T.C.
İSTANBUL ÜNİVERSİTESİ-CERRAHPAŞA
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

YÜKSEK LİSANS TEZİ

MODERN PERAKENDECİLİK SEKTÖRÜNDE VERİ MADENCİLİĞİ TEKNİKLERİNİN UYGULANMASI

Barış YILDIRIM

DANIŞMAN
Dr. Öğr. Üyesi Tuncay ÖZCAN

Endüstri Mühendisliği Anabilim Dalı

Endüstri Mühendisliği Programı

İSTANBUL-2019

Bu çalışma 25.06.2019 Tarihinde aşağıdaki jüri tarafından
Endüstri Mühendisliği Anabilim Dalı, Endüstri Mühendisliği Programı Yüksek Lisans Tezi olarak kabul edilmiştir.

TEZ JÜRİSİ

Dr.Öğr.Üyesi Tuncay ÖZCAN
İstanbul Üniversitesi-Cerrahpaşa
Mühendislik Fakültesi

Prof. Dr. Mehmet Mutlu YENİSEY
İstanbul Üniversitesi-Cerrahpaşa
Mühendislik Fakültesi

Prof. Dr. Umman Tuğba ŞİMŞEK GÜRSOY
İstanbul Üniversitesi
İşletme Fakültesi

20.04.2016 tarihli Resmi Gazete'de yayımlanan Lisansüstü Eğitim ve Öğretim Yönetmeliğinin 9/2 ve 22/2 maddeleri gereğince; Bu Lisansüstü teze, İstanbul Üniversitesi-Cerrahpaşa'nın abonesi olduğu intihal yazılım programı kullanılarak Lisansüstü Eğitim Enstitüsü'nün belirlemiş olduğu ölçütlere uygun rapor alınmıştır.

ÖNSÖZ

Bu çalışma İstanbul Üniversitesi-Cerrahpaşa Lisansüstü Eğitim Enstitüsü Endüstri Mühendisliği Ana Bilim Dalı Yüksek Lisans Programı çerçevesinde hazırlanmıştır.

Veri madenciliği kavramının popülerliği her geçen gün artmaktadır. Bu artışta, veri büyüklüğünün günden güne büyümesi etkin rol oynamaktadır. Bankacılık, sigortacılık, perakendecilik, sağlık gibi birçok sektörde veri madenciliği fonksiyonları yaygın olarak kullanılmaktadır. Bu çalışma önemi giderek artan veri madenciliği uygulamalarının en verimli kullanıldığı sektörlerden biri olan perakendecilik sektöründe kümeleme analizi ve birliktelik kuralları madenciliği ile ilgili uygulamalar yapılmıştır.

Öncelikle veri madenciliği kavramıyla beni tanıştıran ve çalışmalarından sürekli yararlandığım hocam Prof. Dr. U. Tuğba ŞİMŞEK GÜRSOY'a teşekkür ederim. Çalışmamın başından sonuna dek bana yardımcı olan yönlendiren ve ufkumu genişleten değerli tez danışmanım Dr. Öğr. Üyesi Tuncay ÖZCAN'a, ilgili olduğum alanda çalışmama olanak tanıyan Prof. Dr. Alp BARAY'a teşekkürlerimi bir borç bilirim.

Haziran, 2019 Barış YILDIRIM

İÇİNDEKİLER

Sayfa No

ÖNSÖZ ..... **iv**
İÇİNDEKİLER ..... **v**
ŞEKİL LİSTESİ ..... **vii**
TABLO LİSTESİ ..... **viii**
SİMGE VE KISALTMA LİSTESİ ..... **ix**
ÖZET ..... **xi**
SUMMARY ..... **xiii**
1. GİRİŞ **1**
2. GENEL KISIMLAR **5**
2.1 VERİ AMBARI 5
2.2 VERİ MADENCİLİĞİ SÜRECİ 5
2.3 MODERN PERAKENDECİLİK 8
2.4 VERİ MADENCİLİĞİ FONKSİYONLARI 9
2.4.1 Tanımlayıcı Modeller 9
2.4.1.1 İlişki Analizi 9
2.4.1.2 Kümeleme Analizi 15
2.4.2 Tahmin Edici Yöntemler 25
2.4.2.1. Sınıflandırma 25
2.4.2.2. İstatistiksel Tahmin Yöntemleri 29
2.5 LİTERATÜR TARAMASI 30
3. MALZEME VE YÖNTEM **35**
3.1 PROBLEMİN TANIMLANMASI 35
3.2 VERİLERİN TOPLANMASI 36
3.3 VERİLERİN BİRLEŞTİRİLİP TEMİZLENMESİ 38
3.4 VERİLERİN HAZIRLANMASI VE SEÇİLMESİ 41
4. BULGULAR **44**
4.1 MODELİN KURULMASI VE DEĞERLENDİRİLMESİ 44
4.1.1 Kümeleme Analizi 44
4.1.2 Birliktelik Kuralları Madenciliği 48
5. TARTIŞMA VE SONUÇ **54**

KAYNAKLAR **56**
EKLER **60**
EK 1. k-Ortalamalar Algoritmasının ve Kümeleme İndekslerinin Hesaplanmasının R Dili Kodları 60
EK 2. Bulanık c-Ortalamalar Algoritmasının ve Kümeleme İndekslerinin Hesaplanmasının R Dili Kodları 61
ÖZGEÇMİŞ **62**

ŞEKİL LİSTESİ

Sayfa No

Şekil 2.1: Homojen Değişkenlere Sahip Heterojen Kümeler.....22

Şekil 2.2: Yapay Sinir Ağı Mimarisinin Temel Elemanları (Tezcanlar, 2007). 27

Şekil 2.3: Gelişmiş Bir Yapay Sinir Ağı Örneği (Tok, 2002).....28

Şekil 3.1: Önerilen Modelin Süreç Adımları. 35

Şekil 3.2: Ham Veri Seti Alıntısı. 37

Şekil 3.3: Veri Setinde Bulunan Kayıtlı ve Kayıtsız Alışveriş Sayıları.....39

Şekil 3.4: Kümeleme Analizine Hazır Hale Getirilmiş Veri Seti. 43

Şekil 4.1: R 3.6.0 Paket Programında *K*=10 iken Kümeleme Analizi İndeks Değerleri Görünümü.....45

TABLO LİSTESİ

Sayfa No

Tablo 2.1: Örnek Birliktelik Kuralı Sepet İçerikleri.....10

Tablo 2.2: Birliktelik Kuralı Örnek Kodlaması.....13

Tablo 2.3: Karar Ağaçları İle İlgili Örnek Veri Seti (Altıntaş, 2006). 26

Tablo 3.1: Veri Analizinde Kullanılacak Olan Veri Parametreleri ve Açıklamaları.....38

Tablo 3.2: Veri Setinde RFM Parametrelerine Ait Değerler.....39

Tablo 3.3: Veri Seti Müşteri Yaşam Evresi Histogramı.....40

Tablo 3.4: Veri Seti Müşteri Fiyat Duyarlılığı Tablosu. 41

Tablo 4.1: Küme Sayısı Tahminlerine Göre İndeks Değerleri. 45

Tablo 4.2: Kümeleme Analizi İle Küme Adları ve Merkezleri. 47

Tablo 4.3: "Küme-1" Birliktelik Kuralı Değerler Tablosu.....48

Tablo 4.4: "Küme-2, 3, 9" Apriori Algoritması Değerler Tablosu. 51

SİMGE VE KISALTMA LİSTESİ

Simgeler **Açıklama**

a : Bağımsız Değişken değeri "0" iken Oluşan Regresyon Değeri
$a(i)$: : i. Niteliğin Bulunduğu Kümedeki Noktalara Ortalama Mesafesini
$b(i)$: i. Niteliğin Bulunmadığı Kümelerdeki Noktalara Ortalama Uzaklığı
b_n : Denklemdeki Değişkenlerin Katsayısı
C_k : Öğe Aday Kümesi
C_j : Küme Merkezi
$d(i, j)$: i ve j Gözlem Değerleri Arasındaki Uzaklık
e : Denklemdeki Hata Değeri
ϵ : Tolerans Değeri
K : k-Ortalamalar Algoritmasında Küme Sayısı
L_k : Sık Bulunan "k" Öğeli Kümeler
m : Bulanıklık Parametresi(Bulanık C-Ortalamalar Yönteminde)
n : Gözlem Sayısı
P_0 : Gruba Ait Olmama Olasılığı
P_1 : Gruba Ait Olma Olasılığı
R : Korelasyon Katsayısı
R_i : Kümeler Arası Maksimum Karşılaştırma Oranı
U_{ij} : Üyelik Matrisi
x_n : Denklemdeki Değişken Sayısı
δ_i : i. Kümedeki değişkenin Küme Merkezine Uzaklığı

Kısaltmalar **Açıklama**

A.B.D. : Amerika Birleşik Devletleri
AR-GE : Araştırma Geliştirme
BSS : Kümeler Arası Uzaklık Kareler Toplamı
CHI : Calinski-Harabasz İndeksi
CRISP-DM : Veri Madenciliği Süreci (Cross Industry Standart Process for Data Mining)
CRM : Custom Relationship Management, Müşteri İlişkileri Yönetimi
DBI : Davies-Bouldin İndeksi
DNA : Deoksiribo Nükleik Asit
DNI : Dunn İndeksi
KLI : Kryzanowski-Lai İndeksi
MS : Microsoft
NCR : National Cash Register (Amerika Menşeili Teknoloji Şirketi)
PAM : Medoidlerin Etrafında Bölümlendirme (Partitioning Around Medoids)
RFM : Yakınlık, Frekans, Harcama (Recency, Frequence, Monetary)
SAS 8.2 : Statistical Analyse Software Version 8.2, İstatistik Analiz Programı
SI : Silhouette İndeksi
SPSS : Statsitical Package for Social Sciences, İstatistik Analiz Programı

SSE : Hata Karelerinin Toplamı
OLAP : Online Analitik İşleme
VA : Veri Ambarı
VM : Veri Madenciliği
VMUT : Veri Madenciliği Uygulama Teknikleri
WGSS : Küme Merkezi Uzaklıkları Kareler Toplamı
WSS : Kümeler İçi Uzaklık Kareler Toplamı
XBI : Xie-Beni İndeksi

ÖZET

MODERN PERAKENDECİLİK SEKTÖRÜNDE VERİ MADENCİLİĞİ TEKNİKLERİNİN UYGULANMASI

YÜKSEK LİSANS TEZİ

Barış YILDIRIM

İstanbul Üniversitesi-Cerrahpaşa

Lisansüstü Eğitim Enstitüsü

Endüstri Mühendisliği Anabilim Dalı

Danışman : Dr. Öğr. Üyesi Tuncay ÖZCAN

Bilginin kaynağı olan veri, veri madenciliği kavramında işlenerek anlam kazanan bir güç haline gelmiştir. Dünya'da ve Türkiye'de veri madenciliğine yapılan yatırımlar gün geçtikçe artmaktadır. Özellikle perakendecilik sektöründe sıkça kullanılan veri madenciliği yöntemleri işletmelerin vazgeçilmez parçaları haline gelmiştir.

Bu çalışmada, ilk olarak veri madenciliğinin içinde barındırdığı temel kavramlar açıklanmış ve veri madenciliği süreci sıralanmıştır. Daha sonra, veri madenciliğinin işletmelere kattığı pozitif değerlerden biri olan müşteri sadakatini sağlama amacı doğrultusunda ilk olarak yaklaşık 900.000 müşteri alışveriş kaydı kullanılarak müşteri segmentasyonu yapılmıştır. Bu amaçla; k-ortalamalar algoritması, Bulanık c-Ortalamalar algoritması kullanılmıştır. En iyi algoritma ve küme sayısı, kümeleme indekslerinden faydalanılarak belirlenmiştir. Son olarak, oluşturulan kümeler için geri kazandırılma ve sadakati artırma amacı hedefi ile pazar sepet analizi gerçekleştirilmiştir. Bu noktada; Apriori algoritması ile birliktelik analizi yapılmış ve birlikte alınan ürün kategorileri keşfedilmiştir. Çalışmanın sonunda, kurulan modellerin çıkarımları göz önünde bulundurularak hedef kitle ve gelecek çalışmalar için farklı pazarlama stratejileri ve öneriler sunulmuştur.

Haziran 2019, 76 sayfa.

Anahtar kelimeler: Veri Madenciliği, Kümeleme Analizi, k-Ortalamalar, Bulanık c-Ortalamalar, Birliktelik Kuralları Madenciliği, Apriori Algoritması

SUMMARY

APPLICATIONS of DATA MINING TECHNIQUES IN MODERN RETAIL SECTOR

M.Sc. THESIS

Barış YILDIRIM

Istanbul University-Cerrahpasa

Institute of Graduate Studies

Department of Industrial Engineering

Supervisor : Assist. Prof. Dr. Tuncay ÖZCAN

Data, which is the source of information, has become a power that has gained meaning by being processed in the concept of data mining. Investments made in data mining in the world and Turkey are increasing day by day. Data mining methods, which are often and particularly used in the retail sector, has become an essential part of businesses.

In this work, the fundamental concepts included in data mining are explained and the process of data mining is aligned initially. Then, in order to ensure customer loyalty, which is one of the positive values that data mining has contributed to businesses, customer segmentation has been formed from approximately 900.000 customer transaction records. For this purpose, the k-Means algorithm and the Fuzzy c-Means algorithm were used. The best algorithm and the number of clusters were determined by using clustering indices. Finally, market basket analysis was carried out with the aim of recovery and loyalty increase for the clusters that were created. At this point, an association analysis was performed with the Apriori algorithm and the product categories bought together were discovered. At the end of the study, different marketing strategies and suggestions were presented for the target group and future studies by considering the implications of the established models.

June 2019, 76 pages.

Keywords: Data Mining, Clustering Analyse, K-Means, Fuzzy C-Means, Assocation Rules Mining, Apriori Algorithm

1. GİRİŞ

Özellikle günümüzde internetin yaygın kullanımı ve elektronik sistemlere dayalı alışverişler ile başlayan veri oluşumları birçok sektörde verinin büyümesine yol açmıştır. Araştırmacılar, bu veriler arasındaki örüntüleri ve yararlı bilgileri bulabilmek için bir takım analizler oluşturmaya çalışmışlardır. 90'lı yılların başında başlayan bu analizler ile veri madenciliği (VM) kavramının temelleri atılmıştır. VM, özellikle 90'lardan günümüze aktif bir araştırma alanı olmakla birlikte karmaşık doğası nedeniyle henüz tam anlamıyla çözümlenememiştir (Kaspa ve diğ., 2018).

Kelime anlamına bakıldığında, dar anlamda bilgiyi altın bilgiye çevirmek olarak açıklanabilecek veri madenciliği kavramının, Türkçe literatürde tek bir tanımı yer almamaktadır. Geniş bir tanımı yapıldığında, karmaşık bilgi yoğunluğu bulunan dev veri yığınları arasından ilk bakışta keşfedemeyeceğimiz örüntülerin, belirli bir amaç için kurulmuş modeller ile amaca hizmet edecek şekilde ortaya çıkarılması ve sonucunda anlamlı bilgiler elde edilmesi olarak tanımlanır.

VM, Galileo'nun bilimsel yöntem geleneğini takip etmekte ve daha sonraları 19. yüzyıldaki bilimsel metotlara geri dönerek günümüze kadar gelmektedir (Fung ve Mangasarian, 2002). Günümüzde bilgi ve bilgiyi kullanan teknolojilerin önemi oldukça fazladır. Bu sebeple, eskiye göre her geçen gün artan veri sayısı, işletmelerin ellerinde giderek artan bir veri yığını haline gelmiştir. Daha sonra, bu veri yığınları arasında işletmenin tanımlar çıkarmasına ve sonuçlar keşfetmesine yardımcı olacak veriler, veri yığınları arasında bulunamayacak duruma gelmiş ve bu verilerin anlamlandırılıp işlenerek işletmenin amaçlarına hizmet eder bir duruma getirilmesi amaçlanmıştır. İşte bu noktada; VM teknikleri içerdiği makine öğrenmesi esaslı modeller ile veri yığınlarının işlenmesine ve faydalı bilgilerin elde edilmesine imkân vermektedir.

Başka bir ifadeyle, küresel ekonomide bilgiye sahip olmak oldukça değerlidir. Dolayısı ile bilgiye sahip olanın her zaman bir adım önde olduğu bu dönemde işletmeler bulundukları pazarda avantajlı hale gelebilmek için her gün daha da artan bir çaba içine girmektedirler. Tüm bu çabalar sonucu çevrelerinden topladıkları veriler ne kadar fazla olursa olsun işletmeleri henüz diğerlerinden ayıracak bir duruma eriştiklerini söylemek zordur. Veri

büyüklüğünün çok hızlı arttığı bir ortamda, işletmeler ellerinde anlamlı bilgilerin anlamsız bilgiler ile bir arada olduğu karışık bir veri ambarına sahip olmaktadır. Bu noktada; işletmeler ellerindeki veri yığınından anlamlı bilgiler elde edebilmek için verileri nasıl kullanması gerektiğini VM teknikleri ile analiz edebilecektir. VM, işletmeler için bu durumda verilerin düzenlenmesinden, yapılacak analizlerin yöntemlerine kadar bir kılavuz niteliği taşımaktadır.

VM uygulamaları için ana bileşenlerden birisi veri ambarıdır. Veri ambarlarının oluşumu geçmiş yıllarda oldukça uzun zaman almasına rağmen günümüzde dev veri ambarları çok daha kısa sürelerde oluşturulabilmektedir. Veri ambarlarının hızlanarak oluşması veri madenciliği tekniklerinin çoğalmasına ve farklı amaçlar için farklı analizler yapılmasına olanak tanımaktadır. Günümüzün rekabetçi ortamında, yalnızca veriye sahip olan değil onu doğru şekilde işleyen işletmeler ön sırada yer alacaktır.

VM algoritmaları, günümüzde birçok sektörde başarıyla uygulanmaktadır. Bu sektörler arasında; finans, bankacılık, güvenlik, savunma, tıp ve otomotiv sektörlerinde VM kavramı etkin olarak kullanılmakta ve önemi her geçen gün artmaktadır.

Veri madenciliğinin amacı yalnızca verilerden kullanılabilir bilgi elde etmek değildir. VM işletmenin satış politikalarının sağlıklı şekilde ilerlemesi için de kullanılmaktadır. Örneğin; iki ürün arasındaki bağlantının keşfi bizi farklı bir satış politikasına sevk edebilir. Bu sayede satışlar artırılabilir. VM aynı zamanda maliyeti düşürmek amacı ile kullanılabilir. Yüksek stok miktarları ile çalışan bir firma elindeki veriler ile gelecek tahminini oluşturarak atıl stok miktarını azaltarak daha doğru miktarda bir sipariş verebilir. Ayrıca şirketin yeni politikalarının uygulanma başarısı ve Ar-Ge çalışmalarında da kılavuz olarak kullanılabilir. Aşağıda VM'nin kullandığı alanlara ait örnekler verilmiştir.

VM pazarlama alanında;

- Müşteri segmentasyonunun yapılması
- Satın alma alışkanlıklarının belirlenebilmesi
- Gelecek dönemlere yönelik satış ve ciro tahminlerinin yapılması
- Raf dizaynı ve reyon düzenlemesi
- Reklam, promosyon vb. kampanyaların düzenlenmesi
- Müşteri sadakatinin sağlanması ve artırılması

Bankacılık alanında;

- Müşteri riskinin belirlenmesi (Kredi puanı)
- Kredi kartları ile ilgili kişiye veya kişi grubuna özel ayrıcalıklar oluşturulması
- Sahtekârlıkların tespiti
- Çek karnesi verilmesi kredili mevduat hesabı limitinin belirlenmesi, kredi kartı limitinin belirlenmesi
- Vadesiz hesapların kullanılabilir bakiyelerinin belirlenmesi
- Finansal tabloların analizi ve aralarındaki ilişkilerin saptanması

Otomotiv sektöründe;

- Üretim bandında etkinliğin ve verimliliğin artırılması
- Maliyetlerin düşürülmesi

Eğitim sektöründe;

- Uzaktan eğitim sistemlerinde öğrenci profilini belirlenmesi
- Öğrenci başarı kriterlerinin belirlenmesi
- Seçmeli derslerin belirlenmesi

Tıp alanında;

- Gen haritalarının analizi
- DNA'lar arasında gen ve genlere dayalı hastalıkların tespiti

Sigortacılık sektöründe;

- Müşteri ile alakalı risk oranlarının belirlenmesi
- Poliçe takiplerinde ve yenilemelerdeki avantajların belirlenmesi

gibi problemlerin çözümünde veri madenciliği yöntemleri yaygın olarak kullanılmaktadır.

Bu tez çalışmasının bundan sonraki bölümlerinin gelişimi şu şekildedir: İkinci bölümde; genel kısımlar başlığı altında veri ambarı, VM süreci ve modern perakendecilik gibi VM ile ilgili temel kavramlar açıklanmıştır. Üçüncü bölümde; çalışmada kullanılan veri seti açıklanmış ve veri önişleme adımları gerçekleştirilmiştir. Dördüncü bölümde; kümeleme analizi ve birliktelik kuralları madenciliği kullanılarak gerçekleştirilen uygulama çalışmasına yer verilmiştir. Son bölümde ise, çalışmadan elde edilen sonuçlar yorumlanmıştır.

2. GENEL KISIMLAR

Bu bölümde VM ile ilgili temel kavramlar hakkında açıklamalara yer verilmiştir.

2.1 VERİ AMBARI

Veri ambarları; verilerin işlenerek tutarlı duruma getirildiği veri tabanlarıdır. Veri tabanındaki veriler birçok süzgeçten geçirilerek veri ambarlarına aktarılmaktadır. Dolayısı ile veri ambarı (VA) her zaman analizlere ve raporlamalara hazırdır.

VA, farklı kaynaklardan gelen verilerin bir araya getirilmesiyle oluşturulmaktadır. Veri ambarları içerisinde müşteri işlem verileri, müşteri demografik özellikleri gibi veriler bulunabilir. Belirli zaman aralıkları ile eklenen bu veriler çeşitli analizlere hazır şekilde tutulur ve değiştirilemez. VA içinde veri temizlemeyi, verilerin entegrasyonunu ve çevrimiçi analitik sürecini (OLAP) barındırır (Han ve diğ., 2012). "Online Analitik İşleme" ifadesinin kısaltması olan OLAP çok boyutlu bir modelleme aracıdır. İki boyutlu olan ve veri ambarında bulunan verilerin daha fazla boyutta analiz edilmesini sağlar.

VA, birçok verinin birbiriyle ilişkilendirilmesine olanak tanır ve karar vermeye yardımcı olur. İçinde barındırdığı veriler standartlaştırılmış ve güvenilirliği onaylanmıştır. Veri ambarında veriler yeniden oluşturulmaz, belirli dönemlerle ambara yeni veriler eklenir. Bir veri ambarı kendi bütünlüğünü ve esnekliğini koruyabilir. Bir veri ambarının iç bütünlüğü, veri ambar ilkelerine bağlı kalındığı takdirde, veri ambarının bir veri ambarı olmaya devam etmesine izin verir (Silvers, 2011).

İşletmelerde başarıyla kurulduğunda, VA birçok veri madenciliği projesinin gerçekleştirilmesine temel oluşturmaktadır. Bankacılık, sigortacılık, finans ve kamuda sistemlerin işleyişini ve verimliliğini artırmaya yarayan veri madenciliği uygulamalarında veri ambarlarından faydalanılmaktadır.

2.2 VERİ MADENCİLİĞİ SÜRECİ

VM, anlık yapılan bir analizden oldukça farklı bir yapıdadır. Bir yandan bünyesinde birçok analizi barındırırken, diğer yandan bir süreç mantığında işler ve geri bildirimler ile süreci

denetler. Bu sebeple VM için uygulanan kalıplaşmış süreç adımları oluşturulmuştur. Bu sürecin adı CRISP-DM süreci olarak adlandırılmaktadır. "The Cross Industry Standard Process for Data Mining" kelimelerinin kısaltmasıdır ve 1996 yılında henüz çok yeni olan veri madenciliği alanında üç firma tarafından ortaya atılmıştır. VM tekniklerinin kullanılmasında öncü olan Chrysler, 1990 yılında kurulan ve 1994 yılında ilk VM uygulamaları içeren paket programı hazırlayan SPSS ve VM konusunda teknolojik gelişmeleri takip eden ve veri madenciliğinin danışmanlığını yapan NCR şirketinin oluşturduğu konsorsiyum CRISP-DM terimini literatüre kazandırmıştır (Şimşek Gürsoy, 2009).

I. *Problemin tanımlanması:* VM sürecinin en önemli adımı olarak görülebilir. Problemin açık bir şekilde belirlenmesi, kullanılabilir sonuçlar elde etmek için en önemli etkendir. VM sürecinin mantıksal işleyişini sağlamak için belirlenen probleme odaklı bir işletme yapısı ve süreci uygulayanların işletmenin hedefinin farkında olması gerekmektedir. Bu süreçte probleme uygun çözüm yollarını saptayan yine süreci uygulayanlardır.

II. *Verilerin toplanması:* Tanımlanan problemle alakalı olduğu düşünülen verilerin yer aldığı kaynaklardan veri temin edilmesi aşamasıdır. Bu aşamada analizi yapacak kurum veya kişi kendi kaynakları dışında kaynaklardan da veri toplayabilir. Fakat bu aşamada dikkat edilmesi gereken önemli bir husus veri kaynaklarından toplanacak yetersiz miktarda verinin analiz kalitesini düşüreceği, fazla miktarda verinin ise analiz sürecini uzatacağı gerçeğidir. Bu sebeple uygun sayıda verinin varlığı analizin başarılı ilerlemesini sağlayacaktır. Ayrıca güvenilir olmayan veri kaynaklarından yapılan veri tedariki sonuçların geçerliliğini tehlikeye atabilir. Bu nedenle toplanan verilerin güvenilir kaynaklardan alınıp alınmadığı sorgulanmalıdır.

III. *Verilerin temizlenmesi ve birleştirilmesi:* Farklı kaynaklardan toplanan verilerin uyumlu hale getirildiği, aksaklıkların giderildiği ve tek veri tabanında toplandığı aşamadır. Bu aşamada kod farklılıkları, farklı ölçü birimlere sahip verilerin bulunması gibi uyuşmazlıklar giderilir. Hatalı veriler analizden çıkarılır. Uç veriler mantıksal sınamalardan geçirilerek analizden çıkarılıp çıkarılmayacağına karar verilir. Ayrıca bu süreçte titiz davranılması, verilerin düzenli ve uyumlu olmasını sağlayarak ilerideki aşamalarda sorunların yaşanmasına engel olur.

IV. *Verilerin hazırlanması:* Elimizde bulunan verilerin kalitesi analiz sonuçlarının geçerliliği açısından oldukça önemlidir. Bu sebeple verilerin bazı işlemlerden geçirilerek amaca hizmet

edecek kalitede veriler oluşturulması amaçlanır. Veri madenciliğinin en önemli aşamalarından biri olan verinin hazırlanması aşaması, analistin toplam zaman ve enerjisinin %50-%85'ini harcamasına neden olmaktadır (Piramuth, 1999).

Bu aşamada belirlenen problemlerle ilgili değişkenler seçilir ve değişkenlere uygun olarak veriler temizlenir. Bu sırada verilerde yer alan kodlama hataları, aşırı değerler ve eksik veriler düzeltilerek veya çıkarılarak analize hazır duruma getirilir.

V. Verilerin seçilmesi: Kullanılacak modele bağlı olarak veri seçimi yapılır. Örneğin, tahmin edici bir modelde bağımlı ve bağımsız değişkenler belirlenirken, sıra numarası, kimlik numarası gibi anlamsız değişkenler analize sokulmaz.

VI. Modelin kurulması ve değerlendirilmesi: Belirlenen probleme en iyi cevap verdiği düşünülen model kuruluncaya kadar yinelenen süreçtir. Analizi yapan kişi yaptığı analizin sonuçlarını yaklaşık olarak kestirebilmelidir. Model doğru kurulmadığı takdirde istenilen örüntülerin bulunamayacağı unutulmamalıdır. Ayrıca çarpıcı sonuçlar çıkması beklenen bir analizde uygulanan başarısız modeller çıktı kalitesini düşürecektir. Ek olarak kurulan modelin ana kütleyi temsil eden örneklerden kurulu olmasına dikkat edilmelidir. Ana kütleyi temsil etmekten uzaklaşan bir örneklem çıktıları kullanacak birimleri yanıltıcı sonuçlar elde etmesine yol açabilir.

Model kurulurken çok büyük miktarda verinin bulunduğu sistemlerde tüm veri tabanında birden fazla model denenerek uygun modellerin bulunması yerine, temsil gücü yüksek bir ana kütleye aynı modelleri uygulayarak sonuçları görmek, zaman tasarrufu açısından analizi yapan kişiye avantaj sağlayacaktır.

Model kuruluş süreci denetimli ve denetimsiz öğrenmenin kullanıldığı modellere göre farklılık gösterir (Giudici, 2003). Örnekten öğrenme olarak da isimlendirilen denetimli öğrenimde, bir denetçi tarafından ilgili sınıflar önceden belirlenen bir kritere göre ayrılarak her sınıf için çeşitli örnekler verilir (Göral, 2007). Sistemin amacı verilen örneklerden hareket ederek her sınıfa ilişkin özelliklerin bulunması ve bu özelliklerin kural cümleleriyle ifade edilmesidir (Akpınar, 2000).

Denetimli öğrenmede model ikiye ayrılır. Birinci kısımda model eğitilirken ikinci kısımda modelin geçerliliği test edilir.

Denetimsiz öğrenmede ise, örnekler gözlenerek aralarındaki benzerlikler tespit edilir ve benzer özelliklere sahip örnekler aynı sınıfa atanır.

VII. Modelin kullanılması: Kurulan modellerin çalıştırıldığı aşamadır. Belirlenen bir model kendi başına bir hedef çalışma olabileceği gibi, hedeflenen başka çalışmaların altyapısı olarak

da kullanılabilir. Kurulan modeller çeşitli işletme uygulamalarında doğrudan kullanılabilir. Örnek olarak kullanılan model promosyon planlaması simülasyonuna bilgi sağlayabilir veya tahmini envanter düzeyleri yeniden sipariş noktasının altına düştüğünde, otomatik olarak sipariş verilmesini sağlayacak bir uygulamanın içinde de kullanılabilir (Shearer, 2000). Kullanılan modelin zaman içinde izlenerek, meydana gelen değişikliklere ayarlanabilecek yapıda olmasını sağlamak gerekmektedir.

2.3 MODERN PERAKENDECİLİK

Modern perakendecilik kavramı günümüzde neredeyse her bireyin alışveriş yaptığı süpermarketlerin büyümesi ile oluşmuş bir terimdir. Modern perakendeciliğe göre yalnızca ürün satılmaz, ürünün yanında hizmet de satılmaktadır. Teknolojinin kullanımı, veri toplama ve veri işleme mekanizmaları ile perakendecilik sektörüne yeni bir anlam katılmasıyla beraber firmalar bir ürünü değil bütünüyle firmayı satış politikasına dâhil etmiştir. Hedef kitleye uygun mağaza dekorasyonundan kurum logosuna, ürün yerleştirmeden personel davranışına kadar bir ekol yaratma düşüncesine sahip firmalar modern perakendecilik sektörünün oluşmasını sağlamışlardır. Ayrıca ürünün dağıtım kanalları ve satış sonrası gibi hizmetlerinin de sağlandığı düşünüldüğünde modern perakendeciliğin geleneksel yöntemden ne denli farklı olduğu anlaşılmaktadır.

Türkiye'de modern perakendeciliğe geçiş aşamaları aşağıdaki gibidir.

I. 1950- 1960 yılları arasında Türkiye'de dağınık ve düzensiz perakende faaliyetleri ve devletin girişimleri sonucu açılan birkaç mağaza (Migros ve Gima gibi) vardı.
II. 1960'tan 1980 yılına kadar bazı küçük aile girişimleri (Beymen, İGS gibi) ön plana çıkmıştır.
III. 1980 yılından itibaren başlayan liberal ekonomi ve serbest dış ticaret perakendecilikte de beraberinde değişimleri getirmiştir.
IV. 1984 yılından itibaren yabancı markalar Türkiye pazarına girmeye başlamıştır.
V. 1987 yılında modern anlamda ilk alışveriş merkezi (Galleria) açılmıştır.
VI. 1993 yılından itibaren alışveriş merkezleri (Capitol, Akmerkez gibi) perakendecilik sektöründeki yerlerini almaya başlamıştır.
VII. 1996 yılında AB ile Gümrük Birliği anlaşması imzalanmıştır.

VIII. 1996'dan itibaren perakendecilikte önemli yabancı markalar Türkiye pazarına girmeye başlamıştır.

IX. 2000'li yıllardan itibaren, perakendecilikte yeni eğilimlerle birlikte modern mağazacılık anlayışı, alışveriş merkezleri ve cadde mağazacılığı başlamıştır[1].

2.4 VERİ MADENCİLİĞİ FONKSİYONLARI

Veri madenciliğinde kullanılan modeller tanımlayıcı ve tahmin edici modeller olmak üzere ikiye ayrılmaktadır.

2.4.1 Tanımlayıcı Modeller

Tanımlayıcı modellerde; amaç karar vermeyi etkileyecek olan ve veriler arasında yer alan örüntülerin tanımlanmasıdır. Örneğin, X çocuk sayısına sahip B yaş sınıfındaki biri ile Y çocuk sayısına sahip C yaş sınıfındaki birinin Z hastalığına yakalanmaları arasında benzerlik olup olmadığının araştırılması bu modele bir örnektir. *"İlişki Analizi"* ve *"Kümeleme Analizi"* tanımlayıcı modellerin ana bileşenleridir

2.4.1.1 İlişki Analizi

VM modelleri ile ilgili çalışmalarının bir kısmında *"Birliktelik Kuralı"* ve *"Kümeleme Analizi"* ayrı ayrı sınıflandırılırken, çalışmamızda ilişki analizi ana başlığı altında incelenmiştir.

Birliktelik Kuralı Analizi: Günümüzde işletmeler her dakika artan bir veri topluluğuna sahiptirler. Bu sebeple işletmeler bu veriler arasından istenilen stratejileri çıkarmak için değişkenler arasında bağlar kurmaya çalışmaktadırlar. Birliktelik kuralının temelini bu bağ ortaya çıkarmaktadır.

Birliktelik kuralı (aynı zamanda "Pazar-Sepet Analizi" olarak da bilinir), mağazaların işlem veri tabanlarından ilişkilendirmeler veya ortak oluşumlar çıkararak müşteri satın alma modellerini keşfetmenin yararlı bir yöntemidir. Analizden elde edilen bilgiler pazarlama,

---

[1] Anonim, 2008, *Pazarlama ve Perakende-Perakendeciliğin Özellikleri,* http://hbogm.meb.gov.tr/modulerprogramlar/kursprogramlari/pazarlama/moduller/perakendecilik.pdf, [Ziyaret Tarihi: 15 Mart 2013].

satış, servis ve işletme stratejilerinin oluşturulmasında kullanılabildiğinden araştırma konusu üzerine ilgi günden güne artmıştır (Chen ve diğ., 2005).

Market sepetleri, tüketicilerin alışveriş gezilerinden doğar ve birbirlerinden bağımsız olarak sıkça seçilen çoklu kategorilerdeki ürünleri içerir (Boztuğ ve Reutterer, 2008). Bu sepetlerden elde edilen veriler paket programlarda analiz edildiğinde, birliktelik olasılığı (güven oranı) ve destek oranı adı verilen kriterlerle karşılaşılır. Destek adı ile beliren bu unsur, ürünlerin birliktelik kuralının tüm veriler arasındaki yüzdesini simgelemektedir. Örneğin, yazıcı alan bir müşteri %90 olasılıkla A4 kâğıt da alacaktır ve bu birliktelik ana kütlenin %4'ü kadar desteklenmektedir. Her iki kriterin formülü aşağıdaki gibidir (Şimşek Gürsoy, 2009):

A ürün kümesinin bulunduğu bir sepette B ürün kümesinin olması durumundaki birliktelik kuralı kriterleri;

$$\text{Destek } (A \rightarrow B) = \frac{\text{A ve B ürün kümesinin bulunduğu sepet sayısı}}{\text{Toplam Sepet Sayısı}} \tag{2.1}$$

$$\text{Güven } (A \rightarrow B) = \frac{\text{A ve B ürün kümelerinin bulunduğu sepet sayısı}}{\text{A ürün kümelerinin bulunduğu Sepet Sayısı}} \tag{2.2}$$

Aşağıdaki örnek ile destek ve güven değerlerinin hesaplanması gösterilmektedir.

Tablo 2.1: Örnek Birliktelik Kuralı Sepet İçerikleri

Sepet No	Ürünler
1	Et, Süt, Yumurta
2	Peynir, Et, Süt
3	Yumurta, Et
4	Et ,Süt
5	Peynir, Zeytin, Süt, Yumurta
6	Süt, Et, Zeytin

Tablo 2.1'e göre "Süt" bulunan bir sepette "Et" bulunması durumu için destek ve güven oranı aşağıda hesaplanmıştır.

$$\text{Destek(Süt} \rightarrow \text{Et)} = \frac{4}{6} = 0{,}6667$$

$$\text{Güven(Süt} \rightarrow \text{Et)} = \frac{4}{5} = 0{,}8000$$

Yukarıdaki hesaplamalara göre destek oranı %66,67 güven oranı %80'dir. Oluşan birliktelik kuralı; et bulunan bir sepette %80 olasılıkla süt de bulunacağını ifade etmektedir. Tüm veri setinde %66,67 oranında bu birliktelik gözlenmektedir, şeklindedir.

Pazar sepet analizinin pazarlamadaki asıl amacı, satıcıya doğru karar vermede yardımcı olabilecek alıcının satın alma davranışını anlama konusunda bilgi sağlamaktır (Kaur ve Kang, 2016). Bu doğrultuda günümüzde süper marketlerde, orta ve büyük ölçekli alışveriş mağazalarındaki satış noktalarında akıllı satış sistemlerinin kullanımı oldukça yaygındır. Bu satışlardan elde edilen verilerde, işlem tarihi, satın alınan ürünlere ait bilgiler (ürün kodu, miktar, fiyat, iskonto vb.) yer alır ve ayrıca hareket numarası tekildir. Bazı kuruluşlar bu tip bilgileri içeren veri tabanlarını pazarlama alt yapılarının önemli parçalarından biri olarak görmekte ve bu verileri kullanmak için çaba harcamaktadırlar (Sever ve Oğuz, 2002).

Sepet analizi tekniği daha önceden veri tabanlarında tutulmuş olan veri yığınlarından hareketle hangi ürünlerin hangi ürünlerle beraber satıldığını, hangi ürünlerin promosyona girmesi gerektiği gibi bilgileri ortaya çıkarır (Tantuğ, 2002). Örneğin, pazarlama sektöründe bir işletmede pazar sepet analizi yapıldığında,

Yağsız peynir alan bir müşterinin %80 olasılıkla tam tahıllı ekmek alacağı,

Bulaşık deterjanı alan bir müşterinin %70 olasılıkla temizleme bezi de alacağı,

gibi sonuçların ortaya çıkması mümkündür. Bunlar dışında, bilinenden farklı örüntülerin saptanması da muhtemeldir. Örneğin, A.B.D.'de bir şirket, evli ve çocuklu erkeklerle ilgili olarak, bebek bezi ve bira arasındaki satın alma bağını keşfetmiş ve bunu pazarlama stratejisinde bir bilgi olarak kullanmıştır.

Birliktelik kuralında sağlık sektörü ile ilgili olarak, "X" ilacını kullanan bir kişide %30 olasılıkla "Y" yan etkisi görülmektedir şeklinde sonuçlar da ortaya çıkarabilir.

Pazar sepet analizi yapıldıktan sonra firmaya önemli ipuçları verir; fakat bu ipuçlarının yorumlanması için birkaç farklı yöntem vardır. Bu yöntemlerden ilki stok düzenlemedir. Firma, birliktelik kuralı doğrultusunda satışını artıracağını düşündüğü ürünü en iyi biçimde konumlandırarak çalışır. Böylelikle ürün satışı ile beraber stok azalması ve bunun sonucunda müşteri kaybını yaşamaz. İkinci yöntem raf ve reyon düzenlemesidir. İşletme yine birliktelik kuralını dikkate alarak bir arada alınan ürünleri birbiriyle ilişkili stratejik noktalara yerleştirir.

Bu sayede satışların artırılması sağlanır. Bir diğer yöntem ise fiyatlandırma ve indirimdir. Firma satışının artacağını beklediği bir ürünün fiyatında yaptığı bir indirimle rakiplerinin direncini kırma şansına sahip olabilir. Yine birliktelik kurallarına dayanan bir başka yöntem promosyondur. Satın alma alışkanlığının kazanılmasında etkin bir rolü olan promosyon birden fazla ürünün bir araya getirildiği ve tek ürünün bedelinin tüketiciye yansıtıldığı stratejidir. Ayrıca ürünün kullanımına yardımcı ürünler de ana ürünle servis edilerek satışlar artırılabilir.

Tedarikçilerle yapılan işbirliği ile birliktelik kuralı analizleri ürüne sahip tedarikçi firmalar ile paylaşılarak onların strateji geliştirmelerine olanak sağlanabilir. Örneğin A.B.D.' de Wall-Mart, 6 ülkedeki 2900 merkezinde 7,5 TB'lık veri deposuna sürekli olarak alışveriş verilerini kaydetmekte ve 3500 tedarikçisine bu verilere ulaşma ve onları analiz etme imkânı vermektedir. Bu bilgiler bölgesel merkez stoklarını kontrol etme, yeni iş fırsatlarını bulma ve tedarikçilerin perakendeciye bağlı kalmadan promosyon yapmasını sağlamaktadır (Çoban, 2006).

Paket programlarda birliktelik kuralları analizi yapılırken "0" ve "1" rakamları kullanılır. Burada "0" ürünün sepette yer almama, "1" ise yer alma durumunu gösterir. Paket programlar bu kodlamaya göre sepet birlikteliklerini hesaplarlar. Ayrıca birliktelik kuralı analizlerinin geçerliliğini test etmek amacıyla bir takım oransal hesaplamalar yapılarak kuralın anlamlılığını ortaya koyan ölçütler bulunmaktadır. Bu ölçütlerin adları "Kaldıraç" ve "İnanç" olarak tanımlanmıştır.

Kaldıraç (Lift); aralarında birliktelik kuralı bağıntısı yapılan iki ürünün veya ürün gruplarının aralarında bağ olup olmadığını göstersen oransal ölçü birimidir. Kaldırıç değeri pay kısmında değişkenlerin aynı sepette bulunma sıklığı/sepet sayısı yani destek oranı, payda kısmında ise her değişkenin ayrı ayrı destek oranlarının çarpımının oranlanması ile elde edilir. Kaldıraç değeri 1'den büyükse değişkenler arası bağlılık vardır. Fakat iki ürün birbirinden bağımsız ise pay sıfır olacaktır dolayısıyla kaldıraç sıfır değeri alacaktır. Böyle bir sonuçta, birliktelik kuralı oluşturulmaz. Kaldıraç değerinin hesaplanması aşağıdaki örnek ile açıklanabilir:

Tablo 2.2: Birliktelik Kuralı Örnek Kodlaması

SEPET NO	KLAVYE	HOPARLÖR	KULAKLIK	PİL	GÜÇ KABLOSU
1	1	0	1	1	0
2	1	1	0	0	1
3	0	0	1	1	1
4	1	0	1	0	1
5	1	0	1	1	0

Yukarıdaki tablodan yola çıkarak Klavye, Kulaklık →Pil kuralının kaldıraç değeri aşağıdaki gibi hesaplanabilir:

Kaldıraç("KLAVYE, KULAKLIK" →"PİL")

$$\text{Kaldıraç} = \frac{\text{Destek(Klavye, Kulaklık, Pil)}}{\text{Destek(Klavye, Kulaklık) x Destek(Pil)}} \tag{2.3}$$

$$= \frac{0,40}{0,60 \text{ x } 0,60} = 1,1111\ldots$$

Değer 1'den büyük olduğu için Kaldıraç oranına göre birliktelik kuralı oluşturulabilir.

Bir diğer birliktelik kuralı değerleme ölçütü ise "İnanç (Conviction)'tır". Değişkenlerin olmadığı durumdaki oransal sıklık değerini içeren yöntemdir. Bu yönteme göre oran 1 değerinden uzaklaştıkça birliktelik kuralı kuvvetlenecek yakınlaştıkça zayıflayacaktır. Eğer oran 1 değeri ise değişkenler birbirinden bağımsızdır ve birliktelik kuralı geçersiz olacaktır. Yöntemin formülü aşağıda gösterilmiştir. Aynı tablo kullanılarak İnanç değeri aşağıdaki gibi hesaplanabilir:

İnanç("KLAVYE,KULAKLIK"→ "PİL")

$$\text{İnanç} = \frac{\text{1-Destek(Pil)}}{\text{1-Güven [(Klavye, Kulaklık),Pil)]}} \tag{2.4}$$

$$= \frac{1-0,60}{1-0,67} = 1,2121$$

Sonuç "1" değerine yakındır. Birliktelik analizi çok kuvvetli olmamakla beraber kural geçerliliği desteklenmektedir.

Birliktelik kuralı yani bir başka deyişle Pazar-sepet analizi yapılırken bir takım algoritmalar kullanılır. Günümüzde en çok kullanılan ve en etkili olduğu düşünülen iki algoritma aşağıda açıklanmıştır.

1. Apriori Algoritması: Adını "prior" yani bir önceki adımdan alan algoritma iterasyon temeline dayanır. Veri tabanlarında yer alan sıklıkları keşfetmeyi amaçlar. *k*-öğeli kümeler ile *k*-1 öğeli kümelerin birleştirilmesi ile oluşur. İlk etapta bir elemanı olan kümeler oluşturmak için destek oranı hesaplanır ve *k*-1 kümesinde sıklıkla bulunan elemanlar (L_{k-1}) öğe aday kümesini oluşturmaktadır (C_k). Veri tabanı tarandığında ve C_k' ların destek değerleri hesaplanır ve L_k kümeleri oluşturulur. Süreç yinelenerek sıklıkla bulunan elemanların olduğu öğe aday kümesini oluşturamayana kadar devam eder. Bu süreçteki adım L_{k-1} kullanılarak L_k elde edilmesi işlemidir. Bu kısımda birleştirme ve budama adı verilen işlemler uygulanır ve sonuca ulaşılır.
2. ECLAT Algoritması: "Equivalence Class Transformation (eş değer sınıfsal dönüşüm)" teriminin kısaltması olarak literatüre girmiştir. Apriori algoritmasına benzerlikler gösterir, kümelerin kesiştiği noktaları birliktelik kurallarında kullanır. Apriori yöntemine göre daha küçük veri setleri için uygundur. Aksi durumda geçici kümeleme işlemi sebebiyle çok fazla yer kapladığı gibi işlem uzun zaman dilimlerini alabilir.

Ardışık Zamanlı Örüntüler: Ardışık zamanlı örüntüler birbiri ile ilişkili değişkenlerin belirli zaman aralığı içinde gerçekleşmesiyle oluşur. Bu modele birliktelik kuralının zaman boyutu eklenmiş şekli de denebilir.

Birliktelik kuralları ve ardışık zamanlı örüntüleri birbirinden ayıran özellik zaman kavramının uygulamada olmasıdır. Belli bir dönem boyunca nesneler arasındaki birlikteliklerin incelenmesi "ardışık zamanlı örüntü çözümlemesi" olarak da isimlendirilir (Goebel ve Gruenwald, 1999).

Örneğin,

Led Televizyon alan bir kişi 3 ay içinde %70 olasılıkla Netflix üyeliği de alacaktır.

BMW 5.20i araba alan bir kişi 6 ay içinde %80 olasılıkla kar lastiği alacaktır.

2.4.1.2 Kümeleme Analizi

Günümüzde VM tekniklerini kullanmak için gereken veri tabanlarının, boyutları ve veri sayıları giderek artmaktadır. Bazen bu boyutlar öyle büyük olur ki, VM teknikleri bile anlamlı sonuçlar çıkaramaz. Bu noktada, yapılması gereken karmaşık yani heterojen yapıdaki veri tabanını alt gruplara ayırmaktır. Bu yöntem kümeleme analizi olarak adlandırılır. Pazarlama alanında müşteri segmentasyonu kümeleme analizinin en yaygın kullanım alanıdır. Müşteri segmentasyon yaklaşımları başlangıçta coğrafi kriterlere dayanmaktadır, böylelikle şirketler müşterilerini ikamet veya çalışma alanlarına göre kümelere ayırabilmektedir (Miguéis, ve diğ., 2012). Yöntemlerin, müşteri algılarının ve veri ambarlarının gelişip zenginleşmesi ile kriterler gün geçtikçe yenilenmekte ve geliştirilmektedir.

Kümeleme analizinde, alt kümeler belirli bir ortak özelliğe sahip elemanlardan oluşmaktadır. Bu da homojen yapıda kümeler olmalarını sağlar. Kümeler oluşturulurken küme elemanlarının birbirlerine yakınlık derecelerine göre kümeler gruplandırılır. Bu sebeple, bir takım uzaklık ölçüm yöntemleri geliştirilmiştir.

- *Kümeleme Analizi Uzaklık Ölçüm Teknikleri:*

Farklı değişkenler ve gözlemlerden oluşan bir veri setini matris şeklinde gösterilebilir. "*n*" sayıda değişken ve "*p*" sayıda gözlemin yer aldığı "*X*" matrisi aşağıda yer almaktadır.

$$X = \begin{matrix} x_{11} & x_{12} & \ddots & x_{1n} \\ x_{21} & x_{22} & \ddots & x_{2n} \\ x_{31} & x_{32} & \ddots & x_{3n} \\ \ddots & \ddots & \ddots & \ddots \\ x_{p1} & x_{p2} & x_{p3} & x_{pn} \end{matrix}$$

Yukarıdaki matriste birinci gözlemin konumu $(x_{11}, x_{12}, x_{13}, \ldots\ldots x_{1n})$ iken ikinci gözlemin konumu $(x_{21}, x_{22}, x_{23}, \ldots\ldots x_{2n})$ şeklindedir. Bu iki gözlem arasındaki uzaklık da "$d(1,2)$" şeklinde ifade edilir. Bu uzaklıkların oluşturduğu matris de aşağıdaki gibidir (Koçoğlu ve Özcan, 2016).

$$d = \begin{matrix} 0 & & \\ D(2,1) & 0 & \\ D(3,1) & D(3,2) & \\ \ddots & \ddots & \ddots \\ D(P,1) & D(P,2) & 0 \end{matrix}$$

Gözlem verileri arasındaki farklılık veya benzerlikleri keşfedebilmek için gözlemler arası uzaklık değerini saptayacak birçok yöntem geliştirilmiştir. Bunların en önemlilerinden bir kaçı aşağıda açıklanmıştır.

Öklid Uzaklığı Yöntemi: İşlenmiş veriler yerine ham verilerle hesaplanan değerdir. Bu ölçüm yönteminde uç verilerin veri setine eklenmesi ölçümü etkilemez. Yine bu yönteme göre tüm değişkenlerin birbirinden bağımsız olduğu koşulu göz önünde bulundurulur. Ayrıca Öklid uzaklığının uygulanabilirliği değişkenlerin en azından aralıklı ölçekte ölçülmüş olması koşuluna bağlıdır. Yöntemin formülü aşağıda yer almaktadır (Çakmak, 1999).

$$d(i, j) = \sqrt{(|x_{i1} + x_{j1}|)^2 + (|x_{i2} + x_{j2}|)^2 + \ldots + (|x_{ip} + x_{jp}|)^2} \quad (2.5)$$

Manhattan Uzaklığı Yöntemi: Bu yöntemde uzaklık boyutlar arasındaki farka eşittir. Bu yöntemde uç değerlerin etkisi azdır çünkü uzaklık farklarının karesi alınmaz. Formülü aşağıda yer almaktadır.

$$d(i, j) = |x_{i1} + x_{j1}|^2 + |x_{i2} + x_{j2}|^2 + \ldots + |x_{ip} + x_{jp}|^2 \quad (2.6)$$

Formülde de görüleceği üzere gözlemler arası mutlak değer göz önüne alınarak hesaplanır. Gözlemler önem derecesine göre tasarlanmışsa formül; aşağıdaki şekli alır.

$$d(i, j) = \sum_{k=1}^{p} \left(w_k \left| x_{ik} + x_{jk} \right| \right) \quad (2.7)$$

Minkowski Uzaklığı Yöntemi: Temelleri Öklid uzaklık yöntemine benzer; fakat farkı λ ' nın farklı değerlerden hesaplanmasıdır. Minkowski uzaklığı yöntemi, Öklid uzaklığı ve Manhattan uzaklığı yöntemlerini kapsayan bir ifade olarak ele alınabilir. Minkowski uzaklığının formülü şu şekildedir (Çakmak, 1999):

$$d(x_i, j_i) = \left[\sum_{k=1}^{p} |x_{ik} - x_{jk}|^{\lambda} \right]^{\frac{1}{\lambda}} \qquad \lambda \geq 1 \text{ (durumu için)} \quad (2.8)$$

Mahalanobis Uzaklığı Yöntemi: Bu yöntem değişkenler arası korelasyonu da göz önünde bulundurarak varyans-kovaryans matrisini de analize dahil etmiştir. Bu hali ile Öklid ve Minkowski uzaklık yöntemlerinin eksik kısmını tamamlar. Değişkenler arasındaki ilişki bir diğer adı ile korelasyon sıfır ise değer Öklid uzaklığının karesine eşit olacaktır. Bu özelliğiyle

de Öklid uzaklığı yönteminden geliştirilmiş bir yöntem olduğu görülmektedir. Formülü aşağıda yer almaktadır (Akın, 2008).

$$d^2(i,j) = (x_i - x_j)^T \sum^{-1} (x_i - x_j) \quad (2.9)$$

Canbera Uzaklığı Yöntemi: Sıfıra yakınlaşan ve negatif değerlere sahip olmayan veri setlerinde başarılı sonuçlar vermekte olan bir yöntemdir. Formülü aşağıda yer almaktadır.

$$d(x_i, x_j) = \frac{\sum_{k=1}^{p} |x_{ik} - x_{jk}|}{\sum_{k=1}^{p} (x_{ik} + x_{jk})} \quad (2.10)$$

Kümeleme Analizi Algoritmaları: Kümeleme algoritmalarının veri madenciliğinde kullanılabilmeleri için, ölçülebilirlik, çok yönlülük, farklı yapıdaki kümeleri bulabilme, gereken girdi parametrelerinin minimum olması, kayıp verilerden çok etkilenmeme, verilerin sıralarının değişmesi durumuna duyarlı olmama ve çok boyutlu verilerde rahatça çalışabilme gibi özelliklere sahip olması gerekir.

Günümüzde paket programlarda kullanılan bir çok kümeleme algoritması yer almaktadır. Hiyerarşik ve hiyerarşik olmayan olarak iki ana gruba ayrılan bu algoritmalardan yoğunlukla kullanılan birkaç yöntem aşağıda açıklanmıştır.

K-Ortalamalar Yöntemi(K-Means): Veri madenciliği kümeleme algoritmaları arasında en çok kullanılan yöntemdir. Algoritmada bulunan "*k*" değeri küme sayısını temsil eder. Bu yöntem ana hatları ile "*k*" küme sayısının belirlenmesi, küme merkezlerinin bir diğer adı ile küme ortalamalarının belirlenmesi ardından her bir gözlemin küme merkezine uzaklığına göre uygun kümeye atanması ile devam etmektedir. Kümeye eklenen her yeni gözlemden sonra yeniden küme ortalaması belirlenerek süreç tekrarlanır. Önceden belirlenen "*k*" sayısı ile kullanılan bu algoritma gürültülü verilere çok duyarlıdır. Bu yöntemin avantajı olarak sayısal veri setlerinde başarılı sonuçlar vermesi gösterilebileceği gibi olumsuz yönü olarak kategorik verilerde sorunlar yaratması gösterilebilir.

k-Ortalamalar yönteminde amaç fonksiyonu olarak; hata karelerinin ortalamasını kullanılır. Bu değer; her bir elemanın atandığı küme merkezine olan uzaklığını esas almaktadır.

$$SSE = \sum_{i=1}^{k} \sum_{x \epsilon Ci} dis^2(m_i, \chi) \quad (2.11)$$

K-Medoids: "*k*" adet "Temsilci Nesne" bulma temeline dayanan bu yöntemde, temsilci nesne ile gözlemlerin arasındaki uzaklığı en küçüklemek amaçlanır. "*k*" adet temsilci nesne seçildikten sonra her gözlem yakınında olan temsilci nesneye atanarak *k* adet küme oluşturulur. Bu algoritmanın ana adımları aşağıda yer almaktadır.

- Başlangıç medoidleri rastgele seçilir.
- Kalan her nesne en yakın olduğu medoidin kümesine atanır.
- Medoid olmayan rastgele (random) bir nesne seçilir.
- *Oj* medoidi ile O random'ın değişimi için toplam maliyet, *TC* hesaplanır.
- Eğer *TC*< 0 ise *Oj* ve O rastgele değiştirilerek medoid kümesi yenilenir.
- Hiçbir değişim olmayıncaya kadar ikinci adıma geri dönülür.

PAM Algoritması: Partitioning Around Medoids kelimelerinin kısaltması olan bu algoritma tıpkı *K*-means algoritması gibi rastgele seçilen *k* adet elemanı merkez olarak alır. Kümeye eklenen her yeni gözlemi deneyerek kümenin gelişimine en verimli katkı yapacak gözlemi yeni küme merkezi yaparak işleme devam eder.

Clara Algoritması: Büyük veri tabanlarında başarısız sonuçlar verebilen *k*-medoids algoritması bu yöntemlerin geliştirilmesine sebep olmuştur. Büyük bir veri setinden seçilen pilot bir veri setine PAM algoritması uygulanarak oluşturulan yöntemdir.

Clarans Algoritması: CLARANS algoritması ise CLARA'nın sonuçlarını örneklem seçimine bağlı olmaktan kurtarmak amacı ile 1994 yılında bilim dünyasına sunulmuştur. CLARANS örneklem seçimindeki ön yargıyı gidermek için sabit bir örneklem yerine her aşamada değişen örneklem kavramını ortaya atmıştır. Rastgele seçilen noktalar çevresi dikkate alınarak örneklem oluşturulur.

Bulanık C-Ortalamalar (Fuzzy C-Means) Algoritması: k-Ortalamalar algoritmasının bazı ek parametreler ile bulanık mantık kurallarına göre düzenleniş halidir (Akpınar 2014). Bu algoritmada *k*-ortalamalar yönteminden farklı olarak değişkenler birden fazla kümeye dahil olabilmektedir. Değişkenler 0 ile 1 arasında değişen değerler alırlar. Değişkenin merkezine yakın olduğu kümenin üyelik değeri 1'e yakınlık gösterir. Ayrıca değişkenin analizdeki toplam üyelik değeri de 1 olmalıdır. Bulanık *c*-ortalamalar yönteminde amaç aşağıda yer alan fonksiyonun en küçüklenmesidir.

$J_m = \sum_{i=1}^{N}\sum_{j=1}^{C} U_{ij}^{m} \|X_i - C_j\|^2 \quad 1 \leq m < \infty$ (2.12)

U_{ij} : Üyelik değeri (U_{ij}'ler bir U üyelik matrisi oluşturur.)

X_i : *i.* gözlem

C_j : Küme merkezi

Bulanık *c*-Ortalamalar yöntemi aşağıdaki adımlardan oluşur (Koçoğlu ve Özcan, 2016).

Küme Sayısı (c), bulanıklık parametresi (m), durma kriteri (ϵ) belirlenir.

Rastgele üyelik değerleri atanır. Üyelik Matrisi oluşturulur veya rassal küme merkezleri belirlenir.

Aşağıdaki denkleme göre küme merkezleri hesaplanır.

$$C_j = \frac{\sum_{i=1}^{N} U_{ij}^{m} X_i}{\sum_{i=1}^{N} U_{ij}^{m}} \quad (2.13)$$

Aşağıdaki formüle göre yeni üyelik değerleri hazırlanır ve buna bağlı üyelik matrisleri elde edilir.

$$U_{ij} = \frac{1}{\sum_{k=1}^{c} \frac{\|X_i - C_j\|}{\|X_i - C_k\|}} \quad (2.14)$$

Yeni üyelik değerleri ve daha önce belirlenen değerler kıyaslanır. Üyelik matrisleri arasındaki fark ilk belirlenmiş ϵ değerinden küçük oluncaya kadar adımlar tekrar eder. Kümeleme analizinde küme sayısının belirlenmesi problemi en temel sorunlardan biri olarak kabul edilir. Özellikle, k-ortalamalar yöntemi gibi küme sayısının önceden belirlendiği yöntemler, kümeleme analizinin geçerliliğinin test edilmesinin önemini artmıştır. Bu sebeple kümeleme analizi geçerlilik yöntemleri geliştirilmiştir. Fakat bu yöntemler bazı veri setlerinde küme içi uyuşmazlıklar sebebiyle birbiriyle çelişen sonuçlar vermektedir. Temel olarak kümeleme analizi geçerlilik indeksleri iki temel prensibe dayanır.

Yoğunluk: Küme içindeki unsurların birbirine olan yakınlıklarıdır.

Ayrılma: Kümelerin birbirlerine olan uzaklıklarıdır.

Küme öğeleri arasındaki uzaklıkları inceledikten sonra kümeleme analizine dönülecek olursa kümeleme işlemi çoğunlukla bir başka VM uygulaması için bir ilk işlem olarak kullanılır (Tantuğ, 2002). Pazarlama sektöründe müşteri demografik yapısının bulunmasında, hedef müşteri kitlesinin saptanmasında ve müşteri profilinin belirlenmesinde yaygın olarak kullanılmaktadır.

K sayısının saptanması kümeleme analizinde önemli bir önem taşır. *K* sayısı ile ilgili sezgisel yaklaşımlarla beraber yararlanılan en pratik yollardan birisi aşağıda verilen eşitliğin kullanılmasıdır. Hiyerarşik olmayan kümelemede, küme sayısının belirlenmesine gerek yoktur. Çünkü bu teknikte küme sayısı önceden bellidir.

$$k = \left(\frac{n}{2}\right)^{\frac{1}{2}} \tag{2.15}$$

n = gözlem sayısı

k = küme sayısı

Küçük örneklemler için kullanılabilir görülen bu formül örneklem hacminin büyük olması durumunda iyi sonuçlar vermemektedir (Tezcanlar, 2007).

Kümeleme analizini pazarlama alanında özellikle tercih edilen bir analiz yöntemidir. Bu yöntemi müşteri segmentasyonu için kullanmak gerektiğinde RFM (Yakınlık, Frekans, Harcama) kavramı son yıllarda trend olan ve çokça karşılaşılan bir yöntem olarak görülmektedir. Bu sebeple; kümeleme analizinin önemli bir parçası haline gelmiştir. Çalışmanın sonraki adımlarında yapılacak kümeleme analizi içinde de yer alacak olan RFM kavramı aşağıda detaylı bir şekilde açıklanmaktadır.

RFM kavramı ilk olarak 1995 yılında ortaya konulmuştur (Bult ve Wansbeek,1995). Bu terim; Recency, Frequency ve Monetary kelimelerinin baş harflerinden oluşmaktadır. Yakınlık (Recency), müşterilerin ne kadar yakın tarihte alışveriş yaptıklarını belirtir. Yakınlık ölçütü RFM analizi için en temel etkendir. Frekans (Frequency), müşterilerin yaptıkları alışveriş sıklığını göstermektedir. Frekans, belirli bir süre içinde yapılan bir alışveriş sayısıdır ve yüksek olması daha fazla sadakat olduğunu belirtmektedir (Mohammadzadeh ve diğ., 2017). Harcama (Monetary), müşterilerin yaptıkları harcama tutarını ifade etmektedir. Birçok akademik araştırma, RFM'nin pazarlama maliyetlerini düşürdüğünü ve getirileri artırdığını

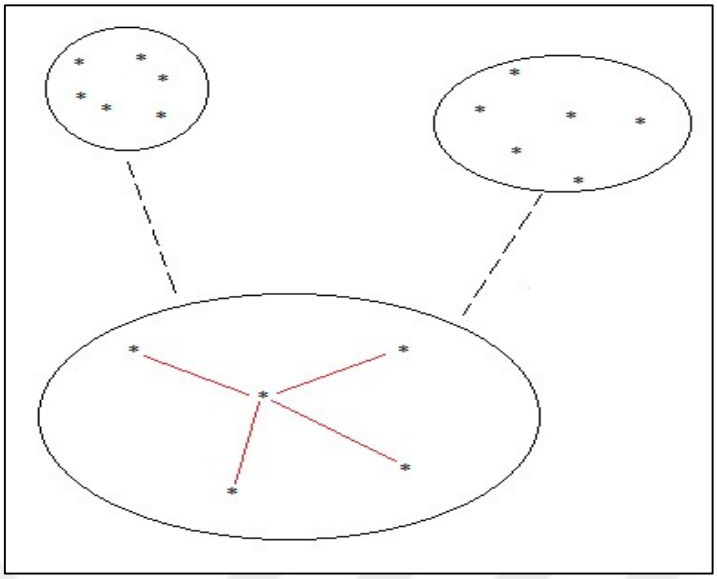
onaylamaktadır. CRM'i uygulamak için en güçlü ve en basit model RFM modeli olabilir (Khajvand ve diğ., 2011).

Bir işletmenin müşterilerinin etkin şekilde bölümlendirilmesi için, müşteriler RFM değerlerine dayalı olarak benzer davranış gruplarına ayrılır. Şirketin işlem verileri belirli bir süre boyunca analiz edilir. RFM ile yapılmış Segmentasyon, müşterilerin ihtiyacını daha iyi anlar ve şirketin potansiyel müşterilerini belirlemeye yardımcı olur (Christy ve diğ., 2018).

CRM'nin önemli ölçüm kaynaklarından biri olan RFM analizi, pazarlamada en çok bilinen ve uygulanan analiz tekniklerinden biridir. Bu yöntemin bir diğer amacı piyasaya sunulacak yeni ürün ve hizmetlere tepki verecek müşteri kitlesini saptamaktır. Bu amaç için RFM analizi, müşterilerin geçmişte yaptıkları alışveriş tarihlerinin yakınlığına, alışveriş miktarına ve harcama tutarına bakılarak müşterilerin olası satın alma trendlerini belirlemektedir. Mevcut müşteriler arasında yeni bir teklife cevap vermeye en hazır olanı hangisidir sorusuna cevap veren bir tekniktir.

Müşterileri tek bir parametreye göre değerlemek yeterli olmayacaktır. Örnek verecek olursak; en yüksek ciroya sahip müşteriler en karlı müşteriler olabilir mi? Peki sadece bir kez satın alma yapmışlarsa veya işlem çok uzun zaman önce gerçekleşmişse. Ürün artık kullanılmıyorsa ya da modası geçmiş ise. Yüksek ciroya sahip müşteriler, en karlı müşteriler olarak kabul edilebilirler mi? Bu sorulara cevap olarak geliştirilen RFM modelinde; üç farklı müşteri özelliği, müşterileri daha iyi analiz etmek için birleştirilmektedir.

RFM'i hesaplamada genellikle iki yöntem kullanılmaktadır. İlk yöntem, RFM ölçütlerine dayanarak müşteri veri tabanından verilerin toplanması, bu verilerin eşit ölçülerde gruplanması ve bunların analizidir. İkinci yöntem ise, R, F ve M için regresyon analizini kullanmak yoluyla ilişkili ağırlıklı ortalamaların (relative weights) hesaplanmasını ve bunun ardından bu ağırlıklı ortalamaların RFM'in bütünleşik etkisini hesaplamak için kullanılmasını kapsamaktadır. RFM bu durumda bir müşteri için R, F ve M'nin ağırlıklandırılmış değerlerinin bir toplamı olarak dikkate alınmaktadır. RFM modelini kapsayan kümeleme analizi ile küme sayıları belirlenerek analiz yapıldığında analizin ne derece doğru ve yararlı kümeler oluşturduğu sorusu sorulabilir.



Şekil 2.1: Homojen Değişkenlere Sahip Heterojen Kümeler.

Literatürde; oluşturulan kümeleme sonucunun küme içi homojen ve kümeler arası heterojenliklerini test eden bazı geçerlilik indeksleri geliştirilmiştir. Önemli indekslerden bazıları şunlardır (Koçoğlu ve Özcan, 2016):

Calinski-Harabasz İndeksi: Calinski ve Harabasz 1974 yılında bu indeksi geliştirmiştir. Calinski ve Harabasz indeksi aşağıdaki gibi hesaplanabilir:

$$\text{CHI}= \frac{\frac{\text{BSS(k)}}{\text{k-1}}}{\frac{\text{WSS(k)}}{\text{n-k}}} \tag{2.16}$$

Buradaki "BSS" kümeler arası uzaklık karelerinin toplamını, "WSS" ise küme içi uzaklık karelerinin toplamını ifade eder. Hesaplanma yöntemleri aşağıda verilmiştir:

$$\text{BSS(k)}=\frac{1}{2}\sum_{i=1}^{k}\sum_{\substack{i\epsilon\, C_i \\ j\notin\, C_i}} d(i,j) \tag{2.17}$$

$$\text{WSS(k)}=\frac{1}{2}\sum_{i=1}^{k}\sum_{j\epsilon\, C_i} d(i,j) \tag{2.18}$$

Burada, kümeler içi kareler toplamı ile kümeler arası kareler toplamı bulunarak CHI değerini en yüksek seviyeye çıkaran "*k*" sayısı uygun küme sayısı olarak kabul edilmektedir.

Krizanowski-Lai İndeksi: Krzanowski ve Lai 1988 yılında KLI indeksini geliştirmiştir. KLI indeksi ile küme içi uzaklık kareler toplamı (WSS) değerinin azalışını dikkate almaktadır. Bu yöntem, hata kareleri toplamı azalışının "*k*" uygun değeri ile karşılaşana kadar hızla azaldığını, uygun değerden itibaren azalmanın yavaşladığını belirtmektedir.

$$\text{KLI(k)} = (k-1)^{\frac{2}{p}}\text{WSS(k-1)-}k^{\frac{2}{p}}\text{WSS(k)} \tag{2.19}$$

$$\text{KLI}= \left| \frac{KLI(k)}{KLI(k+1)} \right| \tag{2.20}$$

Burada "*k*" değeri KLI indeksi için en uygun küme sayısını "*p*" değeri değişken sayısını ifade etmektedir.

Davies-Bouldin İndeksi: Bu indeks bir yandan küme içindeki değişkenlerin kümenin merkezine uzaklığını en küçüklemeyi amaçlarken, diğer yandan kümeler arası uzaklığı en büyüklemeyi amaçlamaktadır.

i=1,2,…,*k* ve *j*=1,2,…,*k* olmak üzere *i*. ve diğer kümeler arasındaki maksimum karşılaştırma oranı her bir küme için R_i ile gösterilen küme indeksi,

$$R_i = max_{i \neq j} \left(\frac{\delta_i + \delta_j}{d_{ij}} \right) \tag{2.21}$$

ile hesaplanmaktadır.

Burada d_{ij} küme merkezlerine olan uzaklığı δ_i ve δ_j *i.* ve *j.* kümedeki nesnelerin küme merkezine olan mesafelerini belirtmektedir. Davies – Bouldin İndeksi;

$$\text{DBI} = \frac{1}{\text{k}} \sum_{i=1}^{k} R_i \tag{2.22}$$

ile tanımlanır ve bu indeksi minimum yapan "*k*" değeri en uygun küme sayısı olarak kabul edilmektedir.

Silhouette İndeksi: Kümelenecek her bir nesnenin kümeye olan uygunluğunu tanımlamak amacıyla oluşturulmuştur.

a(*i*); *i*. nesnenin bulunduğu kümedeki nesnelere olan ortalama mesafesini

b(i); i. nesnenin bulunmadığı kümelerdeki nesnelere olan ortalama uzaklığını göstermektedir.

$$\mathrm{SI}(i) = \frac{b(i)-a(i)}{\max(a(i),b(i))} \tag{2.23}$$

Hesaplanan i. değişkene ait değer 1'e yaklaşıyorsa nesne bulunduğu kümeye uygundur. 0'a veya negatif bir sayıya yaklaşması durumunda nesnenin bulunduğu kümeye uygun olmadığı sonucu ortaya çıkmaktadır. Tüm kümeleme analizinin geçerliliği için hesaplanması gereken Silhouette İndeksi:

$$\mathrm{SI} = \frac{1}{n}\sum_{S_{i\epsilon S}} S(i) \tag{2.24}$$

şeklindedir. Buna göre maksimum değere ulaşılan küme sayısı uygun küme sayısı olmaktadır.

Dunn İndeksi: K_i ve K_j kümeleri arasındaki mesafeyi D(K_i, K_j) gösterir. "x" ve "y" sırasıyla K_i ve K_j kümelerinin elemanları ise aralarındaki minimum uzaklık:

$$\mathrm{d}(\mathrm{K}_i, \mathrm{K}_j) = \min \mathrm{d}(x, y) \rightarrow x \epsilon \mathrm{K}_i,\ y \epsilon \mathrm{K}_j \tag{2.25}$$

ile gösterilir. K kümesinin çapı da küme içinde yer alan değişkenler arasındaki maksimum uzaklık olarak tanımlanır.

$$\mathrm{Çap}(K_i) = \max \mathrm{d}(x, y) \rightarrow x,y \epsilon K_i \tag{2.26}$$

Bu açıklamalarla birlikte Dunn İndeksi;

$$\mathrm{DNI} = \min_{i=1,2,3..k} \left(\min_{j=1,2,3..k} \left(\frac{\mathrm{D}(Ki, Kj)}{max_{i=1,2,3..k}\ \mathrm{Çap}(K_i)}\right)\right) \tag{2.27}$$

şeklinde gösterilir. Dunn İndeksi daha büyük olan küme sayısı en uygun olarak kabul edilmektedir.

Xie-Beni İndeksi: Temelde Bulanık c-ortalamalar yöntemi için kullanılsa da diğer yöntemler için de kullanılabilmektedir. Bu indekste kümeler arası farklılaşma küme merkezleri arası uzaklık karelerinin toplamı minimumu iken, küme içi bütünlük ise her gözlem için bu gözlemin de bulunduğu küme merkezi arasındaki uzaklık karesinin ortalaması şeklinde yer almaktadır (Liu ve diğ., 2010).

$$\delta_1(c_k, c_{k'}) = min_{\substack{i\epsilon\ k \\ j\epsilon\iota_{k'}}} \mathrm{d}(M_i, M_j) \tag{2.29}$$

olmak üzere, denklem 2.30'da ifade edilmiştir.

$$\mathrm{XBI} = \frac{1}{N} \frac{WGSS}{min_{k<k'}\delta_1(C_k, C_{k'})^2} \tag{2.30}$$

Bu indekse göre en iyi küme sayısı indeks değerinin minimum olduğu değer olarak kabul edilmektedir.

2.4.2 Tahmin Edici Yöntemler

Sonuçları önceden bilinen verilerden hareketle yeni tahminler yapılmasını içeren modellerdir. Bu yöntemler eldeki verilerin bir kısmı ile bir model kurarak modelden çıkan sonuçların, sonuçları bilinmeyen veri kümesine uygulanması ve değerlerin tahmin edilmesini kapsamaktadır.

2.4.2.1. Sınıflandırma

Denetimli öğrenmeyi esas alan bir veri madenciliği fonksiyonudur. Bu fonksiyonda, kesikli değere sahip bir sınıf etiketinin tahmini amaçlanmaktadır. Dolandırıcılık tespiti, ayrılan müşteri analizi gibi problemler önemli sınıflandırma problemleri olarak sayılabilir. Karar ağaçları, yapay sinir ağları, genetik algoritmalar, bayes ağları, *k*-en yakın komşu algoritması gibi algoritmalar sınıflandırma problemlerinin çözümünde yaygın olarak kullanılan tekniklerdir. Sınıflandırma problemlerinin çözümünde veri kümesi eğitim ve test verisi olarak ikiye ayrılmaktadır. Eğitim verisi üzerinde sınıflandırma modeli oluşturulmakta, oluşturulan modelin başarımı test verisi üzerinde doğruluk, kesinlik, anma, F-ölçüsü gibi farklı ölçütler kullanılarak denetlenmektedir.

Karar Ağaçları: Karar ağaçları sınıflandırma problemlerinin çözümünde sıklıkla kullanılan bir analizdir. Adını, belirlenen problemin ağaç köküne, ilişkili olayların olasılıklarının gövdeye ve kararların da çoğalan dallara benzemesinden almaktadır. Karar ağaçlarına verilecek en klasik örnek tenis maçı oynama/oynamama kararı Tablo 2.3'de gösterilmektedir.

Tablo 2.3: Karar Ağaçları İle İlgili Örnek Veri Seti (Altıntaş, 2006)

HAVA	SICAKLIK	NEM	RÜZGAR	OYUN
Güneşli	Sıcak	Yüksek	Rüzgarsız	Hayır
Güneşli	Sıcak	Yüksek	Rüzgârlı	Hayır
Bulutlu	Sıcak	Yüksek	Rüzgarsız	Evet
Yağmurlu	Ilıman	Yüksek	Rüzgarsız	Evet
Yağmurlu	Soğuk	Normal	Rüzgarsız	Evet
Yağmurlu	Soğuk	Normal	Rüzgârlı	Hayır
Bulutlu	Soğuk	Normal	Rüzgârlı	Evet
Güneşli	Ilıman	Yüksek	Rüzgarsız	Hayır
Güneşli	Soğuk	Normal	Rüzgarsız	Evet
Yağmurlu	Ilıman	Normal	Rüzgarsız	Evet
Güneşli	Ilıman	Normal	Rüzgârlı	Evet
Bulutlu	Ilıman	Yüksek	Rüzgârlı	Evet
Bulutlu	Sıcak	Normal	Rüzgarsız	Evet
Yağmurlu	Ilıman	Yüksek	Rüzgârlı	Hayır

Tabloda da görüldüğü üzere karar ağaçları değişkenlerin birbirleri ile ilişkileri sonucu oluşan tüm sonuçları göstermektedir. Sıklıkla kullanılan bu yöntem her faktörü dikkate almakta ve olası sonuçların kolaylıkla bulunmasını sağlamaktadır. Karar ağaçları;

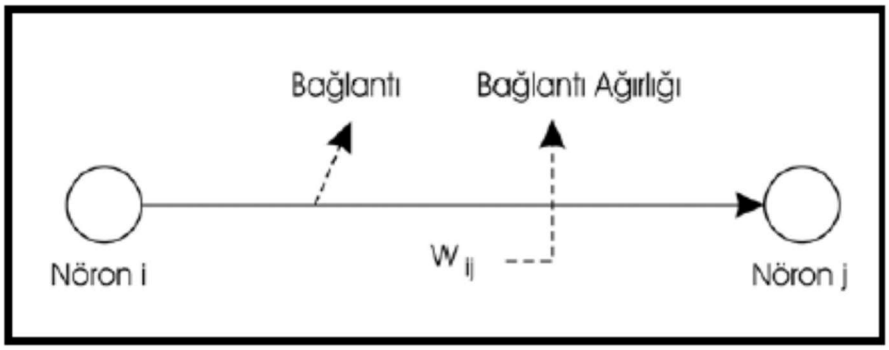
- Kuruluşlarının ucuz olması,
- Yorumlanmalarının kolay olması,
- Veri tabanı sistemleri ile kolayca entegre edilebilmeleri,
- Güvenilirliklerinin daha iyi olması

gibi faktörler ile sınıflandırma modelleri içerisinde en yaygın kullanıma sahiptir.

Perakendecilik sektöründe çapraz satışların artırılması, bankacılıkta kredi puanının belirlenmesi, insan kaynaklarında çalışanların profilinin belirlenmesi ve üretim bandında hataya yol açan değişkenlerin belirlenmesi gibi konularda karar ağaçları yaygın olarak kullanılır (Şimşek Gürsoy, 2009).

Yapay Sinir Ağları: İnsan beyninin işleyiş yapısından esinlenerek oluşturulan modeldir. Sinir hücreleri (nöronlar) çeşitli şekillerde birbirine bağlanarak ağı oluştururlar. İlişkiler kurarak otomatik olarak öğrenmeyi sağlayan bu ağ nöronlardan sinyali alır, birleştirir, dönüştürür ve sayısal sonuç olarak sunar. Şekil 2.1'de gösterildiği gibi her yapay sinir ağında girdi alan (Nöron *j*) ve çıktı veren (Nöron *i*) nöronlar bulunmaktadır. Ek olarak, bu modelde nöronlar

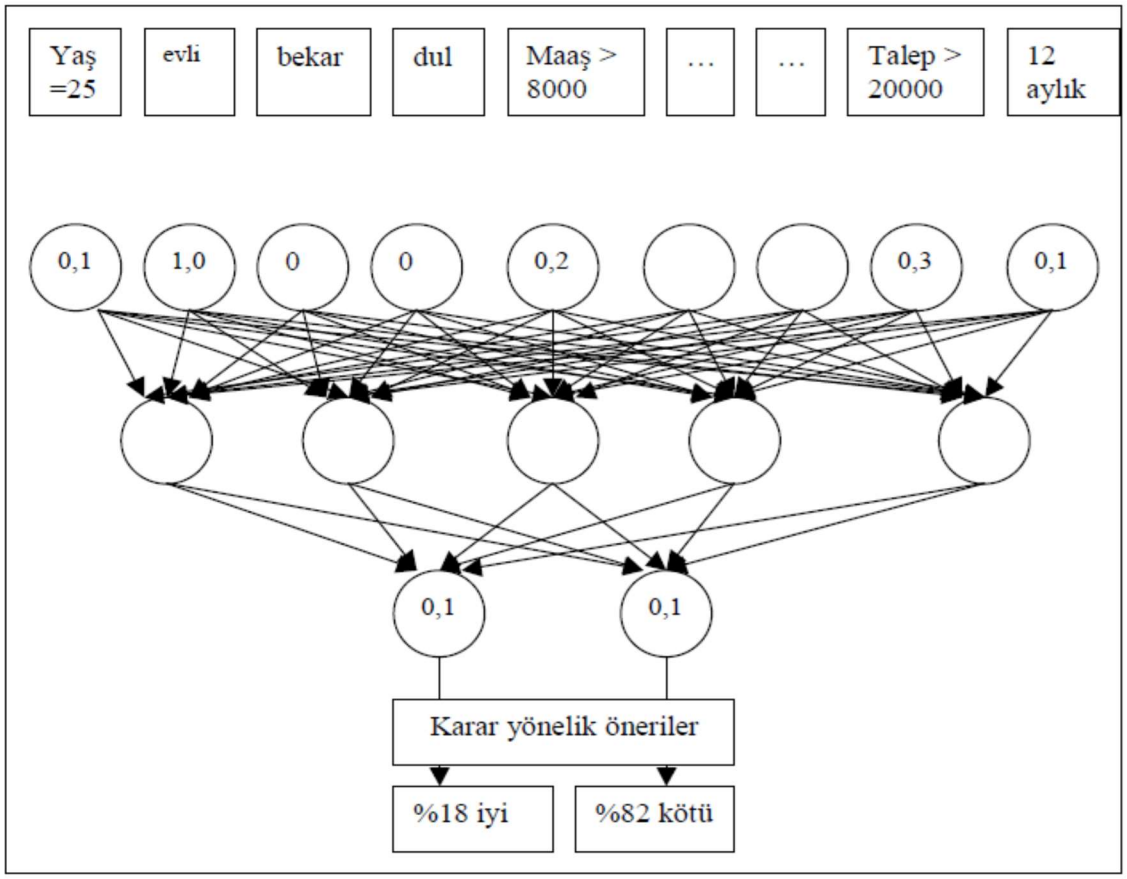
arasındaki bağlantı ve bu bağlantının sağlamlığını gösteren bir ağırlık katsayısı(w_{ij}) da bulunmaktadır.



Şekil 2.2: Yapay Sinir Ağı Mimarisinin Temel Elemanları (Tezcanlar, 2007)

Sinir ağları algoritmaları sayısal veriler üzerinde çalışırlar, denetimli öğrenme yöntemidirler (Akbulut, 2006). Geleneksel yöntemlerin derinliğinin yeterli olmadığı problemlerin çözümlenmesinde kullanılmaktadır. Diğer yöntemlere göre daha zaman alıcı bir yöntemdir. Kesin sonuçlar ortaya çıkarmaz ve üretilen sonuçların açıklaması bulunmaz. Sadece ham veriyle ve iyi bir test sürecinden sonra, ortaya tahminleme yapabilecek durumda olan bir analiz programı çıkartmaktadırlar (Han ve Kamber, 2001).

Yapay sinir ağları tıp sektöründe, ekonomide kredi başvurusu değerlendirmede, hilekârlıkların tespitinde, savunma sanayinde, üretim kalite kontrolünde sıklıkla kullanılan bir yöntemdir. Yapay sinir ağlarının bir sorunla ilgili karar verme durumuna etkisini gösteren örnek Şekil 2.3'te gösterilmiştir.



Şekil 2.3: Gelişmiş Bir Yapay Sinir Ağı Örneği (Tok, 2002)

Genetik Algoritmalar: Endüstri alanında klasik programlama ve yöneylem araştırması teknikleri ile geliştirilen programlar yerini artık yapay zekâ teknikleri kullanılarak geliştirilen çalışmalara bırakmaktadır. Böylece planlanan, üretimi artırmak ve kârı en büyüklemek için; sezgisel parametreleri kullanma, doğru analiz yapabilme ve anında karar verme gibi insana özgü olan yetileri kullanarak karar veren veya tavsiyelerde bulunan sistemlerin geliştirilmesi ile daha hızlı ve gerçekçi çözümler elde edilmektedir (İşçi, 2007). Genetik algoritmalar temelini çevreye uyum sağlayan canlının hayatta kalması felsefesinden almaktadır. Çok zor ve karmaşık problemlerin çözümünde kullanılan genetik algoritmaların açıklanabilir çözümleri vardır. Ayrıca yapay sinir ağları ile birlikte çalışarak başarılı sonuçlar verebilir. Optimizasyonda kullanılabilen bu yöntem veri grubundan özel bir veri kümesinin saptanmasını sağlamaktadır.

2.4.2.2. İstatistiksel Tahmin Yöntemleri

VM uygulamalarında yaygın olarak kullanılan istatistik temelli teknikler regresyon analizi, diskriminant analizi ve lojistik regresyon analizidir.

Regresyon Analizi: Herhangi bir değişkenin bir veya birden fazla değişken ile arasındaki ilişkinin matematiksel temsili olarak tanımlanabilir. Regresyon denklemlerinde bağımlı değişken ile bağımsız değişken(ler) arasındaki bağın kuvvetli olup olmadığını gösteren bir katsayı bulunmaktadır. Bu katsayıya korelasyon katsayısı denir ve "*R*" ile temsil edilir. -1 ve +1 arasında bir değer alan korelasyon katsayısı 0'a yaklaştıkça değişkenler arasındaki ilişkinin zayıfladığı, 1'e yaklaştıkça bu ilişkinin pozitif veya -1'e yaklaştığında negatif yönde olduğu anlaşılmaktadır. Değerler arasındaki ilişki doğrusal olduğunda doğrusal regresyon modeli, doğrusal olmadığında ise doğrusal olmayan regresyon modeli adını almaktadır. Tek değişkenin olduğu bir regresyon modeli basit doğrusal regresyon analizi adını alırken birden fazla değişken olması durumunda çoklu doğrusal regresyon analizi adını almaktadır.

Diskriminant Analizi: Diskriminant analizi iki veya daha fazla sayıda grubun aralarındaki farkları ortaya koyan bir analiz yöntemidir. Analiz iki gruplu ve çok gruplu olmak üzere ikiye ayrılmaktadır. Diskrinminant analizinin kullanım amaçları gruplar arası farkı en iyi belirleyen değişkeni saptamak, analize dâhil edilen bağımsız değişkenler için gruplar arası bir fark olup olmadığını tespit etmek ve gruplara alınacak yeni gözlemlerin hangi gruba alınacağını saptamaktır. Diskriminant analizi iki veya daha fazla grubu en iyi ayırt eden fonksiyonu üretmektedir.

Diskriminant analizi yapılmadan önce sahip olduğu varsayımlar kontrol edilmelidir. Bu yöntemin varsayımları;

Grupların ayrık olup tanımlanabilir özellikte olduğu

Değişkenler arasında çoklu doğrusal bağlantı olmadığı

Bağımsız değişkenlerin çok boyutlu normal dağılıma sahip olduğu

Gruplar arası varyans-kovaryans matrislerinin eşit olduğudur.

Varsayımlar sağlanamazsa değişkenlere gerekli dönüşümler uygulanarak varsayımların geçerliliği tekrar test edilmelidir (Şimşek Gürsoy, 2009).

Lojistik Regresyon Analizi: Lojistik regresyon analizi diskriminant analizinde bulunan varsayımlardan çoklu normallik ve varyans-kovaryans matrislerinin eşitliğinin sağlanamadığı durumlarda sonuç üretebilmek için alternatif olarak geliştirilmiş bir yöntemdir.

Lojistik regresyon analizinde bağımlı değişken kategorik iken bağımsız değişkenler kategorik olmayabilir. Lojistik regresyon analizinin geçerli ve güvenilir olması için gerekli varsayımlar;

0-1 kategorik değişkenlerinin bulunduğu bir analizde, herhangi bir gözlemin 1 ile kodlanan gruba ait olma olasılığı P_1 ise, 0 ile kodlanan gruba ait olma olasılığı $P_0 = 1- P_1$ olmalıdır.

Gözlemler birbirinden bağımsız olmalıdır.

Model bağımlı değişkeni açıklayacak tüm bağımsız değişkenleri içermeli ve gözlem sayısı büyük tutulmalıdır (Şimşek Gürsoy, 2009).

2.5 **LİTERATÜR TARAMASI**

Veri madenciliği, her geçen yıl önemi ile birlikte popülaritesini de artıran bir değere sahiptir. Bu bölümde veri madenciliği uygulamaları ile ilgili literatür çalışmaları detaylandırılacaktır.

Çoban (2006) üretim sektöründe VM çalışmalarının az oluşuna değinmiş ve veri madenciliği yöntemleri ile tedarikçi seçimi uygulamasını SPSS Clementine 9.0 programını kullanarak ortaya koymuştur. Düzenlenmiş veri seti, yapay sinir ağları ve karar ağacı algoritmaları ile çalıştırılmış ve tedarikçi firmaların başarı düzeyleri gecikme ve zamanında teslim kodlamaları ile oluşturmuştur. Sonuç olarak kalite belgesi olan firmaların olmayanlara oranla daha başarılı olduğu VM metotlarıyla ortaya çıkmıştır.

Altıntaş (2006) banka müşterilerinin kredi ödeme durumlarına göre kümeleme analizleri yapmış ve *"K-ortalamalar"* algoritmasını kullanmıştır. *"Weka"* paket programı kullanılarak yürütülen çalışmada müşterileri üç ana gruba ayrılmış olup, aylık gelirler ile kredilerin düzenli ödenmesi arasında oldukça kuvvetli bir bağ olduğu ortaya çıkmıştır.

Aggarwal ve diğ. (2012) VM ile ilgili paket programların bazılarını incelenmiş ve birbirleri arasında karşılaştırmalar yapılmıştır. Sonuç olarak, hiçbir paket programın tüm VM yöntemlerine cevap veremediği sonucu ortaya çıkmıştır.

Jiménez ve diğ. (2018) VM kullanımına örnek olarak, Mallorca kentinde bir okulun 14-18 yaş öğrencileri arasından 9300 kişilik bir grup arasında seçilmiş örneklemden uyuşturucu kullanma nedenlerini araştırmıştır. Karar ağaçları, yapay sinir ağları ve lojistik regresyon kullanılarak yapılan analizlerde öğrencilerin %57'sinin alkol kullandığı %18'inin tütün ürünleri kullandığı %19'unun kenevir ve %1,5'inin kokain kullandığı sonuçlarına ulaşılmıştır. Yine aynı yöntemler ile uyuşturucu kullananların hangi sebeplerle uyuşturucu kullandıkları ile ilgili saptamalar yapılmıştır. Sonuç olarak sorunları unutmak, arkadaş tavsiyesi ve yeni heyecanlar yaşama isteği sonuçlarına ulaşılmıştır.

VM tekniklerinin pazarlama sektöründeki çalışmalarına bakacak olursak ilk olarak, Kütahya ilinde yapılan bir çalışmada tüketici profili belirlenmek istenmiş ve veri setinden örneklem seçilerek anket çalışması yapılmıştır. 4 temel grup belirlenmiş ve 3 grubun markaya yönelmeyen bir grubun ise marka tercihini önemli bulduğu saptaması yapılmıştır (Yılmaz, 2006).

Chen ve diğ. (2005) çoklu depolama sistemlerine sahip işletmelerde birliktelik kuralı analizi üzerine bir araştırma yapılmıştır. Çoklu depo ortamlarının zamansallık ve coğrafi değişkenleri göz ardı eden analizler ile incelenmesi eleştirilmiş, bu durumun dezavantajlarına değinilmiştir. Bu sebeple "Apriori-like" adında bir yöntem geliştirerek birliktelik kurallarına konum ve zaman bilgisi eklenmiştir. Sonuç olarak önerilen yöntemin satış mağazaları büyük, ürün sirkülasyonunun fazla ve dönemselliğin artış gösterdiği durumlarda geleneksel birliktelik kuralı analizlerinden daha çarpıcı bilgiler verdiği sonucuna ulaşılmıştır.

CRM ve Veri madenciliğini bir araya getiren Tezcanlar(2007)'ın çalışmasında MS Analysis Server 2000 programı ile kümeleme ve karar ağaçları analizleri yapılmıştır. Madeni yağ üreticisi bir firmanın 14 aylık satış bilgileri kullanılarak yapılan analiz sadık müşteri kitlesini keşfetmiş alışveriş ölçeği küçük işletmeler için firmanın yaptığı satış artırma politikaların anlamsız olduğu sonucunu ortaya çıkarmıştır.

Perakendecilik sektöründe de veri madenciliği uygulamaları ile ilgili birçok çalışma görülmektedir. Adana'da bir süpermarketin 8 günlük verileri temel alınarak oluşturulan

Dolgun (2006)'un çalışmasında *"SAS 8.2"* paket programı kullanılmış olup, yapılan gruplandırmada yeşillik alanların sebze, soğan alanların yeşillik, bisküvi alanların çikolata aldığı gibi kurallar keşfedilmiştir. Bir diğer pazar sepet analizinin yapıldığı çalışmada Döşlü (2008) süpermarket verileri üzerinde birliktelik kuralları uygulaması gerçekleştirmiştir.

Tang ve diğ. (2008) çoklu mağaza ve depolama sistemi olan ve market zincirlerine benzer işletmelerde, pazar-sepet analizi yapmak için yeni bir yaklaşım önerilmiştir. Çalışmada ortam ve dönemsellik özelliği ile zaman dilimlerine ve lokasyona göre birliktelik kuralları belirlenmesi hedeflenmiştir. Zaman ve mekân hiyerarşileri tanımlanarak türetilen tüm kurallar için destek ve güven oranları hesaplanmıştır. Sonuç olarak önerilen yöntem sayısal değerler ile kullanıldığında, yöntemin diğer pazar sepet analizlerinden daha spesifik ve daha zengin bilgilere sahip olduğu ortaya koyulmuştur.

Boztuğ ve Reutterer (2008) çalışmalarında müşterilerin bölümlendirilerek birliktelik kuralı analizi yapılmasının daha olumlu sonuçlar verdiğini öne sürmüştür. Öncelikle genel müşteri verilerinden yola çıkılarak yapılan pazar-sepet analizi ile müşteriler bölümlendirildikten sonra yapılan analiz arasında çarpıcı farklar ortaya çıkmıştır. Çalışmada, bu sonuçların müşteri bağımlılığını artırmak adına kullanılabileceğini ortaya konulmuştur. Ayrıca çalışmada her müşteri kümesi için ortaya çıkan analizler sonucu ortaya çıkan birliktelik kurallarına uygun stratejiler geliştirilmesi gerektiği vurgulanmaktadır.

Pazar-sepet analizini konu alan Kaur ve Kang (2016) çalışmasında ise birliktelik kurallarının değişen müşteri alışkanlıklarına uyarlanması amaçlanarak periyodik veri madenciliği terimi kullanılmıştır. Uç verilerin fark edilebilmesi için eşik değerlerinin otomatik olarak saptanması gerektiğinden bahsetmektedir. Çalışmanın gelecek çalışmalarda yalnızca pazar sepet analizi verilerine değil daha farklı amaçlarla da kullanılması tavsiye edilmiştir.

Çalışmanın bir diğer bölümünü oluşturan kümeleme analizi ile ilgili literatürde yer alan çalışmalardan biri Miguéis ve diğ. (2012) tarafından gerçekleştirilmiştir. Bu çalışmada son yıllarda özellikle Avrupa'da perakende sektörü ile müşteriler arasındaki ilişkinin boyut değiştirdiği belirtilmektedir. Çalışmada müşterilerin tutundurulması ve müşteri sadakatinin önemi vurgulanmaktadır; fakat işletmelerin veri toplama yatırımlarına verdikleri önemin henüz ileri seviyede olmaması sebebiyle genelde müşteriler bir bütün olarak görülüp müşteri odaklı çalışmalar yapmanın işletmeler tarafından daha çok tercih edildiği belirtilmiştir.

Çalışma müşterilerin bir bütün olarak görülmesi yerine kümelenerek bu stratejilerin belirlenmesinin gün geçtikçe artacağına vurgu yapmaktadır. Çalışmanın uygulama kısmında ise VARCLUS algoritması ile bir süpermarket firmasının ürünleri müşterilerin yaşam tarzlarına göre sınıflandırılmış ve ürünler hakkında tanıtımlar yapılmıştır. Elde edilen sonuçlar, bir takım reklam ve promosyon çalışmaları ve raf düzenlemeleri yapılmasına olanak sağlanmıştır

Hamka ve diğ. (2014) çalışması ise teknolojinin hızla ilerlediği günümüzde akıllı telefon kullanımının artması ile operatör şirketlerinin müşterilerini nasıl segmentlere ayırmaları gerektiği ile ilgili bir çalışma yapılmıştır. Müşterilerin gözlenen davranışları ile ilgili bir segmentasyon yapılması bunların demografik ve psikografik verilere bağlanması amaçlanmıştır. Veri toplama yöntemi olarak müşteri davranışlarını saptama amaçlı olan anketin yanı sıra LOG adı verilen bir uygulama ile veri toplanması ve daha büyük bir veri seti ile çalışılması amaçlanmıştır. Özellikle müşteri davranışları ile ilgili yapılan anketlerin cevaplanmama oranı yüksekliği, örneklemin sınırlı olması ve boyutu küçük bir örneklemin içindeki uç verilerin sonucu çarpıtabileceği belirtilerek, gelecek çalışmalar için bir model önerilmiştir.

Müller ve Hamm (2014) kümeleme çalışmalarında veri setlerindeki zaman kavramının analize etkilerini araştırmıştır. Bu çalışma; zaman değişkenliğinin daha stabil olduğu durumlarda kümeleme analizinde daha verimli sonuçlar alındığını ortaya atan metodolojik bir çalışma olma özelliği taşımaktadır. Analize eleştirel yaklaşan bu çalışmada müşteri tutumlarının teknolojik gelişmelerin veya birçok sebebin zamansal olarak değişerek analizi zorlaştırdığını, mümkün olduğunca içte homojen diğer kümelerle heterojen bir kümeleme oluşturulması gerektiği belirtilmiştir. Sonuç olarak, pazarlama stratejilerinin belirlenen kümelere esnek ve adaptasyon amaçlı sunulması gerektiği keşfedilmiştir.

Akamavi ve diğ. (2015) yolculukları aksamış havayolu müşterilerinin sadakati incelenmiş ve yolcularla iletişime geçen kişilerin kalitesinin sadakati artırmada önemine vurgu yapılmıştır.

İnsani ve Soemitro (2016) telekomünikasyon sektöründe yer alan bir şirkette müşteri segmentasyonu RFM yaklaşımı ile SPSS Modeler Paket Programı ve k-Ortalamalar algoritması kullanılarak oluşturulmuştur. Aynı zamanda müşterilerin kullandığı bazı elektronik ürünler ile ilgili birliktelik kuralı analizi de yapılmıştır.

Özçalıcı (2017) çalışmasında otomotiv sektöründe araçlar ve donanımları ile ilgili apriori algoritması kullanılarak birliktelik kuralı analizi yapılmıştır. Sonuçlara göre fiyat aralıklarına göre hangi donanımların eklendiği ve hangi donanımların birlikte tercih edildiği gibi sonuçlar ortaya çıkarılmıştır.

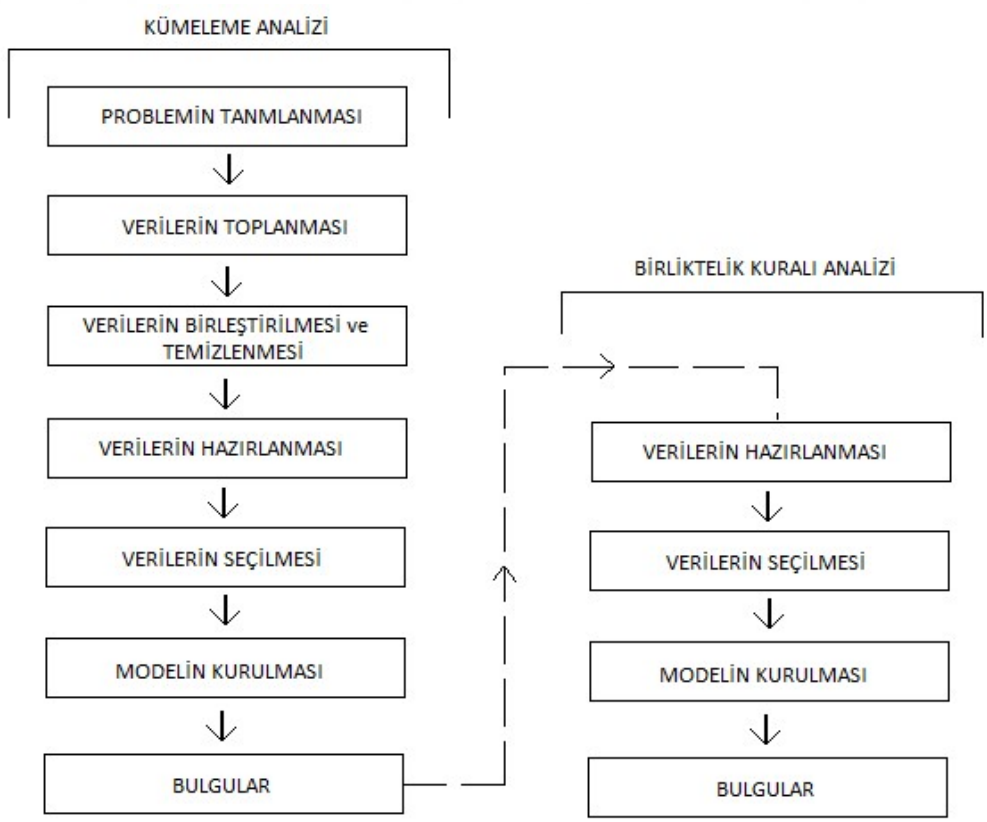
Calvo-Porral ve Lévy-Mangin (2018) çalışmalarında özel gıda mağazaları için müşteri segmentasyonunu incelemiş, bu tür benzer mağazalarının popülerliğinin giderek arttığı, bulundurdukları bazı ürünlerin birçok markette bulunmadığı belirtilmiştir. Yeni oluştuğu öne sürülen bu sektörde müşterilerin kümelere ayrıştırılabilme olanağının olup olmadığı araştırılmış, bu haliyle gelecek çalışmalar için öncü bir çalışma temeli oluşturmuştur. Çalışmanın sonucunda özel gıda mağazaları müşterilerinin tek bir grup olarak görülmemesi gerektiği öne sürülerek müşteriler dört ana grupta toplanmıştır. Çalışma gelecek çalışmalar için yol gösterici olup farklı ülkelerde daha farklı sonuçlar çıkabileceğini ve perakende sektörü için genelleme yapılamayacağını vurgulamaktadır.

Son olarak, Quadadeh ve Abdallah (2018) bir sigorta şirketinin verilerini kullanarak müşteri segmentasyonu çalışması gerçekleştirmiştir. Yüksek boyutlu bir veri kümesinin görselleştirilmesi ve müşteri tanımlamalarının yapılması adına yapılmış olan çalışmanın verileri Hollanda menşeili olup yaklaşık on bine yakın müşteri kaydı ve altı bine yakın eğitim kaydı içermektedir. Her kayıt müşterilerin demografik kayıtlarıyla beraber, müşterilerin satın alma davranışlarını da içermektedir. *K*-ortalamalar yöntemi ve SOM ile bütünleştirilmiş *K*-ortalamalar yöntemini kullanarak kümeleme analizi yapılmıştır. SOM ile kullanılan *K*-ortalamalar yönteminin verdiği sonuçların, yalnızca *K*-ortalamalar yöntemi ile bulunanların aksine, kümeleri daha belirgin oluşturduğu ve daha hızlı sonuç verdiği öne sürülmüştür.

Literatür araştırmalarının gösterdiği gibi birçok yazılım yardımıyla, farklı sektörlerde ve çok sayıda problemin çözümü için VM teknikleri kullanılabilir. Ancak, özellikle Dünya'da ve Türkiye'de satış stratejilerinin geri dönüşünün en çabuk alındığı sektör olan perakendecilik sektörü için VM uygulamalarının önemi oldukça fazladır. Dolayısı ile yapılan çalışmanın ilgi odağı perakendecilik sektörü olmuştur. VM metotlarının birbirinin devamı olacak şekilde uygulanabileceğini ve her metodun birbiriyle ilişkili olduğunu gösterebilmek adına hazırlanan bu çalışmada kümeleme ve birliktelik kuralları madenciliği ile ilgili uygulama gerçekleştirilmiştir.

3. MALZEME VE YÖNTEM

Bu çalışmada; kümeleme analizi ve birliktelik kuralları madenciliği ile ilgili bir uygulama gerçekleştirilmiştir. Bu çalışmada, önerilen modelin süreç adımları VM süreci şeklinde işleyecek, kümeleme analizinden elde edilen hedef müşteri grupları için birliktelik kuralı madenciliği çalışması gerçekleştirilecektir. Şekil 3.1'de çalışmada kullanılacak VM süreç adımları yer almaktadır.



Şekil 3.1: Önerilen Modelin Süreç Adımları

3.1 PROBLEMİN TANIMLANMASI

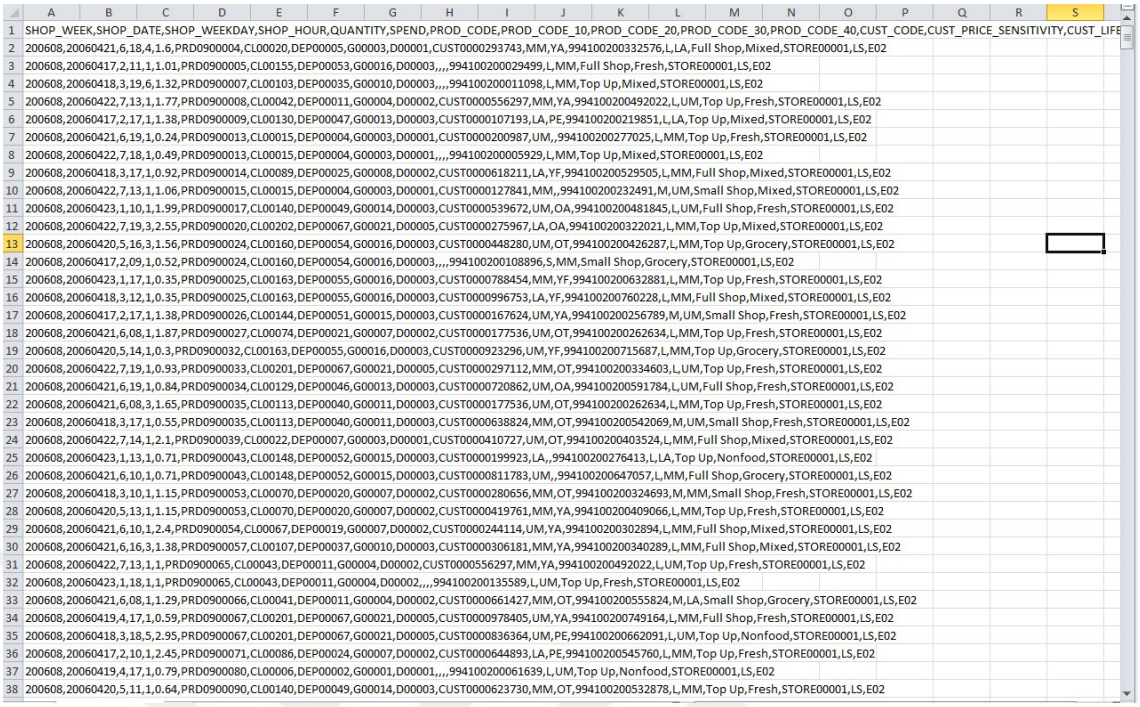
Perakendecilik sektörü gibi rekabetin yüksek olduğu bir sektörde gelişmiş bir veri ambarına sahip olmak firmaları bir adım öne geçirebilir. Bu veri tabanlarının doğru işlenmesi ile elde edilecek sonuçlar işletme için pozitif etkilere sahip olacak stratejilere kaynak olursa başarı

kaçınılmazdır. Bu sebeple, perakendecilik sektöründe müşteri ve sepet kayıtlarının olduğu veri havuzundan hareket ederek hedef müşteri kitlesini belirlemek ve bu kitlenin satın alma kayıtlarından oluşan bir analiz oluşturarak sonuçları müşteri sadakatini sağlamak ve ciro artırımına taşımak, odaklanılan temel problem olmuştur. Bu doğrultuda analiz yapmayı sağlayacak düzenlemeler ve farklı VMUT kullanılarak ana amaca yönelik adımlar gerçekleştirilmektedir. Hedeflerin açıkça belirlenmesi VM sürecinde kritik bir noktada var olmaktadır. Bu çalışmada; ilk olarak hedef müşteri kitlesini belirlemek adına k-ortalamalar ve bulanık c-ortalamalar algoritmaları kullanılarak RFM esaslı müşteri segmentasyonu sağlanacaktır. Daha sonra, hedef müşteri kümelerinin müşteri işlem verileri kullanılarak Apriori algoritması ile birliktelik kuralı analizi yapılacaktır. Bu analizle müşterilerin sepetlerdeki satın alma alışverişleri arasındaki çarpıcı bağların ışığında işletme için yol gösterici fikirler ortaya koyulacaktır.

3.2 VERİLERİN TOPLANMASI

Bu tez çalışmasında kullanılan veri, bir müşteri veri bilimi platformu olan www.dunnhumby.com web sitesinden elde edilmiştir. Bu veri seti, perakendecilik sektöründeki kalıpların tipik bir çoğaltmasından oluşmakta olup, ".csv" uzantılıdır. Söz konusu veri seti, 223 güne ait, 854438 alışveriş kaydından ve 154550 farklı müşteri işlem verisinden oluşmaktadır[2]. Veri setinin genel görünümü Şekil 3.2'de gösterilmektedir.

[2] *Let's Get Sort of Real: Dummy Data to Test Techniques and Algorithms*, https://www.dunnhumby.com/careers/engineering/sourcefiles, [Ziyaret Tarihi: 17.03.2019].



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	SHOP_WEEK,SHOP_DATE,SHOP_WEEKDAY,SHOP_HOUR,QUANTITY,SPEND,PROD_CODE,PROD_CODE_10,PROD_CODE_20,PROD_CODE_30,PROD_CODE_40,CUST_CODE,CUST_PRICE_SENSITIVITY,CUST_LIFE																		
2	200608,20060421,6,18,4,1.6,PRD0900004,CL00020,DEP00005,G00003,D00001,CUST0000293743,MM,YA,994100200332576,L,LA,Full Shop,Mixed,STORE00001,LS,E02																		
3	200608,20060417,2,11,1,1.01,PRD0900005,CL00155,DEP00053,G00016,D00003,,,,994100200029499,L,MM,Full Shop,Fresh,STORE00001,LS,E02																		
4	200608,20060418,3,19,6,1.32,PRD0900007,CL00103,DEP00035,G00010,D00003,,,,994100200011098,L,MM,Top Up,Mixed,STORE00001,LS,E02																		
5	200608,20060422,7,13,1,1.77,PRD0900008,CL00042,DEP00011,G00004,D00002,CUST0000556297,MM,YA,994100200492022,L,UM,Top Up,Fresh,STORE00001,LS,E02																		
6	200608,20060417,2,17,1,1.38,PRD0900009,CL00130,DEP00047,G00013,D00003,CUST0000107193,LA,PE,994100200219851,L,LA,Top Up,Mixed,STORE00001,LS,E02																		
7	200608,20060421,6,19,1,0.24,PRD0900013,CL00015,DEP00004,G00003,D00001,CUST0000200987,UM,,994100200277025,L,MM,Top Up,Fresh,STORE00001,LS,E02																		
8	200608,20060422,7,18,1,0.49,PRD0900013,CL00015,DEP00004,G00003,D00001,,,,994100200005929,L,MM,Top Up,Mixed,STORE00001,LS,E02																		
9	200608,20060418,3,17,1,0.92,PRD0900014,CL00089,DEP00025,G00008,D00002,CUST0000618211,LA,YF,994100200529505,L,MM,Full Shop,Mixed,STORE00001,LS,E02																		
10	200608,20060422,7,13,1,1.06,PRD0900015,CL00015,DEP00004,G00003,D00001,CUST0000127841,MM,,994100200232491,M,UM,Small Shop,Mixed,STORE00001,LS,E02																		
11	200608,20060423,1,10,1,1.99,PRD0900017,CL00140,DEP00049,G00014,D00003,CUST0000539672,UM,OA,994100200481845,L,UM,Full Shop,Fresh,STORE00001,LS,E02																		
12	200608,20060422,7,19,3,2.55,PRD0900020,CL00202,DEP00067,G00021,D00005,CUST0000275967,LA,OA,994100200322021,L,MM,Top Up,Mixed,STORE00001,LS,E02																		
13	200608,20060420,5,16,3,1.56,PRD0900024,CL00160,DEP00054,G00016,D00003,CUST0000448280,UM,OT,994100200426287,L,MM,Top Up,Grocery,STORE00001,LS,E02																		
14	200608,20060417,2,09,1,0.52,PRD0900024,CL00160,DEP00054,G00016,D00003,,,,994100200108896,S,MM,Small Shop,Grocery,STORE00001,LS,E02																		
15	200608,20060423,1,17,1,0.35,PRD0900025,CL00163,DEP00055,G00016,D00003,CUST0000788454,MM,YF,994100200632881,L,MM,Top Up,Fresh,STORE00001,LS,E02																		
16	200608,20060418,3,12,1,0.35,PRD0900025,CL00163,DEP00055,G00016,D00003,CUST0000996753,LA,YF,994100200760228,L,MM,Full Shop,Mixed,STORE00001,LS,E02																		
17	200608,20060417,2,17,1,1.38,PRD0900026,CL00144,DEP00051,G00015,D00003,CUST0000167624,UM,YA,994100200256789,M,UM,Small Shop,Fresh,STORE00001,LS,E02																		
18	200608,20060421,6,08,1,1.87,PRD0900027,CL00074,DEP00021,G00007,D00002,CUST0000177536,UM,OT,994100200262634,L,MM,Top Up,Fresh,STORE00001,LS,E02																		
19	200608,20060420,5,14,1,0.3,PRD0900032,CL00163,DEP00055,G00016,D00003,CUST0000923296,UM,YF,994100200715687,L,MM,Top Up,Grocery,STORE00001,LS,E02																		
20	200608,20060422,7,19,1,0.93,PRD0900033,CL00201,DEP00067,G00021,D00005,CUST0000297112,MM,OT,994100200334603,L,UM,Top Up,Fresh,STORE00001,LS,E02																		
21	200608,20060421,6,19,1,0.84,PRD0900034,CL00129,DEP00046,G00013,D00003,CUST0000720862,UM,OA,994100200591784,L,UM,Full Shop,Fresh,STORE00001,LS,E02																		
22	200608,20060421,6,08,3,1.65,PRD0900035,CL00113,DEP00040,G00011,D00003,CUST0000177536,UM,OT,994100200262634,L,MM,Top Up,Fresh,STORE00001,LS,E02																		
23	200608,20060418,3,17,1,0.55,PRD0900035,CL00113,DEP00040,G00011,D00003,CUST0000638824,MM,OT,994100200542069,M,UM,Small Shop,Fresh,STORE00001,LS,E02																		
24	200608,20060422,7,14,1,2.1,PRD0900039,CL00022,DEP00007,G00003,D00001,CUST0000410727,UM,OT,994100200403524,L,MM,Full Shop,Mixed,STORE00001,LS,E02																		
25	200608,20060423,1,13,1,0.71,PRD0900043,CL00148,DEP00052,G00015,D00003,CUST0000199923,LA,,994100200276413,L,LA,Top Up,Nonfood,STORE00001,LS,E02																		
26	200608,20060421,6,10,1,0.71,PRD0900043,CL00148,DEP00052,G00015,D00003,CUST0000811783,UM,,994100200647057,L,MM,Full Shop,Grocery,STORE00001,LS,E02																		
27	200608,20060418,3,10,1,1.15,PRD0900053,CL00070,DEP00020,G00007,D00002,CUST0000280656,MM,OT,994100200324693,M,MM,Small Shop,Fresh,STORE00001,LS,E02																		
28	200608,20060420,5,13,1,1.15,PRD0900053,CL00070,DEP00020,G00007,D00002,CUST0000419761,MM,YA,994100200409066,L,MM,Top Up,Fresh,STORE00001,LS,E02																		
29	200608,20060421,6,10,1,2.4,PRD0900054,CL00067,DEP00019,G00007,D00002,CUST0000244114,UM,YA,994100200302894,L,MM,Full Shop,Mixed,STORE00001,LS,E02																		
30	200608,20060421,6,16,3,1.38,PRD0900057,CL00107,DEP00037,G00010,D00003,CUST0000306181,MM,YA,994100200340289,L,MM,Full Shop,Mixed,STORE00001,LS,E02																		
31	200608,20060422,7,13,1,1,PRD0900065,CL00043,DEP00011,G00004,D00002,CUST0000556297,MM,YA,994100200492022,L,UM,Top Up,Fresh,STORE00001,LS,E02																		
32	200608,20060423,1,18,1,1,PRD0900065,CL00043,DEP00011,G00004,D00002,,,,994100200135589,L,UM,Top Up,Fresh,STORE00001,LS,E02																		
33	200608,20060421,6,08,1,1.29,PRD0900066,CL00041,DEP00011,G00004,D00002,CUST0000661427,MM,OT,994100200555824,M,LA,Small Shop,Grocery,STORE00001,LS,E02																		
34	200608,20060419,4,17,1,0.59,PRD0900067,CL00201,DEP00067,G00021,D00005,CUST0000978405,UM,YA,994100200749164,L,MM,Full Shop,Fresh,STORE00001,LS,E02																		
35	200608,20060418,3,18,5,2.95,PRD0900067,CL00201,DEP00067,G00021,D00005,CUST0000836364,UM,PE,994100200662091,L,UM,Top Up,Nonfood,STORE00001,LS,E02																		
36	200608,20060417,2,10,1,2.45,PRD0900071,CL00086,DEP00024,G00007,D00002,CUST0000644893,LA,PE,994100200545760,L,MM,Top Up,Fresh,STORE00001,LS,E02																		
37	200608,20060419,4,17,1,0.79,PRD0900080,CL00006,DEP00002,G00001,D00001,,,,994100200061639,L,UM,Top Up,Nonfood,STORE00001,LS,E02																		
38	200608,20060420,5,11,1,0.64,PRD0900090,CL00140,DEP00049,G00014,D00003,CUST0000623730,MM,OT,994100200532878,L,MM,Top Up,Fresh,STORE00001,LS,E02																		

Şekil 3.2: Veri Setinin Genel Görünümü

Yukarıda alışverişin yapıldığı mali hafta, alışveriş tarihi, alışveriş saati, sepette bulunan ürünün miktarı, birim fiyatı, hangi hiyerarşik üretim koduna sahip olduğu, müşteri kodu, müşteri-fiyat duyarlılığı (düşük etkili, orta market, ileri market, sınıfsız), müşteri yaşam seviyesi (Genç yetişkinler, yaşlı yetişkinler, genç aileler, yaşlı aileler, emekliler, diğer ve sınıfsız) bulunmaktadır. Ayrıca sepet numarası, sepet-ücret duyarlılığı (düşük etkili, orta market, ileri market, sınıfsız), sepet tipi, sepetteki baskın ürün grubu(taze, bakkaliye ürünler, karışık, yiyeceksiz, tanımsız), mağaza kodu, mağaza boyutu(Küçük, Orta, büyük, dev) ve son olarak mağaza bölgesi (Doğu-1, Batı-1, Batı-2, Kuzey-3) gibi bilgiler yer almaktadır. Veri setinde yer alan niteliklere ait açıklayıcı bilgiler aşağıda tabloda yer almaktadır.

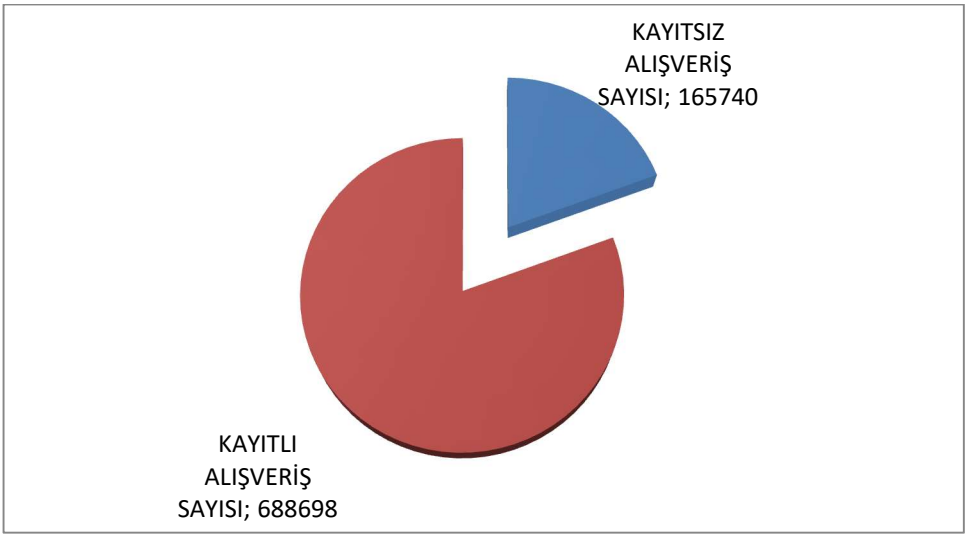
Tablo 3.1: Veri Analizinde Kullanılacak Olan Nitelikler

SÜTUN BAŞLIKLARI	AÇIKLAMALAR	TÜRÜ	ÖRNEK
Alışveriş Tarihi	Alışverişin Yapıldığı Tarih	Zaman	20073011
Miktar	Aynı Sepette Bulunan Aynı Ürünün Adedi	Nümerik	Tam Sayı
Harcama	Satın Alınan Ürünlere İlişkin Harcama Tutarı	Nümerik	İki Ondalık Basamağı Olan Sayılar
Ürün Kodu	Hiyerarşik Ürün Ağacında Ürünün Yer Aldığı Bölüm	Nominal	Ürünkod10, Ürünkod20
Müşteri Kodu	Alışveriş Yapan Müşterilere Atanan Numaralar	Nominal	Müşteri001624, Müşteri002932
Müşteri Fiyat Duyarlılığı	Müşterilerin Fiyata Olan Hassasiyeti	Nominal	LA=Daha Az Zengin, MM=Orta Pazar, UM=Yüksek Market, XX=Sınıflandırılmamış
Müşteri Yaşam Evresi	Müşteri Yaşlarına Göre Yapılan Sınıflandırma	Nominal	YA=Genç Yetişkinler, OA=Daha Yaşlı Yetişkinler, YF=Genç Aileler, OF=Daha Yaşlı Aileler, PE=Emekliler, OT=Diğer,
Sepet Numarası	Alışveriş Sepetlerine Atanan Numaralar	Nominal	99003255, 9900561

Tablo 3.1'de kullanılan nitelikler kısaca açıklanmıştır. Sonraki adımlarda ham veri seti analiz formatına uygun bir şekilde hazırlanıp, analize uygun veriler seçilecektir.

3.3 VERİLERİN BİRLEŞTİRİLİP TEMİZLENMESİ

Veri seti, 32 farklı ".csv" uzantılı dosyadan oluşmaktadır. Bu nedenle, ilk olarak veri seti parçaları birleştirilmiştir. Daha sonra, analizin olumsuz etkilememesi adına müşteri numarası olmayan alışveriş sepetleri analizden çıkarılmıştır. Şekil 3.3'te analizden çıkarılan sepet sayısı ve analize dâhil edilen sepet sayısı belirtilmiştir.



Şekil 3.3: Veri Setinde Bulunan Kayıtlı ve Kayıtsız Alışveriş Sayıları.

Ayrıca kayıtlı müşterilerin alışverişleri ile RFM bileşenlerinin sayısal analizleri Tablo 3.2'de yer almaktadır.

Tablo 3.2: Veri Setinde RFM Parametrelerine Ait Değerler

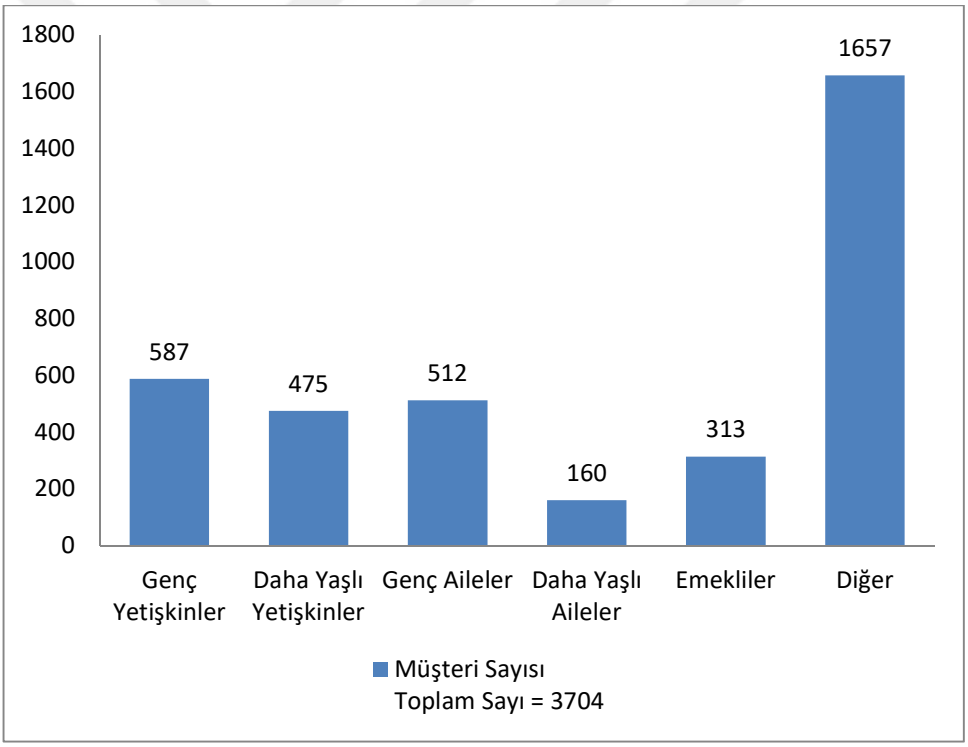
	Yakınlık	Frekans	Harcama
Ortalama	29,4320	163,9240	661,9570
Minimum Değer	0,0000	1,0000	0,0100
Maksimum Değer	223,0000	1605,0000	40044,9500
Varyans	1810,3660	38152,3970	1498163,2510
Standart Sapma	42,5480	195,3260	1223,9950
Ortalamanın Standart Hatası	0,7070	3,2090	20,1110
Medyan	11,0000	86,0000	284,0850
Mod	0,0000	2,0000	0,0100

Tablo 3.2'ye göre müşterilerin ortalama son alışveriş tarihi 29 gün öncedir. 223 günü içeren veri setinde bir müşteri ortalama olarak yaklaşık 164 kez alışveriş yapmış ve bu alışverişlerin

sonunda ortalama 661,9570 para birimi harcamıştır. Minimum ve maksimum değerlere bakıldığında alışveriş frekansında 1605 değeri ile bir müşterinin çok yüksek seviyede alışveriş yaptığı gözlenmektedir. Yine harcamalarda ortalamayı aşağı çeken 0,0100 para birimi değerinde alışveriş kaydına rastlanmaktadır. Standart sapmalara bakıldığında ise özellikle harcama parametresindeki değer dikkat çekicidir. Bu değerin fazla olmasında, müşteri alışverişlerindeki tutar dalgalanmaları arasındaki yüksek farklar sebep gösterilebilir.

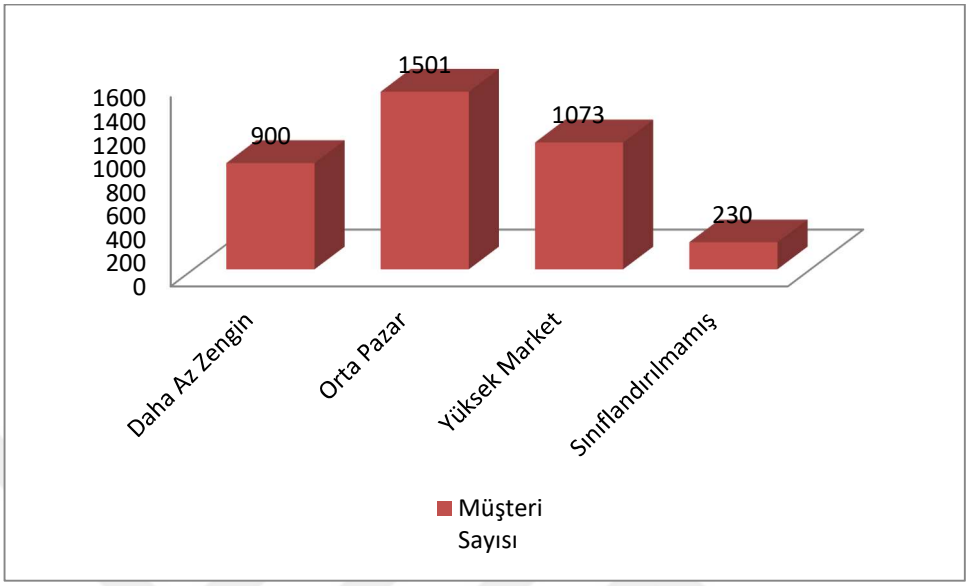
Veri seti içinde yer alan "Müşteri Yaşam Evresi" ve "Müşteri Fiyat Duyarlılığı" ile ilgili tablolar aşağıda yer almaktadır.

Tablo 3.3: Veri Seti Müşteri Yaşam Evresi Histogramı



Tablo 3.3'te görüldüğü üzere yaşam evresi sütununda en çok nüfusa sahip küme 1657 kayıt ile "*Diğer*" olarak atanan kayıtlar adı altında yer almaktadır. "Daha *Yaşlı Yetişkinler*", "*Emekliler*", "*Genç Yetişkinler*" ve "*Genç Aileler*" birbirlerine yakın oranlarda (475, 313, 587, 512) iken "Daha *Yaşlı Aileler*" kümesi nüfus olarak diğer değerlere göre daha az sayıda bulunmaktadır (160).

Tablo 3.4: Veri Seti Müşteri Fiyat Duyarlılığı Tablosu



Müşteri Fiyat hassasiyetine bakıldığında ise "*Orta Market*" adı verilen grup 1501 kişilik nüfusuyla toplam müşteri sayısının %40'ını aşan bir gözlem değerine sahiptir. "*Daha Az Zengin*" ve "*Yüksek Market*" adı verilen sınıflar ise birbirlerine yakın eleman sayılarına sahiptir (900, 1073). Son olarak "*Tanımsızlar*" 230 adet üye ile toplam müşteri sayısı baz alındığında yaklaşık %6'lık bir paya sahiptir.

3.4 VERİLERİN HAZIRLANMASI VE SEÇİLMESİ

VM süreçlerinden verilerin birleştirilmesi kısmına uygun olarak birleştirilen ve temizlenen veri seti, kümeleme analizi için ön hazırlık safhasına getirilmiştir.

Verilerin hazırlanması aşamasında ,analizlere uygun olarak verilerin düzenlenmesi gerekmektedir. Buradan hareketle, kümeleme analizi yapmak için ilk olarak perakendecilik sektöründe sıkça kullanılan ve Bölüm 2.4.1'de de açıklanan RFM modeli kurulması için veri setinden Yakınlık, Frekans ve Harcama değerleri saptanmış analizde kullanmak amacı ile "Ortalama Sepet Harcama Tutarı" da bir sütun olarak veri setine eklenmiştir.

Yakınlık değeri, elimizde bulunan son alışveriş tarihi ile kayıtlı müşterilerin her birinin yaptığı son alışveriş arasındaki fark olarak hesaplanmıştır. Frekans değeri müşterilerin sepet

sayısını ifade etmektedir. Harcama ise ürünlerin birim fiyatı ile müşterinin aldığı ürün miktarının çarpımlarının toplanması ile bulunmuştur. Ortalama sepet tutarı da müşterinin belirtilen tarih aralığında işletmeden yaptığı toplam alışveriş tutarının sepet sayısına bölünmesiyle elde edilmiştir.

RFM modeline ek olarak daha da verimli bir kümeleme analizi yapmak adına müşteri fiyat duyarlılığı ve müşteri yaşam seviyesi parametreleri de kümeleme analizinde kullanılmıştır. RFM modeli ile kümeleme analizi yapılacak veri setine ait temsili görünüm Şekil 3.4'te sunulmaktadır.

Verilerin seçilmesi aşamasında ise RFM modelini temel alarak oluşturulan veriler uygun analiz için seçilmiş ve modelin kurulması adımında kümeleme analizi yapılması için gereken zemin hazırlanmıştır.

A1 MUSTERI_KODU

	A	B	C	D	E	F	G
1	MUSTERI_KODU	MUSTERI FIYAT_DUY	MUSTERI_YASAM_SEV	YAKINLIK	FREKANS	HARCAMA	ORT_SEP_TUTARI
2	CUST0000005874	LA	OA	2	64	581.2200	9.0816
3	CUST0000014549	LA	OA	5	690	1532.9200	2.2216
4	CUST0000015976	LA	OA	108	58	284.9300	4.9126
5	CUST0000028627	LA	OA	78	66	132.0800	2.0012
6	CUST0000033514	LA	OA	40	40	203.2700	5.0818
7	CUST0000041315	LA	OA	13	201	851.8300	4.2380
8	CUST0000044856	LA	OA	1	362	396.4100	1.0951
9	CUST0000045451	LA	OA	24	379	631.4500	1.6661
10	CUST0000046556	LA	OA	100	6	38.2800	6.3800
11	CUST0000049880	LA	OA	1	556	1097.9600	1.9747
12	CUST0000051933	LA	OA	35	71	251.9100	3.5480
13	CUST0000055392	LA	OA	28	7	16.9100	2.4157
14	CUST0000060281	LA	OA	121	4	5.7200	1.4300
15	CUST0000062780	LA	OA	14	46	95.2400	2.0704
16	CUST0000072124	LA	OA	45	63	231.5400	3.6752
17	CUST0000090376	LA	OA	34	61	105.7900	1.7343
18	CUST0000100432	LA	OA	3	62	290.5800	4.6868
19	CUST0000101185	LA	OA	12	54	73.8900	1.3683
20	CUST0000109079	LA	OA	5	160	689.9100	4.3119
21	CUST0000110607	LA	OA	35	42	142.2200	3.3862
22	CUST0000113788	LA	OA	17	81	117.6200	1.4521
23	CUST0000122425	LA	OA	3	14	25.6400	1.8314
24	CUST0000124963	LA	OA	7	154	474.4100	3.0806
25	CUST0000127498	LA	OA	0	321	2087.6100	6.5035
26	CUST0000141094	LA	OA	83	60	684.7900	11.4132
27	CUST0000141657	LA	OA	9	242	1161.3500	4.7990
28	CUST0000145387	LA	OA	20	3	5.6800	1.8933
29	CUST0000154831	LA	OA	4	53	126.3900	2.3847
30	CUST0000155661	LA	OA	111	29	148.7500	5.1293
31	CUST0000156198	LA	OA	6	291	637.8100	2.1918
32	CUST0000168307	LA	OA	62	45	85.9000	1.9089
33	CUST0000176387	LA	OA	2	6	28.0400	4.6733
34	CUST0000184526	LA	OA	1	452	2487.2300	5.5027
35	CUST0000193087	LA	OA	6	72	143.2200	1.9892
36	CUST0000212869	LA	OA	0	356	2292.6300	6.4400
37	CUST0000225324	LA	OA	3	220	208.8600	0.9494
38	CUST0000242508	LA	OA	6	413	1406.4900	3.4055

Şekil 3.4: Kümeleme Analizine Hazır Hale Getirilmiş Veri Seti

Kümeleme analizinden elde edilen veriler ile birliktelik kuralı analizi yapılacağından benzer süreç Şekil 3.1'de gösterildiği gibi birliktelik kuralları analizi için yeniden gerçekleştirilecektir.

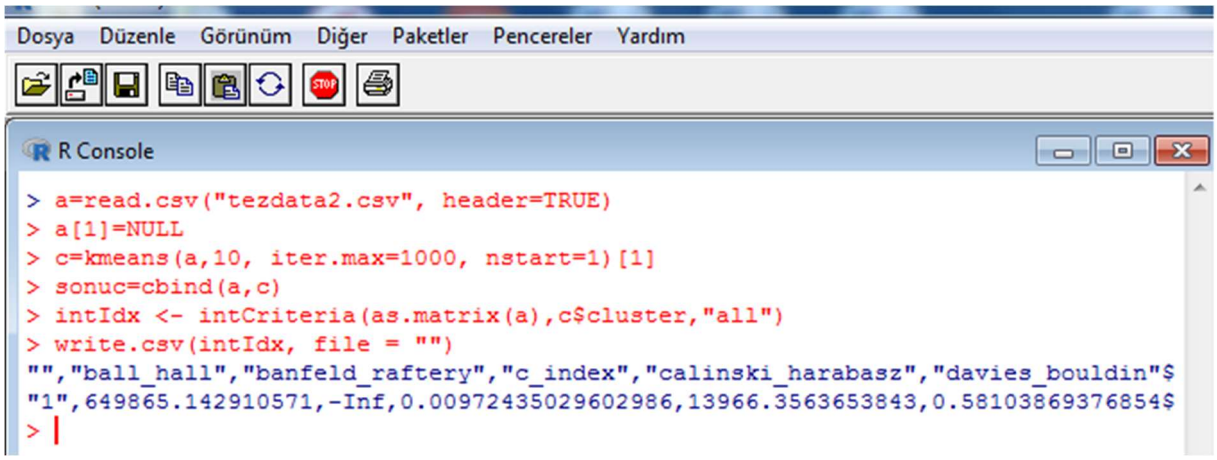
4. BULGULAR

4.1 MODELİN KURULMASI VE DEĞERLENDİRİLMESİ

Bu bölümde, kümeleme analizi ve birliktelik kuralları madenciliği uygulama adımları ve bu uygulamalardan elde edilen bulgular özetlenecektir.

4.1.1 Kümeleme Analizi

Kümeleme analizinden, müşteri segmentasyonunda yaygın olarak faydalanılmaktadır. Bölüm 2.4.1.2.'de açıklandığı üzere kümeleme analizi için geliştirilen birçok algoritma vardır. Bu çalışmada kullanılan k-ortalamalar algoritması ve bulanık c-ortalamalar algoritmaları en yaygın kullanılan algoritmalar arasındadır. Her iki kümeleme algoritmasında, küme sayısı, kümeleme öncesi kullanıcı tarafından belirlenmektedir. Küme sayısını belirlemek adına tek bir sezgisel yaklaşımın kullanılması yerine kümeleme indeksleri ile küme sayısı hakkında bilgi sahibi olmak, aynı zamanda hangi kümeleme algoritmasının tercih edilmesi gerektiğini saptamak mümkündür. Kümeleme analizinde kullanılan bu indeksler Bölüm 2.4.1.2'de açıklanmıştır. Bu indeks değerleri ışığında uygun küme sayısının ve kümeleme algoritmasının belirlenmesi çalışmanın güvenilirliğini artıracaktır. Bu nedenle; k-ortalamalar ve bulanık c-ortalamalar algoritmasının, 5 farklı kümeleme analizi indeksi (Calinski-Harabasz, Davies-Bouldin, Silhouette, Dunn ve Xie-Beni) değeri küme sayısının 5-20 değer aralığı için hesaplanmıştır.



Şekil 4.1: R 3.6.0 ile K-Ortalamalar Algoritmasının k=10 için İşletilmesi ve Kümeleme İndekslerinin Hesaplanması

Kümeleme algoritmalarının uygulanması ve indeks değerlerinin hesaplanması için R programlama dili kullanılmıştır. Bu noktada, R 3.6.0 programının, "Cluster" ve "ClusterCrit" paketlerinden faydalanılmıştır. Şekil 4.1'de k-ortalamalar algoritmasının k=10 değeri ile çalıştırılması sonucunda, kümeleme indekslerinin hesaplanması ile ilgili R dili kodları sunulmaktadır. Kümeleme analizinde kullanılan k-ortalamalar ve bulanık c-ortalamalar algoritmasının, mevcut veri seti için küme sayısının 5-20 aralığında olması durumunda, belirlenen 5 farklı kümeleme indeksinin değerleri Tablo 4.1'de sunulmaktadır.

Tablo 4.1: Kümeleme algoritmalarının indeks değerleri

Algoritma	Küme	İndeksler				
		Calinski-Harabasz	**Davies-Bouldin**	**Silhouette**	**Dunn**	**Xie-Beni**
***K*-Ortalamalar**	5	4223,0952	0,5901	0,0004	**0,4533**	2050,1224
	6	3829,4907	0,6090	0,0002	0,4296	8799,5763
	7	11847,2806	**0,4836**	0,0006		2687,5383
	8	12050,2975	0,5369	0,0007		1763,4334
	9	12165,6222	0,5733	**0,0011**		**681,2207**
	10	13966,3564	0,5810	0,0006		2812,5364
	11	**14415,8573**	0,6085	0,0007		1652,0227
	12	14289,4050	0,6463	0,0009		891,5752
	13	13784,9414	0,6859	0,0007		1665,6816
	14	13098,6863	0,7376	0,0004		3972,8933
	15	12605,3960	0,7400	0,0005		3302,5031
	16	12134,0609	0,7654	0,0003		7636,4533
	17	11622,5905	0,7625	0,0006		1981,6131
	18	11217,0004	0,7737	0,0006		1932,9024

	19	10641,2396	0,8020	0,0003		8045,0331
	20	10136,9608	0,8067	0,0005		2802,4739
Bulanık	5	3964,7226	0,5805	0,0002	0,4532	9483,7349
***C*-Ortalamalar**	6	3685,6299	0,6022	0,0002	0,4418	4809,1553
	7	3481,7438	0,6342	0,0002	0,4181	8867,3593
	8	3147,3525	0,6610	0,0002	0,4020	11558,6815
	9	13699,1184	0,5462	0,0009		1485,1588
	10	13943,1594	0,5871	0,0009		1262,2374
	11	5267,4574	0,6989	0,0002	0,3784	5093,3518
	12	4938,2302	0,7243	0,0002	0,3677	3624,2775
	13	4615,0600	0,7641	0,0002	0,3511	3684,8770
	14	4436,4071	0,7533	0,0001	0,3435	18494,3668
	15	4183,5986	0,8005	0,0002	0,3295	4433,8901
	16	3961,1426	0,8345	0,0002	0,3223	5650,1424
	17	3744,6779	0,8547	0,0001	0,3168	10091,7248
	18	3557,6253	0,8733	0,0002	0,3116	6181,7350
	19	3405,2741	0,8531	0,0001	0,3155	8652,6790
	20	3249,2627	0,8659	0,0001	0,3113	8965,9256

Tablo 4.1'den elde edilen sonuçlar aşağıdaki gibidir:

- Calinski-Harabasz indeksine göre en iyi kümeleme algoritması k-ortalamalar ve en iyi küme sayısı 10'dur.
- Davies-Bouldin indeksine göre en iyi kümeleme algoritması k-ortalamalar ve en iyi küme sayısı 7'dir.
- Silhouette indeksine göre en iyi kümeleme algoritması k-ortalamalar ve en iyi küme sayısı 9'dur.
- Dunn indeksine göre en iyi kümeleme algoritması k-ortalamalar ve en iyi küme sayısı 5'dir.
- Xie-Beni indeksine göre en iyi kümeleme algoritması k-ortalamalar ve en iyi küme sayısı 9'dur.

Bu sonuçlar değerlendirildiğinde; k-ortalamalar algoritmasının Bulanık *c*-Ortalamalar algoritmasına göre tüm indeks değerleri için daha iyi sonuçlara sahip olduğu görülmektedir. Bu nedenle, analizde k-ortalamalar algoritmasının kullanılmasına karar verilmiştir. İki farklı indeks ölçütünde en iyi değere ve diğer indekslerde de kabul edilebilir değerlere sahip olduğu için küme sayısı olarak 9 değeri seçilmiştir.

En uygun kümeleme algoritması ve küme sayısı belirlendikten sonra sıradaki adım analizin gerçekleştirilmesi adımıdır. R 3.6.0 programında k-ortalamalar algoritmasının k=9 değeri ile çalıştırılması sonucunda elde edilen küme merkezleri Tablo 4.2'de sunulmaktadır.

Tablo 4.2: K-Ortalamalar Algoritması (k=9) ile Elde Edilen Küme Merkezleri

Müşteri Kümeleri	Yakınlık	Frekans	Harcama	Ort. Sepet Tutarı	Müşteri Fiyat Duyarlılığı	Müşteri Yaşam Seviyesi	Müşteri Sayısı
Küme - 1	30,2222	413,5556	12250,5878	67,5640	Orta Pazar	Diğer	9
Küme - 2	5,3333	504,3889	2070,4739	10,3608	Orta Pazar	Diğer	216
Küme - 3	8,7500	583,6136	5001,9468	19,3672	Orta Pazar	Diğer	44
Küme - 4	48,7412	28,0825	75,4876	3,3471	Orta Pazar	Diğer	1696
Küme - 5	10,8815	245,3365	817,0984	4,6069	Orta Pazar	Diğer	523
Küme - 6	7,8278	381,7778	1342,9978	4,5960	Orta Pazar	Diğer	360
Küme - 7	2,0000	319,0000	40044,9500	125,5328	Orta Pazar	Diğer	1
Küme - 8	17,0780	130,6640	392,2629	4,9879	Orta Pazar	Diğer	756
Küme - 9	5,6667	569,1616	3019,0171	6,5093	Orta Pazar	Diğer	99
Ortalama	15,1667	352,8423	7223,8691	27,4302			

Küme merkezlerine göre oluşan kümeler aşağıdaki gibi yorumlanabilir:

İlk olarak, harcama tutarı ve alışveriş sıklığı yüksek; fakat son alışveriş tarihi merkezi olarak oldukça büyük değere sahip olan küme "Küme-1" adını almıştır. Bu kümedeki müşteriler kayıp müşteriler olarak değerlendirilebilir. Küme 2'de harcama tutarı ve sepet büyüklükleri düşük, son alışverişleri yakın zamanda gerçekleşmiş ve sıklıkla alışveriş yapan sadık müşteriler yer almaktadır. Küme-3'de ise Küme 2'ye göre harcama tutarı ve sepet büyüklüğü daha yüksek olan sadık müşteriler yer almaktadır. Küme-4 çok düşük frekansa ve harcama tutarına sahip olan ve uzun zamandır alışveriş yapmayan müşterilerden oluşmaktadır. "Küme-5" ise harcama tutarı ve alışveriş sıklığı Küme 4'den yüksek ancak ortalamanın altında olan müşterileri içermektedir. Küme-6'da yer alan müşteriler en az Küme-1 kadar etkin görünmesine karşın, harcama tutarlarında ve sepet ortalamalarında genel ortalamanın altında kalmıştır. Küme-7 harcama tutarı açısından aykırı bir değere sahip tek bir müşteriden oluşmaktadır. Bu durum, kümeleme analizinin aykırı değer tespitinde de başarılı bir şekilde kullanılabileceğini ortaya koymaktadır. Küme-8, alışveriş sıklığı, harcama tutarı ve sepet büyüklüğü oldukça düşük değerlere sahip müşterilerden oluşmaktadır. Bu kümedeki

müşterilerin son alışveriş tarihleri genel ortalamaya yakındır. Küme-9 en sık alışveriş yapan ve son alışverişlerini yakın zamanda gerçekleştirmiş sadık müşterileri içermektedir.

Müşteri sadakatini sağlamak amacı ile oluşturulacak stratejilerin odağında kayıp müşterilerin yeniden kazandırılması ve sadık müşterilerin sadakatini artırmak vardır. Bu oranların işletmenin lehine artırılması için elde edilen kümelerdeki müşterilerin alışveriş verileri incelenerek hangi ürün gruplarını birlikte satın aldıkları ile ilgili çıkarımlar yapılacaktır. Bu stratejiler ışığında; değerli ve kayıp müşteriler kümesi olarak Küme-1 ile sadık ve değerli müşteri kümeleri olarak Küme-2, Küme-3 ve Küme-9 birliktelik kuralları madenciliği için hedef kümeler olarak seçilmiştir. Bu doğrultuda, Apriori algoritması kullanılarak, hem Küme-1 için birliktelik kuralları analizi yapılacak hem de Küme-2, Küme-3 ve Küme-9'un birleşiminden oluşan "Küme-2, 3, 9" kümesi için birliktelik kuralları analizi yapılacaktır.

4.1.2 Birliktelik Kuralları Madenciliği

Pazar sepet analizi için belirlenen kümelerde yer alan müşterilere ait müşteri işlem verileri kullanılmıştır. Birliktelik kuralları, ürün kategorileri (departmanlar) bazında elde edilmiştir. Bu analizlerin gerçekleştirilmesinde Weka 3.8.3 yazılımından faydalanılmıştır.

Hedef kümelerden ilki olan 9 müşteriden ve 322 sepet kaydından oluşan Küme-1 için Apriori algoritmasının uygulanması sonucunda elde edilen ilk 50 kural Tablo 4.3'de sunulmaktadır. Bu kurallar, kaldıraç değerlerine göre sıralanmaktadır. Appriori algoritmasının işletilmesinde minimum güven değeri %50, minimum destek değeri ise %10 olarak alınmıştır.

Tablo 4.3: Küme-1 için Appriori Algoritması ile Elde Edilen Birliktelik Kuralları

Kural No	Kural Açıklaması	Kaldıraç İndeksi	İnanç İndeksi	Güven Oranı	Destek Oranı
1	DEP00011, DEP00076 → DEP00019	2,2300	10,0300	0,9600	0,1574
2	DEP00008, DEP00076 → DEP00019	2,1500	6,5900	0,9300	0,1667
3	DEP00067, DEP00076 → DEP00019	2,1400	6,0200	0,9200	0,1512
4	DEP00022, DEP00052 → DEP00008	1,6500	4,3700	0,9200	0,1667
5	DEP00021, DEP00052 → DEP00008	1,6200	4,2200	0,9100	0,1605
6	DEP00011, DEP00019, DEP00052 → DEP00008	1,6100	3,8700	0,9000	0,1698
7	DEP00019, DEP00022 → DEP00008	1,5900	3,6200	0,8900	0,1574
8	DEP00019, DEP00052 → DEP00008	1,5800	3,3900	0,8800	0,2284
9	DEP00011, DEP00022 → DEP00008	1,5800	3,2200	0,8800	0,1574
10	DEP00011,DEP00019 → DEP00008	1,5600	3,3200	0,8800	0,2623
11	DEP00019, DEP00021, DEP00067 →DEP00008	1,5600	3,0100	0,8700	0,1636

12	DEP00011, DEP00019, DEP00067 → DEP0008	1,5400	2,9100	0,8600	0,1574
13	DEP00019, DEP00076 → DEP00008	1,5400	2,8000	0,8600	0,1667
14	DEP00019, DEP00067 → DEP00008	1,9700	2,8400	0,8600	0,2191
15	DEP00076 → DEP00019	1,5200	3,5000	0,8500	0,1944
16	DEP00010, DEP00019 → DEP00008	1,5100	2,5900	0,8400	0,1667
17	DEP00010, DEP00011 → DEP00008	1,5100	2,5600	0,8400	0,1790
18	DEP00019 → DEP00008	1,4900	2,4900	0,8300	0,3580
19	DEP00010 → DEP00008	1,4800	2,4000	0,8200	0,2469
20	DEP00011, DEP00052 → DEP00008	1,4800	2,3600	0,8200	0,2160
21	DEP00022 → DEP00008	1,4700	2,2700	0,8100	0,2438
22	DEP00021 → DEP00008	1,4600	2,2000	0,8100	0,2099
23	DEP00019, DEP00076 → DEP00011	1,8300	2,7100	0,8100	0,1574
24	DEP00008, DEP00011, DEP00067 → DEP00019	1,8700	2,7500	0,8100	0,1574
25	DEP00052, DEP00067 → DEP0008	1,4400	2,0600	0,8000	0,1821
26	DEP00011 → DEP0008	1,4300	2,1200	0,8000	0,3519
27	DEP00011, DEP00067 → DEP00008	1,4200	1,9800	0,7900	0,1944
28	DEP00008, DEP00011, DEP00052 → DEP00019	1,8200	2,4800	0,7900	0,1698
29	DEP00076 → DEP00008	1,4100	1,9300	0,7800	0,1790
30	DEP00008, DEP00021 → DEP00019	1,8000	2,4100	0,7800	0,1636
31	DEP00019, DEP00076 → DEP00067	1,9800	2,5500	0,7800	0,1512
32	DEP00008, DEP00067 → DEP00019	1,7900	2,3700	0,7700	0,2191
33	DEP00008,DEP00021 → DEP00052	1,7000	2,2000	0,7600	0,1605
34	DEP00054 → DEP00052	1,6900	2,1700	0,7600	0,1667
35	DEP0008, DEP00011 → DEP00019	1,7300	2,1600	0,7500	0,2623
36	DEP00008, DEP00019, DEP00052 → DEP00011	1,6800	2,0700	0,7400	0,1698
37	DEP00011, DEP00067 → DEP00019	1,7100	2,0700	0,7400	0,1821
38	DEP00008, DEP00019 → DEP00011	1,6600	2,0300	0,7300	0,2623
39	DEP00076 → DEP00008,DEP00019	2,0400	2,2600	0,7300	0,1667
40	DEP00021 → DEP00019	1,6800	1,9900	0,7300	0,1883
41	DEP00019, DEP00052 → DEP00011	1,6500	1,9600	0,7300	0,1883
42	DEP00008, DEP00010 → DEP00011	1,6400	1,9400	0,7200	0,1790
43	DEP00067 → DEP0008	1,3000	1,5700	0,7200	0,2840
44	DEP00052 → DEP00008	1,2900	1,5400	0,7200	0,3241
45	DEP00008, DEP00019, DEP00067 → DEP00011	1,6300	1,8900	0,7200	0,1574
46	DEP00011, DEP00052 → DEP00019	1,6600	1,9300	0,7200	0,1883
47	DEP00076 → DEP00011	1,6200	1,8800	0,7200	0,1636
48	DEP00076 → DEP00067	1,8300	2,0500	0,7200	0,1636
49	DEP00052, DEP00067 → DEP00011	1,6200	1,8800	0,7200	0,1636
50	DEP00052, DEP00067 → DEP00019	1,6600	1,9100	0,7200	0,1636

Bu tabloda yer alan ilk 5 kural aşağıdaki gibi açıklanabilir:

Kural 1: DEP00011 ve DEP00076 ürün kategorilerini sepetlerinde bulunduran müşteriler %96 olasılıkla DEP00019 ürün kategorisini de sepetine eklemektedir. Bu üç ürün kategorisi tüm veri setinin %15,74'ünde birlikte görülmektedir.

Kural 2: DEP00008 ve DEP00076 ürün kategorilerini sepetlerinde bulunduran müşteriler %93 olasılıkla DEP00019 ürün kategorisini de satın almaktadır. Bu üç ürün kategorisi müşteri işlemlerinin %16,67'sinde birlikte görülmektedir.

Kural 3: DEP00067 ve DEP00076 ürün kategorilerine ait ürünleri satın alan bir müşteri %92 olasılıkla DEP00019 ürün kategorisine ait ürünü de satın alacaktır. Veri setinin %15,12'sinde bu üç ürün kategorisine ait ürünler birlikte görülmektedir.

Kural 4: DEP00022 ve DEP00052 ürün kategorilerine ait ürünleri satın alan bir müşteri %92 olasılıkla DEP00008 ürün kategorisine ait ürünü de satın alacaktır. Bu üç ürün kategorisinin birlikte yer aldığı müşteri işlem verisi, toplam işlem sayısının %16,67'sidir.

Kural 5: DEP00021 ve DEP00052 ürün kategorisinin bulunduğu bir alışveriş sepetinde %91 olasılıkla DEP00008 ürün kategorisi de bulunacaktır. Veri setinin %16,05'inde bu üç ürün kategorisi birlikte yer almaktadır.

Yukarıda oluşturulan kurallar göz önüne alındığında özellikle DEP00019 ve DEP00008 ürün kategorileri birliktelik kurallarının odağında bulunmaktadır. Küme-1'in son alışveriş zamanının diğer kümelere oranla daha fazla olduğu da göz önüne alındığında DEP00008 ve DEP00019 ürün kategorilerine ait ürünler için fiyat indirimleri, promosyonlu satışlar bu kümeye farkındalık oluşturmak amacı ile kullanılabilir. Yine işletme bağlılığını yeniden kazanmak için yapılan kampanya bildirimleri (e-posta bildirimi, sms bildirimi, Google reklamları vb.) ürünlerin satışlarını artırmaktan da öte, Küme-1'in işletmeye tekrar kazandırılması için stratejilerde olumlu bir rol üstlenebilir. Küme-1 ile ilgili daha farklı sadakat sağlayıcı stratejiler üretmek mümkündür. İşletmelerin sektörüne, lokasyonuna veya müşteri taleplerine göre kurallar değişim gösterebilir. Küme-1'in sadakati sağlamaya yönelik stratejilerin ardından üç sadık müşteri kümesinin birleştirmesi ile oluşan 359 üyeli ve 22785 sepet kaydı bulunan "Küme-2, 3, 9" için Appriori algoritmasının işletilmesi sonucunda elde edilen en güçlü 50 kural Tablo 4.4'de sunulmaktadır. Birliktelik kuralları güven değerlerine göre sıralanmıştır. Appriori algoritmasının Weka 3.8.3 programı ile çalıştırılmasında, minimum güven değeri %50, minimum destek değeri ise %10 olarak alınmıştır.

Tablo 4.4: Küme-2, 3, 9 için Apriori Algoritması ile Elde Edilen Birliktelik Kuralları

Kural No	Kural Açıklaması	Kaldıraç İndeksi	İnanç İndeksi	Güven Oranı	Destek Oranı
1	DEP00011, DEP00021 → DEP00019	1,7900	2,5900	0,7800	0,1066
2	DEP00008, DEP00021 → DEP00019	1,7800	2,5700	0,7800	0,1089
3	DEP00011, DEP00067 → DEP00019	1,7800	2,5500	0,7800	0,1059
4	DEP00008, DEP00067 → DEP00019	1,7200	2,2900	0,7600	0,1132
5	DEP00011, DEP00022 → DEP00019	1,7200	2,2700	0,7500	0,1018
6	DEP00008, DEP00011 → DEP00019	1,7100	2,2400	0,7500	0,1462
7	DEP00008, DEP00022 → DEP00019	1,7100	2,2300	0,7500	0,1091
8	DEP00008, DEP00052 → DEP00019	1,6700	2,0800	0,7300	0,1066
9	DEP00049 → DEP00019	1,6100	1,8900	0,7000	0,1055
10	DEP00011, DEP00019 → DEP00008	1,9400	2,1100	0,7000	0,1462
11	DEP00020 → DEP00019	1,5800	1,8100	0,6900	0,1312
12	DEP00053 → DEP00019	1,5600	1,7700	0,6800	0,1168
13	DEP00019, DEP00067 → DEP00008	1,8800	1,9800	0,6800	0,1132
14	DEP00019, DEP00021 → DEP00008	1,8800	1,9700	0,6700	0,1089
15	DEP00021 → DEP00019	1,5400	1,7200	0,6700	0,1613
16	DEP00019, DEP00052 → DEP00008	1,8500	1,9100	0,6700	0,1066
17	DEP00020 → DEP00008	1,8400	1,8900	0,6600	0,1258
18	DEP00019, DEP00021 → DEP00011	2,0200	1,9800	0,6600	0,1066
19	DEP00019, DEP00022 → DEP0008	1,8000	1,8100	0,6500	0,1091
20	DEP00008 → DEP00019	1,4700	1,5900	0,6500	0,2323
21	DEP00067 → DEP00019	1,4700	1,5800	0,6500	0,1673
22	DEP00011 → DEP00019	1,4700	1,5700	0,6400	0,2097
23	DEP00012 → DEP00019	1,4600	1,5500	0,6400	0,1059
24	DEP00053 → DEP00008	1,7700	1,7500	0,6300	0,1088
25	DEP00019, DEP00067 → DEP00011	1,9400	1,8300	0,6300	0,1673
26	DEP00008, DEP00019 → DEP00011	1,9300	1,8200	0,6300	0,2323
27	DEP00001 → DEP00019	1,4200	1,4900	0,6200	0,1100
28	DEP00001, DEP00022 → DEP00011	1,8500	1,7000	0,6000	0,1686
29	DEP00022 → DEP00019	1,3700	1,4000	0,6000	0,1686
30	DEP00011 → DEP00008	1,6600	1,5900	0,6000	0,1952
31	DEP00021 → DEP00008	1,6100	1,5300	0,5800	0,1392
32	DEP00052 → DEP00019	1,3200	1,3400	0,5800	0,1602
33	DEP00067 → DEP00008	1,6100	1,5200	0,5800	0,1499
34	DEP00021 → DEP00011	1,7400	1,5600	0,5700	0,1360
35	DEP00020 → DEP00011	1,7200	1,5400	0,5600	0,1068
36	DEP00008 → DEP00011	1,6600	1,4700	0,5400	0,1952
37	DEP00019 → DEP00008	1,4700	1,3600	0,5300	0,2323
38	DEP00052 → DEP00008	1,4700	1,3600	0,5300	0,1460
39	DEP00067 → DEP00011	1,6000	1,4100	0,5200	0,1357
40	DEP00022 → DEP00008	1,4400	1,3300	0,5200	0,1458
41	DEP00011, DEP00019 → DEP00021	2,1200	1,5400	0,5100	0,1066

42	DEP00011, DEP00019 → DEP00067	1,9500	1,4900	0,5000	0,1059
43	DEP00008, DEP00019 → DEP00067	1,8800	1,4400	0,4900	0,1132
44	DEP00011, DEP00019 → DEP00022	1,7300	1,4000	0,4900	0,1018
45	DEP00022 → DEP00011	1,4700	1,3000	0,4900	0,1353
46	DEP00052 → DEP00011	1,4700	1,2900	0,4800	0,1324
47	DEP00019 → DEP00011	1,4700	1,2900	0,4800	0,2097
48	DEP00008, DEP00019 → DEP00022	1,6700	1,3600	0,4800	0,1091
49	DEP00008, DEP00019 → DEP00021	1,9500	1,4300	0,4700	0,1089
50	DEP00021 → DEP00022	1,6500	1,3400	0,4600	0,1114

Bu tabloda yer alan ilk 5 kural aşağıdaki gibi açıklanabilir:

Kural 1: DEP00011 ve DEP00021 ürün kategorileri bulunan bir sepette %78 olasılıkla DEP00019 ürün kategoriside bulunacaktır. Bu 3 ürün kategorisinin birlikte bulunduğu gözlemlerin oranı %10,66'dır.

Kural 2: DEP00008 ve DEP00021 ürün kategorilerini alan bir müşteri %78 olasılıkla DEP00019 ürün kategorisini de satın alacaktır. Bu üç ürün kategorisinin birlikte yer aldığı müşteri işlem verisi, toplam işlem sayısının %10,89'udur.

Kural 3: DEP00011 ve DEP00067 ürün kategorilerinin satın alındığı bir alışveriş sepetinde %78 oranında DEP00019 ürün kategorisi de bulunacaktır. Bu kategoriler veri setinin %10,59'unda birlikte yer almaktadır.

Kural 4: DEP00008 ve DEP00067 ürün kategorilerinin bulunduğu bir sepet %76 olasılıkla DEP00019 ürün kategorisini de barındırır ve bu birliktelik veri setinin %11,32'sinde yer almaktadır.

Kural 5: DEP00011 ve DEP00022 ürün kategorilerine ait ürünleri satın alan bir müşteri aynı sepette %75 olasılıkla DEP00019 ürün kategorisine ait bir ürünü de alacaktır. Bu ürün kategorilerin toplam sepet sayısının %10,18'inde birlikte yer almaktadır.

Yukarıda yer alan kurallara hareketle DEP00019 ürün kategorisinin Küme-2, 3, 9'da yer alan müşteriler için kritik bir öneme sahip olduğu açıktır. Stratejiler özellikle bu ürün kategorisi üzerine yoğunlaştırılırsa bu küme için göz önüne alınan ciro artırımı ve müşteri sadakatini sürdürme stratejileri hedefleri gerçekleştirilebilecektir. Tüm kurallar, üçlü bir sepet varlığı

oluşturmaktadır. Bu sebeple üç farklı ürün kategorisine ait ürünlerin bir arada olduğu promosyonlar düzenlenebilir.

Hedef kümeler olan “Küme-1” için oluşturulan çıkarımlar ve “Küme-2, 3, 9” için oluşturulan çıkarımlar birbirine paralellik gösterebilir. Birbiri ile birleştirilebilir veya birbirine zıt durumlar ortaya çıkabilir. Bu durumlarda doğru stratejinin seçilip işletmenin amaçlarına uygun bir şekilde hareket edilmesini sağlayacak adımlar atılması önemlidir. Dolayısıyla, VM sürecinin başına dönülerek problemin tanımlanması kısmında yeni bir problem belirlenerek bu amaca uygun verilerin toplanması ve sürecin işletilmesi ile süreç yeniden başlatılabilir.

5. TARTIŞMA VE SONUÇ

Perakende işletmeleri, her geçen gün gittikçe büyüyen ürün ve müşteri niteliklerini içeren temel verilere ve müşteri işlem verilerine sahiptir. Bu doğrultuda; büyük veri yığınlarını kullanarak müşteri davranışlarının analizi, ayrılan müşteri analizi, öneri ve kampanya sistemlerinin tasarlanması ve müşteri tutundurma stratejilerinin geliştirilmesi için bu verilerden anlamlı bilgiler elde edilmesi kritik önem taşımaktadır. Bu noktada; veri madenciliği algoritmaları perakende işletmelerine önemli fırsatlar sunmaktadır.

Bu çalışmada, ilk olarak veri madenciliğinin temel kavramları açıklanmış ve veri madenciliği fonksiyonları özetlenmiştir. Daha sonra, yaklaşık 900.000 müşteri alışveriş kaydı içeren deneysel bir veri seti kullanılarak müşteri segmentasyonu çalışması yapılmıştır. Bu amaçla; k-ortalamalar algoritması, Bulanık c-Ortalamalar algoritması kullanılmıştır. En iyi algoritma ve küme sayısı, 5 farklı kümeleme indeksi esas alınarak belirlenmiştir. Son olarak, oluşturulan kümeler esas alınarak sadık ve kaybedilen müşteri kümeleri belirlenmiştir. Bu kümelerde yer alan müşteriler için pazar sepet analizi çalışması yapılmıştır. Bu noktada; Apriori algoritması ile birlikte alınan ürün kategorileri keşfedilmiştir.

Bu çalışmada, veri madenciliği algoritmaları, şirketlerin politikalarına nasıl pozitif yönde etki eder sorusuna cevap aramak için kullanılmıştır. Bu doğrultuda, bu çalışmada iki önemli veri madenciliği fonksiyonu olan kümeleme analizi ve birliktelik kuralı analizinden faydalanılmıştır. İlk olarak, müşteri segmentasyonu yapılması için kümeleme analizi yapılmasına karar verilmiştir. Bu doğrultuda, veri madenciliği yöntemlerinden olan kümeleme analizinin hangi algoritma ile yapılacağı ve kaç adet kümeden meydana geleceği sorunu ile karşılaşılmış, kümeleme analizi geçerlilik indeksleri ile k-ortalamalar algoritması seçilmiş ve aynı indeksler ışığında küme sayısı 9 olarak belirlenmiştir. Kümeler, veri madenciliği sürecinin ilk adımında belirlenen problemin çözümüne olumlu katkı yapacağı düşünülen ve modern perakendecilik sektöründe yaygın olarak kullanılan RFM modeline göre düzenlenmiş ve analizler bu model ışığında gerçekleştirilmiştir. Kümeleme analizi ile küme merkezleri belirlenmiş kümeler RFM değerlerine göre yorumlanmıştır. Oluşan 9 kümeden, veri madenciliği süreç adımında belirlenen, müşteri sadakatini artırma, yeniden kazanma ve ciro artırımı için uygun olduğu düşünülen “Küme-1” adı verilen küme, değerli ancak kaybedilen

müşterilerden oluşmaktadır. Bu nedenle, ilk olarak bu küme için pazar-sepet analizi yapılmış ve hangi ürünlerin birlikte satın alındığı tespit edilmiştir. Pazar-sepet analizi Apriori algoritması kullanılarak geliştirilmiş, elde edilen bulgulara paralel olarak bu küme üyelerinin yeniden kazandırılması için stratejiler geliştirilmiş ve açıklanmıştır.

Kümeleme Analizinin bir diğer aşamasında, RFM modeli odaklı kümeleme analizinde birbirine yakın değerlere sahip üç kümenin (Küme-2, 3, 9) birleştirilmesine karar verilmiştir. Bu kümeler satın alma alışkanlıkları bir arada incelenmek amacıyla, sadakati ön planda olan ve bu doğrultuda ciro artırım hedefine uygun stratejiler geliştirilecek olan "Küme-2, 3, 9" grubunu oluşturmaktadır.

Çalışmada, Weka 3.8.3 yazılımı ve R 3.6.0 programı uygulama adımlarının gerçekleştirilmesi sırasında kullanılmıştır.

Gelecek çalışmalarda, müşterilere ait niteliklerin de yer aldığı veri setlerinin RFM analizi ile birlikte kullanılması kümeleme kalitesini arttıracaktır. Bu sayede, daha çok niteliğe sahip kümeler için daha farklı stratejiler belirlenmesi ve uygulanması mümkün olacaktır. Ayrıca, kümeleme analizinde k-ortalamalar ve bulanık c-ortalamalar algoritmasının dışında yer alan algoritmaların başarımının sınanması ve bu algoritmaların parametrelerinin optimizasyonu bir diğer geliştirme alanını oluşturmaktadır. Benzer şekilde, birliktelik kuralları madenciliği Appriori algoritması dışında yer alan algoritmalar kullanılarak gerçekleştirilebilir.

KAYNAKLAR

Aggarwal, N., Kumar, A., Khatter, H., Aggarwal, V., 2012, Analysis The Effect Of Data Mining Techniques On Database, *Advances In Engineering Software,* 47(1),S 164-169.

Akamavi, R. K., Mohamed, E., Pellmann, K., & Xu, Y., 2015, Key Determinants of Passenger Loyalty In the Low-Cost Airline Business, *Tourism Management*, 46, S 528-545.

Akın, Y.K., 2008, *Veri Madenciliğinde Kümeleme Algoritmaları ve Kümeleme Analizi*, Yüksek Lisans Tezi, Marmara Üniversitesi Sosyal Bilimler Enstitüsü Ekonometri Anabilim Dalı İstatistik Bilim Dalı.

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

Akpınar, H., 2014, *Data Veri Madenciliği Veri Analizi*, Papatya Yayıncılık, İstanbul.

Altıntaş, T., 2006, *Veri Madenciliği Metotlarından Olan Kümeleme Algoritmalarının Uygulamalı Etkinlik Analizi,* Yüksek Lisans Tezi, Sakarya Üniveristesi Fen Bilimleri Enstitüsü.

Boztuğ, Y., Reutterer, T., 2008, A combined approach for segment-specific market basket analysis, *European Journal of Operational Research,* 187 (2008), S 294–312.

Bult, J.R., Wansbeek, T., 1995, Optimal Selection for Direct Mail, *Marketing Science*, 13 (4), S 343-466.

Calvo-Porral, C.,Lévy-Mangin, J.P., 2018, From "foodies" to "cherry-pickers": A clustered based segmentation of specialty food retail customers, *Journal of Retailing and Consumer Services,* 43 (2018), S 278–284.

Chen, Y., Tang, K., Shen, R., Hu, Y., 2005, Market basket analysis in a multiple store environment, *Decision Support Systems,* 40 (2005), S 339– 354

Christy, A.J., Umamakeswari, A., Priyatharsini, L., Neyaa, A., RFM Ranking – An Effective Approach to Customer Segmentation, 2018, *Journal of King Saud University – Computer and Information Sciences*, xxx (2018), xxx–xxx(Baskıda).

Çakmak, Z., 1999, Kümeleme Analizinde Geçerlilik Problemi ve Kümeleme Sonuçlarının Değerlendirilmesi, *Dumlupınar Üniversitesi Sosyal Bilimler Dergisi*, 3 (2) S 190-191

Çoban, A., 2006, *İmalat Sanayinde Veri Madenciliği Destekli Tedarikçi Seçimi Uygulaması,* Doktora Tezi, Sakarya Üniversitesi Fen Bilimleri Enstitüsü.

Dolgun, M.Ö., 2006, *Büyük Alışveriş Merkezleri için Veri Madenciliği Uygulamaları*, Yüksek Lisans Tezi, Hacettepe Üniversitesi Fen Bilimleri Enstitüsü.

Döşlü, A., 2008, *Veri Madenciliğinde Market Sepet Analizi ve Birliktelik Kurallarının Belirlenmesi*, Yüksek Lisans Tezi, İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü.

Fung, G., Mangararian, O.L., 2002, Incremental Support Vector Machine Classification, *Proceedings of the Second SIAM Conference on Data Mining*, Arlington,Virginia,11-13 April 2002,SIAM, Philadelphia, S 247-260.

Goebel, M., Greunwald L., 1999, A Survey of Data Mining And Knowledge Discovery Software Tools, *SIKKDD Explorations,* June 1999, Newsletter, S 20-33.

Göral, M.A., 2007, *Kredi Kartı Başvuru Aşamasında Sahtecilik Tespiti İçin Bir Veri Madenciliği Modeli,* Yüksek Lisans Tezi, İ.T.Ü. Fen Bilimleri Enstitüsü.

Guidici, P., 2003, *Applied Data Mining: Statistical Methods For Business And Industry*, John Wiley and Sons Ltd., London England, ISBN 0-470-84678-X.

Han J., Kamber M., Pei, J., 2012, *Data Mining: Concepts and Techniques*, 225 Wyman Street Waltham MA 02451, USA, Morgan Kaufmann Publishers Inc., ISBN: 978-0-12-381479-1.

Hamka, F., Bouwman, H., Reuver, M., Kroesen, M., 2014, Mobile customer segmentation based on smartphone measurement, *Telematics and Informatics*, 31 (2014), S 220–227

Insani, R., & Soemitro, H. L., 2016, Data Mining For Marketing In Telecommunication Industry, *2016 IEEE Region 10 Symposium (TENSYMP)*, Bali,New Jersey 9-11 May 2016, IEEE, ISBN -13 : 978-1-5090-0931-2, S 179-183.

İşçi, Ö., 2007, Genetik Algoritma Yaklaşımı ve Yöneylem Araştırmasında Bir Uygulama, Celal Bayar Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi Yönetim ve Ekonomi, 10(2),S 191-208.

Jiménez, R., Anupol, J., Cajal, B., Gervilla, E., 2018, Data mining techniques for drug use research, *Addictive Behaviors Reports,* 8 (2018), S 128–135.

Kaspa, L.P., Akella, V.N.S.S., Chen Z., Shi Y., 2018, Towards Extended Data Mining: An Examination of Techniqual Aspects, *Procedia Computer Science*, 139(2018), S 49–55.

Kaur, M., Kang, S., 2016, Market Basket Analysis: Identify The Changing Trends of Market Data Using Association Rule Mining, *Procedia Computer Science,* 85 (2016), S 78 – 85

Khajvand, M., Zolfaghar, K., Ashoori, S., Alizadeh, S., 2011, Estimating Customer Lifetime Value Based on RFM Analysis of Customer Purchase Behavior: Case Study, *Procedia Computer Science*, 3 (2011), S 57–63.

Koçoğlu, F., Özcan, T., 2016, Veri Madenciliğinde Kümeleme Algoritmaları ile Müşteri Segmentasyonu, R ile Veri Madenciliği Uygulamaları, Balaban M.E.(ed.), Kartal, E.(ed.), 8, Çağlayan Kitabevi, Beyoğlu-İstanbul, ISBN: 978-975-436-093-6, S 187-215

Liu, Y., Li, Z., Xiong, H., Gao, X., Wu, J., 2010, Understanding of Internal Clustering Validation Measures, *2010 International Conference on Data Mining*, 13-17 Aralık, University of Technology Sydney, Sydney, NSW, IEEE, ISBN: 2374-8486, S 911-916.

Miguéis, V.L, Camanho, A.S., Junha, J.F., 2012, Customer Data Mining for Lifestyle Segmentation, *Expert Systems with Applications,* 39 (2012), S 9359–9366.

Mohammadzadeh, M., Hoseini, Z.Z., Derafshi, H., 2017, A data mining approach for modeling churn behavior via RFM model in specialized clinics Case study: A public sector hospital in Tehran, 2017, *Procedia Computer Science*, 120 (2017), S 23–30.

Müller, H., Hamm, U., 2014, Stability of market segmentation with cluster analysis – A methodological approach, *Food Quality and Preference*, 34 (2014), S 70–78.

Piramuth, U.S., 1999, Evaluating Feature Selection Methods For Learning In Data Mining Applications, *Thirty-First Annual Hawaii International Conference On System Sciences 5, 6-9 January 1998 Kohala Coast, Hawaii U.S.A.,* IEEE Computer Society, S 294.

Qadadeh, W., Abdallah, S., 2018, Customers Segmentation in the Insurance Company (TIC) Dataset, *Procedia Computer Science*, 144 (2018), S 277–290.

Özçalıcı, M, 2017, Veri Madenciliğinde Birliktelik Kuralları ve İkince El Otomobil Piyasası Üzerine Bir Uygulama, *ODÜ Sosyal Bilimler Araştırmaları Dergisi (ODÜSOBİAD),* 7(1), S 45-58

Sever, H., Oğuz, B., 2002, Veri Tabanlarında Bilgi Keşfine Formal Bir Yaklaşım, *Bilgi Dünyası,* 3, S 173-204.

Shearer, C., 2000, The CRISP-DM Model: The New Blueprint For Data Mining, *Journal of Data Warehousing,* 5(4), S 13-23.

Silvers, F., 2011, *Data Warehouse Designs*, CRC Press, Boca Raton-Florida, ISBN: 978-1-4398-7076-1.

Şimşek Gürsoy, U.T., 2009, *Veri Madenciliği Ve Bilgi Keşfi,* Pegem Akademi, Yenişehir-Ankara, ISBN: 978-605-4282-01-2

Şimşek Gürsoy, U.T., 2010, *Uygulamalı Veri Madenciliği Sektörel Analizler,* Pegem Akademi, Yenişehir-Ankara, ISBN: 978-605-364-040-0.

Tang, K., Chen, Y., Hu, H., Context-based market basket analysis in a multiple-store environment, 2008, *Decision Support Systems,* 45 (2008), S 150–163.

Tantuğ, A.C., 2002, *Veri Madenciliği ve Denetleme,* Yüksek Lisans Tezi, İ.T.Ü. Fen Bilimleri Enstitüsü.

Tezcanlar, P., 2007, *Müşteri İlişkileri Yönetimi Veri Madenciliği ve Bir Uygulama,* Yüksek Lisans Tezi, İ.Ü. Sosyal Bilimler Enstitüsü.

Tok, A., 2002, *Müşteri İlişkileri Yönetimi ve Veri Madenciliği,* Yüksek Lisans Tezi, Uludağ Üniversitesi Sosyal Bilimler Enstitüsü.

Yılmaz, E., 2006, *Kütahya İlinde Sosyal Sınıfların Belirlenmesi ve Veri Madenciliği İle Tüketici Profilinin Çıkarılmasına Yönelik Bir Uygulama,*Yüksek Lisans Tezi, Dumlupınar Üniversitesi Sosyal Bilimler Enstitüsü.

EKLER

EK 1. k-Ortalamalar Algoritmasının ve Kümeleme İndekslerinin Hesaplanmasının R Dili Kodları

```
#veri okuma

a=read.csv("tezdata.csv", header=TRUE)

#müşteri kodunun silinmesi

a[1]=NULL

#k-ortalamalar algoritmasının çalıştırılması

c=kmeans(a,10, iter.max=1000, nstart=1)[1]

#elemanlara atanan kümelerin veri seti ile birleştirilmesi ve yazdırılması

sonuc=cbind(a,c)

write.csv(sonuc, "sonuc.csv")

#kümeleme indekslerinin hesaplanması ve yazdırılması

intIdx <- intCriteria(as.matrix(a),c$cluster,"all")

write.csv(intIdx, "indeks.csv")
```

EK 2. Bulanık c-Ortalamalar Algoritmasının ve Kümeleme İndekslerinin Hesaplanmasının R Dili Kodları

```
#veri okuma

a=read.csv("tezdata.csv", header=TRUE)

#müşteri kodunun silinmesi

a[1]=NULL

#bulanık c-ortalamalar algoritmasının çalıştırılması

c=cmeans(a,5, iter.max=1000)[3]

#elemanlara atanan kümelerin veri seti ile birleştirilmesi ve yazdırılması

sonuc=cbind(a,c)

write.csv(sonuc, "sonuc.csv")

#kümeleme indekslerinin hesaplanması ve yazdırılması

intIdx <- intCriteria(as.matrix(a),c$cluster,"all")

write.csv(intIdx, "indeks.csv")
```

ÖZGEÇMİŞ

Kişisel Bilgiler	
Adı Soyadı	Barış Yıldırım
Doğum Yeri	Bakırköy / İSTANBUL
Doğum Tarihi	30.11.1988
Uyruğu	☑ T.C. ❒ Diğer:
Telefon	0536 916 11 30
E-Posta Adresi	baris_yildirim88@hotmail.com
Web Adresi	



Eğitim Bilgileri	
Lisans	
Üniversite	İSTANBUL ÜNİVERSİTESİ
Fakülte	İŞLETME FAKÜLTESİ
Bölümü	İŞLETME BÖLÜMÜ
Mezuniyet Yılı	2010

Yüksek Lisans	
Üniversite	İstanbul Üniversitesi-Cerrahpaşa
Enstitü Adı	Lisansüstü Eğitim Enstitüsü
Anabilim Dalı	Endüstri Mühendisliği Anabilim Dalı
Programı	ENDÜSTRİ MÜHENDİSLİĞİ PROGRAMI

Makale ve Bildiriler