Veri örnekleme sırasında, büyük veri kümelerinin yerine onun içerdiği alt kümesi benzeri küçük veri kümeleri oluşturulması amaçlanır. Veri genellemesi verilerin bireysel olarak değil genel itibariyle ifade edilmesini sağlar.

Şekil 3.3: Veri indirgeme yöntemleri



3.3.4 Verinin Dönüştürülmesi

Verileri bazı durumlarda çözümlemelere olduğu şekilde eklemek doğru olmayabilir. Değişkenlerin varyansları ve ortalamaları birbirlerinden farklı olduğu takdirde büyük varyansa ve ortalamaya sahip olanların diğer değişkenler üzerindeki etkisi daha fazla olur ve onların veri içerisindeki önemini büyük oranda azaltır. Bu durumdan dolayı dönüşüm yöntemlerinden biri uygulanarak değişkenlerin normalleştirilmesi veya standartlaştırılması yapılması gerekmetedir.

3.3.4.1 Min-Max normalleştirmesi

Bir veri kümesinin en büyük ve en küçük değerleri belirlenir. Bu iki değer dışındaki bütün veriler, iki değere göre normalleştirilir. Normalleştirmeye çalışmaktaki amaç seçilen değerler arasındaki büyük değer ile küçük değeri 1 ve 0 değerlerini alacak şekilde normalleştirmek ve seçilen veriler dışındakileri 0-1 aralığına yaymaktır.

$$Xnormal = \frac{X - Xmin}{Xmax - Xmin} \quad \textbf{(3.1)}$$

3.1'deki formüle göre veri kümesindeki değişkenlerin normalleştirme değeri hesaplanır. Örnek bir normalleştirme örneği aşağıda verilmiştir.

3, 8, 9, 11, 20, 22, 24, 25, 27, 30

Yukarıdaki sayıların normalleştirilmiş halleri aşağıda verilmiştir:

$$X = \frac{3-3}{30-3} = 0$$

Tablo 3.1: Min-Max Normalleştirme dönüşümü sonucu elde edilen değerler

X	Xnormal
3	0
8	0.185
9	0.22222222
11	0.29622962
20	0.629629
22	0.703703
24	0,7777777
25	0,814814
30	1

Yukarıda örnekten anlaşılacağı gibi küçük ve büyük değer 0 ve 1 olacak şekilde normalleştirilmiştir. Bunun dışındaki değerler ise 0 ile 1 aralığında değerler almıştır.

3.3.4.2 Z-score standartlaştırma

Verilerin ortalaması ve standart hatası göz önüne alınarak yeni değerlere dönüştürülmesi esasına dayanmaktadır.

$$X^* = \frac{X - \mu}{\sigma x} \quad \textbf{(3.2)}$$

3.2'deki formüle göre, bir değerin standartlaştırılabilmesi için, bu değerin veri kümesi içerisinde hesaplanan ortalama değere (μ) olan uzaklığı hesaplanır ve hesaplanan uzaklık standart sapma(σ) değerine bölünür. Ortalama değer 3.3'deki formül kullanılarak bulunur.

$$\mu = \frac{1}{n}\sum_{i=1}^{n} Xi \qquad \textbf{(3.3)}$$

Standart sapma aşağıdaki formül ile bulunur.

$$\sigma x = \sqrt{\frac{\sum_{i=1}^{n}(Xi-\mu)^2}{n-1}} \qquad \textbf{(3.4)}$$

Tablo 3.2: Z-score dönüşümü sonucu elde edilen değerler

X	X*
30	-1,1735
36	-0,6912
45	0,0321
50	0,4340
62	1,3985

3.3.5 Veri Madenciliği Algoritmasının Uygulanması

Veri madenciliği algoritmalarının uygulanabilmesi için veri madenciliği süreçlerinden verinin durumuna göre uygun görülen işlemler yapılır. İşlemler yapıldıktan ve verinin hazır hale geldiğinden emin olduktan sonra yapılmak istenen işleme göre veri madenciliği algoritmaları uygulanır. Bu algoritmalar tezin ileriki aşamalarında daha detaylı bir şekilde anlatılacaktır. Söz konusu algoritmalar sınıflandırma, kümeleme ve birliktelik kurallarıdır.

3.3.6 Algoritmanın Sonuçlarının Sunum ve Değerlendirilmesi

Veriler üzerinde istenilen algoritma uygulandıktan sonra, elde edilen sonuçlar ilgili birimler ile paylaşılır. Sonuçlar çoğu kez grafiklerle desteklenir.

3.4 YÖNTEMLERİ

Veri madenciliği yöntemleri tahmin edici ve tanımlayıcı olmak üzere ikiye ayrılır.

Tahmin edici yöntemler, sonuçları daha önceden bilinen veriler üzerinden model çıkarıldıktan sonra, çıkarılan bu model örnek alınarak sonuçları daha önceden bilinmeyen kümeler için sonuç tahmini yapılması planlanmıştır. Örneğin facebookta insanın hangi oyunları oynadığı verisi olabilir. Bu veriden yola çıkarak yeni sunulan oyunlardan hangilerini oynayıp oynayamayacağı verisi tahmin edilebilir.

Tanımlayıcı yöntemler ise elimizde bulunan verilerin örüntülerinin tanımlanmasını sağlayıp, karar vermeye rehberlik etmede kullanılabilir. Örnek olarak geliri x-y aralığından daha düşük ve geliri x-y aralığında olan iki ailenin satın alma alışkanlıklarının birbirine benzediğinin belirlenebilmesi tanımlayıcı yöntemdir.

Veri madenciliği yöntemleri işlevlerine göre,

a) Sınıflama ve Regresyon Modelleri
b) Kümeleme
c) Birliktelik Kuralları ve Ardışık Zamanlı Örüntüler

üç ana başlık altında incelenebilir. Kümeleme ve birliktelik kuralları ve ardışık zamanlı örüntüler tanımlayıcı, sınıflama ve regresyon modelleri tahmin edici modellerdir.

3.4.1 Sınıflama ve Regresyon Modelleri

Eldeki verilerden yola çıkılarak tahminler yürütülmesinde faydalanılan ve en yaygın kullanılan teknik sınıflama ve regresyon modelleridir. Sınıflama ve regresyon modelleme yapılırken 5 teknik kullanılır.

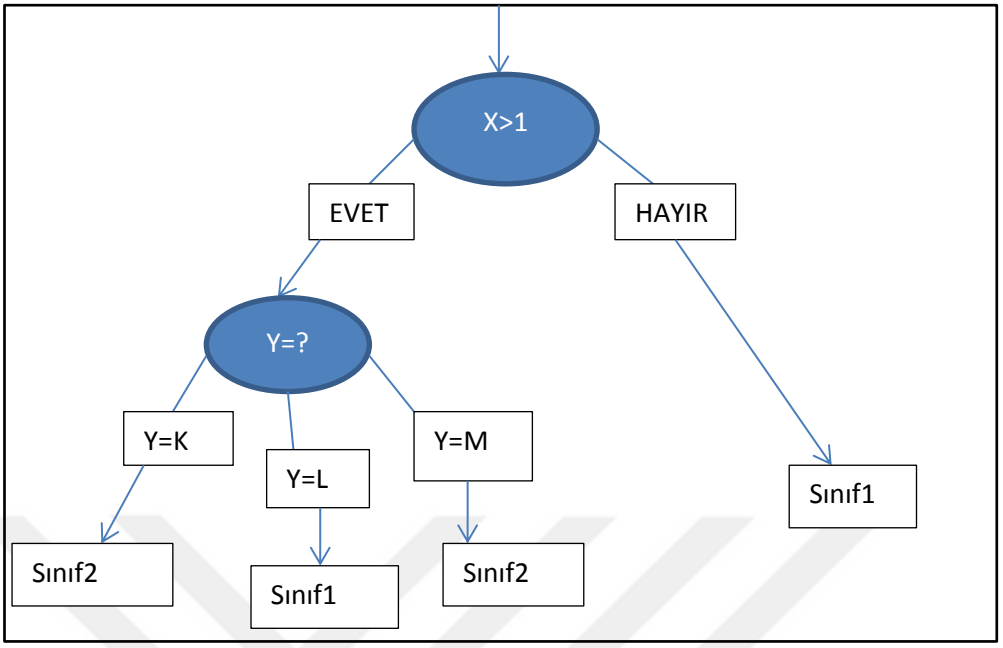
a) Karar Ağaçları Tekniği
b) Yapay Sinir Ağları Tekniği
c) Genetik Algoritmalar Tekniği
d) K-En Yakın Komşu Tekniği
e) Bellek Temelli Nedenleme Tekniği

3.4.1.1 Karar ağaçları

Verinin sınıflandırma işlemi yapılması sırasında en çok kullanılan algoritmalardan biridir. Ağaç görünümünde olan tahmin edici bir yöntemdir. Veritabanındaki her kayıt bu ağaç içerisinde dallandırılır ve sonuca göre de bu kayıt sınıflandırılır. Karar ağaçlarında ağacın her dalı bir sonraki aşamada evet veya hayır olacak şekilde minimum iki dala ayrılır. Ağaç üzerindeki bütün dalların bir olasılığı mevcuttur. Bütün dalların olasılığı mevcut olduğundan dolayı bütün dalların birbirlerine olan olasılıkları hesaplanabilir. Bütün ağacın hata oranları ve hesaplamanın verimliliği artırılması için ağacın bazı dallarının kesilmesi yani çok fazla faydası olmayan kurallar çıkarılarak arttırılabilir.

Örnek olarak X ve Y'den oluşan basit karar ağacı aşağıdaki şekil 3.4 üzerinde görülmektedir. Y'nin değerini gözönüne almadan X<=1 ve Y=K ve Y=M koşulunu sağlayan örnekler sınıf2'de; Y=L koşulunu sağlayan örnekler sınıf1'de sınıfında yer almaktadır.

Şekil 3.4: X ve Y değişkenleri üzerinden uygulanan karar ağacı



Verinin sınıflandırılması karar ağacı kullanılarak iki aşamalı bir işlemde yapılır. İlk aşama öğrenme ikinci aşama sınıflandırma aşamasıdır. Öğrenme aşamasında daha önceden elimizde bulunan test verisi, bir model oluşturabilmek için sınıflama algoritmaları kullanılarak çözümlenmeye çalışır. Çözümlenen model, karar ağacı ya da sınıflandırma kuralları şeklinde gösterilebilir. Sınıflandırma aşamasında eğitim verisi, karar ağacının ya da sınıflandırma kurallarının doğru olduğunu belirleyebilmek için kullanılır. Sonucun doğru olduğu kabul edilebilir oranda ise, oluşan kurallar eldeki yeni veri kümesini sınıflama yapılabilmesi için kullanılır.

Karar ağaçları, kurulum bedelinin ucuz olması, güvenilir ve yorumlanabilir olması ve veritabanı sistemine kolay bir şekilde entegre olabilme yetenekleri ile sınıflandırma modelleri içerisinde en yaygın kullanılandır.

3.4.1.2 Yapay sinir ağları

Biyolojik sinir sisteminden baz alınarak örneklemesi yapılan programlama yaklaşımıdır. Örneklemesi yapılan sinir hücreleri farklı şekilde birbirleriyle iletişim kurarak ağ oluştururlar. Bu ağlar verileri hafızaya alabilir, aralarındaki ilişkileri ortaya çıkarabilir ve öğrenme yetenekleri vardır.

Yapay sinir ağları; sınıflandırma ve kümeleme ile kullanılabilecek güçlü bir tekniktir. Yapay sinir ağları veri madenciliğinin uygulanabildiği birçok alanda kullanılabilmektedir.

Tantuğ'a göre veri madenciliği açısından yapay sinir ağlarının güçlü olduğu tarafları aşağıda belirtilmiştir.(2002)

a) Geniş kapsamlı sorunlara çözüm üretebilmesinde kullanılabilir.
b) Karışık durumlarla karşılaşıldığın da dahi güzel sonuçlar vermektedir.
c) Sayısal ve kategorileşmiş verilerle işlem gerçekleştirebilirler.

Veri madenciliği açısından yapay sinir ağlarının zayıf tarafları aşağıda belirtilmiştir.

a) 0 ile 1 arasında değer içeren veriler olmak zorundadır.
b) Ortaya çıkan sonuçları açıklayamazlar.
c) Ortaya çıkan sonucun en iyi sonuç olduğuna dair bir garanti yoktur.

3.4.1.3 Genetik algoritmalar

Genetik algoritmalar, biyolojik teorileri modelleyerek geliştirilmiş yöntemlerden biridir ve sonuçları açıklanabilir olma özelliğine sahiptir. Verileri işleyebilme özelliği olan algoritmalar optimizasyon sebebi ile kullanılır. Genetik algoritmaları ve yapay sinir ağları birlikte kullanılırsa başarılı sonuçlar ortaya çıkmaktadır. Genetik algoritmalar, bellek tabanlı yöntemler için kombinasyon metodunun ortaya çıkarılması ve yapay sinir ağlarının eğitilmesi işlemlerinde kullanılmışlardır (Tantuğ, 2002).

Genetik öğrenmenin aşamaları şu şekilde tarif edilmektedir. Rastgele oluşturulmuş kurallardan oluşan ilk popülasyon oluşturulur. Oluşan bütün kurallar bit dizisi halinde gösterilir.

Bütün pozitif özelliklerine rağmen genetik algoritmalarda birkaç sorun ortaya çıkmaktadır. Bunlardan en görünen özelliği karışık problemlerin genetik algoritmalarla kodlanmasının zorluğudur. Kodlanmasından sonra dair iyi bir sonuç üretildiğiyle ilgili bir garantisi yoktur (Tantuğ, 2002).

3.4.1.4 K-En yakın komşu

Veri kümelerinde birbirlerine tip olarak benzer olan kayıtlar, birbirlerine komşu konumundadırlar. Bu doğrultuda, anlaması basit ama kendisi kuvvetli olan en yakın komşu algoritması ortaya çıkarılmıştır. Yeni gelen bir verinin davranışları üzerinden tahmin yapılmak istenirse, veri kümesinde gelen veriye benzerlik açısından yakın olan örnek olarak 5 verinin davranışları incelenir. Bu 5 verinin davranışlarının ortalaması hesaplanarak, araştırılan veri için hesaplanan değer tahmini bir değer olarak kabul edilir. K - en yakın komşu algoritmasındaki k harfi komşu sayısının toplamı. Örnek olarak, k - en yakın komşu algoritmasında en yakın 10 komşuya bakılır. Adriaans ve Zantinge (1996).

3.4.1.5 Bellek temelli nedenleme

İnsanlar genellikle kararlarını yaşadıkları tecrübelere göre verirler. Benzer şekilde bellek tabanlı yöntemler de deneyimleri kullanmaktadır. Bellek temelli nedenleme yönteminde verileri içeren bir veritabanı oluşturulur ve yeni bir veri geldiğinde bu veriye yakın olan diğer veriler belirlenir ve elde bulunan veriler sayesinde tahmin yapılır veya sınıflama işlemlerinden herhangi biri uygulanır. Veriyi olduğu şekliyle kullanabilme yeteneği bellek tabanlı yöntemlerin en önemli özelliğidir. Bellek tabanlı yöntemler diğer veri madenciliği yöntemlerinden farklı olarak, veriler arasındaki uzaklıkları hesaplayan uzaklık fonksiyonu ve komşu verilere göre sonuç bulan kombinasyon fonksiyonu ile ilgilenir. (Tantuğ, 2002).

Bellek tabanlı yöntemlerin güçlü olduğu noktalar şunlardır:

a) Test veri kümesi oluşturulması kolaydır.
b) Rastgele alınan, birbiri ile alakasız olan verilerde bile kullanılabilir,
c) Kolay anlaşılabilir sonuç verebilir,
d) Çözümleme alanının fazla olduğu durumda dahi etkileyici olacak şekilde çalışabilir,

Bellek tabanlı yöntemlerin zayıf olduğu noktalar şunlardır;

a) Test veri kümeleri fazla miktarda alana ihtiyaç duyar,
b) Elde edilen sonuçlar, komşu sayısının miktarına, seçilen kombinasyon ve uzaklık fonksiyonuna göre ortaya çıkar.
c) Sınıflama ve tahmin işlemlerinde kullanılırsa işlemin maliyeti çok yüksek olabilir. (Tantuğ, 2002).

3.4.2 Kümeleme

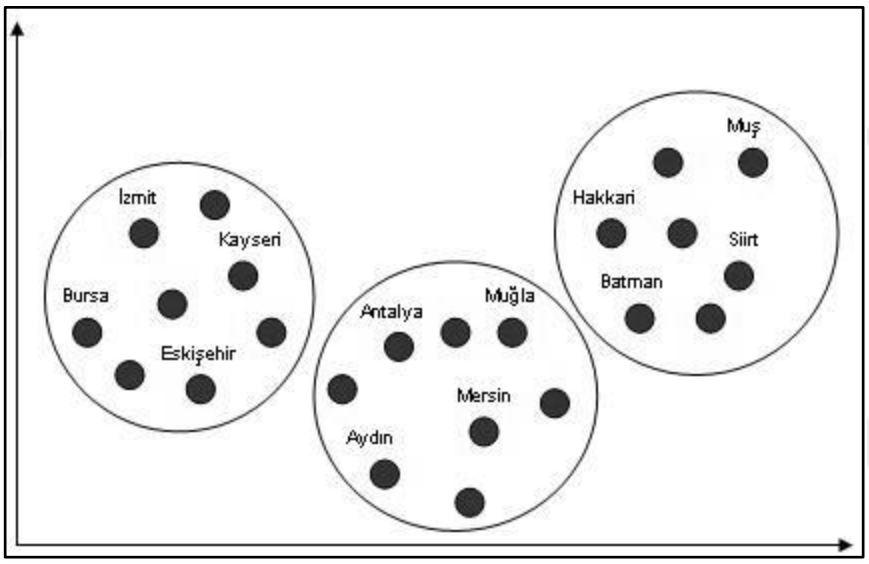
Kümeleme, veri içerisindeki birbirine benzeyen kayıtları gruplandırmayı sağlayan bir yöntemdir. Hangi yöntem ile yapılırsa yapılsın, bütün verilerin mevcut kümeler ile karşılaştırması yapılır. Veri kendisine en uygun olabilecek kümeye atanarak, atandığı kümenin değerini değiştirir. En doğru çözüm ortaya çıkıncaya kadar veriler tekrar atanır ve kümelerin merkezleri bu duruma göre ayarlanır (Akbulut, 2006).

Kümeleme analizi, büyük parçaları kendinden daha küçük parçalara ayırarak, "böl ve yönet" prensibini edinmiştir. Kümenin elemanları birbirlerine benzeyen ama aynı özellikleri barındırmayan üyelerdir. Başlangıçta veritabanındaki verilerin hangi kümeye ayrılacağı veya kümelemenin nasıl yapılacağı bilinmemekte, konu hakkında uzmanlığı bulunan bir veya birden fazla kişi tarafından incelenerek kümelerin neler olabileceği hakkında tahmin yürütülmektedir.

Kümeleme analizi birçok alanda kullanılabilir. Örnek olarak, Türkiye'deki şehirlerde yaşayan insan profilini belirlemeyebilmek amacıyla yapılacak bir araştırma için tarım ve sanayi bazlı gelir sistemine göre şehirleri kıyaslamak doğru sonuçlar vermeyebilir. Aynı

şekilde nüfusu milyonlarla ve yüzbinlerle ölçülebilen şehirleri karşılaştırmak da doğru olmayabilir. Karar verdiğimiz kriterler üzerinden benzer özellikleri barındıran şehirler aynı grupta toplanır ve birbiri arasında analiz yapılır. Örneğin Şekil 3.5'te görüldüğü gibi Van'ı İstanbul ile karşılaştırmak yerine benzer profil özellikleri gösterebilecek Siirt, Batman, Muş gibi şehirlerle karşılaştırmak daha doğru ve güvenilir olabilecek sonuçlara sahip olmamızı sağlayacaktır

Şekil 3.5: Kümeleme Analizi Örneği



Başlıca kümeleme algoritmaları şu şekilde sıralanabilir;

a) Model Tabanlı Yöntemler
b) Bölme Yöntemleri
c) Yoğunluk Tabanlı Yöntemler
d) Hiyerarşik Yöntemler
e) Izgara Tabanlı Yöntemler

3.4.3 Birliktelik Kuralları ve Ardışık Zamanlı Örüntüler

Bir işlem sırasında genellikle birlikte bulunan nesneleri barındıran kurallar topluluğudur. Zaman kavramının sadece uygulamada olması birliktelik kuralları ile ardışık zamanlı örüntülerini ayıran en önemli özellitir (Dolgun, 2006).

Birliktelik kuralları ile birlikte pazar sepeti analizi yapılabilmektedir. Pazar sepeti analizinde, ürünler müşterilerin satın aldığı, işlem ise birden fazla ürünü içinde barındıran bir defa yapılan satın alma durumudur. Pazar sepeti analizi sırasında sık sık birlikte alınan ürünler üzerinde çalışılır ve ürünlerin birbirleriyle ne çeşit bir ilişkisi olduğu bilgisine ulaşılmaya çalışılır (Dolgun, 2006).

Birliktelik kuralları analizi yapılırken karşılaşılan en büyük problem, bir eşik değerin belirlenmesidir. Eşik değerini belirleyebilmek ve ortaya çıkabilecek birliktelik kuralları içinden değerli olabilecek bilgiye ulaşabilmek çok kolay değildir. Elde edilen birliktelik kuralları içinden, ilgi çekici olmayanları çıkarabilmek amacıyla ölçüt değerlerinin belirli olması gerekir. Bu ölçüt değerleri destek ve güven değerleridir. Zantinge ve Adriaans (1996).

Birliktelik kuralları aşağıdaki örneklerde incelendiğinde eş zamanlı gerçekleşen ilişkilerin tanımlanmasında kullanılır.

a) Bira alan müşteriler, patates cipsini de yüzde 75 ihtimalle alırlar,
b) Yağ oranı düşük peynir ve yoğurt alan müşteriler, yağsız sütü de yüzde 85 ihtimalle satın alır.

Ardışık zamanlı örüntüler ise aşağıdaki örnekler incelendiğinde birbirleri arasında ilişki olan fakat birbirini takip eden zamanlarda gerçekleşen olayların ilişkilerinin açıklamasında faydalanılır.

a) Bir kişi x ameliyatı olduğunda, yüzde 45 ihtimalle 15 gün içinde Y enfeksiyonuna yakalanacaktır,
b) Bir kişi çekiç satın alıyorsa, yüzde 15 ihtimalle ilk üç ay içerisinde, yüzde 10 ihtimalle ilk üç aydan sonraki üç ay içerisinde çivi satın alacaktır.

4. KÜMELEME ANALİZİ

Araştırılan değişkenlerin birbirleri arasındaki benzerliklere göre sınıflandırma yapmayı, değişkenleri ortak noktalarına göre belirlemeye ve belirlenen sınıfların tanımlanmasını sağlayan yöntem kümeleme analizidir. Şahin ve Hamarat(2002). Kümeleme analizi sonucunda elde edilen kümeler kendi içerisinde homojen fakat birbirleri arası heterojenlik gösterirler (Sharma 1996).

Kümeleme analizi, küme sayısıyla veya küme yapısıyla ilgili herhangi bir tahminde bulunmaz. Herhangi bir veri kümesi içerisinde bulunan ve gruplamaları kesinlikle bilinemeyen değişkenleri veya birimleri birbirleri ile ortak yönleri olan alt kümelere ayrılmasını sağlayan yöntemler olarak adlandırılır. Kümeleme analizi, birim sayısına göre hesaplanan ve birimlerin benzerliklerini hesaplayabilmek için kullanılan bazı değerler ile birimleri homojen gruplar haline getirmek için kullanılır.

Kümeleme analizinin uygulanacağı veri kümesindeki her veri nesne olarak adlandırılır ve kümeleme analizi, benzerlik esasına göre birbirine benzeyen nesneleri aynı küme altında birleştirir. Nesnelerin kendi aralarındaki uzaklık ölçülerine benzerlikler denir. Nesnelerin kendi arasındaki benzerlikler, uzaklık matrisi D olarak gösterilir. D matrisinin her bir elemanı d_{ij} şeklinde ifade edilir. Birimlerin birbirleri ile olan benzerlik seviyeleri, benzerlik matrisi adı verilen Sim, Sim matrisinin her bir elamanı da Sim_{ij} şeklinde ifade edilir. Birimlerin benzerliklerini hesaplayan formül aşağıdaki gibidir.

$$\text{Simij} = 100\ (1 - \frac{\text{dij}}{\max \text{dij}}) \qquad \textbf{(4.1)}$$

Kümeleme analizi aşağıdaki şekilde aşamalandırılır;

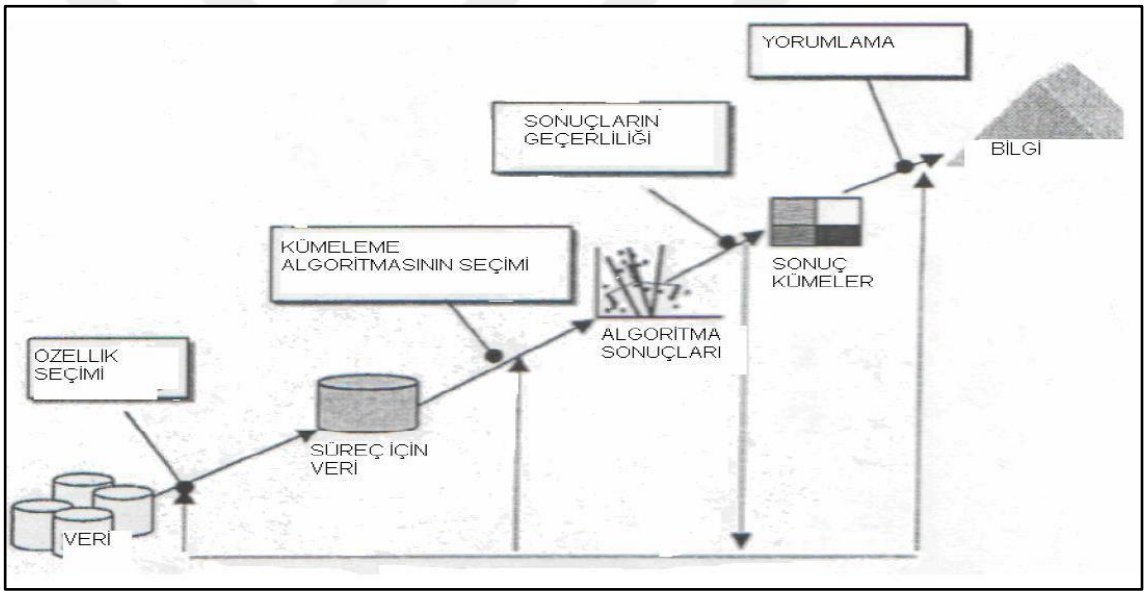
Veri matrisinin belirlenmesi: Daha önceden herhangi bir gruplandırma bilgisi bulunmayan veri kümeleri ile ilgili değişkenlerin kendileri ile ilgili gözlemlerin ortaya çıkarılmasıdır.

Benzerlik ya da farklılık matrisinin belirlenmesi: Değişkenlerin birbiri arasındaki benzerlik veya farklılığı gösteren benzerlik ölçüsü ile değişkenlerin kendi aralarındaki uzaklıkların hesaplanmasıdır.

Kümelere ayırma: Kümeleme yöntemlerinden uygun olanın kullanılması ile benzerlik ve uzaklık matrisleri göz önüne alınarak değişkenlerin uygun miktarda kümeye ayrılması işlemidir.

Yorumlama: Kümeleme analizi sonucu ortaya çıkan kümelerin doğruluğunun kanıtlanması için gereken analitik yöntemler topluluğudur.

Şekil 4.1: Kümeleme sürecinin adımları(Güler 2006)



Şekil 4.1‘ de de görüldüğü üzere kümeleme sürecinin adımları aşağıdaki gibidir;

Özellik Seçimi: Amaç, ilgilenilen konuda mümkün olduğu kadar çok bilgiyi kodlayabilen, kümeleme ile ilgili özellikleri doğru dürüst bir şekilde seçmektir. Bu yüzden, verilerin kümeleme adımlarından önce işlenmesi gerekli olabilir.

Kümeleme Algoritması: Bu adım, veri seti için iyi bir kümeleme tasarımının tanımından ortaya çıkan algoritmanın seçimiyle ilgilidir. Yakınlık ölçüsü ve kümeleme

kriteri çoğunlukla, veri setinin yapısına uygun kümeleme tasarımını tanımlamak için oldukça hızlı ve verimli çalışan kümeleme algoritmasını karakterize eder.

Sonuçların Geçerliliği: Kümeleme algoritmasının sonuçlarının doğru olup olmadığı uygun kriter ve tekniklerle test edilebilir. Kümeleme algoritmaları önceliği bilinmeyen kümeleri tanımladığından, kümeleme metotlarına bakılmaksızın, verinin sonuç bölünmesi çoğu uygulamada bazı değerlendirmeler gerektirir.

Sonuçların Yorumu: Birçok durumda, uygulama alanındaki uzman kişiler doğru karara varmak için diğer deneysel kanıtları da göz önüne alarak küme sonuçlarını değerlendirmek zorundadır.

4.1 UZAKLIK ÖLÇÜLERİ

Kümeleme analizi, gözlenen bireylerin aralarındaki benzerlik, yakınlık veya uzaklığı elde etmektir. Benzerlik, anlam olarak uzaklığın tersidir ve değerin büyük çıkması nesnelerin birbirine yakın olduğunu, değerin küçük çıkması nesnelerin birbirine uzak olduğunu göstermektedir. Kümeleme analizinde nesnelerin n değişken miktarına göre aralarındaki uzaklıkları hesaplayabilmek için birçok çeşit uzaklık ölçü birimi kullanılmıştır (Özdamar, 1999).

N sayıda değişken baz alınarak aralarındaki uzaklık değerlerini hesaplamak için kullanılan uzaklık birimine Minkowski uzaklık ölçüsü denir. Minkowski uzaklık ölçüsünün formülü aşağıdaki gibidir:

$$d(x,y) = [\sum_{i=1}^{n} |Xi - Yi|]^{\frac{1}{m}} \quad \textbf{(4.2)}$$

Benzerlik veya uzaklık ölçüleri veri matrisi üzerinde bulunan parametrelerin ölçü birimlerine göre değişkenlik gösterebilmektedir. Eğer parametrelerin değerleri aralıklı ölçekle veya oransal olarak elde edildiğinde uzaklık ölçülerinden faydalanılır. Yapılan ölçümler sayısal değer olacak şekilde yapılmışsa phi kare veya ki kare uzaklık ölçüsü tercih edilir.

Minkowski uzaklığının hesaplama şekli Öklid uzaklığının hesaplanması olarak bilinir. n*p boyutlu matrisin i. ve j. elamanları arasındaki öklid uzaklık hesabı şekil 4.2'deki gibi hesaplanır.

$$\mathrm{d(i,j)} = \sqrt{\sum_{k=1}^{p}(Xik - Xjk)^2} \qquad \textbf{(4.3)}$$

Özdamar (1999) öklid uzaklığının değişkenler ve birimlerin birbirleri arasındaki benzerlikleri ve uzaklıkları hesaplamada kullanılan doğru bir yöntem olduğu ve sonuçlarının tutarlı olduğu düşünülmektedir şeklinde açıklamıştır.

4.2 KÜMELEME YÖNTEMLERİ

Kümeleme yöntemleri; uzaklık veya benzerlik matrisinden faydalanarak değişkenlerin bulunduğu küme içerisinde homojen ve oluşan kümelerin birbirleri arasında heterojen gruplar oluşturmayı hedefler. Kümeleme yöntemleri, grupları oluştururken belirledikleri yöntemler ikiye ayrılmaktadır.

i. Hiyerarşik kümeleme
ii. Hiyerarşik olmayan kümeleme

4.2.1 Hiyerarşik Kümeleme

Hiyerarşik kümeleme yöntemlerini bağlantı yöntemleri olarak da bilinir. Birimleri farklı zamanlarda bir araya getirerek, arda ardana gelecek şekilde kümeler oluşturmayı ve bu kümelere girmesi gereken elemanların uzaklık veya benzerlik seviyesinde küme elemanı olduğunun belirlenmesine yönelik işlemlerdir (Doğan,2008)

Hiyerarşik kümelemede sırasında, her bir küme için verilerin bütününü içeren bir iletişim kurulur. Hangi yöntem kullanılırsa kullanılsın kümeler birbirlerine benzeyen özellik gösteren elemanlardan oluşturulur. Kümeler bu şekilde oluşturulduğunda içerisinde benzer özelliği barındıran elemanlar bulundurmuş olur. Hiyerarşik kümeleme işlemi, henüz aynı küme içerisinde bulunmayan birbirine en çok benzeyen iki değişkenin ve bu iki değişkenin şu an için içinde bulunduğu kümeleri belirlenmesidir.

Bu doğrultuda, kümeleme analizinin başlangıç aşamasında bütün değişkenler ayrı bir küme oluncaya kadar devam edilir. Daha sonra tek tek küme olan değişkenler birleşip tek küme haline gelinceye kadar bütün tek elemanlı kümeler ikişer kümenin bir araya gelmesi şeklinde tekrarlanır (Şimşek, 2006).

Selanik (2007) hiyerarşik kümeleme tekniklerinin uygulandığı sırada sonucunda kaç tane küme oluşacağının bilinmediğini belirtmektedir. Kümeleme yapılırken sürecin başlangıcında her birey kendi başına küme olarak kabul edilir, sürecin sonunda ise bütün bireyler aynı küme içerisinde toplanır.

Kümeleme sürecinin aşamaları aşağıdaki şekilde sıralanmıştır:

a) Öncelikle birey sayısı eşittir küme sayısı olacak şekilde işlem aypılır.
b) Birbirine en yakın iki küme birleştirilir.
c) Küme sayısında birer birer eksiltme yapılarak uzaklık matrisi bulunur.
d) b ve c adımları n-1 defa tekrar edilir.

Hiyerarşik kümelemenin özellikleri aşağıdaki şekilde belirtilmiştir.(Altıntaş,2006)

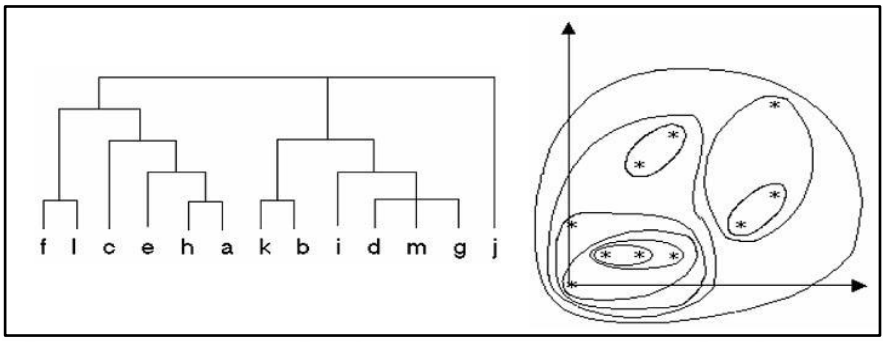
a) Veritabanını birden fazla kümeye ayrıştırır.
b) Ayrıştırma işlemi ağaç yapısı (dendogram) şeklinde uygulanır.
c) Bu ağaç, gövdeden yapraklara veya yapraktan gövdeye şeklinde kurulabilir.

Hiyerarşik kümeleme yöntemleri ikiye ayrılmıştır. Birleştirici ve ayrıştırıcı yöntemler. Birleştirici kümeleme yöntemleri başlangıç aşamasında bütün birimlerin farklı kümeler oluşturduğunu varsayarak, n tane birimleri sırası ile n, n-1, n-2, n-r, ..., 3, 2, 1 kümeye yerleştirmeye amaçlar.

Ayrıştırıcı kümeleme yöntemleri birleştirici yöntemim tam zıttı olarak başlangıç aşamasında bütün birimlerin ayrı birer küme oluşturduğunu varsayarak, birimleri kademeli şekilde sırası ile 1, 2, 3, ..., n-r, n-2, n-1, n kümeye yerleştirmeye amaçlar (Özdamar, 1999).

Şekil 4.2'de birleştirici kümeleme yöntemi baz alınarak hazırlanmış dendogram ve bu sayede oluşturulmuş kümenin grafiksel gösterimi bulunmaktadır.

Şekil 4.2: Örnek bir dendogram ve küme grafiği



Birleştirici kümeleme algoritmalarında Tek Bağlantı Yöntemi, Merkezi Kümeleme Yöntemi, Ortalama Bağlantı Yöntemi, Tam Bağlantı Yöntemi ve Ward Yöntemi benzeri yaklaşımlar, ayırıcı kümeleme algoritmalarında ise Otomatik Etkileşim Dedektörü Yöntemi, Bölünmüş Ortalamalar Yöntemi benzeri yaklaşımlar uygulanmaktadır (Şimşek, 2006).

4.2.2 Hiyerarşik Olmayan Kümeleme

Küme sayısı ile ilgili bilginin bilindiği veya araştırmayı yapan kişinin küme sayısının anlamlı olduğunu düşündüğüne karar verdiği zamanda, uygulama aşaması fazla süren hiyerarşik yöntemlerden ziyade, hiyerarşik olmayan yöntem tercih edilir. Hiyerarşik olmayan kümeleme yöntemi, hazırlanmış prototip sayesinde alt nüfusun parametrelerini tahmin etmeyi amaçlayan yöntemlerdir. (Doğan, 2008). Bu yöntemler kurumsal dayanakları çok iyi olduğundan dolayı daha fazla tercih edilmektedir (Selanik 2007).

Hiyerarşik olmayan yöntemlerde değişkenlerin birbiriyle eşleştikleri küme içerisinde bir araya gelmeleri ve n değişkenin k miktarda kümeye ayrılması hedeflenmiştir. Değişkenlerin ayrılacakları küme miktarı belirlendiği zaman, kümelerin belirlenme kriterleri baz alınarak değişkenlerin hangi kümelerin içine girmesi gerektiğine karar verilir ve girmesi gereken küme içerisine ataması yapılır (Özdamar ,1999).

Hiyerarşik olmayan teknikleri, düğüm yöntemi şeklinde adlandırabiliriz. Veri setleri öncesinde belirlenen sayıdaki kümelere ayrılır. Bu kümelerin merkezleri, düğüm noktaları hesaplanmaktadır. Gözlemler, veri setleri herhangi bir küme içerisine atanıncaya kadar devam etmektedir. Hiyerarşik kümeleme yöntemlerinde gözlemler herhangi bir küme içerisine girdikten sonra yer değişikliği yapamaz. Hiyerarşik olmayan yöntemlerde ise küme sayısının adeti başlangıç aşamasında belirlendiği için, gözlemler farklı bir kümeye atanabilmektedir (Doğan, 2008).

Hiyerarşik olmayan kümeleme yöntemlerinde veri, x adet farklı gruba ayrılır ve bu gruplar bir kümeyi belirtir. Bundan dolayı hiyerarşik kümelemenin tam tersi olarak, kümelerin toplam sayısı daha öncesinde bilinmemektedir. Hiyerarşik olmayan kümeleme analizi alttaki adımlar ile yapılır (Sharma, 1996).

a) Analistler tarafından küme sayısı kadar küme ortalaması belirlenir
b) Yeni gelen eleman, gözle inceleme yapılarak hangi küme ortalamasına daha yakınsa, o kümeye eklenir.
c) Yeni gelen elemanın eklenmesinden sonra kümelerin ortalama değerleri tekrar hesaplanır.
d) Hesaplanan ortalama değer sonucunda küme elemanları içerisinde hiç bir değişiklik olmuyorsa işlem sonlandırılır. Fakat değişiklik varsa b adımı tekrar edilir.

Hiyerarşik olmayan yöntemler arasında metoid kümeleme, k-ortalamalar yöntemi, yığma ve bulanık kümeleme benzeri yöntemler yer almaktadır. Fakat bu yöntemlerin içinden en çok k-ortalamaları yöntemi kullanılmaktadır (Şimşek, 2006).

5. VERİ MADENCİLİĞİ PROGRAMLARI

Veri Madenciliği çalışmaları yapabilmek için bilgisayar programları kullanmak gerekir. Bu doğrultuda birçok yazılım geliştirilmiştir. Bu bölümde tezin geliştirilmesinde kullanılan WEKA ve SPSS veri madenciliği programlarından bahsedilmiştir.

5.1 WEKA

Weka dünyada birçok insan tarafından kullanılan veri madenciliği uygulamalarını geliştirmek için kullanılan programdır. WEKA, Waikato Üniversitesi tarafından açık kodlu bir program olarak java platformu üzerinde geliştirilmiştir. Weka programı içerisinde sınıflandırma ve regresyon, veri işleme, kümeleme, ilişki kuralları ve görüntüleme araçları içerir.

Weka Java veritabanı bağlantı sistemiyle SQL veritabanına erişim sağlayabilir ve sorgunun çalıştırılmasından elde edilen sonucu işleyebilir. İlişkisel veri madenciliği yapamaz fakat bir koleksiyona bağlı veritabanı tablosunu tek tabloya dönüştürebilen ayrı bir yazılımı vardır.

Weka aşağıdaki özelliklere sahiptir:

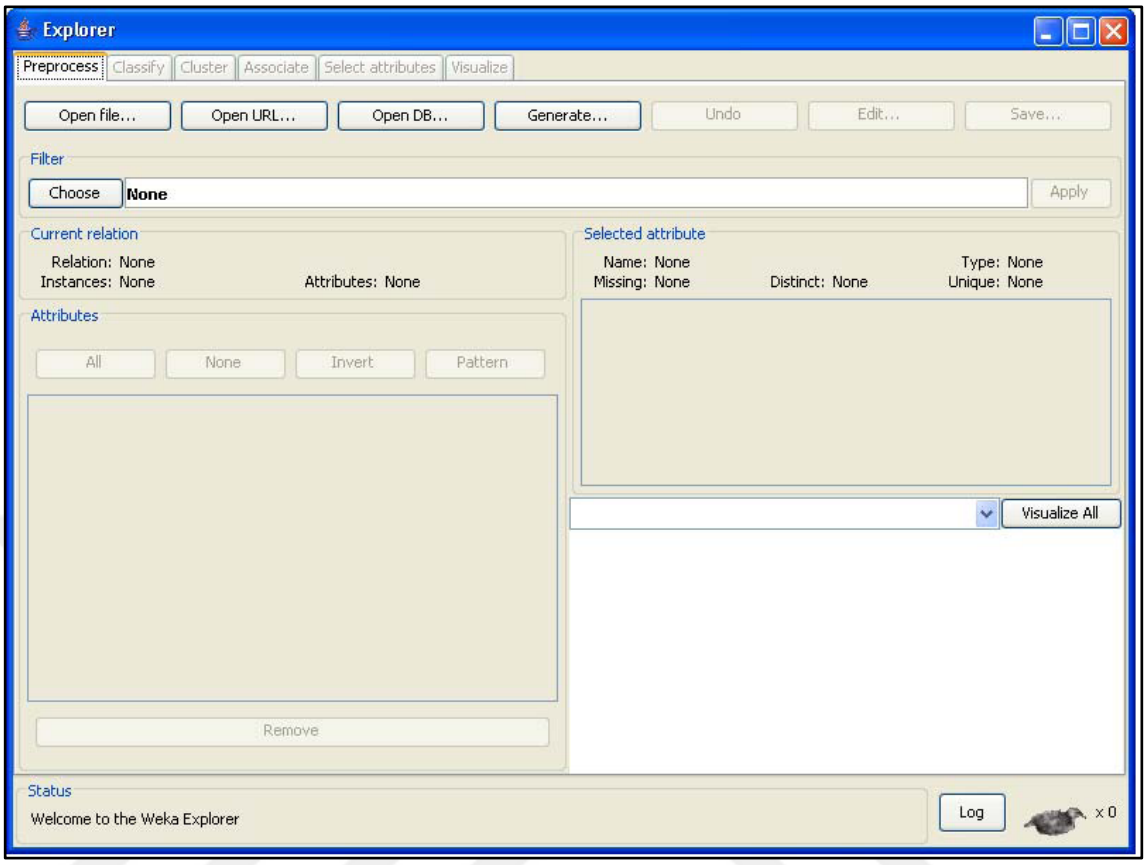
a) Seçilen algoritmanın performansını tahmin etmek için bir test yöntemine karar verilmesi
b) Sınıflandırma için kullanılacak muhtemel özelliklerin çıkarılması
c) Seçilen veri seti için mümkün sapmaların araştırılması ve etkisinin nasıl önlenebileceği.
d) Örnek alt setin seçilmesi , örneğin makine öğrenme baz alınarak yapılan kayıtlar.
e) Veritabanındaki analiz ve önizleme özelliklerinin ve verinin doğruluğunu değerlendirme.
f) Öğrenme işleminde kullanılması için özelliklerin bir alt set olarak seçilmesi
g) Örnek setlerin uygun sınıflara bölünüp sınıf niteliklerinin tanımlaması

h) Öğrenme işlemi için sınıflandırma algoritması programı

Şekil 5.1: Weka uygulama seçim ekranı

WEKA çalıştırıldıktan sonra şekil 5.1'de görüldüğü gibi uygulama menüsü listelenmektedir. Bunlar uygulamalar içerisinde komut olarak çalıştırmayı sağlayan Simple CLI, projeyi görsel ortamda çalıştırmayı sağlayan Explorer ve sürükle bırak yöntemiyle çalıştırmayı sağlayan KnowledgeFlow seçenekleridir. Explorer seçeneği ise üzerinde çalışılacak verilerin seçilmesi, bu verilerin temizleme ve dönüştürme işlemlerinin gerçekleştirilebilmesini sağlayan görünümü Şekil 5.2'deki gibi olan ekran ile karşılaşılmaktadır.

Şekil 5.2: Weka data giriş ekranı



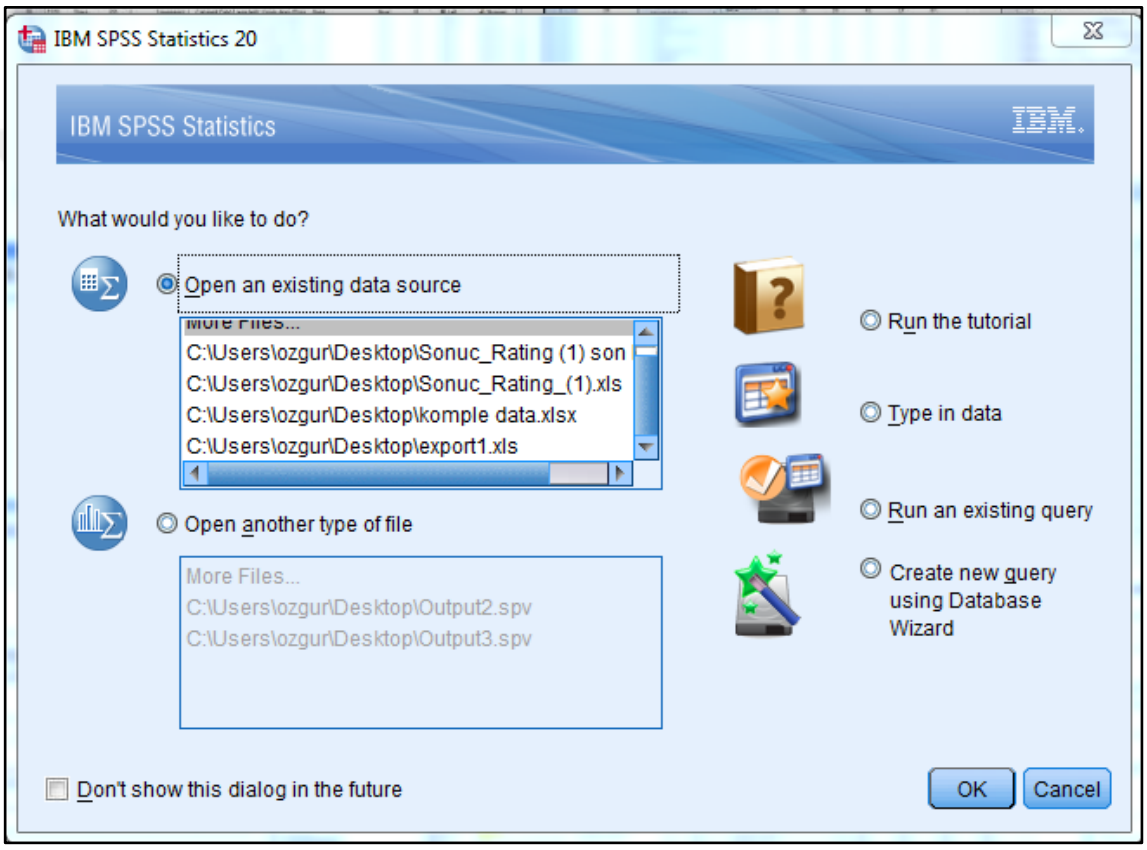
WEKA ile Arff, Csv, C4.5 formatında bulunan dosyalar import edilebilir. Text formatındaki dosyaları WEKA ile işlemek imkansızdır. Bundan sonrasında yapılacak işlemler projenin amacına göre (Sınıflandırma, Kümeleme, İlişkilendirme) uygun algoritma veya algoritmalar seçilerek veriler üzerine uygulanabilmektedir.

5.2 SPSS

SPSS istatiksel analizler yapabilmek için kullanılan bir programdır. Verilerin hızlı bir şekilde görüntülenmesini, hipotezlerin formüle edilmesini ve değişkenler arasında ilişkilerin netleştirilmesini, kümeler oluşturulmasını, eğilimlerin belirlenmesi ve tahminler yürütülmesini sağlayan bir yazılımdır. SPSS'i dünyada genellikle pazar araştırması ve sağlık araştırması yapanlar, anket firmaları, devlet kurumları ve eğitim araştırması yapan firmalar kullanılır. SPSS en çok aşağıdaki alanlarda kullanılır.

a) Kalitenin artırılması
b) İnsan kaynakları ve kaynak kullanımı
c) Akademik araştırmalar
d) Anket ve market araştırması
e) Rapor yazma ve karar verme
f) Planlama ve ileri öngörü

Şekil 5.3: SPSS programı giriş sayfası

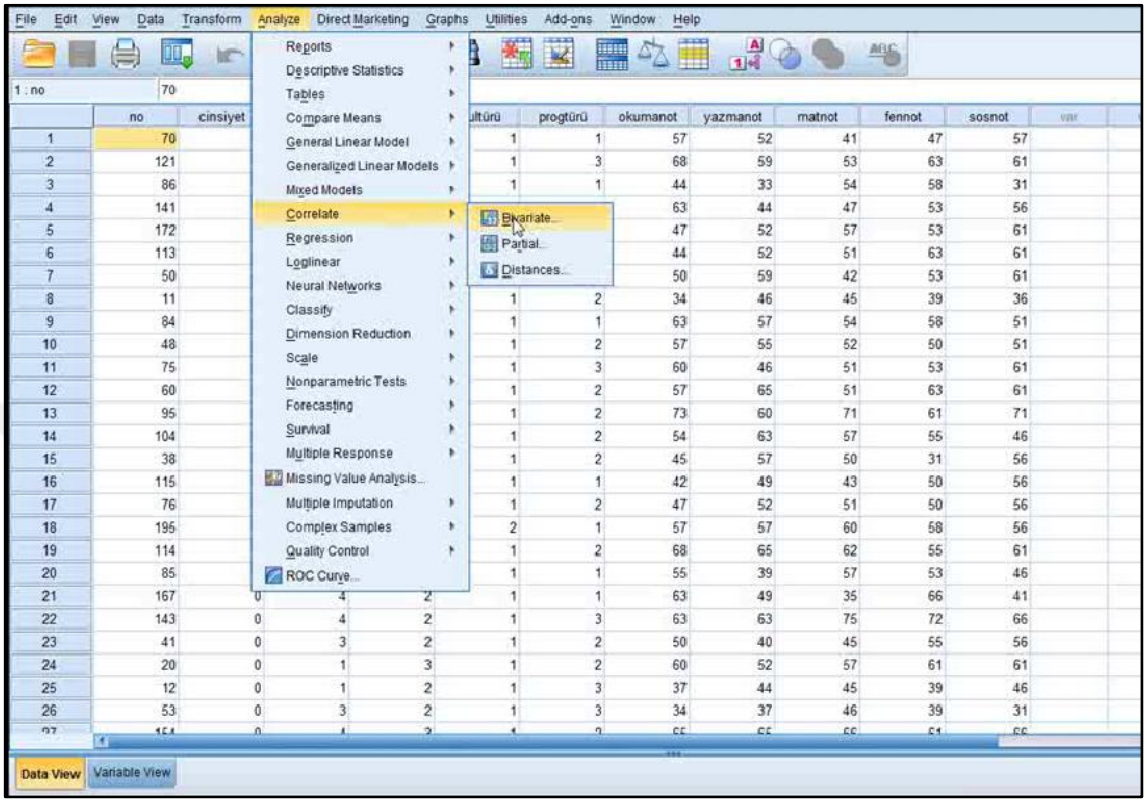


SPSS programı genel yetenekleri

a) Karmaşık verilerin doğal özelliklerine uygun olacak şekilde gelişmiş istatistiksel yöntemler sunar.
b) Verilere daha fazla model uygulanmasını sağlar.
c) Verilerin doğruluğunun kesin olmamasından dolayı analistlerin olası sonuçları modellemesini sağlayarak karar alma ve risk analizi sürecini hızlandırır.

d) Verilerin kolayca anlaşılmasını ve sonuçların farklı kimseler için hızlıca alınmasını sağlar.

Şekil 5.4: SPSS programı genel görünümü ve analiz çeşitleri



6. İÇERİK İZLENME ORANLARINA GÖRE MÜŞTERİ KÜMELEME

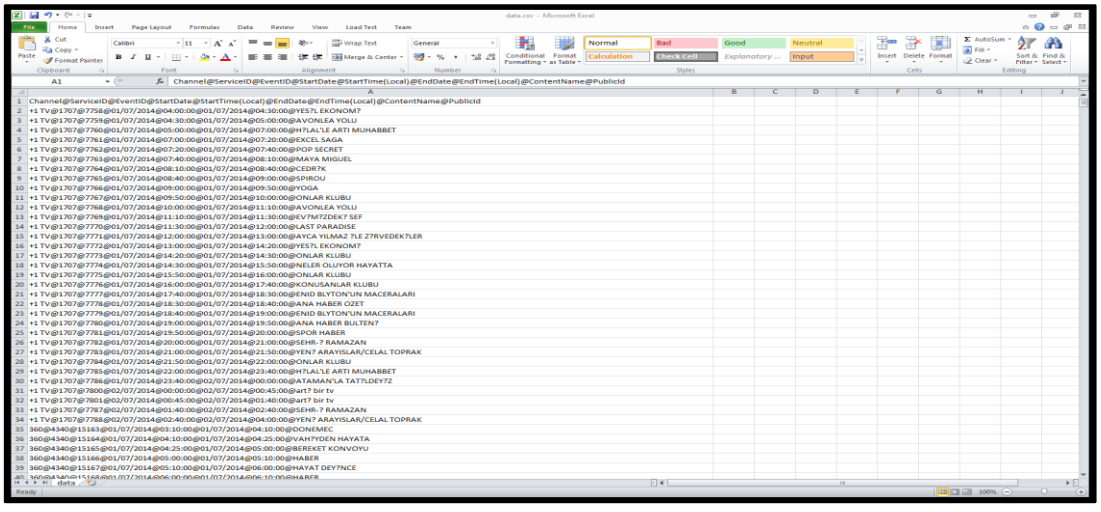
Günümüzde farklı sektörlerdeki firmaların ihtiyaçlarına göre veri analizleri yapılmaya çalışılmaktadır. Bu tez kapsamında dijital yayın yapan bir firmanın yayın içeriklerinden yola çıkılarak, müşterilerin bu yayınları izleme oranlarına göre kümeleme işlemi yapılmıştır. Kümeleme işlemi yapılmadan önce kullanılacak data üzerinde işlemler yapılmıştır. Datanın kullanılabilir hale gelmesi için yapılan işlemler veri hazırlanması kısmında bahsedilmiştir.

6.1 VERİ HAZIRLANMASI

Öncelikle verilerin hazırlanış aşamasında geliştirilen bütün uygulamalar tez cdsi içerisinde uygulamalar klasörü altında eklenmiştir.

Veri analizi yapabilmek için öncelikle verinin analiz edilebilebilir duruma getirilmesi gerekiyor. Üzerinde çalışmamız gereken datanın oluştulması için birkaç farklı datadan birleştirmeler yapılarak uygun bir veri hazırlanmıştır. Bu verilerin örnek şablonları Şekil 6.1 ve Şekil 6.2'de gösterilmiştir. Data hazırlanırken en önemli nokta, firmanın kendi kanalları üzerinden tutulan verinin üzerinde çalışılmış olmasıdır. Örneğin data içindeki Public Id dolu olanlar dikkate alınmıştır.

Şekil 6.1: Zaman bazlı kanal içerikleri



Channel => Yayınlanan kanal

Service Id => Kanal id

Event Id => Verilen yayının idsi

StartDate => Başlangıç tarihi

StartTime => Başlama zamanı(Yerel saate göre)

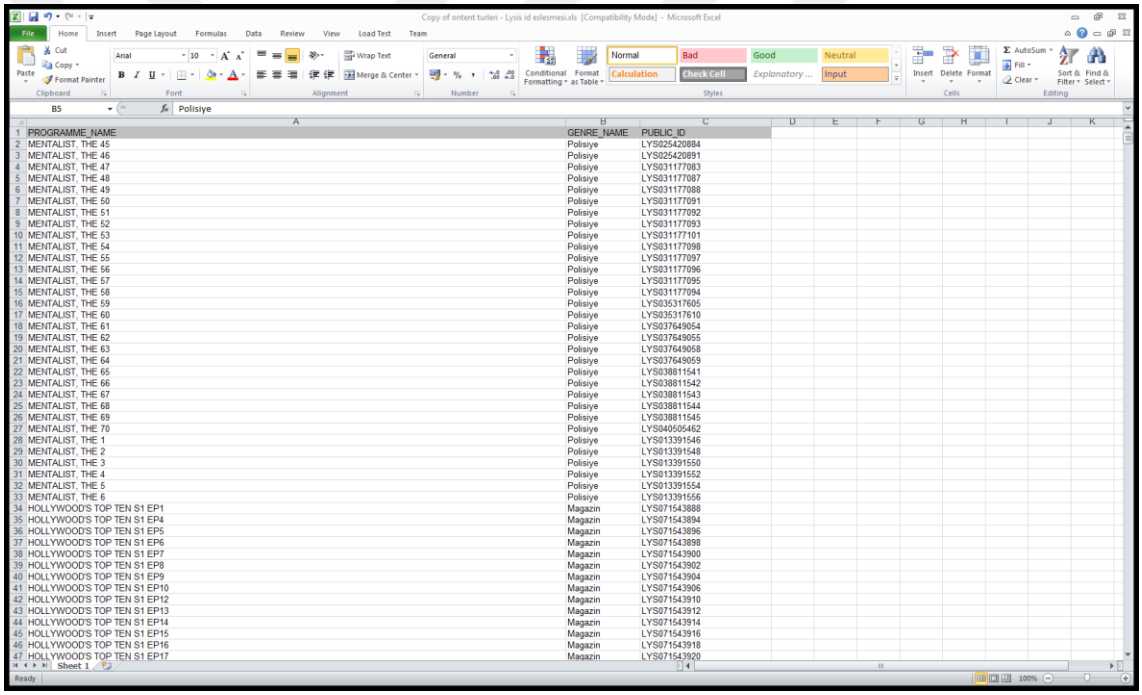
EndDate => Bitiş tarihi

EndTime => Bitiş tarihi(Yerel saate göre)

ContentName => İçerik ismi

Public Id => İçeriğin alındığı dış kaynağın id'si

Şekil 6.2: Kanal içerik türleri



PROGRAMME_NAME	GENRE_NAME	PUBLIC_ID
MENTALIST, THE 45	Polisiye	LYS025420884
MENTALIST, THE 46	Polisiye	LYS025420891
MENTALIST, THE 47	Polisiye	LYS031177083
MENTALIST, THE 48	Polisiye	LYS031177087
MENTALIST, THE 49	Polisiye	LYS031177088
MENTALIST, THE 50	Polisiye	LYS031177091
MENTALIST, THE 51	Polisiye	LYS031177092
MENTALIST, THE 52	Polisiye	LYS031177093
MENTALIST, THE 53	Polisiye	LYS031177101
MENTALIST, THE 54	Polisiye	LYS031177098
MENTALIST, THE 55	Polisiye	LYS031177097
MENTALIST, THE 56	Polisiye	LYS031177096
MENTALIST, THE 57	Polisiye	LYS031177095
MENTALIST, THE 58	Polisiye	LYS031177094
MENTALIST, THE 59	Polisiye	LYS035317605
MENTALIST, THE 60	Polisiye	LYS035317610
MENTALIST, THE 61	Polisiye	LYS037649054
MENTALIST, THE 62	Polisiye	LYS037649055
MENTALIST, THE 63	Polisiye	LYS037649058
MENTALIST, THE 64	Polisiye	LYS037649059
MENTALIST, THE 65	Polisiye	LYS038811541
MENTALIST, THE 66	Polisiye	LYS038811542
MENTALIST, THE 67	Polisiye	LYS038811543
MENTALIST, THE 68	Polisiye	LYS038811544
MENTALIST, THE 69	Polisiye	LYS038811545
MENTALIST, THE 70	Polisiye	LYS040505462
MENTALIST, THE 1	Polisiye	LYS013391546
MENTALIST, THE 2	Polisiye	LYS013391548
MENTALIST, THE 3	Polisiye	LYS013391550
MENTALIST, THE 4	Polisiye	LYS013391552
MENTALIST, THE 5	Polisiye	LYS013391554
MENTALIST, THE 6	Polisiye	LYS013391556
HOLLYWOOD'S TOP TEN S1 EP1	Magazin	LYS071543888
HOLLYWOOD'S TOP TEN S1 EP4	Magazin	LYS071543894
HOLLYWOOD'S TOP TEN S1 EP5	Magazin	LYS071543896
HOLLYWOOD'S TOP TEN S1 EP6	Magazin	LYS071543898
HOLLYWOOD'S TOP TEN S1 EP7	Magazin	LYS071543900
HOLLYWOOD'S TOP TEN S1 EP8	Magazin	LYS071543902
HOLLYWOOD'S TOP TEN S1 EP9	Magazin	LYS071543904
HOLLYWOOD'S TOP TEN S1 EP10	Magazin	LYS071543906
HOLLYWOOD'S TOP TEN S1 EP12	Magazin	LYS071543910
HOLLYWOOD'S TOP TEN S1 EP13	Magazin	LYS071543912
HOLLYWOOD'S TOP TEN S1 EP14	Magazin	LYS071543914
HOLLYWOOD'S TOP TEN S1 EP15	Magazin	LYS071543916
HOLLYWOOD'S TOP TEN S1 EP16	Magazin	LYS071543918
HOLLYWOOD'S TOP TEN S1 EP17	Magazin	LYS071543920

ProgramName=> Yayınlanan programın ismi

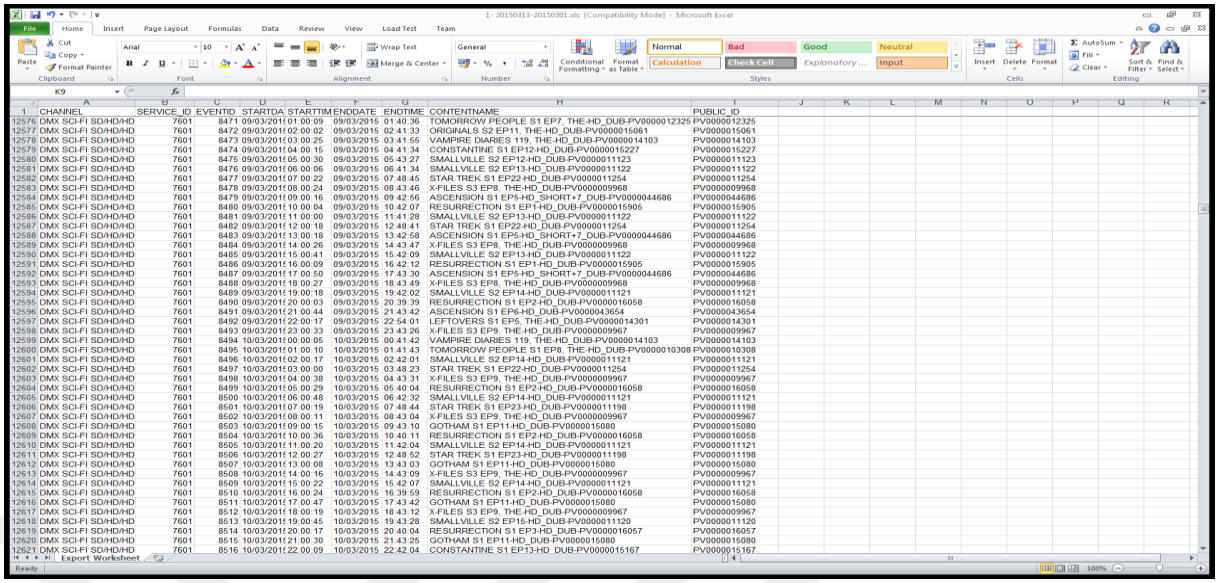
GenreName=> İçerik Türü

Public id=> İçeriğin alındığı dış kaynağın id'si

Eldeki datanın hazırlanması aşamasındaki adımlar aşağıdaki gibidir.

1. Öncelikle data excel formatına uygun hale getirilmiştir. Excel formatına uygun hale getirilirken Visual Studio üzerinde .NET uygulaması yazılmıştır. Uygulama sonucunda datanın formatlanmış hali Şekil 6.3'deki gibi oluşmuştur.

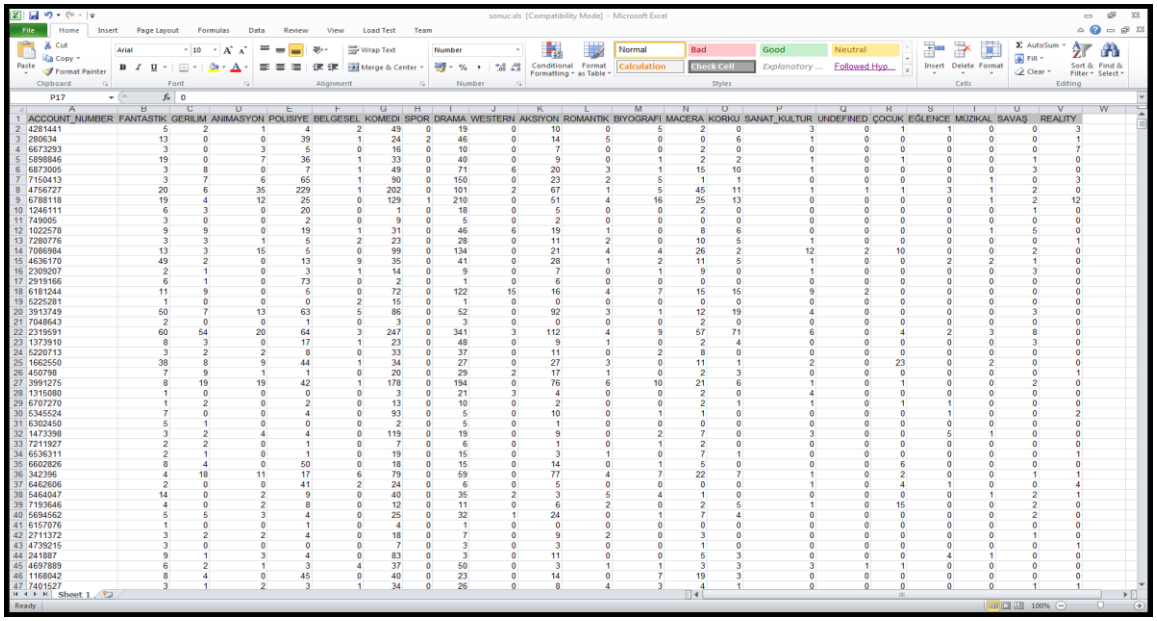
Şekil 6.3: Datanın formatlanmış hali



2. Daha sonra ORACLE veritabanı kullanılarak, tablodaki data ile tür içeriklerinin bulunduğu data arasında ilişki kurularak üyelerin izlediği içerikleri, içeriklerin tarihlerinden bağımsız olarak kaç defa izlediklerinin hesaplamasını yapan sql scripti yazılmış ve data son istenilen halini almıştır. Sql scripti ….adıyla cd içerisindedir.

Oluşan datanın içerisinde yaklaşık 25895 üyenin içeriklere göre izleme sayılarının verisi vardır. Oluşan datanın örnek şablonu Şekil 6.4'deki gibidir.

Şekil 6.4: Veri analizi yapılacak datanın son hali



Oluşan veri öncelikle üyelerin içerikleri izleme sayılarına göre hazırlanmıştır. Daha sonra izlenme sayıları binary olarak tutulmuştur. Kısacası bundan sonra üyenin o içeriği kaç defa değil, izleyip izlemediği incelenmiştir. Bu bilgiler dışında verinin içerisine üyelerin takım bilgileri, yaşadığı şehirler ve yaşları eklenmiştir. Üyelerin yaşadığı şehirleri İstanbul, Ankara, İzmir, Bursa, Adana ve Diğer şehirler olarak gruplandırılmıştır. Üyelerin tuttuğu takımlar şampiyon takımlar ve diğer olarak gruplandırılmıştır. Son olarak üyelerin yaşları veri içerisindeki üye sayısına eşdeğer gelecek şekilde 3 kategoride gruplandırılmıştır. Bunlar 18-37, 38-46, 47-100 yaş kategorileridir.

6.2 VERİ MODELLEME AŞAMALARI

Öncelikle veri kullanılabilecek duruma geldikten sonra toplam 25895 üyenin içerikleri izleyip izlemediği, takım bilgileri, yaşadığı şehir ve yaş bilgileri elde edilmiştir. Bu bilgiler doğrultusunda veri güvenilirlik analizi yapılmış, daha sonra verilerin birliktelik analizi tekniklerinden biri olan Apriori algoritması kullanılmıştır.

6.2.1 Veri Güvenilirlik Analizi

Veri modelleme işlemine başlamadan önce eldeki verinin güvenilirlik analizinden geçmiş olması gerekmektedir. Güvenilirlik analizi ortaya çıkan datanın birbirleri ile olan yakınlık derecesini ortaya çıkarmak için yapılır. Bu analiz yapılırken önemli olan alfa değeridir.

a) $\alpha < 0.40$ ise güvenilir değil
b) $0.40 < \alpha < 0.60$ ise düşük güvenilirlikte
c) $0.60 < \alpha < 0.80$ ise oldukça güvenilir
d) $0.80 < \alpha < 1.00$ ise yüksek güvenilirlikte

Tez aşamasında güvenilirlik analizi için SPSS programı kullanılmıştır ve 2 farklı şekilde yapılmıştır. Birinci analiz üyelerin içerikleri toplam izledikleri sayı üzerinden yapılmıştır. İkinci analiz ise üyelerin içerikleri izleyip izlememesi (binary olarak)üzerinden yapılmıştır. İki analizin sonucu da aşağıda belirtilmiştir. Şekil 6.5

içeriklerin izlenme sayılarına, şekil 6.6 içeriklerin izlenip izlenmediğine göre güvenilirlik analiz sonucunu göstermektedir.

Şekil 6.5: İçeriklerin izlenme sayılarına göre güvenilirlik analizi

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
,750	,865	16

Şekil 6.6: İçeriklerin izlenip izlenmemesine göre güvenilirlik analizi

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
,786	,778	16

16 değişken üzerinden iki farklı güvenilirlik analizi yaptığımızda, Cronbach's Alpha değerinin ilk analiz için 0.75, ikinci analiz için 0.786 olduğunu görüyoruz. Alpha değerinin iki durum içinde 0.70 (dünya genelinde bu değer kullanılır) üzerinde olması itibariyle verinin güvenilir olduğunu söyleyebiliriz. Fakat verinin binary şekliyle data güvenilir olduğunu görmekteyiz. Bu sebepten dolayı, bundan sonraki analiz aşamasında datanın binary haliyle devam edilmiştir.

Güvenilirlik analizi yapılırken ikinci düşünülmesi gereken kısım elimizde olan bütün alanların güvenilirlik analizinden geçirilip geçirilmemesi kararanın verilmesidir. Bu karar, Şekil 6.7'daki Cronbach's Alpha değerinin alan bazlı silinmesi halinde ne olduğunun hesaplanmasıyla verilebilir.

Şekil 6.7: Cronbach's Alpha değerinin alan silinmesine göre değerleri

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
FANTASTIC_BINARY	3,98	7,654	,469	.	,768
GERILIM_BINARY	3,95	7,523	,501	.	,765
ANIMASYON_BINARY	4,04	8,019	,368	.	,777
POLISIYE_BINARY	3,90	7,543	,447	.	,770
BELGESEL_BINARY	4,07	8,314	,242	.	,784
KOMEDI_BINARY	3,38	7,874	,334	.	,779
SPOR_BINARY	4,16	8,754	,137	.	,789
DRAMA_BINARY	3,40	7,780	,360	.	,777
AKSIYON_BINARY	3,70	7,216	,503	.	,765
ROMANTIK_BINARY	4,03	7,893	,427	.	,772
BIYOGRAFI_BINARY	4,05	7,919	,433	.	,772
MACERA_BINARY	3,83	7,252	,528	.	,762
KORKU_BINARY	3,98	7,715	,443	.	,771
SANAT_KULTUR_BINARY	3,96	7,964	,304	.	,782
COCUK_BINARY	4,03	8,440	,133	.	,792
SAVAS_BINARY	4,00	7,848	,392	.	,774

Grafikteki değerler incelendiğinde alanlar içerisinden herhangi birinin çıkarılması halinde Cronbach's Alpha değerinin ne kadar yükseleceği bilgisi verilmiştir. Buradaki değerlerin, bütün alanların güvenilirlik analiziyle eşdeğer olmasından dolayı analiz esnasında herhangi bir alan çıkarılmasına gerek duyulmamaktadır.

Şekil 6.8: Alan bazlı istatistik

Item Statistics

	Mean	Std. Deviation	N
FANTASTIC_BINARY	,19	,389	25894
GERILIM_BINARY	,21	,409	25894
ANIMASYON_BINARY	,12	,330	25894
POLISIYE_BINARY	,26	,439	25894
BELGESEL_BINARY	,10	,296	25894
KOMEDI_BINARY	,78	,411	25894
SPOR_BINARY	,00	,068	25894
DRAMA_BINARY	,76	,425	25894
AKSIYON_BINARY	,46	,499	25894
ROMANTIK_BINARY	,13	,337	25894
BIYOGRAFI_BINARY	,12	,324	25894
MACERA_BINARY	,33	,471	25894
KORKU_BINARY	,18	,386	25894
SANAT_KULTUR_BINARY	,20	,401	25894
COCUK_BINARY	,13	,337	25894
SAVAS_BINARY	,17	,375	25894

Şekil 6.8'de 16 değişkenin standart sapma(dtd. deviation) ve ortalama(mean) değerlerine baktığımızda, KOMEDI_BINARY değişkeninin 0.78 değeriyle en yüksek ortalama değeri aldığını, fakat DRAMA_BINARY değerinin güvenilirlik analizine standart sapma değeriyle en yüksek etkiyi yaptığını görüyoruz.

6.2.2 Verilerin Birliktelik Kurallarının Çıkarılması

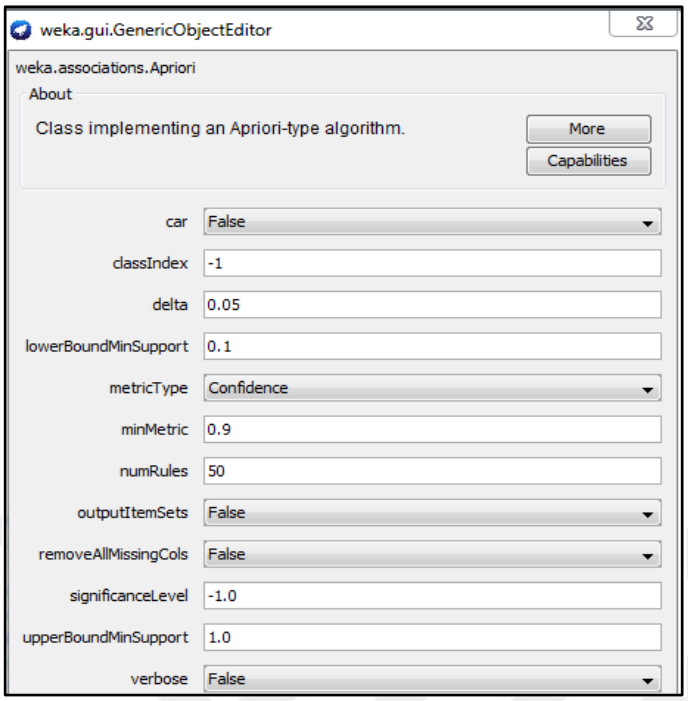
Kümeleme analizi yapılmadan önce veri içerisinde birliktelik kuralları çıkarılmıştır. Birliktelik kuralları çıkarılabilmesi için WEKA üzerindeki Apriori algoritması kullanılmıştır.

6.2.2.1 Apriori algoritması

Apriori algoritması veriler arasındaki ilişkinin çıkarılmasında kullanılır. Algoritma işleyiş olarak veri içerisinde bir elemanın diğer bütün elemanlarla olan ilişkilerini ortaya çıkarmaya çalışır. Veri içerisindeki tüm elemanlar için çalışmasını search algoritmalarına benzetebiliriz. Algoritmayı kullanmadan önce kullanılan parametreler Şekil 6.9'da gösterilmiştir.

LoverBoundMinSupport => Minimum destek sayısı
UpperBoundMinSupport => Maksimum destek sayısı
NumRules => Kural sayısı
MetricType => Kuralların sınıfı (default değeri confidence)
MinMetric => Minimum metric type değeri
Delta => Maksimum destek sayısından azalma oranı

Şekil 6.9: Apriori algoritması parametre değerleri



6.2.2.2 Birliktelik kurallarının çıkarılması

1) Birliktelik analizi ilk olarak apriori algoritmasının parametre değerleri değiştirilmeden, uygulama çalışmaya başladığındaki default değerleri ile yapılmıştır. Çıkan sonuçlar destek sayısı 19000-25000 arasında değerlerdir.

a) FantasticBinary=N 21082 ➔ BelgeselBinary=N 19536 conf:(0.93)

b) FantasticBinary=N 21082 ➔ BiyografiBinary=N 19465 conf:(0.92)

   a. Üyelerin yüzde 84,32'si fantastik içeriğini izlemiyorken, fantastik içeriğini izlemeyen üyelerin yüzde 78.14'ü belgesel ve yüzde 77.86'sı biyografi türünde içerikleri de izlemiyor.

c) AnimasyonBinary=N 22685 ➔ BelgeselBinary=N 20953 conf:(0.92)

   a. Üyelerin yüzde 90.74'ü animasyon içeriğini izlemiyorken, animasyon içeriği izlemeyen üyelerin yüzde 82.37'si belgesel türünde içeriği de izlemiyor

d) BiyografiBinary =N 22800 ➔ BelgeselBinary=N 21013 conf:(0.92)

a. Üyelerin yüzde 91.20'si biyografi içeriğini izlemiyorken, biyografi içeriği izlemeyen üyelerin yüzde 84.05'i belgesel türünde içeriği de izlemiyor

e) KorkuBinary =N 21174 ➔ BelgeselBinary=N 19510 conf:(0.92)

f) KorkuBinary =N 21174 ➔ BiyografiBinary =N 19439 conf:(0.92)

a. Üyelerin yüzde 84.69'u korku içeriğini izlemiyorken, korku içeriği izlemeyen üyelerin yüzde 78.04'ı belgesel ve yüzde 77.75 biyografi türünde içerikleri de izlemiyor

g) RomantikBinary=N 22519 ➔ BelgeselBinary =N 20722 conf:(0.92)

h) RomantikBinary=N 22519 ➔ BiyografiBinary =N 20544 conf:(0.91)

a. Üyelerin yüzde 90.07'i romantik içeriğini izlemiyorken, romantik içeriği izlemeyen üyelerin yüzde 92.01'i belgesel ve yüzde 91.22'si biyografi türünde içerikleri de izlemiyor

i) SavasBinary=N 21497 ➔BelgeselBinary =N 19706 conf:(0.92)

j) SavasBinary =N 21497 ➔ BiyografiBinary =N 19617 conf:(0.91)

a. Üyelerin yüzde 85.98'i savaş içeriğini izlemiyorken, savaş içeriği izlemeyen üyelerin yüzde 91.67'si belgesel ve yüzde 91.25'i biyografi türünde içerikleri de izlemiyor

Çıkarılan ilk kurallardan da anlaşılacağı gibi destek sayısı ne kadar çok yüksek tutulursa birliktelik kuralları o kadar üye bilgilerinden eksik biçimde oluşacaktır. Bundan sonraki kural çıkarımlarında destek sayıları ve confidence değeri değiştirilerek farklı şekilde kurallar çıkarılması sağlanmıştır.

2) Confidence = 1.0 ile 0.5 aralığı arasında

Maksimum Support = 0.7

Rule sayısı = 200

a) Üyelerin yüzde 72.08'i gerilim ve korku içeriğini izlemiyorken aynı anda bu üyelerin yüzde 93.96'sı biyografi içeriğini de izlemezler.

b) Üyelerin yüzde 72.92'si fantastik ve korku içeriğini izlemiyorken aynı anda bu üyelerin yüzde 93.14'ü romantik içeriğini de izlemezler.

c) Üyelerin yüzde 77.29'u animasyon ve savaş içeriğini izlemiyorken aynı anda bu üyelerin yüzde 92.92'si biyografi içeriğini de izlemezler.
d) Üyelerin yüzde 76.54'ü polisiye içeriğini izlemiyorken bu üyelerin yüzde 93.15'i biyografi içeriğini ve yüzde 93.58'i belgesel içeriğini de izlemiyor.
e) Üyelerin yüzde 69.16'sı macera içeriğini izlemiyorken bu üyelerin yüzde 93.72'si romantik içeriğini ve yüzde 93.31'i animasyon içeriğini izlemiyor.
f) Üyelerin yüzde 67.12'si polisiye ve korku içeriğini izlemiyorken aynı anda bu üyelerin yüzde 94.36'sı biyografi içeriği de izlemezler.
g) Üyelerin yüzde 79.14'ü drama içeriğini izliyorken bu üyelerin yüzde 89.36'sı belgesel içeriğini izlemiyor.
h) Üyelerin yüzde 81.23'ü komedi içeriğini izliyorken bu üyelerin yüzde 86.42'si çocuk içeriğini izlemiyor.
i) Diğer şehirlerde yaşayan üyelerin yüzde 90.16'sı belgesel içeriğini ve yüzde 87.99'u biyografi içeriğini izlemiyor.

3) Confidence = 1.0 ile 0.5 aralığı arasında

Maksimum Support = 0.6

Rule sayısı = 500

a) Üyelerin yüzde 61.56'sı komedi içeriğini izleyip animasyon ve biyografi içeriğini izlemiyorken, bu üyelerin yüzde 93.14'ü belgesel izlemiyor.
b) Üyelerin yüzde 67.34'ü drama içeriğini izleyip animasyon içeriğini izlemiyorken, bu üyelerin yüzde 92.01'i belgesel içeriğini izlemiyor.
c) Diğer şehirlerde yaşayan üyelerin yüzde 83.03'ü savaş içeriğini izlemiyorken, bu üyelerin yüzde 91.61'i belgesel içeriğini ve yüzde 91.21'i biyografi içeriğini izlemiyor.
d) Üyelerin yüzde 76.52'si diğer şehirlerde yaşarken, bu üyelerin yüzde 78.30'u komedi içeriğini izliyor.
e) Üyelerin yüzde 55.17'si drama ve komedi içeriğini izlerken animasyon içeriğini izlemiyorlar ve bu üyelerin de yüzde 91.17'si belgesel içeriğini izlemiyor.

f) Üyelerin yüzde 40.56'sı diğer şehirlerde yaşayıp komedi, drama izleyip ve biyografi içeriğini izlemiyor ve bu üyelerin yüzde 90.58'i belgesel içeriğini izlemiyor.

Bundan sonraki birliktelik kuralları üyelerin yaş ve takım bilgilerini de elde edebilmek amacıyla confidence değerini yüksek tutup destek değerini ara ara düşürerek hesaplanmıştır.

4) Confidence = 1.0 ile 0.9 aralığı arasında

Maksimum Support = 0.4 – 0.1

Minimum Support = 0.1 – 0.01

Rule sayısı = 40000

a) Yaş aralığı 18-37 olan üyelerin yüzde 91.64'ü belgesel, yüzde 90.70'i çocuk, yüzde 90.45'i biyografi içeriğini izlemiyor.
b) Üyelerin yüzde 20.88'i 18-37 yaş aralığında olup drama izleyip belgesel ve animasyon izlemiyor ve bu üyelerin yüzde 92.62'si çocuk izlemiyor.
c) Fenerbahçe takımı taraftarı olan üyelerin yüzde 90.62'si belgesel izlemiyor.
d) Yaş aralığı 38-46 olan ve animasyon izlemeyen üyelerin yüzde 92.57'si belgesel izlemiyor.
e) Yaş aralığı 47-100 olan ve biyografi izlemeyen üyelerin yüzde 91.45'i animasyon izlemiyor.
f) Diğer takım taraftarı olan üyelerin yüzde 73.60'ı polisiye izlemiyorken, bu üyelerin yüzde 93.86'sı belgesel de izlemiyor.
g) Galatasaray takım taraftarı olan üyelerin yüzde 71.31'i romantik ve gerilim izlemiyorken, bu üyelerin yüzde 93.84'ü biyografi içeriğini de izlemiyor.
h) Üyelerin yüzde 19.06'sı drama, aksiyon ve macera izliyorken belgesel izlemiyor ve bu üyelerin yüzde 97.81'i komedi izliyor.
i) Üyeler yüzde 20.58'i diğer şehirlerde yaşayıp 47-100 yaş aralığında olup korku izlemiyorlar ve bu üyelerin yüzde 90.72'si animasyon izlemiyor.
j) Üyelerin yüzde 22.20'si 47-100 yaş aralığında olup drama izleyip biyografi izlemiyor ve bu üyelerin yüzde 90.79'u çocukta izlemiyor.
k) Üyelerin yüzde 21.01'i 47-100 yaş aralığında olup komedi izleyip belgesel ve animasyon izlemiyor ve bu üyelerin yüzde 92.10'u çocuk izlemiyor.

l) Üyelerin yüzde 21.50'si diğer şehirlerde yaşayıp 38-46 yaş aralığında olup romantik izlemiyor ve bu üyelerin yüzde 90.88'i biyografi de izlemiyor.

m) Üyelerin yüzde 21'i 38-46 yaş aralığında olup drama izliyor ve çocuk izlemiyor ve bu üyelerin yüzde 90.86'sı belgesel de izlemiyor.

a) Üyelerin yüzde 20.97'si 38-46 yaş aralığında olup komedi izliyor ve fantastik izlemiyor ve bu üyelerin yüzde 92.65'i belgesel izlemiyor.

b) Üyelerin yüzde 9.6'sı animasyon ve aksiyon izliyorken, bu üyelerin yüzde 97.30'u drama da izliyor.

c) Üyelerin yüzde 10.13'ü animasyon, aksiyon ve sanat kültür izliyorken, bu üyelerin yüzde 97.11'i drama da izliyor.

d) Üyelerin yüzde 42.59'u aksiyon ve drama izliyorken, bu üyelerin yüzde 94.10'u komedi de izliyor.

e) Üyelerin yüzde 10.50'si Beşiktaş takım taraftarı olup fantastik ve gerilim izlemiyor ve bu üyelerin yüzde 94.48'i belgesel izlemiyor.

f) Üyelerin yüzde 10.49'u Fenerbahçe takım taraftarı olup drama izleyip aksiyon izlemiyor ve bu üyelerin yüzde 94.15'i romantik izlemiyor.

5) Data içerisinde İstanbul'da yaşayan üyeler üzerinden yapılan birliktelik kuralları

a) İstanbul da yaşayan üyelerin yüzde 50'si komedi izlerken, polisiye ve romantik izlemiyor ve bu üyelerin yüzde 95.04'ü belgesel izlemiyor.

b) Üyelerin yüzde 45.88'i animasyon ve romantik izlemiyorken komedi ve drama izliyor ve bu üyelerin yüzde 92.75'i belgesel izlemiyor.

c) Üyelerin yüzde 46.81'i komedi ve drama izliyorken fantastik izlemiyor ve bu üyelerin yüzde 92.51'i belgesel izlemiyor.

d) Üyelerin yüzde 44.31'i komedi izleyip fantastik, sanat kültür ve çocuk izlemiyor ve bu üyelerin yüzde 92.33'ü romantik izlemiyor.

e) Üyelerin yüzde 49.74'ü drama izliyorken fantastik ve savaş izlemiyor ve bu üyelerin yüzde 92.30'u biyografi izlemiyor.

f) İstanbul'da yaşayıp 18-37 yaş aralığında olan üyelerin yüzde 54.44'ü gerilim ve aksiyon izlemiyorken bu üyelerin yüzde 97'si biyografi izlemiyor.

g) 18-37 yaş aralığında olan üyelerin yüzde 54.80'i aksiyon ve savaş izlemiyorken, bu üyelerin yüzde 97'si biyografi izlemiyor.

h) 18-37 yaş aralığında olan üyelerin yüzde 53.81'i savaş, macera, gerilim ve fantastik izlemiyorken bu üyelerin yüzde 97.25'i romantik izlemiyor.
i) 38-46 yaş aralığında olan üyelerin yüzde 44.3'ü polisiye, macera, sanat ve kültür ve savaş izlemiyorken, bu üyelerin yüzde 96.83'ü biyografi izlemiyor.
j) 38-46 yaş aralığında olan üyelerin yüzde 43.7'si savaş, aksiyon ve gerilim izlemiyorken, bu üyelerin yüzde 96.79'u biyografi izlemiyor.
k) 38-46 yaş aralığında olan üyelerin yüzde 54.8'i savaş, korku, polisiye ve animasyon izlemiyorken, bu üyelerin yüzde 96.53'ü biyografi izlemiyor.
l) İstanbul'da yaşayıp Fenerbahçe taraftarı olan üyelerin yüzde 51.15'i korku, macera, animasyon ve fantastik izlemiyorken, bu üyelerin yüzde 97.84'ü biyografi izlemiyor.
m) Fenerbahçe takımı taraftarı olan üyelerin yüzde 44.23'ü fantastik, gerilim, aksiyon ve romantik izlemiyorken, bu üyelerin yüzde 97.73'ü biyografi izlemiyor.
n) Fenerbahçe takımı taraftarı olan üyelerin yüzde 44.33'ü korku, belgesel, gerilim, fantastik izlemiyorken komedi izliyor ve bu üyelerin yüzde 97.51'i romantik izlemiyor.
o) Fenerbahçe takımı taraftarı olan üyelerin yüzde 43.63'ü çocuk, romantik, polisiye izlemiyorken drama izliyor ve bu üyelerin yüzde 96.78'i belgesel izlemiyor.
p) Fenerbahçe takımı taraftarı olan üyelerin yüzde 47.74'ü romantik ve polisiye izlemiyorken drama izliyor ve bu üyelerin yüzde 96.63'ü belgesel izlemiyor.

6) Data içerisinde Bursa'da yaşayan üyeler üzerinden yapılan birliktelik kuralları
a) Bursa'da yaşayan üyelerin yüzde 46.55'i gerilim, animasyon ve çocuk izlemiyorken drama izliyor ve bu üyelerin yüzde 96.95'i belgesel izlemiyor.
b) Üyelerin yüzde 47.77'si drama izleyip animasyon, korku ve çocuk izlemiyorken, bu üyelerin yüzde 96.61'i belgesel izlemiyor.
c) Üyelerin yüzde 47.36'sı fantastik, animasyon, romantik ve çocuk izlemeyip komedi izliyorken, bu üyelerin yüzde 96.15'i belgesel izlemiyor.
d) Üyelerin yüzde 47.57'si fantastik ve polisiye izlemiyorken komedi izliyor ve bu üyelerin yüzde 95.31'i animasyon izlemiyor.

e) Üyelerin yüzde 50.40'ı macera izlemiyorken komedi izliyor ve bu üyelerin yüzde 78.57'si romantik ve biyografi izlemiyor.
f) Üyelerin yüzde 50.20'si romantik izlemeyip komedi ve drama izliyor ve bu üyelerin 92.74'ü belgesel izlemiyor.
g) 18-37 yaş aralığında olan üyelerin yüzde 50.25'i aksiyon, biyografi ve sanat kültür izlemeyip, bu üyelerin yüzde 98.98'i korku izlemiyor.
h) 18-37 yaş aralığında olan üyelerin yüzde 44.67'si savaş, gerilim, fantastik izlemeyip drama izliyor ve bu üyelerin yüzde 98.86'sı biyografi izlemiyor.
i) 18-37 yaş aralığında olan üyelerin yüzde 44.16'sı savaş, gerilim, fantastik izlemeyip komedi izliyor ve bu üyelerin yüzde 98.85'i biyografi izlemiyor.
j) 18-37 yaş aralığında olan üyelerin yüzde 42.63'ü aksiyon, macera ve sanat kültür izliyor ve bu üyelerin yüzde 98.80'i korku izlemiyor.
k) 38-46 yaş aralığında olan üyelerin yüzde 48.73'ü savaş, macera ve polisiye izlemiyorken, bu üyelerin yüzde 100'ü animasyon da izlemiyor.
l) 38-46 yaş aralığında olan üyelerin yüzde 48.22'si fantastik, macera ve polisiye izlemiyorken, bu üyelerin yüzde 100'ü animasyon da izlemiyor.
m) 38-46 yaş aralığında olan üyelerin yüzde 46.19'u macera ve savaş izlemiyorken komedi izliyor, bu üyelerin yüzde 100'ü animasyon da izlemiyor.
n) 47-100 yaş aralığında olan üyelerin yüzde 72.26'sı romantik ve animasyon izlemiyorken, bu üyelerin yüzde 98.83'ü belgesel de izlemiyor.
o) 47-100 yaş aralığında olan üyelerin yüzde 70.58'i biyografi ve animasyon izlemiyorken, bu üyelerin yüzde 98.80'i belgesel de izlemiyor.
p) Bursaspor takımı taraftarı olan üyelerin yüzde 24.44'ü gerilim, aksiyon ve macera izlemiyorken, bu üyelerin yüzde 99'u animasyon izlemiyor.
q) Bursaspor takımı taraftarı olan üyelerin yüzde 24.20'si romantik, aksiyon ve macera izlemiyorken, bu üyelerin yüzde 98.98'i animasyon izlemiyor.
r) Bursaspor takımı taraftarı olan üyelerin yüzde 21.76'sı animasyon, fantastik ve korku izlemeyip drama izliyor, bu üyelerin yüzde 98.87'si biyografi izlemiyor.

6.3. BULGULARIN DEĞERLENDİRİLMESİ

Birliktelik kurallarının çıkarılması aşamasında, veri setine apriori algoritması uygulanır. Apriori algoritmasının parametre değerlerinde değişiklikler yapılarak çok daha fazla ve farklı şekilde sonuçlar vermesi sağlanır. Algoritmanın parametrelerinde değişiklikler yapılmasının sebebi, algoritmanın default değerleri ile çalışması sırasında çıkan sonuçların, algoritmanın parametrelerinin değiştirilmediği müddetçe aynı şekilde çıkmasıdır. Algoritma üzerinde yapılan değişiklikler sonucunda birçok birliktelik kuralı ortaya çıkarılır. Birliktelik kurallarından yola çıkarak;
Verinin bütünü incelediğinde üyelerin yüzde 90'dan fazlasının belgesel ve biyografi içeriğini çok fazla tercih etmedikleri,yüzde 80'den fazlasının komedi ve drama türüne ilgi duydukları, diğer türlere ise ortalama yüzde 70 ile yüzde 80 arasında ilgi duymadıkları anlaşılır. Drama ve komedi içeriğini tercih eden üyelerin belgesel ve biyografi türünü tercih etmedikleri anlaşılır.

Diğer şehirlerde yaşayan üyelerin, komedi ve drama türüne ilgi duydukları, iki türü birlikte tercih eden üyelerin yüzdelerinin ise daha düşük olduğu anlaşılır. Bütün veri içerisinde belgesel ve biyografiyi en fazla tercih etmeyenlerin yüzdesel olarak diğer şehirlerde yaşayan üyeler olduğu anlaşılır.

Yaşı 18-37 aralığında olan üyelerin diğer üyelerden farklı olarak biyografi türünü çocuk türüne göre daha fazla tercih ettikleri, 18-37 yaş aralığında drama türünü tercih eden üyeler arasında çocuk türünü sevmeyenlerin oran olarak daha fazla olduğu anlaşılır.

Yaşı 38-46 aralığında olan üyeler arasında drama içeriği komedi içeriğinden daha fazla tercih edilir.

Yaşı 47-100 aralığında olan üyeler içinde biyografi içeriğini izlemeyenler aynı zamanda animasyon içeriğini de tercih etmezler. Ayrıca drama ve komedi içeriğini tercih eden üyeler arasında çocuk içeriği tercih edilmez.

İstanbul ilinde yaşayan üyelerin yaklaşık yarısı drama ve komedi içeriğini tercih eder. Romantik, belgesel ve biyografi en çok tercih etmedikleri içerik türleridir. İstanbul da biyografi içeriğini izlemeyen 18-37 yaş grubu 38-46 yaş grubuna göre daha fazladır.

Bursa ilinde yaşayan üyelerin, drama içeriğini tercih ettiklerinde animasyon, çocuk ve belgesel içeriğini tercih etmedikleri anlaşılır. Komedi içeriğini tercih eden üyelerin, macera ve fantastik içeriklerini tercih etmedikleri anlaşılır. Komedi ve drama içeriğini tercih eden üyelerin ise romantik ve belgesel içeriğini tercih etmedikleri anlaşılır. Bursaspor takım taraftarı olan üyelerin aksiyon ve macera içeriğini tercih etmiyorken aynı zamanda animasyon içeriğini tercih etmedikleri anlaşılır.

7. SONUÇ

Veri madenciliğinin en yaygın kullanım alanlarından biri CRM ve Müşteri segmentasyonudur. Şirketler, veri madenciliğini müşteri segmentasyonu aşamasında kullanarak, yeni müşterilere ulaşma, mevcut müşteriyi elde tutma ve müşteri ilişkileri yönetimi gibi pek çok alanda kullanmaktadır.

Bu tezde, dijital yayıncılık sektöründe hizmet veren bir firmanın kendi kanallarının yayın içeriklerinden yola çıkarak, müşterilerin izledikleri içerik türlerinin oranlarına göre kümelenmesi hedeflenmiştir.

Eldeki veri yazılan .NET uygulaması ile temizlenip istenilen formata dönüştürülmesinin ardından, SPSS üzerinde eldeki verinin güvenilirlik analizi yapılmıştır. Verinin güvenilirlik analizinin istenilen sonuçta çıkması sonucu WEKA üzerinden birliktelik kurallarının çıkarılması ile kümeleme yapılmıştır.

Eldeki veri içerisinde 25869 üye vardır. Birliktelik kuralları çıkarılırken verinin içerisinde en fazla komedi ve drama içeriğini tercih edildiği için, kurallar oluşurken bu komedi ve drama içeriklerinin kurallar üzerinde birçok etkisi olmuştur. Birliktelik kuralları sonucunda bazı kümeleme sonuçları ortaya çıkarılmıştır.

Genel olarak üyelerin yüzde 80'den fazlasının komedi ve drama içeriğini tercih ettiği, yüzde 90'dan fazlasının belgesel ve biyografi içeriğini tercih etmediği, diğer türlere ise yüzde 70 ile yüzde 80 arasında ilgi duymadıkları anlaşılır.

Veri içerisinde üyelerin yaşları 3 kategoriye indirgenmiş ve bu kategoriler üzerinden kurallar çıkarılmıştır. Bütün yaş kategorileri düşünüldüğünde drama ve komedi içeriği tercih edilirken çocuk içeriğinin tercih edilmediği görülmüştür.

Veri içerisinde üyelerin yaşadıkları şehirler nüfusun en fazla olduğu 5 şehir ve diğer şehirler olarak 6 kategoriye ayrılmıştır. Şehirler bazında çıkarılan kurallar İstanbul, Bursa ve diğer şehirler olarak üçe indirgenmiştir. Diğer şehirlerde yaşayan üyeler 5

şehirde yaşayan üyelere göre belgesel ve biyografi içeriğini daha fazla tercih etmedikleri anlaşılmıştır. Diğer şehirlerde yaşayan üyelerin komedi ve drama içeriğini yüzdesel olarak İstanbul ve Bursa da yaşayan üyelerden daha fazla tercih ettikleri anlaşılmıştır. İstanbul da yaşayan üyelerin yüzde 90'dan fazlasının romantik, belgesel ve biyografi içeriğini tercih etmediği anlaşılmıştır.

Çıkarılan birliktelik kuralları ile çok daha fazla sonuç çıkarmak mümkündür. Elde edilen sonuçlardan birçok yorum yapılabilir. Fakat eldeki veri kümesinin belirli bir zaman aralığı belirlenerek alınması, veri içerisinde içerik türlerinin izlenme oranlarında birbirlerine çok uzak aralıklı olmalarına sebep olmuştur. İleride daha detaylı yapılabilecek analizlerde daha kesin sonuçlar elde edilebilmesi için verilerin daha tutarlı alınması sağlanabilir. Verinin daha tutarlı olması çok daha güzel kurallar çıkabilmesi anlamına gelmektedir.

KAYNAKÇA

Kitaplar

Adriaans, P. ve Zantinge D. 1996. *Data Mining*, USA: Addison-Wesley Professional.

Cabena, P., Hadjinian, P., Stadler, R., Verhees, J., & Zanasi, A. 1998. *Discovering data mining: from concept to implementation*. Prentice-Hall, Inc.

Han, J. & Kamber, M. ,2001. *Data mining concept and techniques*. Morgan Kaufmann Publishers Inc.

Özkan, Y. (2013). *Veri Madenciliği Yöntemleri.* Papatya Yayıncılık Eğitim.

Özdamar, K. (1999). Paket Programlar İle İstatistiksel Veri Analizi, İkinci Baskı, Kaan Kitabevi, Eskişehir.

Sharma, S. 1996, *Applied Multivariate Techniques*, John Wiley & Sons, Inc., NewYork,

Silahtaroğlu, G. (2013).*Veri Madenciliği Kavram ve Algoritmaları*. Papatya Yayıncılık Eğitim.

Süreli Yayınlar

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

Brachman, R. and Anand, T., 1996 . "The Process of Knowledge Discovery in Databases: A Human-Centered Approach" Advances in Knowledge Discovery and Data Mining, ed. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy, *AAAI/MIT Press*.

Bransten, L., 1999. Technology–Power Tools–Looking for Patterns: Data Mining Enables Companies to Better Manage the Ream of Statistics They Collect; the Goal: Spot the Unexpected. *Wall Street Journal*, **27**(12): pp.16-20.

Güler, N. (2006), Kümeleme Analizi, Yüksek Lisans Tezi, Muğla Üniversitesi Fen Bilimleri Enstitüsü, Muğla, 157s

Hand, D. J., 1998. Data mining: statistics and more?. *The American Statistician*, **52**(2), 112-118.

Hamarat, B. (1998), *Türkiye'de Sağlık Açısından Homojen İl gruplarının Belirlenmesine İlişkin İstatiksel Bir Yaklaşım, Y. Lisans Tezi*, Anadolu Üniversitesi Fen Bilimleri Enstitüsü, Eskişehir, 75s

Kitler, R. and Wang, W., 1998 The Emerging Role of Data Mining, *Solid State Technology*, **42** (11): pp. 45

Piatetsky- Shapiro. G., 1989. Knowledge Discovery in Real Databases. A Report on the IJCAI- 89 Workshop, *Al Magazine*, **11**(5)5, pp. 68-70.

Şahin, M. ve Hamarat, B. (2002), G10 – Avrupa Birliği ve OECD Ülkelerinin Sosyo-Ekonomik Benzerliklerinin Fuzzy Kümeleme Analizi İle Belirlenmesi, ODTÜ Uluslarası Ekonomi Kongresi VI, Ankara.11-14 Eylül 2002, s. 1-19.

Diğer Yayınlar

Akbulut, S. (2006). Veri Madenciliği Teknikleri İle Bir Kozmetik Markanın Ayrılan Müşteri Analizi ve Müşteri Segmentasyonu. (Yayınlanmamış Yüksek Lisans Tezi). Gazi Üniversitesi Fen Bilimleri Enstitüsü, Ankara.

Altıntaş, T. (2006). Veri Madenciliği Metotlarından Olan Kümeleme Algoritmalarının Uygulamalı Etkinlik Analizi. (Yayınlanmamış Yüksek Lisans Tezi). Sakarya Üniversitesi, Fen Bilimleri Enstitüsü, Sakarya.

Brin, S., Motwani, R., Silverstein, C. (1997). Beyond Market Baskets: Generalizing Association Rules to Correlations (http://eecs.ceas.uc.edu/~mazlack/dbm.sp2008/Silverstein_Craig.pdf)

Davis B., 1999. Data Mining Transformed. Information Week , 751: 86

Doğan, B. (2008). Bankaların Gözetiminde Bir Araç Olarak Kümeleme Analizi: Türk Bankacılık Sektörü İçin Bir Uygulama, (Yayınlanmamış Doktora Tezi). Kadir Has Üniversitesi Sosyal Bilimleri Enstitüsü, İstanbul.

Dolgun, M. Ö. 2006. Büyük Alışveriş Merkezleri İçin Veri Madenciliği Uygulamaları, (Yayınlanmamış Yüksek Lisans Tezi).Hacettepe Üniversitesi Fen Bilimleri Enstitüsü, Ankara.

Güçdemir, H. (2013). Bulanık AHP ve Kümeleme Tabanlı Bir Müşteri Segmentasyonu Yaklaşımı: Uluslararası Bir Tv İmalar Firmasında Uygulama.(Yayınlanmamış Yüksek Lisans Tezi) Dokuz Eylül Üniversitesi Fen Bilimleri, İzmir.

Pavel Berkhin, 2002. Survey Of Clustering Data Mining Techniques.(**http://www.cc.gatech.edu/~isbell/reading/papers/berkhin02survey.pdf)**

Selanik, M. (2007). Türk Tarımının Avrupa Birliği İçindeki Yerinin Kümeleme Analizi İle Belirlenmesi. (Yayınlanmamış Yüksek Lisans Tezi). Gazi Üniversitesi, Sosyal Bilimler Enstitüsü, Ankara.

Şimşek, U.T. (2006). Veri Madenciliği ve Müşteri İlişkileri Yönetiminde Bir Uygulama. (Yayınlanmamış Doktora Tezi). İstanbul Üniversitesi Sosyal Bilimler Enstitüsü, İstanbul.

Tantuğ, A. C. 2002. Veri Madenciliği ve Demetleme. (Yayınlanmamış Yüksek Lisans Tezi) İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü, İstanbul.

Turan, E. S.(2010). Bir Telekomünikasyon Firmasında Müşteri Segmentasyonu (Yaynlanmamış Yüksek Lisans Tezi). Beyken Üniversitesi Fen Bilimleri Enstitüsü, İstanbul.