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

BANKACILIK SEKTÖRÜNDE MÜŞTERİLERİN DAVRANIŞSAL SEGMENTASYONU İÇİN VERİ MADENCİLİĞİ UYGULAMASI

YÜKSEK LİSANS TEZİ

Kamil Balıkçı

Tez Danışmanı
Doç. Dr. Turgay Tugay Bilgin

İSTANBUL – 2018

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

BANKACILIK SEKTÖRÜNDE MÜŞTERİLERİN DAVRANIŞSAL SEGMENTASYONU İÇİN VERİ MADENCİLİĞİ UYGULAMASI

YÜKSEK LİSANS TEZİ

Kamil Balıkçı

Tez Danışmanı
Doç. Dr. Turgay Tugay Bilgin

İSTANBUL – 2018

T.C. Maltepe Üniversitesi
Fen Bilimleri Enstitüsü Müdürlüğüne,

28.12.2017 tarihinde tezinin savunmasını yapan Kamil BALIKÇI' ya ait "Bankacılık Sektöründe Müşterilerin Davranışsal Segmentasyonu İçin Bir Veri Madenciliği Uygulaması" başlıklı çalışma, Jürimiz Tarafından Fen Bilimleri Bilgisayar Mühendisliği Anabilim Dalı, Bilgisayar Mühendisliği Tezli Yüksek Lisans Programında Yüksek Lisans Tezi Olarak **Oy Birliği/~~Oy Çokluğuyla~~** Kabul Edilmiştir.

.....

Doç.Dr. Turgay Tugay BİLGİN
(Başkan)
(Danışman)

.....

Yrd. Doç. Dr. Volkan TUNALI
(Üye)

.....

Yrd. Doç. Dr. Mehmet Ali Aksoy TÜYSÜZ
(Üye)

YEMİN METNİ

28 /12/2017

Yüksek Lisans tezi olarak sunduğum "Bankacılık Sektöründe Müşterilerin Davranışsal Segmentasyonu için bir veri madenciliği uygulaması" adlı çalışmanın, proje safhasından sonuçlanmasına kadar olan bütün süreçlerinde bilimsel ahlak ve geleneklere aykırı düşecek bir yardıma başvurulmaksızın tarafımca yazıldığını ve yararlandığım bütün eserlerin "Kaynakça"da gösterilenlerden oluştuğunu, "Kaynakça"da yer alan bu eserlerden metin içinde atıf yaparak yararlanmış olduğumu belirtir ve onurumla doğrularım.

Öğrenci Numarası : 15 14 02 119
Adı-Soyadı: Kamil BALIKÇI

BANKACILIK SEKTÖRÜNDE MÜŞTERİLERİN DAVRANIŞSAL SEGMENTASYONU İÇİN VERİ MADENCİLİĞİ UYGULAMASI

ÖZET

Yüksek Lisans Tezi, Bankacılık sektöründe müşterilerin davranışsal segmentasyonu için veri madenciliği uygulaması, T.C. Maltepe Üniversitesi, Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Anabilim Dalı.

Günümüz rekabet koşullarının çetin olduğu ve hızla değiştiği bankacılık sektöründe, müşterinin beklentisini, ihtiyaçlarını, alışkanlıklarını, davranış biçimini, sosyo-ekonomik durumunu göz önünde bulundurarak müşteri profilleri oluşturmak ve bu müşteri profillerine uygun pazarlama yaklaşımları geliştirmek büyük önem kazanmıştır. Bunu başarmak için veri madenciliği yöntemlerinden yararlanılarak müşterilerin farklı özelliklerine göre gruplara bölme ve sınıflandırma çalışmaları yapılmaktadır.

Bu tez çalışmasında, müşterilerin davranış özelliklerine dayalı olarak gruplara ayırma işlemi yapılmıştır. Özel bir Türk bankasının bireysel müşterilerinin ürün kullanımı, kanal tercihi ve işlem tercihlerini gösteren nitelikler incelenerek, bankanın tüm müşterilerinin davranış profilleri veri madenciliği yöntemleri ile ortaya çıkarılmaya çalışılmıştır. Müşterileri gruplara ayırma işleminde; karışık veri tipli ve çok boyutlu veri kümelerini işleyebilen, birebir küme ataması yapabilen, çalışma süresi diğer yöntemlere göre daha kısa olan k-prototip algoritması kullanılmıştır. Çalışmada veri kümesinin oluşturulması, temizlenmesi, dönüştürülmesi, k-prototip algoritmasının uygulanması, sonuçların yorumlanması ve kaydedilmesini sağlayan bir uygulama geliştirilmiştir.

Bu tez çalışması, bankacılık sektöründe etkin kampanya yönetimi yapılabilmesi ve müşterilere doğru ürünü, doğru kanaldan sunulabilmesine önemli katkı sağlayacağından, bu alanda çalışacak araştırmacılar için referans niteliği taşıyacaktır.

Anahtar Kelimeler: Veri Madenciliği, Kümeleme Analizi, Davranışsal Segmentasyon, k-prototip algoritması.

A DATA MINING CASE STUDY FOR BEHAVIORAL SEGMENTATION OF CUSTOMER IN BANKING SECTOR

ABSTRACT

Master's Thesis, A case study of data mining for behavioral segmentation of customers in Banking Sector, T.C. Maltepe University, Institute of Sciences, Department of Computer Engineering.

In the banking sector, where today's competitive conditions are difficult and rapidly changing, it has become very important to create customer profiles taking into consideration customer expectations, needs, habits, behaviors and socio-economic situation, and developing appropriate marketing approaches to these customer profiles. For achieving these, data mining methods are used to divide and classify groups according to different characteristics of the customers.

In this dissertation, customers are divided into groups according to the behavioral characteristics. In a private Turkish bank, by examining the attributes of their individual customers' product usage, channel and transaction preferences, the behavior profiles of all the customers of the bank have been tried to be revealed by data mining methods. In the process of separating customers into groups k-prototype algorithm has been used. This algorithm can process multidimensional and large data sets with mixed data types. Also, the working time of k-prototype algorithm is shorter than the other comparable methods. In this study, an application has been developed which is capable of cleaning and transformation of data set, interpretation, and storing the results.

This thesis will be a reference for researchers who will work on this area because it will make an important contribution to effective campaign management in the banking sector and to present the correct product to the customer from the right channel.

Keywords: Data Mining, Clustering Analysis, Behavioral Segmentation, k-prototype algorithm.

ÖNSÖZ

Tez konumu seçmeme yardımcı olan, çalışmaya teşvik eden, bu süreçte yol göstericiliğini ve bilgisini benden esirgemeyen değerli danışman hocam Doç. Dr. Turgay Tugay Bilgin'e, sonsuz teşekkürlerimi sunarım.

Yüksek lisans dönemi boyunca eğitimime verdikleri destek ve yardımları için çalışma arkadaşlarıma, sevgisi, fedakârlıkları ve ilgisiyle attığım her adımda arkamda olan öncelikle eşim Serpil'e ve aileme saygı, sevgi ve sonsuz teşekkürlerimi sunarım.

Yüksek lisans dönemi boyunca ders seçimlerimde, seminer dersimde desteklerini esirgemeyen değerli hocalarım Yrd. Doç. Dr. Ali Akman ve Yrd. Doç. Dr. Erdal Güvenoğlu'na teşekkürlerimi sunarım.

Bugünlere gelmemi sağlayan, eğitimim ve ileri yürümemde bana devamlı destek olan, sevgili anne ve babama teşekkür ederim.

Ocak – 2018 Kamil Balıkçı

İÇİNDEKİLER

ŞEKİLLER DİZİNİ

TABLOLAR DİZİNİ

DENKLEMLER DİZİNİ

SİMGELER DİZİNİ VE KISALTMALAR

AGNES : Agglomerative NESting

ACO : Ant Colony Optimization (Karınca Kolonisi Optimizasyonu)

ATM : Automatic Teller Machine (Bankamatik)

BIRCH : Balanced Iterative Reducing And Clustering Using Hierarchies (Hiyerarşik Kümeleme Algoritması)

CHAMELEON: Hierarchical Clustering Using Dynamic Modelling

CLARA : Clustering LARge Applications

CLARANS : Clustering Large Applications based upon RANdomized Search

CLIQEU : CLustering In QUEst

CLUSTER : Küme

CLV : Customer Lifetime Value (Müşteri Yaşam Değeri)

COBWEB : Incremental system for hierarchical conceptual clustering

CRM : Customer Relationship Management (Müşteri İlişkileri Yönetimi)

CURE : Clustering Using Representatives

DBSCAN : Density Based Spatial Clustering of Applications with Noise

DENCLUE : DENsity Based CLUstering

EM : Self-Organizing Maps (Beklenti En Çoklama) Algoritması

HAC : Hierarchical Aggloramative Clustering (Birleşmeli Hiyerarşik Kümeleme)

KNIME : Konstanz Information Miner v.3.3.2 (Açık Kod veri madenciliği aracı)

OPTICS : Ordering Points To Identify The Clustering Structure

PAM : Partioning Around Medoids

R : R (Programlama Dili)

RFM : Recency, Frequency, Monetary (Yenilik, Sıklık, Parasal)

RFML : Recency, Frequency, Monetary, Length (Yenilik, Sıklık, Parasal, Uzunluk)

ROCKS : RObust Clustering using linKs

ROI : Return On Investment (Yatırım Getirisi)

SOM	: Self-Organizing Maps (Öz düzenleyici Haritalar)
SPSS	: Statistical Package for the Social Sciences (Bilgisayar Uygulaması)
SQL	: Structured Query Language (Yapılandırılmış Sorgu Dili)
SSIS	: SQL Server Integration Services (SQL Sunucu Bütünleştirme Servisleri)
STING	: STatistical INformation Grid
SVM	: Support Vector Machine (Destekçi Vektör Makinesi)
TMCM	: Two-step Method for Clustering Mixed (Karışık veri türleri kümeleme için iki adım yöntemi)
WaveCluster	: Clustering Using Wavelet Transformation

1. GİRİŞ

Giriş bölümünde; tez çalışmasının amacı, problemin tanımı, bu çalışmanın katkıları ve tezin düzeni hakkında bilgilere yer verilmiştir.

1.1. Tez Çalışmasının Amacı

Müşteriler, ticari işletmenin en önemli varlıklarıdır. İyi ilişkiler kurulmuş, bu ilişkileri geliştirilmiş, memnuniyeti sağlanmış müşteriler olmadan hiçbir şirketin ticari bir geleceği olmayacaktır. Bu yüzden firmalar, ürün odaklı pazarlama yerine müşteri merkezli pazarlama stratejileri geliştirmeye başlamıştır. Ticari işletmeler, müşterilerine nasıl davranacaklarına dair açık bir strateji planlamalı ve işletmelidirler. Bunu yürütebilmek için CRM (Customer Relationship Management-Müşteri İlişkileri Yönetimi) kavramı ortaya çıkmıştır. CRM; müşterinin uzun süre işletme ile çalışması ve işletmeye sadık müşteriler olması için müşteri ilişkilerinin inşa edilmesi, yönetilmesi ve sürdürülmesi stratejisidir. Bu stratejinin kapsamı; müşterinin farklı ihtiyaçlarını, tercihlerini, davranışlarını belirleyerek ve anlayarak, her müşterinin birbirinden farklı olarak ele alınması ve değerlendirilmesidir [1].

CRM' in iki ana amacı vardır:

1. Müşteri memnuniyeti sağlayarak müşteriyi kaybetmemek,

2. Müşteri iç görüsü sağlayarak müşteriyle ilişkileri geliştirmektir.

CRM stratejilerini geliştirmek için kullanılan en önemli ve temel araç, müşterilerin segmente (bölümlere ayrılması) edilmesidir. Müşteri segmentasyonunun amacı, farklı ihtiyaç ve tercihleri olan müşterileri farklı gruplar hâlinde ele alıp, bu gruplara özel pazarlama stratejileri geliştirmek için ön koşul olan müşteri gruplarını tespit etmektir.

Tez çalışmasında, veri madenciliği yöntemlerini kullanarak müşterilerin davranışlarını baz alan bir segmentasyon çalışması yapılmıştır.

Tez, bankaların pazarlama stratejileri geliştirirken ihtiyaç duyacakları; hedef kitle seçimi, ürün/hizmet fiyatlandırması ve sundukları ürün/hizmetlerin müşteriye özel olarak kişiselleştirilmesine yardımcı olacaktır. Bununla beraber segmentasyon çalışması yürütenlere de referans niteliği taşıyacaktır.

1.2. Problemin Tanımı

Günümüzde her müşteriye tek bir ürün sunumu yapılan ürün odaklı pazarlama anlayışı (mass marketing), yerini müşteri merkezli pazarlama anlayışına bırakmıştır. Müşteri merkezli pazarlama anlayışını gerçekleştirmek için şirketlerin müşteri iç görüsünü (customer insight) elde etmesi gerekmektedir. Müşteri iç görüsü; müşterinin hangi ürünü ve hizmeti aldığının tespit edilmesinden daha ileri bir seviye olan müşteri hangi ürünü neden aldığı bu ürünle ne tür ihtiyacını karşıladığını, seçtiği ürünü hangi dinamiklere göre seçtiğinin kavranmasıdır. Müşteri iç görüsünü elde eden, sürekli ölçen, analiz eden ve bu doğrultuda pazarlama stratejileri geliştirebilen firmalar rekabette bir adım öne geçmektedir. Müşteri iç görüsünü elde ettikten sonra doğru müşteriye doğru kanaldan doğru ürünü sunmak gerekmektedir.

Müşteriler aşağıdaki amaçlara ulaşmak için gruplara (segmentlere) ayrıştırılırlar:

1. Müşteri iç görüsünün elde edilmesi, doğru müşteriye doğru kanaldan ve doğru ürünün sunulması için müşterilerin ihtiyaç, tercih, gereksinim, kullanım, davranış gibi karakteristiklerini ortaya çıkarmak.

2. Benzer özellikli müşteriler için ayrı ayrı özel ürün/hizmet tasarlamak ve sunmak, özel fiyatlamalar/oranlar tespit ve teklif etmek, müşteri grubuna özgü iletişim kanalları oluşturmak.

3. Oluşturulan iletişim kanalları vasıtasıyla iletişimi sürdürmek, çeşitli ödüller ve teşviklerde bulunmak, firma kaynaklarını etkin yönetmek.

Firmaların, bu amaçlara ulaşabilmesi için müşterilerini gruplara ayırma ve organizasyonlarını şekillendirme ihtiyaçları bulunmaktadır [1]. Bu amaçları karşılayabilmek adına müşteriler davranışsal segmentlere ayrılacaktır.

1.3. Tez Çalışmasının Katkıları

Müşteri merkezli pazarlama stratejilerini hayata geçirebilmek adına müşterilerin segmente edilmesi gerekmektedir. Pazarlama stratejilerini gerçekleştirmede müşterilerin farklı özelliklerine dayalı farklı segmentasyon türleri kullanılabilmektedir.

Tez çalışmasında, özel bir Türk bankasının bireysel müşterilerinin ürün kullanım sıklığı, kullanım yeri, kullandığı ürün türü vb. davranışlarını temel alan müşteri veri tabanının tümüne uygulanan davranışsal segmentasyonu ele alınmıştır. Çalışma sonunda bankanın tüm bireysel müşterilerinin davranış segmentleri tespit edilmiş ve her bireysel müşteri tespit edilen bu davranış segmentlerine atanmıştır.

Çalışma sayesinde banka ilgili davranışsal segmentlerine özel kampanyalar tasarlayabilecektir. Yapılan kampanyalara müşterilerin olumlu cevap verme oranı artarken, odaklanılmış bir müşteri grubu üzerinde kampanyalar tasarlandığından kampanya daha az bir maliyetle sonuçlanacaktır. Böylece banka, etkin bir kampanya yönetimi imkânına kavuşacaktır.

Bunun yanında banka, davranışsal segmentlere ayrılmış müşterilerinin ihtiyaçlarına ve davranışlarına uygun özel fiyatlandırma stratejileri de belirleyebilecektir.

1.4. Tez Düzeni

Bu tez çalışması, beş ana bölümde sunulmaktadır; Bölüm 1' de tezin amacı, problemin tanımı ve tezin katkılarının bulunduğu Giriş kısmı yer almıştır. Bölüm 2'de Müşteri İlişkileri Yönetimi ve Segmentasyon kavramı, segmentasyon türleri, bankacılıkta kullanım alanları hakkında bilgiler sunulmuştur. Bölüm 3'te yapılan Kümeleme Analizi ve Karışık Veri Tipli Kümelerde kullanılan teknikler hakkında bilgi verilmiştir. Bölüm 4'te Davranışsal Segmentasyon hakkında yapılmış ilgili çalışmalar hakkında bilgiler sunulmuştur. Bölüm 5'te Davranışsal Segmentasyonun yapıldığı Veri Madenciliği Uygulamasına yer verilmiştir. Bölüm 6'da yapılan tez çalışmasının sonuçları ve katkıları değerlendirilmektedir.

2. MÜŞTERİ İLİŞKİLERİ YÖNETİMİ ve SEGMENTASYON

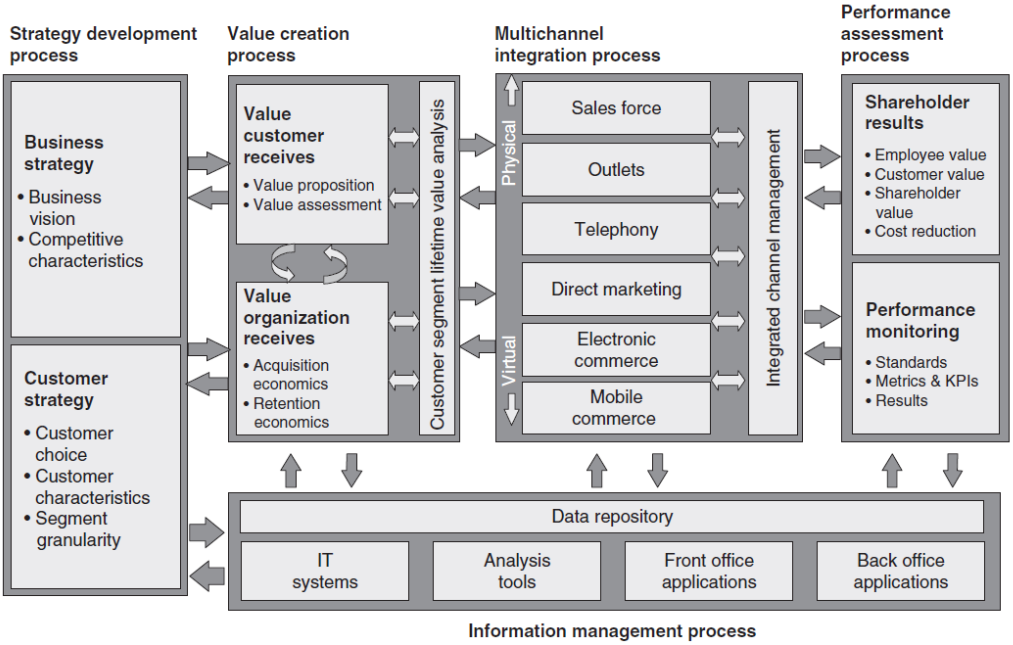
Bu bölümde; müşteri ilişkileri yönetimi, veri madenciliği ve segmentasyon arasındaki ilişki incelenmiş, segmentasyon çeşitleri açıklanmıştır.

2.1. Müşteri İlişkileri Yönetiminin Tanımı

Müşteri İlişkileri Yönetimi; (CRM) şirket kârlılığını artırma ve müşteri değerini maksimize etmek için özenli bir şekilde müşteri ilişkisi oluşturma ve oluşturulan müşteri ilişkilerini geliştirmenin yollarını arayan bir yönetim yaklaşımıdır. Ayrıca CRM, ilişki yönetim stratejilerini yerine getirmek için bilgi teknolojilerinin kullanımı ile de ilgilidir.

CRM tanımı 1990'ların ortalarında IT tedarikçileri ve bu ürünleri kullanan çevre tarafından ortaya atılmıştır. Bu terim, satış otomasyonu gibi teknoloji tabanlı müşteri çözümlerini tanımlamak için kullanılmıştır. Akademik çevrede ilişki yönetimi ve CRM sıklıkla birbirlerinin yerine kullanıldı [2]. Ancak bu kavram daha yaygın şekilde teknolojik çözümler bağlamında kullanıldı ve CRM terimi Ryals ve Payne tarafından bilgi ile geçerlilik kazanan ilişki pazarlaması olarak tanımlandı [3].

Stratejik bakış açısı ile CRM'in yalnızca müşteri tabanını büyütmek ve yeni müşteri kazanmaya yarayan bir IT çözümü olmadığını aksine çoklu kanal içinde müşteri değerinin kurum bütününde anlaşılması, uygun bilgi yönetimi ve CRM uygulamalarının kullanımını, yüksek kalitede operasyonların ve hizmetlerin yerine getirilmesini sentezleyen stratejik vizyon olarak ifade edilmiştir. Bu bağlamda 2005 yılında Payne ve Frow, CRM'i kritik müşteri ve müşteri segmentleri ile uygun ilişkileri geliştirerek tüm paydaşlardan (müşteri, personel, tedarikçi vb.) değer yaratmayı amaçlayan stratejik bir yaklaşım olarak tanımladı [4]. Bunun için stratejik bir çerçeve oluşturdular. Oluşturulan stratejik çerçeve Şekil 1'de yer almaktadır.



Şekil 1. CRM'in Stratejik Çerçevesi [4].

CRM yaklaşımının bir firmada yerine getirilebilmesi için CRM'in bileşenleri olan süreç, teknoloji ve insan bileşenlerinin strateji doğrultusunda bir bütün şeklinde entegre çalışabiliyor olması gerekmektedir.

Stratejik çerçevedeki çalışmaların yapılabilmesi için firmaların süreçlerini bu stratejik hedefleri gerçekleştirebilmesi için düzenlemesi, CRM'i operasyonel sistemlerine taşıması gerekmektedir. Bu operasyonların ve süreçlerin işlevlerini yerine getirebilmesi için müşteriye ait bilgilerin analitik olarak incelenmesi ve sonuçlarının operasyonel sistemlerde kullanılması sağlanmalıdır.

CRM'in özellikle analitik kısımlarındaki çalışmalarda; veri madenciliği, makine öğrenmesi algoritmaları kullanılmaktadır. Analitik çalışmaların sonuçları, operasyonel ve pazarlama faaliyetlerini şekillendirmektedir.

2.2. CRM ve Veri Madenciliği

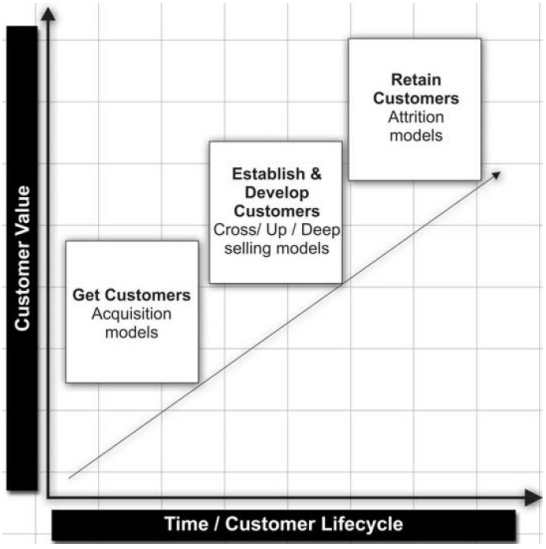
Firmalar, yoğun rekabet ve küresel pazarda hayatta kalmak için müşteri ile daha yakın bir ilişki içinde olması gerektiğini fark etti. İyi müşteri ilişkileri tesis etmek, firma kârlılığını üç şekilde etkileyip, arttırabilmektedir [5]:

1. Uygun ve değerli müşterileri firmaya çekerek (Get Customers) maliyeti azaltmak.

2. Çapraz, Yukarı ve Derin satış (Establish & Develop Customers) aracılığıyla kâr oluşturmak.

3. Mevcut Müşterilerini Tutundurma (Retain Customers) ile kârlılığı büyütmek.

Müşteri ile firma arasındaki ilişkinin yaşam döngüsü ve firmanın müşteriden elde edeceği değeri gösteren grafik Şekil 2. Müşteri Yaşam Döngüsü, Değeri ve Pazarlama Modelleri adlı şekilde yer almaktadır. Müşteri yaşam döngüsünün her safhasında uygulanması gereken veri madenciliği modelleri yine aynı şekil üzerinde gösterilmiştir.



Şekil 2. Müşteri Yaşam Döngüsü, Değeri ve Pazarlama Modelleri [1].

CRM; satış, pazarlama ve müşteri hizmeti gibi müşteri merkezli süreçleri geliştirerek daha yüksek yatırım getirisini (ROI-Return On Investment) vadeder. Veri madenciliği müşteri yaşam döngüsü boyunca müşteri ihtiyaçlarını tahmin ederek ve belirleyerek kârlı müşteri ilişkilerini oluşturmakta firmalara yardımcı olur [5].

2.3. Veri Madenciliği ve Kampanya Yönetimi İlişkisi

Pazarlamacılar yüksek gelir potansiyeli olan müşteri gruplarının satın alma kararlarını etkileyen ve sağlayan kampanyalar tasarlar ve hayata geçirirler. Pazarlamacılar bunu yapmak için veri madenciliği çıktılarını belirli pazar veya müşteri segmentine odaklanan kampanya yönetim yazılımlarına girdi olarak sunarlar. Burada veri madenciliğinin CRM faaliyetlerini desteklediği 3 yol vardır.

1. **Veri tabanı Pazarlama:** Veri Madenciliği; veri tabanı pazarlamacılarının, müşterinin tutum, arzu, istek ve ihtiyaçlarına daha uygun kampanyalar oluşturmasına yardımcı olur. Eğer gerekli bilgiler veri tabanında tutuluyorsa, veri madenciliği müşteri aktivitelerini modelleyebilir. Burada kritik amaç, iş problemi ile ilgili örüntüleri (patterns) belirlemektir. Örneğin veri madenciliği "Otomatik Fatura Ödeme talimatını iptal etmesi muhtemel müşterilerim hangisidir?" ve bir müşterinin "A şubesine 1,000 TL'den daha yüksek para yatırma olasılığı nedir?" şeklindeki sorulara cevap bulunmasına yardımcı olabilir. Bu tür soruların cevabını bulmak müşteri tutundurma ve kampanya cevaplanma oranlarınızı arttırabilir. Bu da nihayetinde satışların ve yatırım getirinizin artması ile sonuçlanır.

2. **Müşteri Kazanımı:** Firmaların büyüme stratejileri çoğunlukla yeni müşteri kazanımlarına bağlıdır. Bunun içinde çeşitli ürün ve hizmetlerden habersiz veya rakiplerinizden ürün veya hizmet alan insanları bulmanız gerekir. Veri madenciliği, müşteri edinim kampanyalarına cevap oranlarını arttırmak için birçok segmentasyon çözümü sunabilmektedir.

Pazarlamacılar veri madenciliği aracılığıyla belirlenen müşterilere yeni ve ilgi çeken teklifler sunma imkânına kavuşmaktadır.

3. **Kampanya Optimizasyonu:** Çoğu pazarlama bölümlerinin var olan ve potansiyel müşterileri ile temas kurmasını sağlayan çeşitli yöntemler vardır. Bir pazarlama kampanyasını optimize etme süreci, firmanın sunduğu teklif ile kampanyanın karakteristiği, kısıtları, kullanılacak pazarlama kanalı vb. parametrelere cevap verebilen müşteri kümesinin eşleştirilmesi ile sağlanır. Veri madenciliği, müşterilerin pazarlama tekliflerine kanala özgü cevaplarını modelleyerek kampanya optimizasyon süreçlerinin etkinliğini arttırabilir.

Kampanya yönetimi ve veri madenciliğinin yakın bir şekilde kullanımı daha iyi iş sonuçlarına ulaşılmasını sağlamaktadır. Örneğin; Kampanya yönetim uygulamaları hedef müşterileri belirlemek için veri madenciliği skorlarını kullanılabilir. Böylece kampanya etkinliğini ve cevaplanma oranlarını arttırılabilir [5].

Veri madenciliği firmalara müşteri bilgilerini yönetmek, müşterilerini elde tutmak, yeni müşteri kazanmak ve kârlı müşteri ilişkilerini sürdürmek için yöntem sunan modern bir teknolojidir.

2.4. Segmentasyon

Veri madenciliği yöntemleri içinde en yoğun ve yaygın kullanılan tekniklerden biri müşteri/pazar segmentasyonu çalışmalarıdır.

Pazarlama açısından segmentasyon; müşteri/pazarın karakteristiği, farklı ihtiyaçlarına ve davranışlarına göre farklı pazarlama stratejileri geliştirmek adına ayrı ve homojen gruplara bölümleme sürecidir [6].

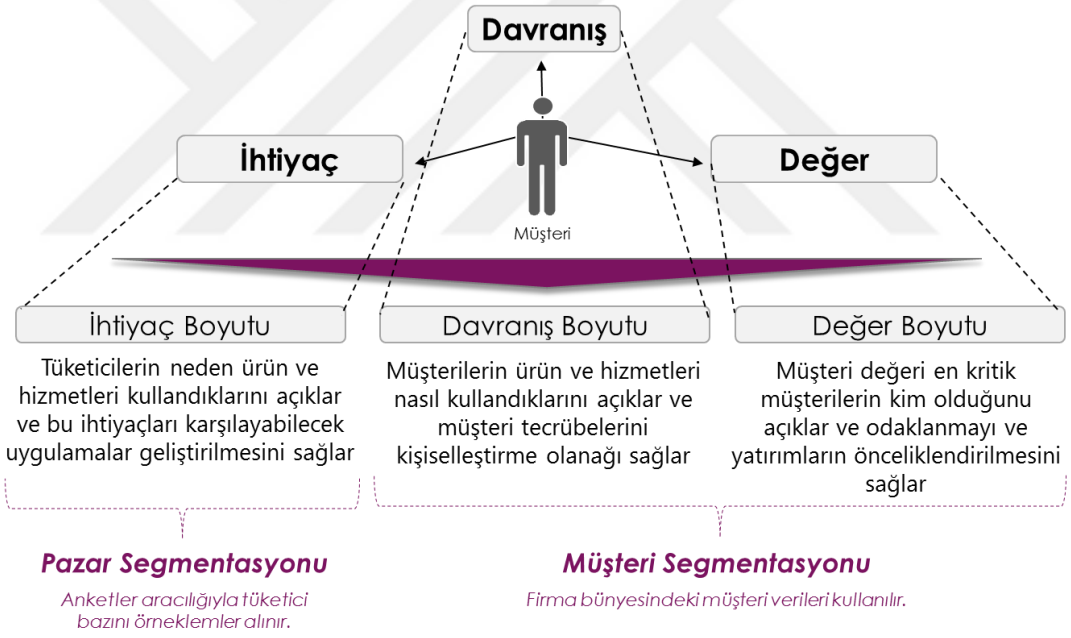
Segmentasyonun amacı, ürün, hizmet ve pazarlama faaliyetlerini her segmente özel bir şekilde uyarlamaktır.

Pazarlamada perspektifi göre 2 tür segmentasyon kavramı kullanılmaktadır.

1. Müşteri Segmentasyonu

2. Pazar Segmentasyonu

Müşteri segmentasyonunda bir firma kendi müşterilerini tanımak ve uygun pazarlama stratejileri geliştirmek için segmentasyona tabii tutarken, pazar segmentasyonunda ise potansiyel pazarda yer alan alıcı, tüketici ihtiyaçlarını görmek ve bu ihtiyaçları karşılayabilecek uygulamalar ve stratejiler geliştirmek için tüketicileri segmente etmeye çalışır. Şekil 3'te segmentasyon perspektifine yer verilmiştir.



Şekil 3. Segmentasyon Perspektifi.

Pazar segmentasyonu çalışmalarında veriler genellikle anket yolu ile toplanırken, müşteri segmentasyonuna ait veriler genellikle firmaların bünyesinden temin edilmektedir.

2.4.1. Segmentasyon Türleri

Pazarlamacılar; bir pazarı/müşteriyi segmentlere bölmek için kişi, organizasyon ve grupların özelliklerini kullanırlar.

Kullanılan segmentasyon kriterlerine göre segmentasyon türleri ortaya çıkmıştır. Müşteriler/tüketiciler özellikle değerleri, sosyoekonomik, yaşam evresi, davranış, ihtiyaç ve tutumlarına göre segmente edilmektedir.

Tez konusu içinde yer alan bireysel müşteriler/tüketiciler tüketici pazarını teşkil etmekte olup, bu pazardaki tüketicilerin segmentasyonunda müşterinin/tüketicinin ihtiyaç, istekleri ve gereksinimleri göz önüne alınmaktadır. Tüketici pazarlarındaki segmentasyon türleri segmentasyon temelini oluşturan tüketici özelliklerine göre yapılan ayrıştırmaya göre çeşitlenmektedir [7].

Aşağıdaki segmentasyon türleri tüketici piyasalarında yoğun şekilde kullanılmaktadır.

2.4.2. Değer Bazlı (Value-based) Segmentasyon

Müşteriler, değerlerine göre gruplanmaya çalışılır. En değerli müşterileri belirlemek ve zamanla değerlerindeki değişiklikleri izlemek ve takip etmek için en yaygın kullanılan segmentasyon türlerinden biridir. Pazarlama girişimlerinde kaynak tahsisini optimize etmek ve hizmet sunma stratejilerini farklılaştırmak için de kullanılır [6].

Bu tür segmentasyonda müşterinin değeri belirlenmeye çalışılır ve bu değer müşterinin şu andaki değeri ve potansiyel değerinin toplamından oluşur. Bu segmentasyona ait yapılan çalışmalarda müşterinin yaşam döngüsü değeri (CLV-Customer Life Cycle Value) kavramının kullanıldığı görülmektedir. Bu sayede müşterinin firmadaki değeri ile aslında toplam değerini izleme şansına sahip olunup, firma ile ilişkileri potansiyelinden daha az olan müşteri grubu tespit edilerek, bu segmentteki müşterilerin üzerine yoğunlaşacak pazarlama stratejileri geliştirilebilir.

2.4.3. Coğrafi Segmentasyon

Coğrafi segmentasyon, bütün Pazar/müşterilerin farklı ülkeler, eyaletler, bölgeler, ilçe, kasaba veya sokaklar gibi farklı coğrafi bölge birimlerine bölünmesidir. Firmalar tüm bölgelerde iş yapmak ya da birkaç bölgede faaliyetlerini sürdürmeye karar verebilir, fakat farklı bölgelerdeki farklı ihtiyaç ve tercihlere çok daha fazla dikkat etmelidirler. Farklı bölgelerin farklı gelenek, görenek ve alışkanlıkları vardı. Bu yüzden bu yerel farklılıklara göre pazarlama stratejilerini sürdürmesi gerekmektedir.

Günümüzde çoğu firma bölge, şehirlerin ihtiyaçlarına uyacak şekilde ürünlerini, reklamlarını, promosyonlarını yerleştirme isteğindedirler [7]. Bu amaçlara ulaşabilmek için coğrafi segmentasyon kullanılabilmektedir.

2.4.4. Sosyo-demografik Segmentasyon

Sosyo-demografik segmentasyonda, Pazar/müşteri yaş, cinsiyet, gelir, meslek, eğitim, din, dil, ırk, millet, yaşam evresi gibi özelliklere göre farklı gruplara ayrıştırılır. Bu değişkenler tüketici/müşteri gruplarını ayrıştırmak için en yaygın kullanılan özelliklerdir. Diğer segmentasyon türlerinde kullanılan değişkenlere göre sosyo-demografik segmentasyon değişkenlerini ölçmek daha kolaydır [7].

Bu tür segmentasyon yaşam evresi pazarlama desteklemenin yanında yaşam evresine dayalı belirli ürünleri desteklemek için de uygundur [6].

2.4.5. Eğilim (Propensity-based) Segmentasyon

Müşteriler, kaybetme skorları (churn scores), çapraz satış skorları (cross selling scores) gibi eğilim skorlarına göre gruplanırlar. Bu eğilim skorları diğer segmentasyonlarla birleştirilerek te kullanılabilmektedir. Örneğin; riske maruz değer (value at risk) segmentasyonu, elde tutma aksiyonlarını önceliklendirmek amacıyla değer

segmentasyonu (value-based segmentation) ile kaybetme eğilimini (churn propensity) birleştirerek geliştirilir [6].

2.4.6. Sadakat Bazlı (Loyalty-based) Segmentasyon

Bu segmentasyon türünde müşterinin sadakat durumunun araştırılması yapılır. Sadık (loyal) ve göç etmiş (migrator) şeklinde segmentler belirlenir. Sadık olan müşterilere ürün teklifinde bulunulurken, sadık olmayan profilli yüksek değerli müşterilere odaklanan elde tutma aksiyonları alınabilir [6].

2.4.7. İhtiyaçlar/Tutumlar (Needs/Attitudinal) Segmentasyon

Bu segmentasyon türü pazar araştırmalarında kullanılır. Müşteri/tüketicilerin ihtiyaç, istek, tutum, tercih ve firmanın ürün ve hizmetlerine bağlı algılarına göre segmentlere ayırır. Kritik ürün özelliklerini ve marka imajını saptamak ve yeni ürün geliştirmeyi desteklemek için kullanılır [6].

2.4.8. Davranışsal (Behavioral) Segmentasyon

Bu segmentasyon türü; Müşterinin/Tüketicinin ürüne verdiği cevaplar, ürün sahipliği, ürün kullanımı, ürüne karşı tutumları temelinde yapılan bölümleme çalışmasıdır.

Bu segmentasyon için gerekli olan ürün sahipliği ve kullanım verileri firmaların genellikle veri tabanlarında mevcut olduğundan dolayı verileri elde edilmesi zor olmamasından kaynaklı olarak yaygın bir şekilde kullanılan segmentasyon türüdür. Bu tür segmentasyon müşteriye özel ürün, teklif, yeni ürün ve müşteri sadakatini geliştirmek için kullanılmaktadır [6].

Bu tez çalışmasında, bireysel müşteri/tüketiciler segmente edilmiştir. Müşteri/tüketicilerin davranışsal niteliklerine dayanan davranışsal segmentasyon çalışması yürütülmüştür.

2.5. Profilleme

Bir firma, muhtemel tüm müşterileri eşit olarak hedeflemek ya da herkese aynı teşvik edici teklifleri sunmak yerine, müşterilerinin kişisel ihtiyaçlarına ve satın alma özelliklerine göre kârlı bir grubu hedefleyebilir. Bireylerin demografik özelliklerine, yaşam tarzlarına, geçmiş davranışlarına göre ileri dönem değerini tahmin eden bir model yapılarak bu başarılabilir. Oluşturulan model en kârlı müşteri grubunu korumak ve arttırmak için müşteriyi hatırlama ve kaydetmeye odaklanır. Buna müşteri davranışı modelleme ya da müşteri profilleme denilmektedir. Müşteri profilleme pazarlama birimleri için müşterilerinin özelliklerini daha iyi tanımayı ve anlamayı sağlayan araçtır. Müşteri profilleme, pazarlamacıların mevcut müşterilerine daha iyi bir servis sunabilmesi ve bu müşterileri ellerinde tutabilmesi için temel araçtır [8].

Müşteri profilleme, müşterileri yaş, gelir ve yaşam tarzı gibi niteliklerine göre tanımlamaktır. Profilleme demografik ve davranışsal müşteri bilgisinin bir araya getirilmesiyle oluşturulur. Müşteri profilleme, mevcut veriye bağlı olarak yeni müşterileri tahmin etmekte veya mevcut kötü müşterileri belirlemede kullanılabilir. Aynı zamanda müşteri profili çıkartılması ile benzer satın alma özelliklerine sahip müşterilerin gruplara ayrılması gerçekleştirilebilir. Amaca bağlı olarak, projeye uygun olan profil seçilmelidir [9].

Müşteri profillemenin uzun dönemde getirisi; müşteri tanıma işlemini otomatik bir etkileşime dönüştürmesindedir. Bu iş için günümüz pazarı birçok alanda süreçlere ve bilgi teknolojisi araçlarına ihtiyaç duymaktadır. Bu araçlar veriyi bir araya getirmede, pazarla ilgili bilgiye ulaşma işlemini basitleştirmede ve pazarlama kampanyalarını planlamakta kullanılmaktadır [9].

3. KÜMELEME ANALİZİ

Bu bölümde tezde uygulanan kümeleme (clustering) analizi ve uygulanması muhtemel olan kümeleme analizi türleri hakkında bilgiler sunulmuştur.

Kümeleme eğitimsiz bir öğrenme (unsupervised learning) yöntemidir. Eğitimsiz öğrenmede örneklerin/nesnelerin gözlenmesi ve bu örneklerin/nesnelerin özellikleri arasındaki benzerliklerden hareketle sınıfların tanımlanması yapılmaktadır [29]. Heterojen olan nesneler kümesinden homojen özellikli nesneleri barındıran alt grupların elde edilmesi işlemidir.

Veri madenciliğinde kullanılan teknikler veri türü ve elde edilen sonuçların kullanım amacına göre 2'ye ayrılabilir. Bunlar Tahmin Etme (predictive) ve Tanımlayıcı (descriptive) modellerdir.

Kümeleme analizi ise uzun zamandan beri pek çok alanda yaygın olarak kullanılan veri içerisindeki grupları tespit etmeye yarayan bir tanımlama (description) çalışmasıdır. Kümeleme Analizi, "Veri içerisindeki grupları bulma sanatıdır" [30].

Kümeleme analizinin temel amacı örnekleri sahip oldukları karakteristik özellikleri temel alarak gruplandırmaktır. Bu yöntem özellikle bilim ve iş alanında birçok durumda uygulanabilen, en kolay yorumlanabilen ve en etkili olan yöntem olma özelliğini taşır. Bu nedenle hemen hemen tüm bilim alanlarında bu yöntemden yararlanılmaktadır [31].

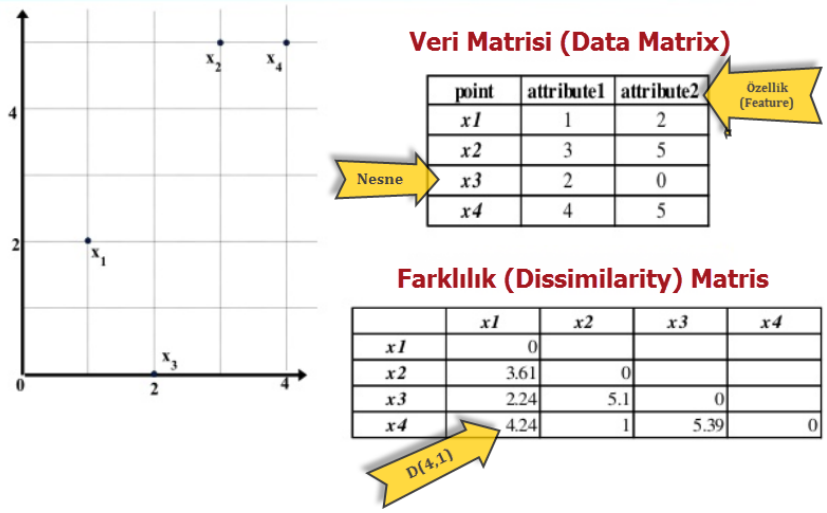
Kümeleme analizinin, örüntü tanıma, metin madenciliği, değişken seçimi (feature selection), müşteri segmentasyonu, pazar araştırmaları, veri özetleme, çoklu ortam veri analizi (multimedia data analysis), biyolojik veri analizi, diğer veri madenciliği ve makine öğrenmesi çalışmalarını kolaylaştırıcı ara işlemlerde kullanımı, sosyal ağ analizi vb. birçok yaygın uygulama alanı bulunur [32].

3.1. Kümeleme Analizinde Kullanılan Veri Yapısı

Kümeleme analizine başlamadan önce veriler matris formuna getirilir. Matris formu bilgisayar ortamında hesaplama yapabilmek için en uygun veri yapısıdır. Kümeleme işleminde temel olarak iki matris grubu kullanılır.

- **Veri Matrisi:** Veriler/nesneler çok boyutlu uzayda bir nokta olarak temsil edilirler. m adet nesnenin p adet özelliği (feature) mxp boyutlu bir matris hâlinde getirilmiş olan veri kümesidir.

- **Yakınlık Matrisi:** Kümeleme algoritmalarının birçoğu (*D*) farklılık matrisini (dissimilarity matrix) veya (*S*) benzerlik matrisini (similarity matrix) kullanır. Her 2 matris de nesneler arasında uzaklığı içerdiğinden bu 2 matrisin genel ismi olarak (*P*) yakınlık matrisi ismi kullanılmıştır. Bu matris nesneler arasındaki yakınlığı temsil ettiğinden *m* adet nesne için mxm büyüklüğünde köşegenleri 0 veya 1 olan üçgen bir matris teşkil eder.



Şekil 4. Veri yapıları [35].

Veri madenciliğinde çoğunlukla veri matrisi ve farklılık matrisi kullanılır. Kümeleme algoritmaları veri matrisi ve farklılık matrisini girdi olarak kullanır. Şekil 4'te veri ve farklılık matrislerinin bir örneği yer almaktadır.

3.2. Kümeleme Analizinde Kullanılan Veri Türleri

Kümeleme analizinde kullanılan değişenler/özellikler sürekli (continuous), ayrık (discrete) veya ikili (binary) olarak sınıflandırılabilir [33].

Değişkenlerin sınıflandırılmasında kullanılan diğer bir yaklaşım, sayıların anlamını yansıtan ölçüm düzeyleridir. Ölçüm düzeyleri, en düşükten en yükseğe sınıflayıcı (nominal), sıralı (ordinal), aralık ölçekli (interval-scale) ve oran ölçekli (ratio-scale) olarak sıralanırlar [34].

Kümeleme Analizinde, Jain ve Dubes'ın tanımının dışında sınıflayıcı değişkenlerin özel bir hâli olan ikili değişkenler tanım olarak eklenir.

- **İkili (Binary) Değişkenler:** En basit şekliyle 0 veya 1, doğru veya yanlış, hatalı veya hatasız vb. şekilde 2 duruma sahip olan değişkenlerdir. Kümeleme analizinde 0 ve 1 şeklinde kodlanırlar.

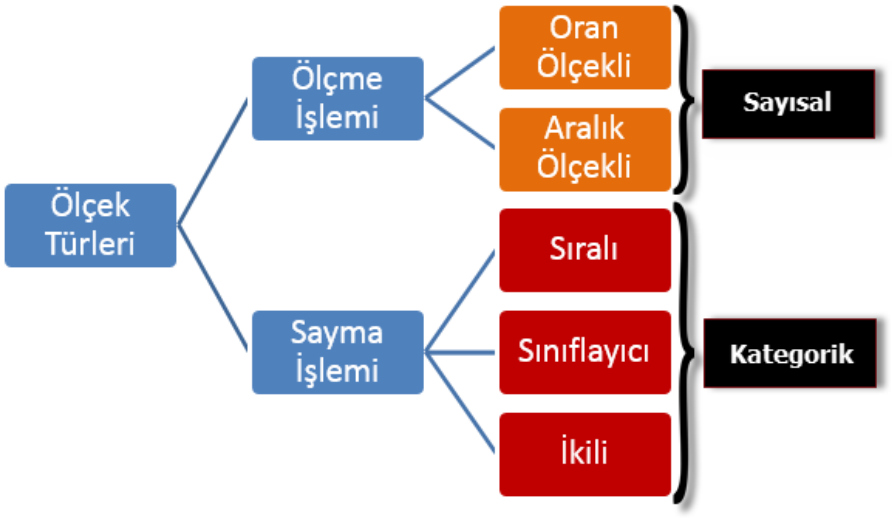
- **Sınıflayıcı, Sıralayıcı Ölçekli Değişkenler:** Sınıflayıcı değişkenler, ikili değişkenlerin genelleştirilmiş hâlidir. Saç rengi, meslek, posta kodu vb. sınıflayıcı değişkenler kümeleme analizinde sayısal verilerle birlikte kullanılabilmesi için 0 ve 1 şeklinde ayrı değişkenler oluşturacak şekilde kodlanırlar.

  Sıralayıcı değişkenler değişken değerleri arasında anlamlı bir sıralamanın veya derecelendirmenin olduğu değişkenlerdir. Rating notu, mezuniyet derecesi vb.

- **Aralık Ölçekli Değişkenler:** Bu değişkenlerin değişken değerleri arasında eşit aralıklar olup, değişken değerleri pozitif, sıfır veya negatif değer alabilen sayısal değişkenlerdir. Bu ölçek türünde sıfır değeri değişkenin gerçekten de olmadığı anlamını taşımaz. Sıcaklık değişkeni aralık ölçeğe güzel bir örnek olup, sıcaklığına sıfır olması sıcaklığın olmadığını göstermemektedir [35].

- **Oran Ölçekli Değişkenler:** Gerçek bir sıfır noktası olan sayısal değişkenlerdir. Para miktarı, boy uzunluğu vb. bu tür ölçekli değişkenlerdir [35].

Şekil 5'te değişken ölçek türlerinin toplu gösterimine yer verilmiştir.



Şekil 5. Değişken Ölçek Türleri.

Değişkenlerin ölçeklerine göre uygulanacak kümeleme analizi farklılaşır. Her algoritma her ölçekteki değişkene uygulanamamaktadır.

3.3. Kümeleme Analizi Türleri

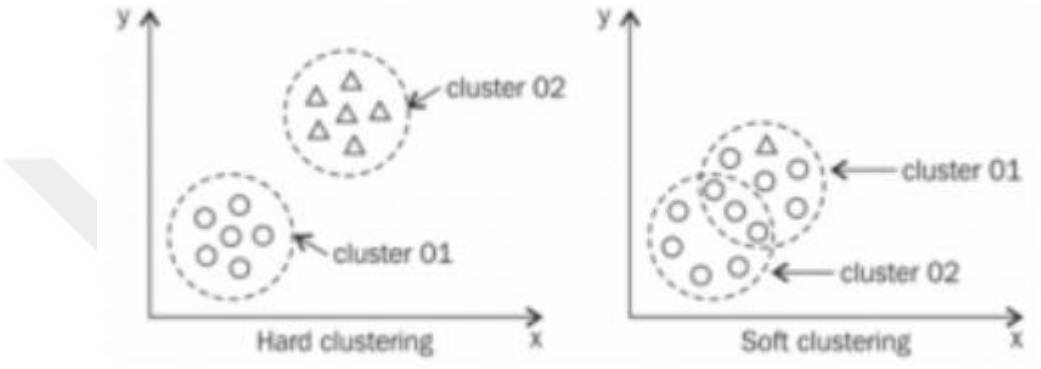
Kümeleme algoritmaları, analize tabii tutulan nesneleri kümelere atama şekli itibarıyla iki kısma ayırılır.

Katı Kümeleme (Hard Clustering): Kümeleme sonucunda nesnelerin sadece bir kümeye atanabildiği kümeleme algoritmaları kullanılarak yapılan kümeleme analizi hard veya crisp clustering olarak adlandırılır. Oluşan kümeler birbirinden ayrık bir yapıdadır.

Esnek Kümeleme (Soft Clustering): Kümeleme sonucunda nesnelerin birden fazla kümeye dâhil olabildiği çoğunlukla olasılıksal sonuç üreten yapıdaki

kümeleme algoritmaları kullanılarak yapılan kümeleme analizine esnek kümeleme (soft clustering) denmektedir. Bu tür kümelemede küme grupları birbiriyle örtüşen (overlapping) bir yapıdadır [36].

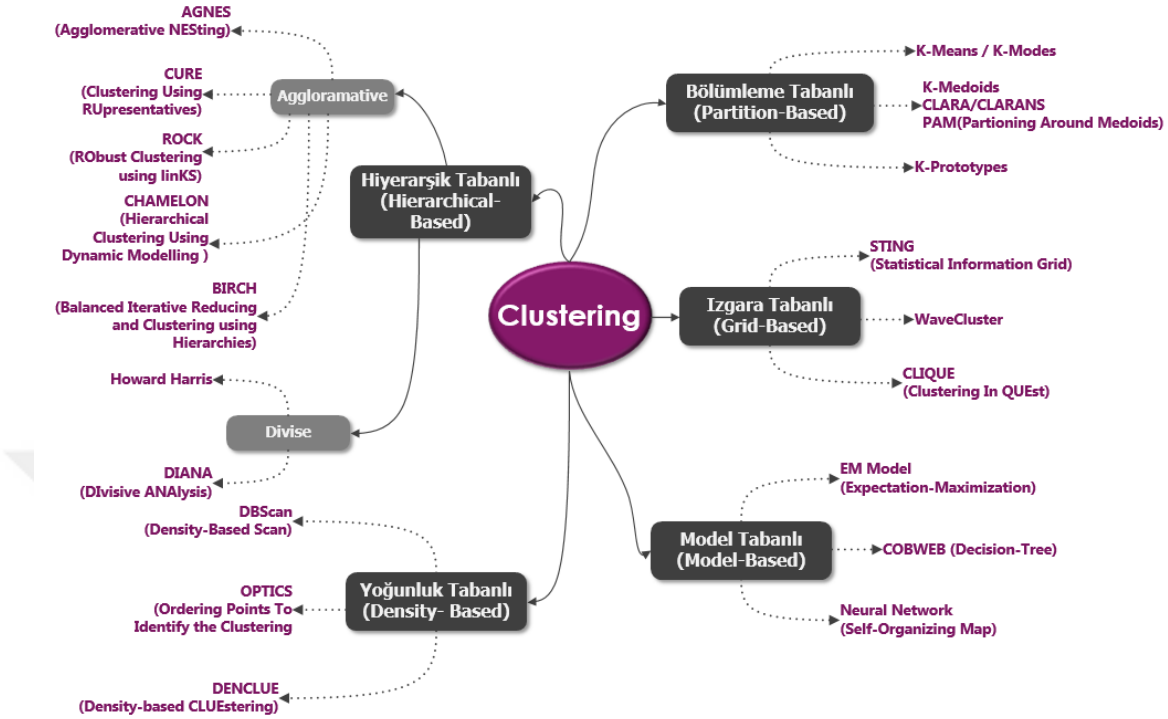
Katı ve Esnek Kümeleme sonuçlarının görselleşmiş bir örneği Şekil 6' da yer almaktadır.



Şekil 6. Katı ve Esnek Kümeleme Dağılım Örneği [66].

Literatürde birçok kümeleme algoritması bulunmasına karşın bu algoritmaları kullandığı yöntemlere göre başlıca beş bölümde sınıflamak mümkündür.

- Hiyerarşik Tabanlı Kümeleme,
- Bölümleme Tabanlı Kümeleme,
- Izgara Tabanlı Kümeleme,
- Model Tabanlı Kümeleme,
- Yoğunluk Tabanlı Kümeleme.



Şekil 7. Kümeleme Analiz Türleri ve Başlıca Algoritmaları.

Bu sınıflandırmaya göre en yaygın ve bilinen başlıca kümeleme algoritmaları Şekil 7'de gösterilmiştir.

3.3.1. Hiyerarşik Tabanlı (Hierarchical-based) Kümeleme Yöntemleri

Bu yöntemler nesneleri yukarıdan aşağı veya aşağıdan yukarı tarzda yinelemeli bir şekilde ayırarak kümeleri oluştururlar. Nesneler arasındaki uzaklıkların bulunduğu uzaklık matrisini (distance matrix) girdi olarak kullanırlar.

Birleşmeli Hiyerarşik Kümeleme (Agglomerative Hierarchical Clustering): Her nesne başlangıçta bir kümeyi temsil eder. Daha sonra yinelemeli bir şekilde kümeler birleştirilerek tek bir küme elde edilir.

Ayrılmalı Hiyerarşik Kümeleme (Divisive Hierarchical Clustering): Tüm nesneler başlangıçta bir kümeye aittir. Daha sonra bu küme yinelemeli bir şekilde alt kümelerine ayrıştırılır [37].

En yaygın kullanılan birleşmeli hiyerarşik kümeleme algoritmalarından bazıları AGNES [30], CURE [38], ROCK [39], CHAMELEON [40], BIRCH [41] ayrılmalı hiyerarşik kümeleme algoritmalarından en bilineni DIANA [30] olarak verilebilir.

3.3.2. Bölümleme (Partition-based) Tabanlı Kümeleme Yöntemleri

n adet nesne ve *p* adet değişkeni (feature) *k* adet kümeye ayıran kümeleme yöntemleridir. $k<n$ olmak zorundadır. Bu kümeleme yöntemlerinde, kümeler içi benzerlik maksimum, kümeler arası benzerlik ise minimum yapmak amaçlanır.

p boyutlu uzayda centroid, medoid veya prototype adı verilen merkez noktaları küme sayısı kadar belirlenir ve kümeleme bu merkez noktası etrafında oluşur. k-means [42], k-medoids [43], k-modes [44], k-prototypes [45], PAM [43], CLARA [30], CLARANS [46] en çok kullanılan bölümleme tabanlı kümeleme yöntemleridir. Bu tür algoritmalarda ayrıştırılmak istenen küme sayısı, algoritmaya giriş parametresi olarak verilmek zorundadır.

3.3.3. Izgara (Grid-based) Tabanlı Kümeleme Yöntemleri

Izgara tabanlı metotlar nesne uzayını ızgara formunda sonlu sayıda hücrelere bölerler. Tüm kümeleme operasyonları bu ızgara yapısı üzerinde yapılır. Özellikle konumsal veri madenciliği (spatial data mining) çalışmalarında etkili bir kümeleme yöntemidir. STING [47], WaveCluster [48] ve CLIQUE [49] başlıca ızgara tabanlı kümeleme algoritmalarıdır. STING ızgara hücrelerinde olan istatistiksel bilgiyi dayalı olarak kümeleme yapar. CLIQUE nesne uzayını alt uzaylara (subspaces) böler. Bu alt uzaylar

daha sonra ızgara formunda hücrelere bölünür. Hücre içerisindeki nesne yoğunluğuna göre kümeleme yapar. WaveCluster dalga dönüşüm (wavelet transform) yöntemi kullanarak nesneleri kümeler [47].

3.3.4. Model Tabanlı Kümeleme Yöntemleri

Modele dayalı kümeleme yöntemleri, kullanılan algoritmalar ile veri arasındaki matematiksel model uyumunu optimize ederek işlemektedir. Bu yöntemlerde, genellikle verilerin olasılık dağılımları tarafından oluşturulduğu varsayılır. Model esaslı kümeleme teknikleri istatistiksel yaklaşım kullanımında karışım modelleri (mixture models) [50], yapay sinir ağı kullanımında SOM [51], karar ağacı kullanımında COBWEB [52] ile yapılmaktadır.

3.3.5. Yoğunluk Tabanlı Kümeleme Yöntemleri

Bu yöntemler her nesnenin çevresindeki komşu nesneler ile yakınlığı hesaplanarak kullanılır. Yakınlık hesaplamasında veri türüne göre farklı uzaklık ölçütleri kullanılır. Yeterli komşu olmayan nesneler tespit etmekte etkilidir. Yoğunluk tabanlı algoritmalar düzgün şekilli olmayan kümeleri bulma, gürültülü ve istisnalardan etkilenmeyen yöntemler olduğu için bu tür problemi barındıran kümelemelerde etkin yöntemlerdir. DBSCAN [53], DENCLUE [54], OPTICS [55] en yaygın kullanılan yoğunluk tabanlı kümeleme algoritmalarıdır.

Yukarıda belirtilen kümeleme analizleri dışında makine öğrenmesi ve istatistik disiplinlerinin kümeleme analizinde kullanımından doğan pek çok kümeleme analizi türü ortaya çıkmıştır.

- Yüksek boyutlu verileri kümelemek için Alt Uzay Kümeleme (Subspace Clustering) algoritmaları, yüksek boyutlu metinlerin veri madenciliğinde yaygın kullanılan küresel k-means [65] algoritması,
- Hem nesneleri hem de değişkenleri alt uzaylarına ayırarak kümeleme analizi yapılabilen ikili kümeleme (BiClustering) algoritmaları,

- Esnek kümeleme yapan Bulanık Kümeleme (Fuzzy Clustering) algoritmaları, kümelemede Destek Vektör Makinelerinin (SVM-Support Vector Machine) kullanımı ile yapılan SVM Kümeleme algoritmaları,
- Öz düzenleyici Haritalar (SOM-Self-Organizing Maps) tekniğinin kullanıldığı Yapay Sinir Ağ Tabanlı Kümeleme algoritmaları,
- Karar ağaçlarının kümeleme amacı ile kullanımından doğan COBWEB algoritmaları, k-means tekniğinin türevi olan k-windows, kernel k-means, algoritmaları diğer başlıca kümeleme teknikleridir.

3.4. Karışık Veri Tiplerinde Kümeleme Analizi

Kümeleme analizinde kullanılacak gerçek dünya veri kümeleri farklı ölçekteki birçok değişkeni içinde barındırır. Bu tür değişkenler literatürde karışık veri tipleri (mixed types) olarak adlandırılır.

Bir veri kümesi hem aralık hem sınıflayıcı hem de ikili değişken tiplerini barındırabilmektedir. Bu tür veri kümelerine karşı kümeleme analizi uygulamak için iki tür çözümü uygulanır.

- **Veri kümesindeki tüm değişkenleri, tek tip ölçekli değişkene dönüştürmek**: Kümeleme algoritmalarının büyük bir çoğunluğu ve parametrik istatistiksel dağılımlar aralık veya oran ölçekli sayısal veri türlerini işleyebilirler. Kümeleme analizinin türüne bağlı olarak veri kümelerindeki kategorik değişkenlerin hepsi sayısal verilere dönüştürülmeli ya da sayısal değişkenlerin hepsi kategorik ölçeğe dönüştürülmelidir. Örneğin; k-means algoritması sadece sayısal verileri işleyebildiğinden kümelemeye tabii tutulacak veri matrisinde yer alan kategorik değişkenler kategori değişkendeki değer adedince yapay değişkenlere çevrilip 0 ve 1 şeklinde kodlanmalıdır. Bu çözüm sayısal bir değeri almayan değişkenleri sayı hâline getirilmesi nedeniyle kümeleme algoritmalarının sağlam ve düzgün çözüm üretmesini engelleyebilmektedir. Bu kategorik veriyi sayısal değerlere dönüştüren

geleneksel yaklaşım kategorik alanın sıralı olmadığı durumlarda anlamlı sonuçlar üretmeyecektir [44].

Tam tersi bir durum olan k-modes algoritması sadece kategorik verileri kullanabildiği için veri matrisinde yer alan sayısal değişkenlerin hepsi frekans tabloları yardımıyla kategorik veri ölçeğine dönüştürülür. Bu işleme veri paketleme (data bining or data discretization) denir. Sayısal değişkenlere ölçek dönüşümü uygulayarak kümeleme analizine tabii tutulması değişkenin açıklama düzeyinde kayıplar yaşanmasına neden olmaktadır.

- **Karışık veri tiplerini girdi olarak alabilen algoritmaları kullanmak**: İlk maddede yer alan çözümün sakıncaları karışık veri türlerini barındıran veri kümelerini de başarılı şekilde kümeleme analizine tabii tutulmasını sağlayan algoritmaların geliştirme ihtiyacını doğurmuştur [56].

Karışık veri tiplerinde kullanılan kümeleme yöntemleri aşağıda bölümlerde açıklanmıştır.

3.4.1. Gower Benzerlik Katsayısı (Gower similarity coefficient / Gower Index)

Farklı ölçeklerdeki değişkenlere sahip olan nesnelerin arasındaki uzaklığı hesaplamak için 1971 yılında J.C. Gower tarafından kendi adını taşıyan Gower benzerlik katsayısı tanıtılmıştır [57].

Karışık veri tipleri için en popüler yöntem yakınlık matrisi oluşturmaktır. Veri kümesindeki her nesne çiftleri arasında benzerlik katsayısı hesaplanır. Bu katsayılardan *n* adet obje için *nxn*'lik bir benzerlik matrisi oluşur. Bu matris PAM, AGNES gibi algoritmalara girdi teşkil eder. Bu algoritmalar birlikte kullanılabilmektedir [58]. Öklid, Manhattan, Minkowski vb. uzaklık ölçütleri kayıp veri ile çalışamaz iken Gower benzerlik katsayısı, uzaklık hesaplamasında kayıp değerlerden etkilenmez. Kayıp değerleri içeren veri kümeleri içinde skor hesaplaması

yapabilir. Gower benzerlik katsayısı hesaplamasında kullanılan formüle, Denklem 1' de yer verilmiştir.

$$GOW_{jk} = \frac{\sum_{i=1}^{n}(W_{ijk} * S_{ijk})}{\sum_{i=1}^{n}(W_{ijk} * S_{ijk})}$$

Denklem 1. Gower Benzerlik Katsayısı [57].

Yukarıda verilen denklemde;

W_{ijk} : j ve k nesnelerinin i değişkeni için karşılaştırılıp karşılaştırılamayacağını gösterir. X_{ij} veya X_{ik} en den en az biri bilinmiyorsa $Wijk$=0 olur.

S_{ijk} : i değişkeni için X_{ij} ve X_{ik} arasındaki farktır.

n : Değişken sayısı

ifade etmektedir.

3.4.2. K-Prototip Algoritması

Bu algoritma Zhexue Huang tarafından 1997 yılında Singapur'da yapılan "Knowledge Discovery and Data Mining" konferansında ortaya çıkmıştır. Büyük veri kümelerini etkin şekilde homojen kümelere ayırmak veri madenciliği alanında temel bir problemdir. Hiyerarşik Kümeleme yöntemlerinin hesaplama verimsizliğinden dolayı, bu yöntemler büyük veri kümelerinin işlenmesinde bir çözüm olamamaktadır. K-means tabanlı metotlar büyük veri kümelerini işleme konusunda etkindirler. K-means tabanlı yöntemlerin kullanımı ise sayısal veri türleri ile sınırlıdır. K-prototip algoritması k-means'in büyük veri kümelerini işlemedeki etkinliğini kaybetmeden sayısal veri işleme sınırını ortadan kaldırır [45].

Bu algoritma, sayısal ve kategorik değişkenli nesneleri k-means algortimasına benzer şekilde kümeler. Nesneler k tane ortalama yerine k tane prototipe göre kümelenir. $X = \{X_1, X_2, \ldots, X_N\}$ kümesi X_i $(i=1,2,\ldots,n)$ n adet veri kümesini temsil etmektedir.

X_i $(1<=i<=n)$ $A_1,A_2,A_3,\ldots,A_p,A_{p+1},\ldots,A_m$ özdeğerleri olarak tanımlanır. Aynı zamanda A_i $(1<=i<=p)$ veri kümesinin sayısal değişkenleri, A_i $(p+1<=i<=m)$ kategorik değişkenleridir. $Dom(Aj)$, Aj değişkeninin rankı demektir.

$$Dom(A_j) = \{ a_j^{(1)} , a_j^{(2)} , \ldots, a_j^{(n_j)} \}$$

Denklem 2. A_j değişkeninin rankı [59].

n_j; A_j kategorik değişkenin, değişken değerlerinin sayısını gösterir. X in elemanı olan X_i, m adet değişkenli bir vektör olarak $X_i = \{x_{i1}, x_{i2}, \ldots, x_{ip}, x_{i(p+1)}, x_{i(p+2)}, \ldots, x_{im}\}$ ifade edilir. Kümeleme merkezi Z_j olarak gösterilir. X in ait olduğu C_l veri kümesi K-prototip algoritmasında bir kümedir. Z_l ise C_{lh} ' ın küme merkezidir. X nesnesi ile Z_l küme merkezi arasındaki uzaklık metriği olarak Denklem 3' de verilen K-Prototip Uzaklık Metriği kullanılır.

$$d(X_i, Z_l) = d_r(X_i, Z_l) + \gamma\, d_c(X_i, Z_l)$$

Denklem 3. K-Prototip Uzaklık Metriği [59].

Uzaklık metriği iki kısımdan oluşur. d_r sayısal değişkenlerinin küme merkezi ile arasındaki uzaklığı, d_c ise nesnenin kategorik değişkenlerinin küme merkezi arasındaki uzaklığını temsil eder.

$$d_r(X_i, Z_l) = \sum_{j=1}^{p} \left(X_{ij} - z_{lj}\right)^2$$

Denklem 4. Sayısal Değişkenlerin Öklid Uzaklığı [59].

Nesnenin sayısal değişkenlerinin küme merkezine uzaklıkta Öklid uzaklığı kullanılır. Hesaplama formülü, Denklem 4. Sayısal Değişkenlerin Öklid Uzaklığında belirtildiği şekildedir.

$$d_c(X_i, Z_l) = \sum_{j=p+1}^{m} \delta(X_{ij} - z_{lj})$$

$$\delta(X_{ij} - z_{lj}) = \begin{cases} 1 & X_{ij} \neq z_{lj} \\ 0 & X_{ij} = z_{lj} \end{cases}$$

Denklem 5. Kategorik Değişkenlerin Simple Matching Uzaklığı [59].

Nesnenin kategorik değişkenlerinin küme merkezine uzaklığında Simple Matching uzaklığı kullanılır. Hesaplama formülü, Denklem 5. Kategorik Değişkenlerin Simple Matching Uzaklığında belirtildiği şekildedir. K-prototip algoritmasının amaç fonksiyonuna Denklem 6'da yer verilmiştir.

$$F(W, Z) = \sum_{l=1}^{k} \sum_{i=1}^{n} w_{li}\, d(X_i, Z_l)$$

$$w_{li} \in \{0,1\},\ 1 \leq l \leq k,\ 1 \leq i \leq n$$

$$\sum_{i=1}^{k} w_{li} = 1,\ 1 \leq i \leq n;$$

$$0 < \sum_{i=1}^{n} w_{li} < n,\ 1 \leq l \leq k$$

Denklem 6. K-Prototip Amaç Fonksiyonu [59].

Yukarıdaki denklemde $w_{li} = 1$, i nesnesinin l. kümeye ait olduğunu gösterir. $w_{li} = 0$, *i* nesnesinin *l.* kümeye ait olmadığını gösterir [59]. K-Prototip Algoritması; Denklem 6' da F olarak gösterilen amaç fonksiyonunun çıktı değerini minimize etmeye çalışır. Algoritmaya ait adımlar aşağıdaki şekildedir.

1. Adım: *k* adet küme için bir tane başlangıç küme merkezi/prototipi rasgele seç.
2. Adım: *X* deki her nesneyi Denklem 3'te verilen k-prototip uzaklık metriğine göre en yakın olan küme prototipine ata ve her atama sonrasında küme prototipini güncelle.
3. Adım: Tüm nesneler bir kümeye atandıktan sonra oluşan prototipler ile nesnelerin benzerliğini test et. Bir nesnenin atanmış olduğu kümesi/prototip

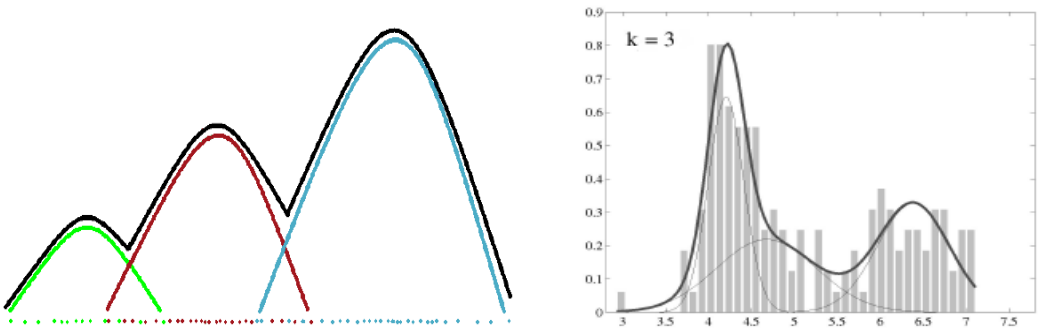
dışında başka bir kümeye/prototipe yakın ise nesneyi yakın olduğu kümeye/prototipe ata ve her iki prototipi güncelle.

4. Adım: *X'* in tüm nesnelerinin kümesi değişmeyene kadar 3. adımı tekrar et [45].

3.4.3. Karışım Modelleri (Mixture Models)

Veriler, birkaç bileşenden (components) oluşan bir karışımla modellenebilir. Bu bileşenlerin her biri normal, bernoulli, poisson, gamma, weibull gibi parametrik bir dağılımdır.

Karışım modelleri, birden fazla yoğunluk fonksiyonun birleşiminden oluşurlar. Kümeleme uzayı kaç kümeye ayırılacak ise o kadar yoğunluk fonksiyonu kullanılır. Şekil 8'de 3 kümeye ayrılmak istenen ve kümelerin normal dağıldığını varsayan tek değişkenli bir gauss karışım modeli yer almaktadır.



Şekil 8. Gauss Karışım Modeli [64].

Karışım Modelleri ile kümeleme; veriye bir karışım modeline giydirme, karışım model bileşenleri ile her kümeyi tanımlama yani verideki her nesneyi bir karışım model bileşenine atayarak kümeleri teşkil etme işlemidir.

Karışım Modelleri ile kümeleme Beklenti En Çoklama (EM-Expectation Maximization) algoritması ile yapılır. Bu algoritmanın iki adımı vardır.

- **Beklenti Adımı:** Karışım modelinin ilk tahminleri için her bileşim dağılımındaki her veri noktasının kısmi üyeliği, her nesne x_j ve karışım modeli Y_i için üyelik değeri $y_{i,j}$ dir. Üyelik değerini tespit eden formül Denklem 7' de yer almaktadır.

$$\boldsymbol{y_{ij}} = \frac{\boldsymbol{\alpha_i f_y(X_j; \theta_i)}}{\boldsymbol{f_x(X_j)}}$$

Denklem 7. Nesne Üyeliklerinin Ataması (Beklenti Adımı) [60].

- **En Çoklama Adımı:** Bu adım nesnelerin ait oldukları üye değerlerine göre karışım modelinin parametrelerinin yeniden hesaplanması adımıdır. $\boldsymbol{\theta_i}$ Karışım modeli parametreleri üyelik değerleri kullanılarak ağırlıklandırılan, x_j veri noktalarını kullanan beklenti maksimizasyonu ile hesaplanır [60]. Örneğin; $\boldsymbol{\theta}$ bir μ ortalamasıysa; μ ortalamaya ait beklenti maksimizasyon formülü, Denklem 8' de gösterildiği şekilde olacaktır.

$$\boldsymbol{\mu_i} = \frac{\sum_j \boldsymbol{y_{i,j}\, x_j}}{\sum_j \boldsymbol{y_{i,j}}}$$

Denklem 8. Yeni Parametre Tahmini (En Çoklama Adımı) [60].

Beklenti ve En Çoklama Adımları, model parametreleri birbirine yaklaşana veya belirlenen yineleme sayısına dek tekrarlanırlar. Karışım modelleri sonucunda her nesne kesin bir kümeye atanmaz. Nesnelerin her bir kümenin üyesi olma olasılıklarına ulaşılır.

3.4.4. İki Adımlı Kümeleme Yöntemi (Two-step Clustering Method)

Hiyerarşik ve bölümleme bazlı kümeleme algoritmalarını beraber uygulayabilen, karışık veri türlü veri kümelerine de işleyebilen bir kümeleme yöntemidir. Bu

algoritma kategorik değişkenlerin birlikte olma sayılarını dikkate alır. Algoritmanın adımları aşağıdaki şekildedir.

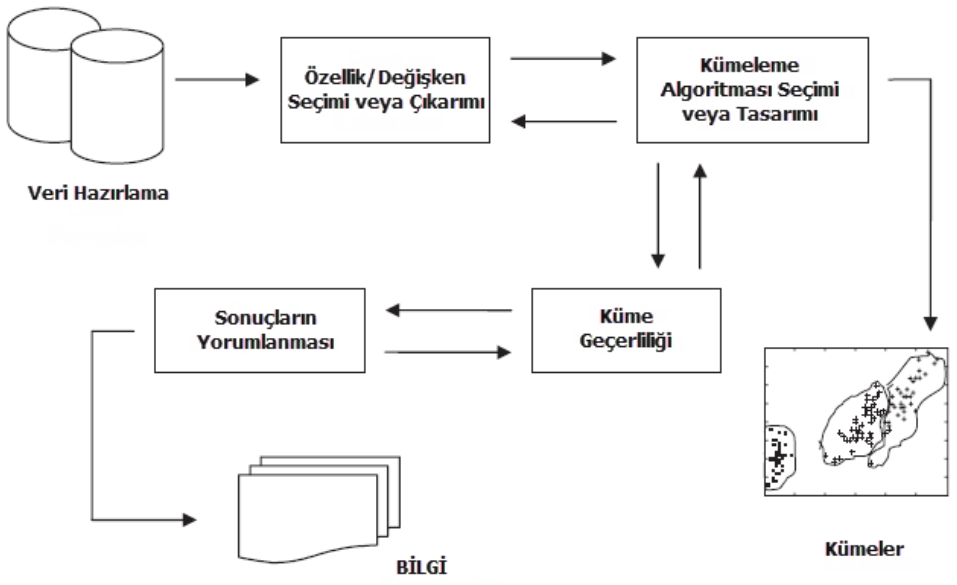
1) Veri ön işleme safhası:
   a) Veri kaynağından okunur. Sayısal değişkenler 0-1 aralığında normalize edilir.
   b) En çok değere sahip değişken, *A* tespit edilir. Bu değişken temel değişken olur. Bu temel değişkendeki değerler temel kalemler olarak belirlenir.
   c) Her temel kalem ile her kategorik değişken değerlerinin birlikte oluşma sıklığı sayılır ve bu bilgi *M* adlı bir matriste saklanır.
   d) Bu *M* matrisi kullanılarak her kategori ve temel kalem arasında benzerlik hesaplanır ve bu bilgi *D* matrisinde saklanır.

2) Sayısal değerleri kategorik kalemlere atama safhası:
   a) Temel kaleme göre grup varyansını minimize eden sayısal değişken bulunur.
   b) Her temel kalemin sayısal değerlerini bulmak için (2.a) adımında bulunan sayısal değişkenle eşleşen değerlerin ortalaması her temel kaleme verilerek kategorik veriler sayısallaştırılırlar.

3) Kümeleme safhası
   a) Veri kümesine Birleşmeli Hiyerarşik Kümeleme (HAC-Hierarchical Agglomerative Clustering) Algoritması uygulanır.
   b) Her oluşan kümenin orta noktasını (centroid) hesaplanır, her kategorik kalem orta noktaya değişken olarak eklenir. Eklenen bu yeni değişken kümede bu kategorik kalem değerine sahip nesnelerin sayısıdır.
   c) İstenilen küme sayısına göre (3.a) ve (3.b) adımda oluşturulan veri kümesine tekrar k-means algoritması uygulanır [61].

3.5. Kümeleme Analizi İşlem Adımları

Her kümeleme analizinde Şekil 9'da belirtilen işlem adımları takip edilir. Öncelikle, problemi çözmek için gereken veriler veri matrisi şeklinde elde edilir. Temin edilen veriye ait değişkenlerin açıklama düzeyi, birbiri ile olan bağımlılıkları, kayıp değerleri incelenir. Verilerin tanımlayıcı istatistikleri (ortalama, varyans, min, max, mod, medyan, basıklık, çarpıklık vb.) incelenir. Kümeleme analizine katkıda bulunacak, hesaplama karmaşıklığını azaltacak değişkenler seçilir. Değişken seçiminde problem konusunda alan uzmanı kişilerin bilgisine başvurularak değişken seçimi gerçekleştirilebilir. Var olan değişkenler birleştirerek veya oranlayarak yeni değişkenler oluşturulur. Çözülecek probleme, değişken yapısına, verinin büyüklüğüne uygun kümeleme algoritması seçilir. Bu algoritma veri matrisine veya uzaklık matrisine uygulanır.



Şekil 9. Kümeleme Analizi İşlem Adımları [63].

Kümeleme algoritması sonrası oluşan sonuçlar görselleştirme, geçerlilik testleri veya alan uzmanı görüşü alınarak geçerliliği sınanır. Geçerlilik sınamasında başarılı olması durumunda kümeleme analizin sonucunu yorumlanır ve veride saklı olan örüntü, bilgiye dönüştürülür.

4. İLGİLİ ÇALIŞMALAR

Sexton ve Kathryn' nin 1980'de yaptığı çalışmada sahne sanatları (müze, bale, opera, tiyatro, modern, dans vb.) izleyicilerinin davranışlarına göre segmentasyonu incelenmiştir. Çalışmada kültürel kuruluşların katılımcılarına odaklanılmıştır. Kültürel kuruluşların pazarlama stratejilerini geliştirmesine bir temel teşkil etmesi için sektör segmentlerini belirlenerek katılımcı profilleri incelenmiştir. Çalışma 30,000 izleyici üyenin katıldığı anketlerle yapılmış, bu anketlerde izleyicilerin demografik bilgileri, son 12 ay boyunca çeşitli kültürel olaylara katılım sıklıkları, ücret gibi değişkenlere karşı davranışları, seçimlerini etkileyen özelliklere ilişkin sorular sorulmuştur. Çalışmada kümeleme analizinde kullanılmak için ilk önce bağımsız değişkenlerin alt kümesini bulmak için Faktör Analizi uygulanmıştır. Kümelemenin dayandığı değişkenlerin arasındaki değişkenlerin bağımsız olduğu varsayılarak Howard Harris Clustering algoritması uygulanarak kümeleme analizi yapılmıştır. Kümeleme sonucunda; katılımcılar Düşük Katılımcılar (Ligths), Müze Hayranları (Museum Fans), Çok Yönlüler (All-Rounders) ve Uzmanlar (Specialists) şeklinde 4 kümeye ayrılmışlardır. Bu sonuçlar sahne sanatları kuruluşları tarafından direkt iletişim aracı olarak kullanılabilecektir. Örneğin; Çok Yönlülere (All-Rounders) e-posta ile farklı türde sahne sanatları için ortak abonelikler sunulabilecektir. [10].

T. Vardar'ın 2010 yılındaki çalışmasında, bir bankanın tüzel müşterileri finansal büyüklükleri (ciro, bilanço büyüklüğü, faaliyet kârı, öz kaynak) ve niteliksel özellikleri baz alınarak segmente edilmiştir. Araştırmanın kapsamı, banka tüzel müşterileri ve bir yıllık verileri ile sınırlandırılmıştır. SPSS paket programı kullanılarak BIRCH algoritması kullanılmıştır. Yapılan müşteri segmentasyonu bu tez çalışması ile benzerlik göstermesinin yanında yapılan segmentasyon çalışması tüzel müşterinin değeri ve büyüklüğünü göz önüne alınarak yapılmış segmentasyon olup, değere dayalı segmentasyon türünde bir çalışma yürütülmüştür [11]. Bu tez çalışmasında ise müşterilerin ürün kullanım şekline göre davranış profillerini çıkaran bir çalışma ortaya konulmuştur.

E. Akarsu'nun 2010 yılındaki çalışmasında, Müşteriyi Kaybetme Analizi (Customer Churn Analysis) yaparken banka müşterilerinin kanal kullanımı davranışlarını ilgilendiren değişkenleri ve RFM (Recency, Frequency, Monetary) bilgilerini birlikte kullanmıştır. Bu değişkenlere SPSS paket programı aracılığıyla Karınca Kolonisi Kümeleme algoritması (ACO-Ant Colony Optimization) uygulamıştır. Akarsu çalışmasında banka müşterilerinin ürüne karşı davranışlarını değişken olarak ele almış ve bu değişkenlerden müşteriyi kaybedip, kaybetmeyeceğini tespit etmeye çalışmıştır [12].

Shili'nin 2009 yılındaki makalesinde bahsedilen pazar bölümlendirme (Market Segmentation) kavramı ilk olarak 1950'lerin ortalarında Wendell R. Smith tarafında ortaya atıldı. Pazar Bölümlendirme tanım olarak sektörde; farklı ürünlere gereksinim duyan, farklı ihtiyaç, davranış ve karaktere sahip satın alıcıların gruplara ayrılması olarak tanımlandı [13]. Bunun dışında birçok farklı pazar bölümlendirme tanımı yapılmıştır. Pensilvanya Eyalet Üniversitesi üyelerinden Peter D. Bennet daha belirgin ve detaylı bir tanım yaptı. Bennet'a göre Pazar Bölümlenmesi; pazarın, benzer ihtiyaca sahip ve aynı şekilde davranan müşterilerini net kümelere ayrılma süreci olarak tanımlamıştır. Harvard Business School kıdemli dekanı Steven C. Wheelwright pazar bölümlenmesini; firmanın potansiyel müşterilerinin birbirinden açık bir şekilde farklı olan gruplara ayrılması ama grup içindeki potansiyel müşterilerin de büyük bir benzerlik göstermesi şeklinde tanımlamıştır. Farklı kelimelerle birçok pazar bölümlenmesi tanımı olmasına rağmen, pazar bölümlenmesinin özü pazardaki ürünü algılama ve değerlendirme şekli, satın alma davranışı ve ürünü kullanım şekli aynı olan potansiyel müşteriler kümesidir. Makalede; pazar bölümlendirme çalışmaları, segmentasyonda kullanılan değişkenlere göre Coğrafik Bölümleme, Demografik Bölümleme, Psikografik Bölümleme ve Davranışsal Bölümleme şeklinde sınıflandırılmıştır [7].

B. Mehdi'nin 2010 yılındaki çalışmasında, bankaların gelirini arttırması için POS cihazlarından daha fazla işlem geçmesini sağlamak istenmiştir. Bunu sağlamak içinde

bankanın POS üye işyeri müşterilerinin davranışlarını bilmesi gerekmektedir. Müşterilerin davranışlarını çıkarmak ve pazarlama, müşteri ilişkileri yönetimi için bir temel oluşturmak için yaygın ve meşhur yaklaşım olan RFM modeli kullanılmıştır. Recency (R); müşterinin en son ne zaman işlem yaptığı, Frequency (F); hangi sıklıkla işlem yapıldığı, Monetary (M); ne kadar para harcamış bilgilerin ölçülmesi yaklaşımı ile üye işyeri gruplarını daha iyi anlamakta, her üye işyeri için bir RFM skoru üretmektedir. RFM skoru müşteri davranışları daha iyi anlamayı, üye işyeri segmentlerine davranışsal anlam içeren etiketleri atanmasını sağlamaktadır. Daha sonra kural bazlı sınıflandırma algoritması ile kümelerin (clusters) tanımlayıcı kuralları çıkarılmıştır. Bu tanımlayıcı kurallar her küme için RFM parametrelerinin sınırlarını belirlemektedir. Sınıflandırma algoritması ile oluşturulan kurallar daha sonra yeni üye işyerlerinin skorlanması ve davranış etiketlerinin sınıfının tahmin edilmesinde kullanılmıştır. Bu kurallar davranışsal kurallar diye adlandırılmıştır [14].

Zhou, Miao ve Guangcan'nın 2009 yılındaki çalışmalarında kablosuz içerik hizmetlerini sunan şirket müşterilerinin tüketim karakteristiklerine dayalı olarak segmente edilmesi incelenmiştir. Karınca Koloni Optimizasyon (ACO) ve k-means algoritmaları kullanılarak segmentasyon yapılmış, algoritmada müşterinin demografik, coğrafik, tutum ve davranışsal verileri kullanılmıştır. Çalışma sonucunda müşteriler 5 adet kümeye ayrılmıştır. Çalışmada k-means algoritmasının etkin olduğu ancak başlangıç değer (küme sayısı seçimi) seçiminin dezavantajları nedeniyle ACO ve k-means kümeleme algoritmalarının birleşiminden oluşan melez bir yaklaşımla segmentasyon çalışması yapıldığı belirtilmiştir [15].

Garima ve Pooja'nın 2014 yılındaki çalışmasında çevreye duyarlı turist gruplarını tespit etmek için çevreye yanlısı turist davranışları segmentasyon kriterleri olarak kullanılmıştır. Araştırma verileri Hindistan yerel turistlerinden anket yoluyla toplanmıştır. Alınan toplam 510 adet anket cevabına kümeleme analizi uygulanmış olup, anketler geçen yıl en az bir kere tatile çıkmış ve 18 yaşından büyük olanlarla

kısıtlanmıştır. Kümeleme analizinde su, enerji ve diğer doğal kaynaklar koruma, bitki ve hayvanların korunması, yerel toplumun gelişimi ve eğitimi için para yardımları vb. çevre yanlısı davranış değişkenleri kullanılmıştır. Çalışmanın ana amacı, çevre yanlısı davranışların temelinde çevreci turist segmentlerini belirlemeyi amaçlamıştır. Çalışmada Ward metodunu kullanarak hiyerarşik kümeleme analizi tekniği kullanılmıştır. Sonuç olarak çevreci turistler "Çevre yanlı davranışı Yüksek Turistler", "Çevre yanlı davranışı Orta Turistler" ve "Çevre yanlı davranışı Düşük Turistler" şeklinde 3 kümeye ayırılmıştır [16].

B. Birkhead'in 2000 yılındaki makalesinde, sektörde yaygın bir şekilde kullanılan davranışsal segmentasyon ve kümeleme metotları ele alınmıştır. Veri Ambarlarının yaygınlığı, sonuç kalitesi, müşteri ürün ve işlem verilerinin zenginliği davranışsal segmentasyon sistemlerinin gelişiminde kümeleme analizi gibi çok değişkenli metotların kullanımını arttırdığı ifade edilmiştir. Segmentasyonu gerçekleştirme yöntemleri olarak tam sistemler ve kümeleme tabanlı sistemler olmak üzere iki tür farklı sistemin kullanıldığı belirtilmiştir. Tam sistemler; çok az veriye dayanan (Örneğin; RFM) geleneksel segmentasyon olarak tanımlanmıştır. RFM gibi tam sistemler az sayıda değişkenin aynı değerini paylaşan müşterileri aynı segmente atarlar. Kümeleme Tabanlı sistemler ise geniş bir davranışsal değişkenler kümesine dayalı olduğu ve aynı küme içindeki müşterilerin benzerliklerine dayalı gruplandığı belirtilmiştir [17]. Yazarlar, segmentasyon için önemli olan değişken sayısı arttıkça tam sistemlerin segment sayısının aşırı derecede arttığını gözlemişlerdir. Bu durumun tam sistemleri kullanışsız hale getirdiği belirtilmiştir. 3 değişkene bağlı bir tam sistemin 27 segment ürettiği; 3'ten çok daha fazla değişkenli bir kümeleme tabanlı sistemin ise 10'dan daha az segment sayısı ürettiği karşılaştırma olarak sunulmuştur [17].

Makalede, davranışsal segmentasyon implementasyonu 8 adımlık bir süreç olarak tanımlamıştır. İlk adımda segmente edilecek tüm müşterilerin içinde tüm özellikleri barındıran bir müşteri grubunun tamamlanır. 2. Adımda seçilen müşteri grubu için

davranışsal tüm değişkenler ve değerlerinin seçimi ve gözden geçirilmesi yapılır. 3. Adımda davranışsal değişkenlere temel bileşenler analizi uygulanır. Segment tespitine güçlü etkisi olan davranışsal değişkenler seçilir. 4. Adımda Kümeleme Analizi uygulanır. 5. Adımda Kümeleme analizi sonucunda çıkan her kümeyi detaylı bir şekilde tanımlamak için tüm davranışsal değişkenler kullanılır. 6.Adımda segmentlere atama kurallarını tespit etmek için diskriminant analizi uygulanır. 7. Adımda tespit edilen kurallara göre tüm müşterilerin segment ataması yapılır. 8. Adımda geliştirmede kullanılan segment profilleri ile tüm müşterilerin segment profilleri istatistiksel olarak karşılaştırılarak, atama kurallarının geçerliliği onaylanır. Makalede davranışsal segmentasyon sonuçlarının pazarlama performansı üzerinde etkisi, her bir segmente özgü pazarlama hedeflerinin belirlenmesi konuları da yer almıştır. Davranışsal Segmentasyon sistemlerinin, pazarlama planlayıcılarının; müşteri davranışı çeşitliliğinden dolayı oluşan karışıklığın üstesinden gelmesine olanak sağladığı belirtilmiştir [17].

Kasturi, Moriarty ve Swartz'ın 1992 yılındaki makalesinde, sanayi piyasasında müşterilerin yalnızca büyüklük veya ürünün faydaları üzerinden segmentasyon yapılmasının nadiren yeterli olduğu söylenmiş, fiyat ve hizmet maliyeti arasındaki denge açısından müşteri davranışlarının önemli bir kriter olduğunun üzerinde durulmuştur. Makalenin yazarı; endüstriyel müşterilerin satın alma davranışına yönelik mikro segmentasyon için bir çerçeve sunmuştur. Kurduğu çerçeve, fiyat ve hizmet maliyeti üzerinedir. Segmentasyonun konusu, müşterilerin fiyat ve hizmet maliyetlerine göre satın alma davranışlarının kümelere ayrılmasıdır. 12 satın alma davranış değişkeninin katıldığı Hiyerarşik Kümeleme Analizi ile davranışsal segmentasyon gerçekleştirilmiştir. Hiyerarşik Kümeleme Analizinde; kümeler arasındaki heterojenliği, kümelerin kendi içindeki homojenliği maksimum yapmaya çalışan Ward metodunu kullanmıştır. Araştırma sonucunda 4 davranışsal segment bulunmuştur. Araştırmada kullanılan Signode Company şirketinin kârını arttırmak için

davranışsal segment çalışması sonuçlarını kullanmış, bu sonuçlara göre kampanya kaynaklarını düzgün yönlendirdiği belirtilmiştir [18].

Van Raaij ve Verhallen'in 1991'deki makalelerinde, Pazar (Market) Segmentasyon çalışmalarında kullanılan nesnel değişkenleri (gelir, yaş, davranış) ve anket, mülakat gibi yöntemlerle belirlenebilen öznel değişkenleri (yaşam tarzı, tuttum, ilgi alanları, niyet) sınıflandırmanın 1. boyutu olarak ayırmışlardır. Segmentasyon sınıflandırmasının 2. boyutu da değişkenlerin genellik seviyesine göre genel, alan/ürün/sınıfa özel, markaya özel şeklinde 3'e ayrıştırmışlardır. Değişkenlerin ölçülme ve genellik boyutlarına göre 3x2'lik bir matris ile sınıflandırma yapılmıştır. Şirketin hedefine göre hangi market segmentasyonun yapılacağı ve bu segmentasyona hangi uygun değişkenlerin (karakteristik mi? davranışsal mı? tutumlar mı?) seçilmesi gerektiğine oluşturulan matristen ulaşılması sağlanmıştır. Alana/Ürüne/Sınıfa Özel Segmentasyon sınıfı özel olarak incelenmiş, Alana/Ürüne/Sınıfa Özel Segmentasyon için karakteristik ve davranışsal değişkenlerin market segmentasyonunda kullanımını önermişlerdir. Bu öneriyi eş zamanlı segmentasyon olarak belirtmişlerdir [19].

Valters ve Roberts 2011 yılındaki makalelerinde, kriz sonrasında rekabet avantajı elde etmek için Litvanya alkol piyasasındaki tüketici davranışları ve tüketim kalıpları analiz edildi. Bunu başarmak içinde davranışsal tüketici segmentasyon modelini önerdiler. En sıklıkla uygulanan segmentasyon yaklaşımlarını fiziksel özelliklere dayalı (coğrafik, demografik vb.), psikografik özelliklere dayalı (yaşam tarzı, ilgi alanları, kişisel değerleri, toplum içi davranışı) ve teklife ilişkin davranışa dayalı (Ürün/Hizmetten beklenen fayda, ürün/hizmet kullanım durumu ve yoğunluğu, marka sadakati) şeklinde sınıflandırdılar [20]. Davranışsal tüketici segmentasyonu için ise tüketicilerin psikografik özellikleri kullanıldı. Tüketicinin psikografik özellikleri ile yapılan segmentasyonun, coğrafik ve demografik segmentasyonlara göre daha müşteri merkezli ve güçlü olduğu belirtildi. Bu tez çalışmasında, Valters ve Roberts' ın

makalelerinde yaptıkları sınıflandırmaya dayanan teklife ilişkin davranışsal özellikler (Propositional-Related Behavioral Attributes) kullanılmıştır.

B. Nakıpoğlu'nun 2007 yılındaki makalesinde işletmelerin pazar bölümlendirme amaçları ve tüketicilerin çevreci tutumlarına uygun davranışlar sergileme eğilimi içinde olmaları, çevreci tutum ve davranışlarına göre hedef pazarların belirlenmesi gerektirdiği ifade edilmiştir. Çalışmada bu amaca uygun olarak tüketicilerin çevreci tutumları faktör analizi ile belirlenmiş, kümeleme analizi aracılığıyla da tüketiciler çevreci tutum düzeylerine göre bölümlendirilmeye çalışılmıştır. Elde edilen bölümleme sonuçlarına göre ortaya çıkan birbirlerinde farklı tüketici gruplarının, çevreci konulara ve ürünlere verdikleri tepkilerinde birbirlerinden farklı oldukları ortaya konulmuştur. Pazar bölümlendirme faaliyetlerini genel olarak coğrafi, demografik, psikografik ve davranışsal kriterlere dayandırılarak yapıldığı Kotler'den alıntılanmıştır [21]. Tüketicilerin çevresel tutum ve davranışlarına göre bölümlendirilmeleri daha etkin pazarlama stratejileri için önemli bir yol haritası niteliği taşıdığı vurgulanmıştır. Çevreci tüketici özelliklerinin belirlenmesinde ve bu özelliklere göre tüketicileri bölümlendirmede ise sosyal statü, yaşam tarzı ve kişilik gibi değişkenlerden oluşan psikografik kriterler ve ürün veya ürün grubuna karşı tutum, kullanım oranı, bağlılık vb. değişkenlerden oluşan davranışsal kriterler kullanılmıştır. Bu davranışsal kriterlere göre tüketiciler Koyu Yeşiller, Yeşiller, Filizler, Şikâyetçiler ve Kahverengiler şeklinde bölümlere ayrıldığı belirtilmiştir. Çalışmanın faydası olarak davranışsal segmentasyon ile çevreci pazarlama faaliyetlerinin hedef kitlesinin belirlenmesi ve farklı bölümlere özel pazarlama stratejileri geliştirilmesi mümkün olabileceği vurgulanmıştır. Çalışmanın verileri, Adana'nın en büyük alışveriş merkezinde anket yapılarak 400 kişilik örneklem oluşturulmuştur. Bunun neticesinde öncelikle örnek gruptaki bireylerin demografik özellikleri tanımlanmış, tüketicileri çevreci davranışlarına göre bölümlendirmede kullanılacak ölçeğin ortaya çıkartılması için faktör analizi uygulanmıştır. Uygun ölçek kullanılarak hiyerarşik kümeleme uygulanarak küme sayısı tespit edilmiş daha sonra

bu küme sayısına göre k-means kümeleme yöntemi kullanılarak, tüketiciler çevreci davranışlarına göre farklı bölümlere ayrılmıştır. Her homojen grup, çevreci tutum, çevreci satın alma ve demografik özellikler açısından tanımlanmıştır. Ortaya çıkan grupların istatistiksel olarak birbirlerinden farklı olup olmadıklarının anlaşılabilmesi için de grup ortalamalarının farkları varyans analizi ile test edilmiştir [22].

Hosseini ve Mohammadzadeh'in 2016 yılındaki, çalışmalarında İran Sağlık Sektöründe hasta sadakatini, memnuniyetini arttıran ve hastane kârlılığını maksimize etmeyi amaçlayan bir çalışma yapılmıştır. Bunun, hastane veri tabanlarında yer alan verilerden bilgi keşfi ile yapılacağını düşünerek hem CLV müşteri değeri üzerinden hem de genişletilmiş RFML (Recency, Frequency, Monetary, Length) değişkenleri üzerinden kümeleme yaparak sonuçları karar ağacı ile sınıflandırmış, bu iki sınıflandırmanın doğruluğu karşılaştırılmıştır. Karşılaştırma sonucu kümeleme bazlı sınıflandırmanın müşteri değerine göre sınıflandırmadan daha doğru sonuçlar ürettiği hükmüne varılmıştır [23].

Braun, Geurten ve Egelhaaf 2010'daki çalışmalarında hayvanların art arda gerçekleşen davranışlarından prototip davranışsal bileşenler oluşturmak istemiştirler. Bu prototip davranış bileşenleri ile hayvanların üreme, avlanma gibi görevleri, bu görevlerin altında yatan mekanizmaları anlamak amaçlanmıştır. Prototip davranış bileşenlerini tespit etmek için hayvan davranışlarına dayalı kümeleme çalışması yapılmıştır. Kümeleme çalışmasının sağlaması, belirlenmiş prototiplerin tüm davranışları temsil edebiliyor olmasına göre değerlendirilmiştir. Çalışmada kurt sineğinin seyir uçuşu esnasındaki prototip baş hareketlerini tanımlamak için kümeleme yaklaşımı kullanılmıştır. Metot olarak ilk önce değişken (feature) seçimi sonrasında kümeleme ve değerlendirme adımları izlenmiş. Değişken (feature) seçiminde Temel Bileşenler Analizi uygulayarak davranışsal değişkenler seçilmiş, bu değişkenlere K-means kümeleme tekniği uygulayarak kümelere ayrıştırmıştır. Her küme merkezinin

(centroid) kalite ve kararlık kriterleri altında çeşitli veri kümeleri kullanılarak kümeleme analizinin doğruluk sağlaması (validation) yapılmıştır [24].

Juković, Pejić Bach, Dumičić ve Šarlija 2012 yılındaki çalışmalarında Hırvatistan'da bir bankanın tüzel müşterilerini Öz Örgütlemeli Harita Ağları (SOM) tekniği ile segmentlere ayırmıştır. Çalışmalarında SOM-Ward algoritmasını kullanılmışlardır. Bankanın tüzel müşterileri segmentlere bölümlendirilmiştir [25]. Çalışmada bankacılıkta geleneksel segmentasyon yaklaşımı olan ve çoğu müşteri segmentasyon çalışmasında kullanılan müşterinin değerine dayalı yaklaşım seçilmemiştir. Bunun yerine veriyi şirketin bir varlığı olarak düşünen ve karar alma sürecine temel oluşturan müşterinin demografiği, işlemleri, memnuniyetleri ve davranışlarını kullanmayı gerektiren iş zekâsı yaklaşımını önermişlerdir [26].

Sinha ve Uniyal Prasad'ın 2005 yılındaki çalışmalarında Hindistan perakende sektöründe alışveriş yapan kişileri segmente etmek için alışveriş davranışları kullanılmıştır. Alışveriş yapanlar, farklı mağazalarda muhbirler tarafından gözlemlenmiş ve davranışsal ipuçlarına göre 6 davranış segmenti (Choise Optimiser, Pre-Meditated, Economiser, Support Seeker, Recreational, Low Information Seeker) tespit edilmiştir. Çalışmada; segmentlerin, mağazalarda satılan ürün türleri ve mağazaların formatı temelinde farklılaştığı bulunmuştur. Çalışmaya ait veriler alışveriş yapan kişilerin muhbirler tarafından gözlemlenmesi ve bu gözlemlemelerin kayıt altına alınması ile toplanmıştır. Davranış gözlemleri bakkal, kozmetik, kitap, hazır giyim, ev eşyaları, ayakkabı gibi 20 mağazada yapılmış. Bu gözlem sonuçlarından Tabakalı Rastgele Örnekleme yöntemi ile örneklem seçilmiş, seçilen örneklerdeki gözlem bulgularının analizi için Gömülü Teori (Grounded Theory) yaklaşımını kullanmıştır [27].

Thomas ve Pickering 2003'teki, çalışmada Yeni Zelanda Şarap sektöründe uygulanmış çeşitli pazar segmentasyonu çalışmaları incelenmiş, bunların

karşılaştırması yapılmış ve bir davranışsal segmentasyon çalışması yürütülmüştür. Davranışsal Segmentasyon çalışmasında veriler 10 sorudan oluşan bir anket ile yapılmış, bu anketler 1144 potansiyel anketçiye e-posta ile iletilmiştir. Araştırmaya 320 anket dâhil edilmiştir. Anket sorularının bazıları ankete cevap verenlerin davranışsal örüntülerinin yakalanabilmesi için şarap satın alma aktivitelerini tespit eden şarap satın alınan mağazalar, aylık satın alınan şişe ve fiçı sayısı, şişe başına ödenen ücret gibi soruları içermiştir. Toplanan veriler SPSS uygulaması yardımıyla betimleyici istatistiksel analiz teknikleri kullanılarak analiz edilmiştir. Satın alma davranışına göre Light Purchaser, Medium Purchaser ve Heavy Purchaser adında 3 davranış segmentine ulaşılmıştır [28].

5. DAVRANIŞSAL SEGMENTASYON UYGULAMASI

Davranışsal segmentasyon yapmak için hazırlanmış olan veri madenciliği uygulaması bu bölümde aktarılmıştır.

5.1. Geliştirme Ortamı ve Kullanılan Araçlar

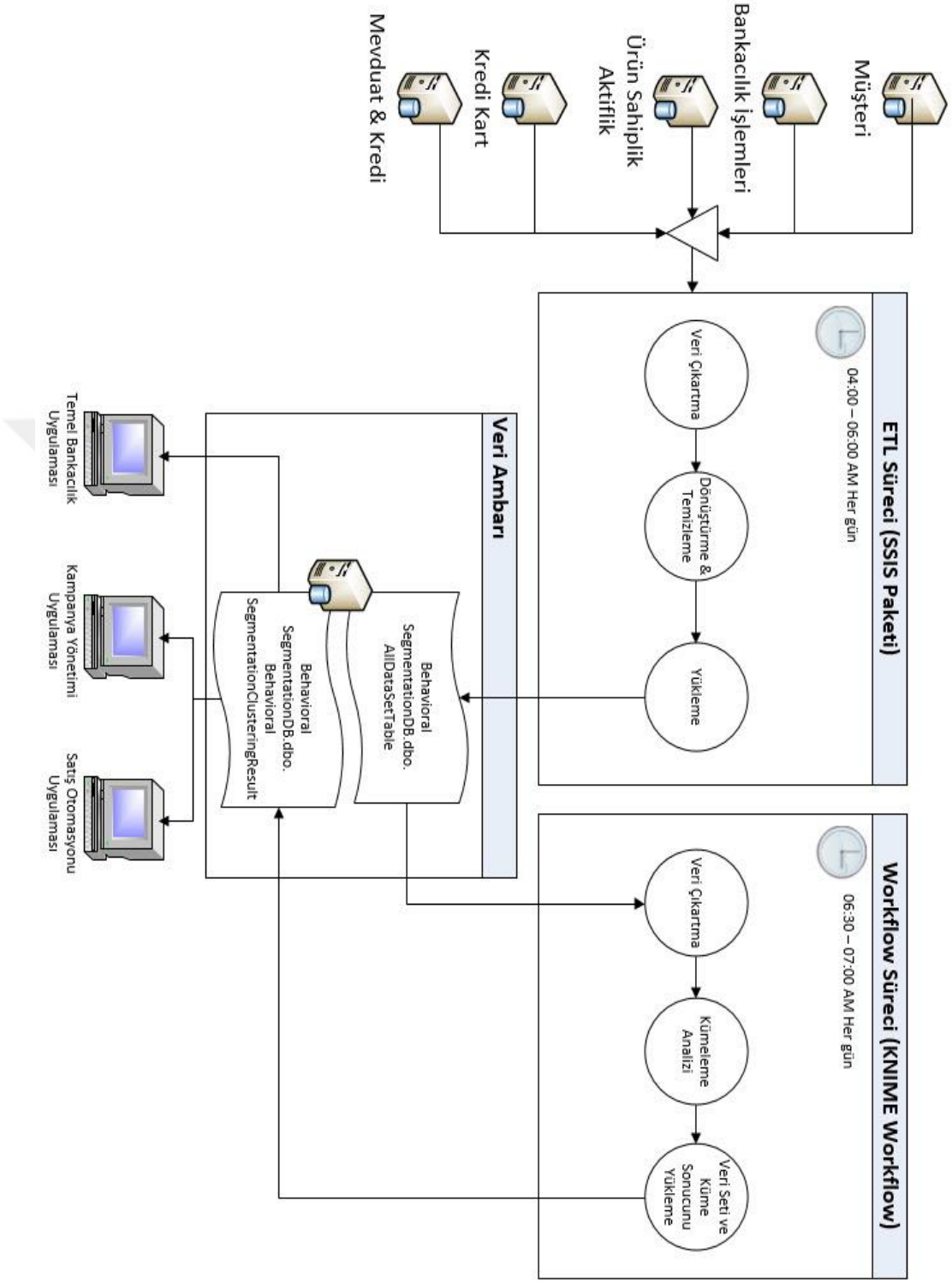
Uygulamada kullanılan veriler "Microsoft SQL Server 2014" veri tabanı sunucusunda depolanmaktadır. Veri tabanından verileri alıp, işleyip, dönüştürme, temizleme ve birleştirme (ETL– Extract Transform Load) işlemleri gerçekleştirmek için "Microsoft Data Tools for Visual Studio 2013" uygulaması kullanılmıştır. Değişkenlerin dönüştürülmesi, kümeleme analizlerinin gerçekleştirilmesi, görselleştirme ve nihai sonuçların tekrar veri tabanına yazılması için "KNIME" platformu kullanılmıştır. "KNIME" platformu içerisinde "R" programlama dili kullanılmıştır.

5.2. Uygulama Genel Akışı

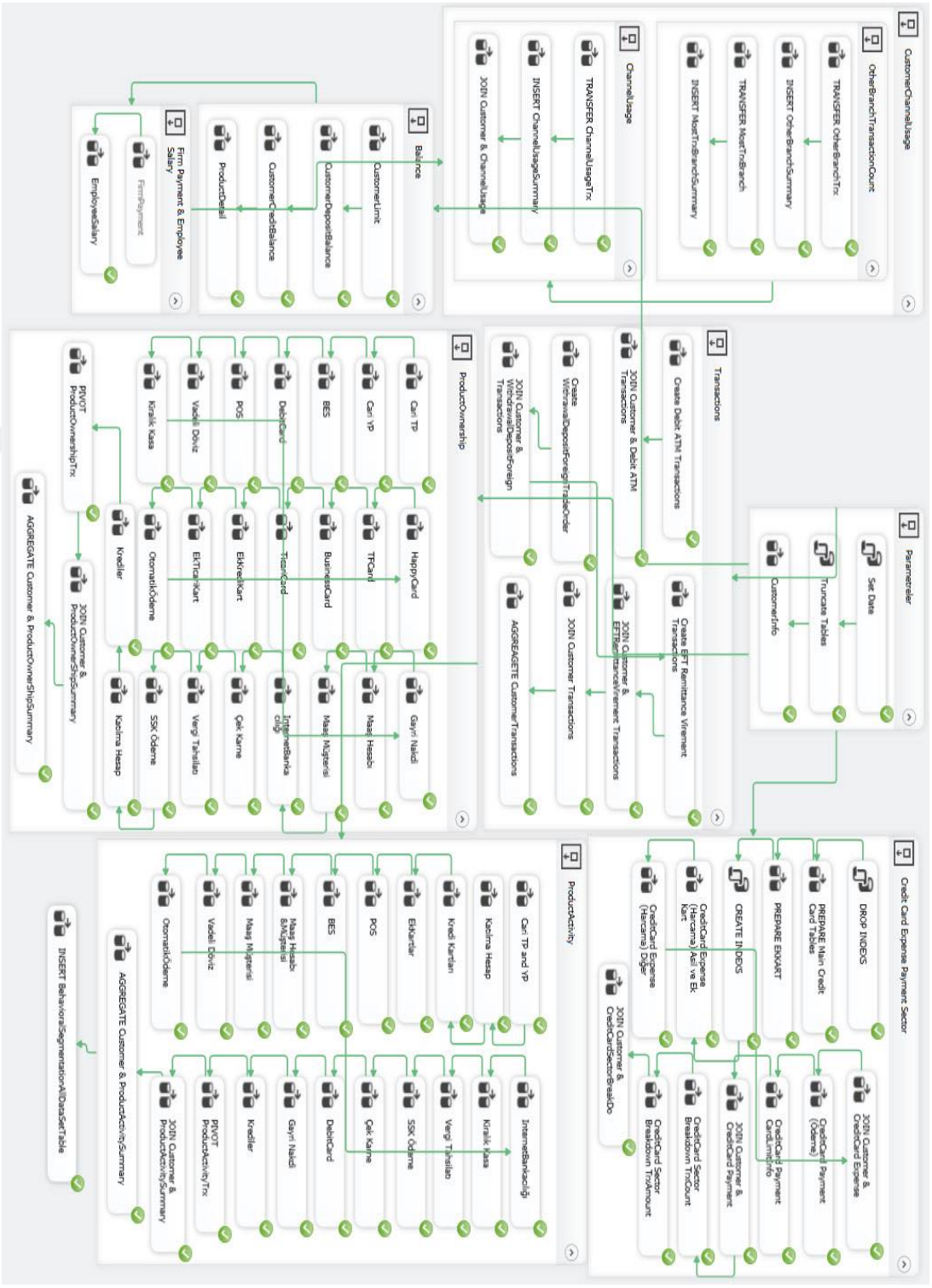
Uygulama iki bölümden oluşmaktadır:

1. **ETL Süreci:** Bu aşamada ilgili veri tabanından değişkenler toplanmakta, birleştirilmekte, veri tutarsızlıkları ortadan kaldırılmakta, veriler temizlenmekte ve değişken seçimi yapılıp, seçilen değişkenler veri tabanında yeni bir tabloya yazılmaktadır. Şekil 9. Kümeleme Analizi İşlem Adımlarında verilmiş olan akış diyagramında kümeleme analizine kadar yapılan işlemler bu süreçte gerçekleştirilmektedir. ETL sürecini yürüten SSIS paketi Şekil 11'de yer almaktadır.
2. **Workflow Süreci:** Bu süreçte seçilmiş olan değişkenlerin kümeleme analizi gerçekleştirilmekte, kümeleme sonuçları görselleştirilmekte ve kümeleri belirlenmiş olan nesneler veri tabanında bir tabloya yazılmaktadır. Bu tablo başka uygulamaların kullanımı için hazır hale getirilmektedir.

ETL Süreci tamamlanmadan Workflow süreci başlamayacağından süreçler arasında öncül ve ardıl şekilde bir sıralama olacaktır. Uygulamaya ait genel akış Şekil 10'da görülmektedir.



Şekil 10. Uygulama Akışı.



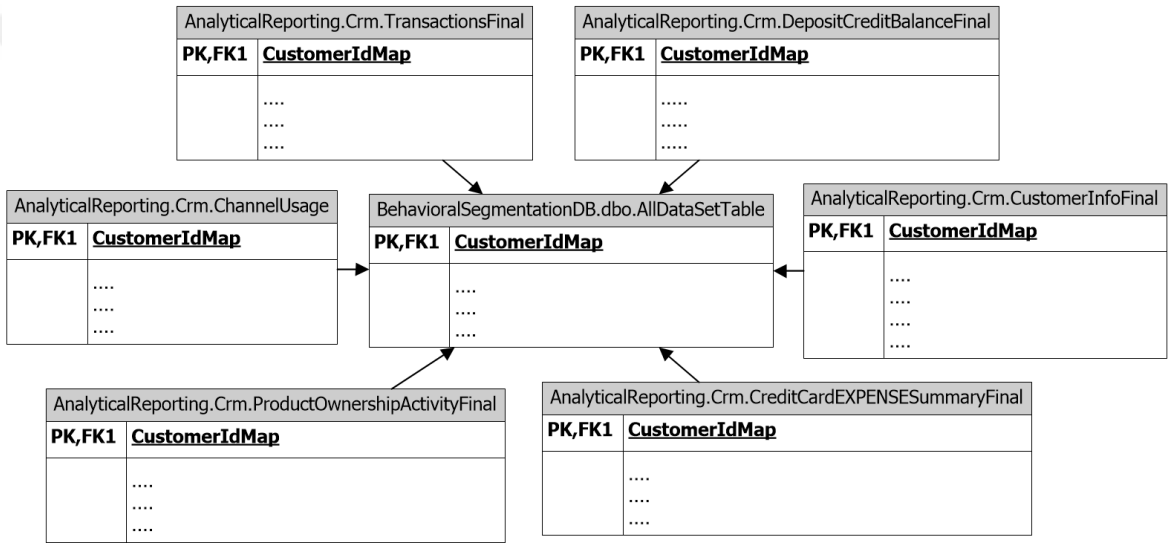
Şekil 11. SSIS Paket Görüntüsü.

5.3. Veri tabanı Tasarımı

Uygulamada 2 ana veri tabanı tablosu yer almaktadır. Bunlar:

- "BehavioralSegmentationDB. dbo.AllDataSetTable"
- "BehavioralSegmentationDB.dbo.BehavioralSegmentation ClusteringResult"

tablolarıdır. Veri tabanı tasarımının ana hatlarının yer aldığı kavramsal şema Şekil 12' de yer almaktadır. "AllDataSetTable" tablosu ETL süreci sonrasında üretilmiş olup kümeleme analizinde kullanılabilecek tüm değişkenleri içeren tablodur.



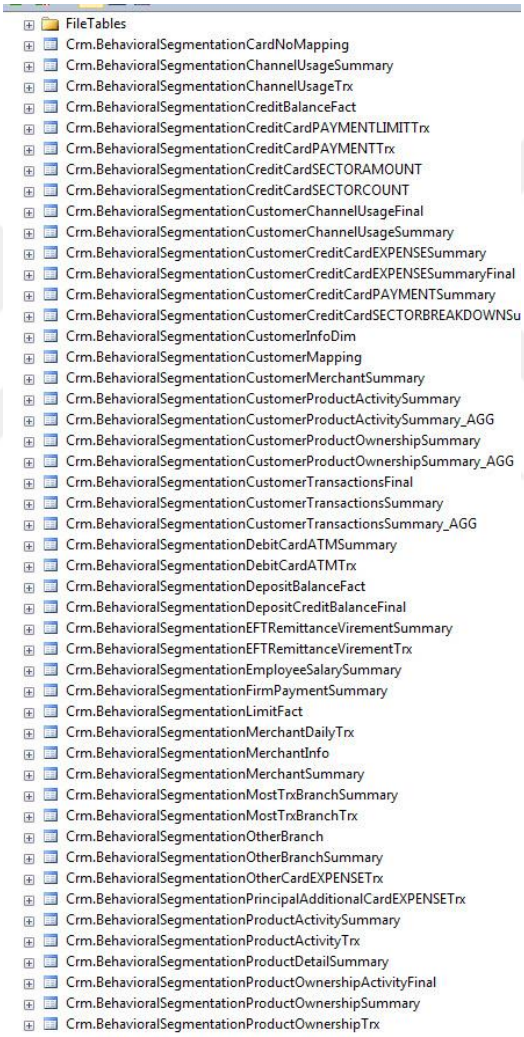
Şekil 12. Veri Tabanı Kavramsal Şeması.

Şekil 12'de görülen tabloların tanımları şöyledir:

- **AnalyticalReporting.Crm.CustomerInfoFinal** : Müşterilerin demografik bilgilerinin yer aldığı tablo
- **AnalyticalReporting.Crm.TransactionsFinal** : Müşterilerin Havale, EFT, Fatura Ödeme vb. işlemlerinin tutulduğu tablo.
- **AnalyticalReporting.Crm.DepositCreditBalanceFinal** : Müşterilerin mevduat ve kredi bakiyelerinin bulunduğu tablo
- **AnalyticalReporting.Crm.ChannelUsage** : Müşterilerin yaptığı işlemleri hangi kanaldan yaptıklarının kaydedildiği tablo

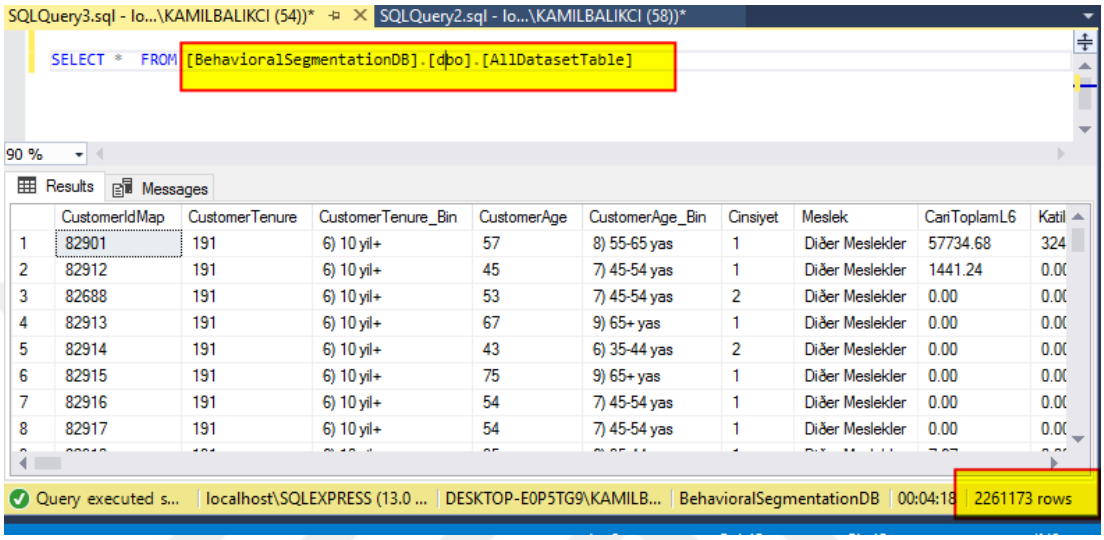
- **AnalyticalReporting.Crm.ProductOwnershipActivityFinal** : Müşterilerin hangi bankacılık ürünlerine sahip ve bu ürünlerden hangilerini hâlâ aktif olarak kullandığının, tutulduğu tablodur.
- **AnalyticalReporting.Crm.CreditCardEXPENSESummaryFinal** : Müşterilerin kredi kartı limitleri, kart harcama tutarlarının tutulduğu tablodur.

AllDataSetTable tablosu yukarıda tanımlarını yapılmış ara tabloları birleştirir. ETL Süreci boyunca üretilmiş tüm ara tablolar Şekil 13'te listelenmiştir.



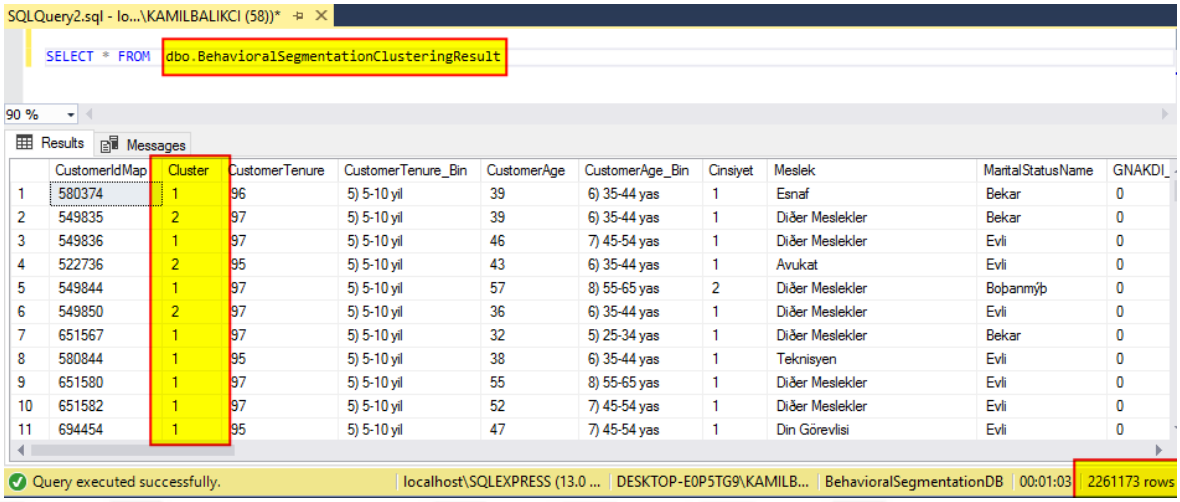
Şekil 13. ETL Süreci Üretilen Ara Tablolar.

“AllDataSetTable” tablosunun fiziksel hâlinin dökümü Şekil 14. ETL Sonuç Tablosunda verilmiştir. Şekil 14’te sağ alt kırmızı çerçevede görüleceği üzere müşteri sayısı kadar satır sayısı oluşmuştur.



Şekil 14. ETL Sonuç Tablosu.

“BehavioralSegmentationClusteringResult” tablosu ise kümeleme analizi sonucunda oluşan tüm müşterilerin hangi kümeye ait olduklarını gösteren tablodur. Bu tablo workflow sürecinin sonucunda oluşturulmaktadır. Tablonun dökümü Şekil 15. Kümeleme Analizi Sonuç Tablosunda verilmiştir. Şekil 15’ in sol alt tarafında yer alan kırmızı çerçeve içinde müşterilerin ait olduğu küme numaraları (Cluster) yer almaktadır. Her müşteriye sadece bir küme değeri ataması yapılmıştır. Aynı şeklin sağ alt köşesinde kırmızı çerçeve içinde çalışma kapsamında küme değerleri tayin edilen toplam müşteri sayısı gösterilmiştir. Bu sonuca göre tüm müşteriler kümelenmiş, küme tayini yapılmamış müşteri kalmamıştır.



	CustomerIdMap	Cluster	CustomerTenure	CustomerTenure_Bin	CustomerAge	CustomerAge_Bin	Cinsiyet	Meslek	MaritalStatusName	GNAKDI_
1	580374	1	96	5) 5-10 yil	39	6) 35-44 yas	1	Esnaf	Bekar	0
2	549835	2	97	5) 5-10 yil	39	6) 35-44 yas	1	Diğer Meslekler	Bekar	0
3	549836	1	97	5) 5-10 yil	46	7) 45-54 yas	1	Diğer Meslekler	Evli	0
4	522736	2	95	5) 5-10 yil	43	6) 35-44 yas	1	Avukat	Evli	0
5	549844	1	97	5) 5-10 yil	57	8) 55-65 yas	2	Diğer Meslekler	Boþanmýþ	0
6	549850	2	97	5) 5-10 yil	36	6) 35-44 yas	1	Diğer Meslekler	Evli	0
7	651567	1	97	5) 5-10 yil	32	5) 25-34 yas	1	Diğer Meslekler	Bekar	0
8	580844	1	95	5) 5-10 yil	38	6) 35-44 yas	1	Teknisyen	Evli	0
9	651580	1	97	5) 5-10 yil	55	8) 55-65 yas	1	Diğer Meslekler	Evli	0
10	651582	1	97	5) 5-10 yil	52	7) 45-54 yas	1	Diğer Meslekler	Evli	0
11	694454	1	95	5) 5-10 yil	47	7) 45-54 yas	1	Din Görevlisi	Evli	0

Şekil 15.Kümeleme Analizi Sonuç Tablosu.

"BehavioralSegmentationClusteringResult" tablosu segmentasyon işleminin sonucunu gösteren tablodur. Bu tablo, Kampanya Yönetimi, Satış Otomasyonu, CRM gibi uygulamalar tarafından kullanılabilir.

5.4. Veri Kümesi ve Veri Hazırlama

Çalışmada, bir Türk bankasının müşteri tabanında yer alan 2.261.173 adet bireysel müşteri verisi kullanılmıştır. Müşterilerin kanal kullanımı, ürün sahipliği, ürün aktifliği, yapılan bankacılık işlemleri ve müşteri bilgilerini içeren değişkenler kullanılmıştır. Başlangıç için 528 adet değişken (feature) oluşturulmuştur. Değişkenler yıllık veya 6 aylık dönemler için toplanmıştır. Yıllık değişkenler için Mart 2016-Şubat 2017 arası, 6 aylık değişkenler için Eylül 2016-Şubat 2017 arası banka verileri kullanılmıştır. Para miktarını gösteren değişkenlerde tüm yabancı para cinsi değerleri ay sonundaki TL kurundan Türk Lirasına dönüştürülmüştür. Böylece para miktarını gösteren tüm değişkenler TL cinsinden standartlaştırılmıştır.

Segmentasyon çalışması müşteri bazında yapılan bir işlem olduğu için veri kümesi müşteri bazında oluşturulmuş ve veriler müşteri numarası bazında tekilleştirilmiştir. Davranışsal segmentasyon değişkenleri, müşterinin bankacılık ürünlerini kullanımını modellemeyi amaçladığından belirli bir andaki veriler yerine belirli bir dönem

aralığındaki veriler kullanılmıştır. Çalışmada müşterilerin davranışlarını modelleyebilmek için yeterli bir süre olan 6 aylık dönem belirlenmiştir. 528 adet değişkeninin büyük çoğunluğu;

- ATMKullanimAdetAy1,
- ATMKullanimAdetAy2,
- ATMKullanimAdetAy3,
- ATMKullanimAdetAy4,
- ATMKullanimAdetAy5,
- ATMKullanimAdetAy6
- …

şeklinde 6 aylık periyotlar hâlinde oluşturulmuştur.

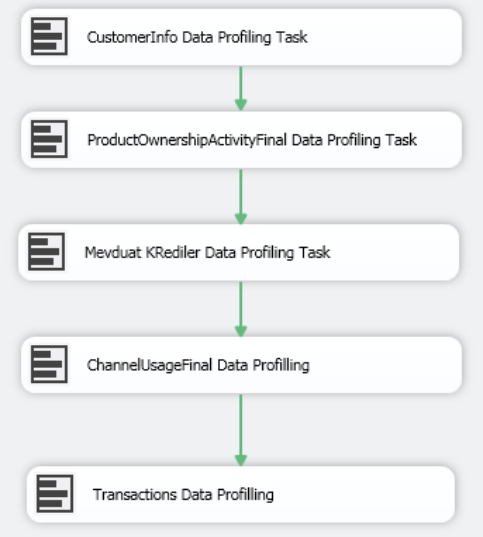
Uygulamanın ETL sürecinde bankanın çeşitli sistemlerinde dağıtık olan veriler konularına göre

- AnalyticalReporting.Crm.CustomerInfoFinal,
- AnalyticalReporting.Crm.TransactionsFinal,
- AnalyticalReporting.Crm.DepositCreditBalanceFinal,
- AnalyticalReporting.Crm.ChannelUsage,
- AnalyticalReporting.Crm.ProductOwnershipActivityFinal,
- AnalyticalReporting.Crm.CreditCardEXPENSESummaryFinal

ara tablolarına doldurulmuştur

5.5. Veri Temizleme

Veri eksiklik ve anomalileri SSIS içinde “Data Profiling Task” kullanılarak tespit edilmiştir. Uygulamada yer alan “Data Profiling Task” görüntülerine Şekil 16. Data Profiling Task Akışı ve Şekil 17. Data Profiling Gösterici şekillerinde yer verilmiştir.



Şekil 16. Data Profiling Task Akışı.

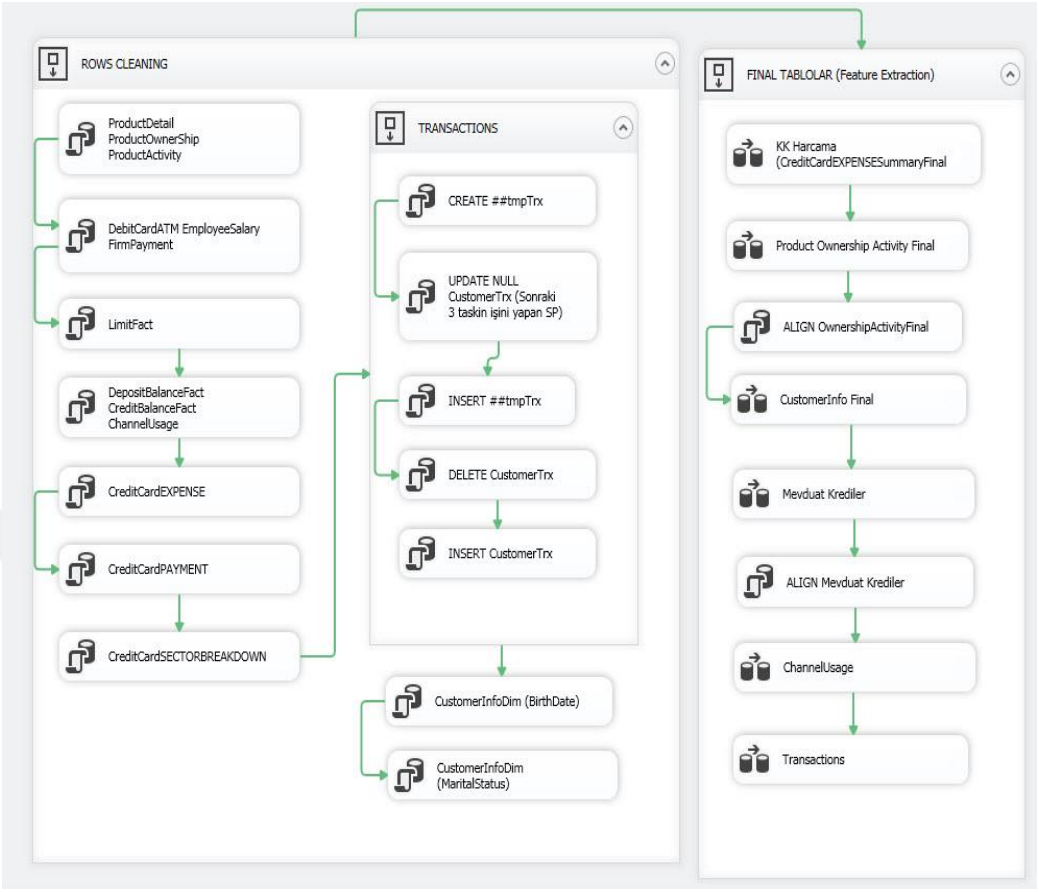
Şekil 17'de kırmızı çerçeveye alınan kısımda görüldüğü üzere minimum değeri sıfır olması gereken değişken değerlerinin veri anomolilerine dikkat çekilmiştir.

Column Statistics Profiles - [Crm].[BehavioralSegmentationCustomerChannelUsageFinal]

Column	Minimum	Maximum	Mean
MobilSubeKullanimAdetAy8	0	3441	0.08757400467...
MobilSubeKullanimAdetAy9	0	2627	0.09219221417...
MobilSubeKullanimTutarAy1	-5.00	5121015.00	268.156732611...
MobilSubeKullanimTutarAy10	-3.00	4723700.28	277.628739457...
MobilSubeKullanimTutarAy11	-6.00	4871824.91	193.253862676...
MobilSubeKullanimTutarAy12	-1.00	800000.00	10.8501170360...
MobilSubeKullanimTutarAy2	-3.00	6391756.55	311.225189997...
MobilSubeKullanimTutarAy3	-4.00	7126841.91	299.188920279...
MobilSubeKullanimTutarAy4	-3.00	16556445.32	300.844917244...
MobilSubeKullanimTutarAy5	-2.00	8093730.80	250.514075952...
MobilSubeKullanimTutarAy6	-3.00	10147076.67	237.896113062...
MobilSubeKullanimTutarAy7	-8.00	7117572.52	275.221691226...
MobilSubeKullanimTutarAy8	-2.00	7000003.30	264.296896451...
MobilSubeKullanimTutarAy9	-4.00	6913454.07	252.171956404...
MobilSubeRasyoL6	0.00	1.00	0.01066088063...
MobilSubeToplamAdetL6	0	11612	0.59509860382...

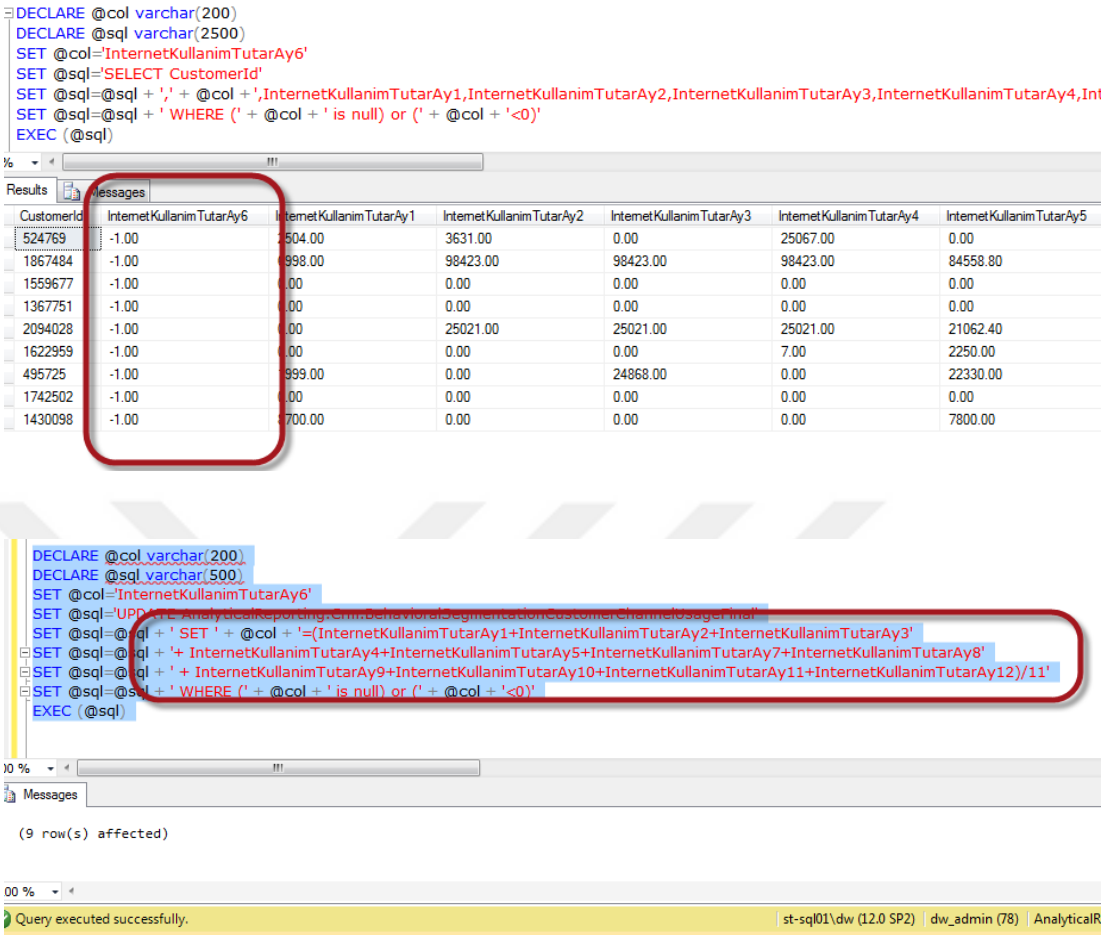
Şekil 17. Data Profiling Gösterici.

Veri eksiklik ve anomalilerinin ortadan kaldırılması işlemleri SSIS paketinin içine kodlanan "SQL" scriptleri ile sağlanmıştır. SQL scriptlerinin ve bu scriptleri çalıştıran SSIS görevlerinin görüntülerine, Şekil 18. SSIS Veri Temizleme Görevleri ve Şekil 19. Veri Temizleme Script Örnekleri şekillerinde yer verilmiştir.



Şekil 18. SSIS Veri Temizleme Görevleri.

Şekil 17. Data Profiling Göstericisinde bir örneği verilen veri anomalileri, Şekil 18'de yer alan SSIS görevlerinin sıra ile çalışması sonucunda ortadan kaldırılmıştır. Şekil 19'da SSIS görevleri içerisinde çalıştırılan SQL scripti örneği yer almaktadır. Şekil 19'da kırmızı çerçeve içerisinde anomali oluşturan yani değeri sıfırdan küçük olan ya da "null" olan değişken değeri yerine ortalama değerinin kullanımı sağlayan SQL script örneği yer almaktadır.



Şekil 19. Veri Temizleme Script Örnekleri.

5.6. Özellik/Değişken Seçimi veya Çıkarımı

528 adet değişken veri madenciliği algoritmalarının işleyebilmesi ve sonuç üretebilmesi açısından çok fazladır. Ayrıca değişkenler arasında birbirine bağımlı değişkenler olduğundan 6 aylık ölçümler şeklindeki değişkenlerden yeni değişkenler oluşturulmuş ve bu sayede ver tabanı boyutları indirgenmiştir.

Ticari Kart, Ek Ticari Kart, Kurumsal Finansman Desteği, 0 faizli kredi, Leasing, Karar-Zarar Ortaklığı değişkenlerinin bazıları bireysel müşterilerin kullanımına izin verilmeyen ürünler olması, bazıları da tüm değerlerinin sıfır olması nedeniyle değişken kümesinden çıkarılmıştır.

Mevduat hareketlerinin ve bakiyenin azalıp azalmadığını tespit için MEVDUAT_TREND_L6 değişkeni oluşturulmuştur. 6 aylık mevduat bakiyelerinden tek değişkenli doğrusal regresyon modeli kurularak eğim bulunmuştur. Bu değişken, mevduatın azaldığını, sabit kaldığını veya arttığını tespit etmeye olanak sağlamakta olup (-1) ve (1) arasında değer almaktadır. MEVDUAT_TREND_L6 değişkenin üretilmesinde kullanılan regresyon denklemi ve β katsayısının hesaplama formüllerine, Denklem 9. Tek Değişkenli Doğrusal Regresyon Modeli ve Denklem 10. Doğrusal Regresyon β formülü başlıklarında yer verilmiştir.

$$y_i = \beta\, x_i + \alpha$$

Denklem 9. Tek Değişkenli Doğrusal Regresyon Modeli.

x_i : Her ayın gösterir.

y_i : Her ayın Mevduat bakiyesi

n : Toplam Ay sayısı

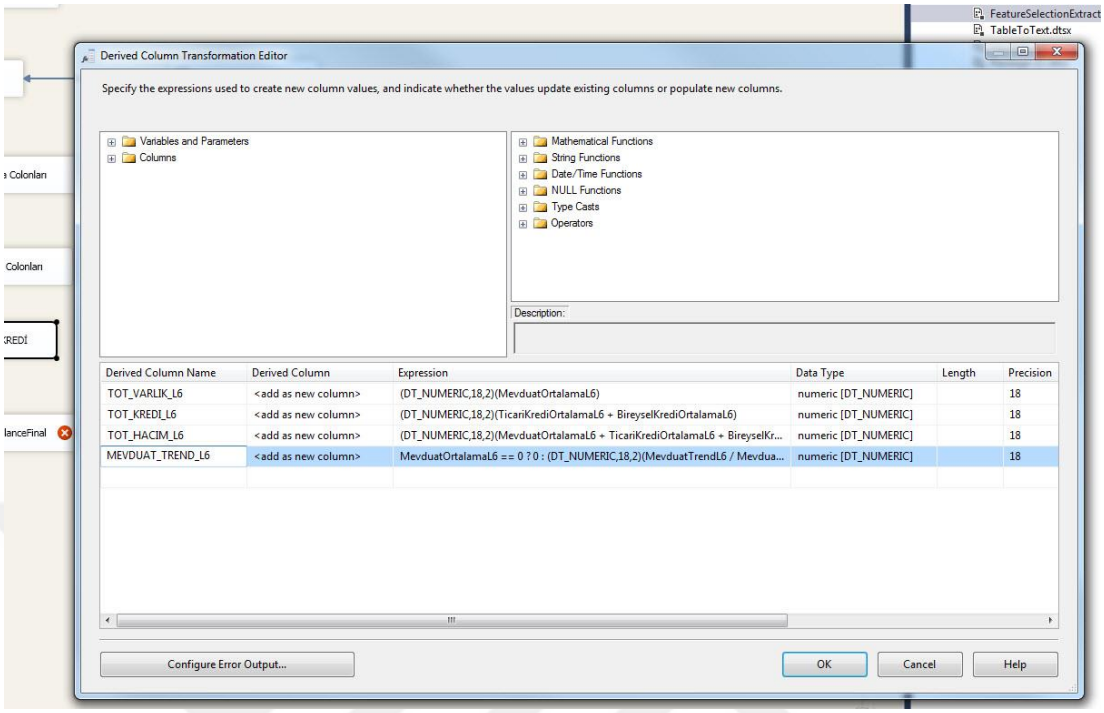
β : Artış ve azalışı gösteren eğimdir.

Buna göre a eğim bilgisi MEVDUAT_TREND_L6 değişkenin değeridir.

$$\beta = \frac{\left(n \sum_{i=1}^{n} x_i\ y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)\right)}{n\,(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2}$$

Denklem 10. Doğrusal Regresyon β Katsayı formülü.

MEVDUAT_TREND_L6 dışında kümeleme analizine girecek tüm değişkenler 528 adet değişkenden türetilmiştir. Kümeleme Analizinde kullanılan tüm değişkenler EK-A'da yer almaktadır. Değişken türetme işlemleri SSIS paket görevlerinin (task) içerisinde yapılmıştır. Değişken çıkarımlarına ait örnek bir görev Şekil 20'de yer almaktadır.



Şekil 20. SSIS Değişken Çıkarım Örneği.

5.7. Kümeleme Algoritması Seçimi ve Tasarımı

Problemin yapısı, ulaşılmak istenen hedef, veri kümesinde kullanılan değişkenlerin türleri, algoritma çalışma süresi, nesnelerin kümelere atanma biçimi olarak ifade edilen kriterler uygulanacak algoritmanın seçimine etki etmektedir. Bütün kriterlere karşılık gelecek mükemmel bir kümeleme algoritması yoktur. Bu yüzden seçilen kümeleme algoritması çalışmanıza en uygun olan algoritma olmalıdır. Problemin yapısı ve ulaşılmak istenen hedef kriterlerine göre müşterinin davranışsal segmentasyonunda müşterilerin mutlaka bir kümenin üyesi olması gerekiyor.

Kümeleme algoritması seçimi sırasında aşağıdaki kriterler dikkate alınmalıdır:

- **Veri kümesinde kullanılan değişken türü:** seçilecek algoritmanın karışık veri türleri ile çalışabiliyor olması gerekiyor. Çünkü, müşteri davranışsal segmentasyon değişkenlerine ait veri kümesi ikili, oransal değişken türlerini içermektedir.

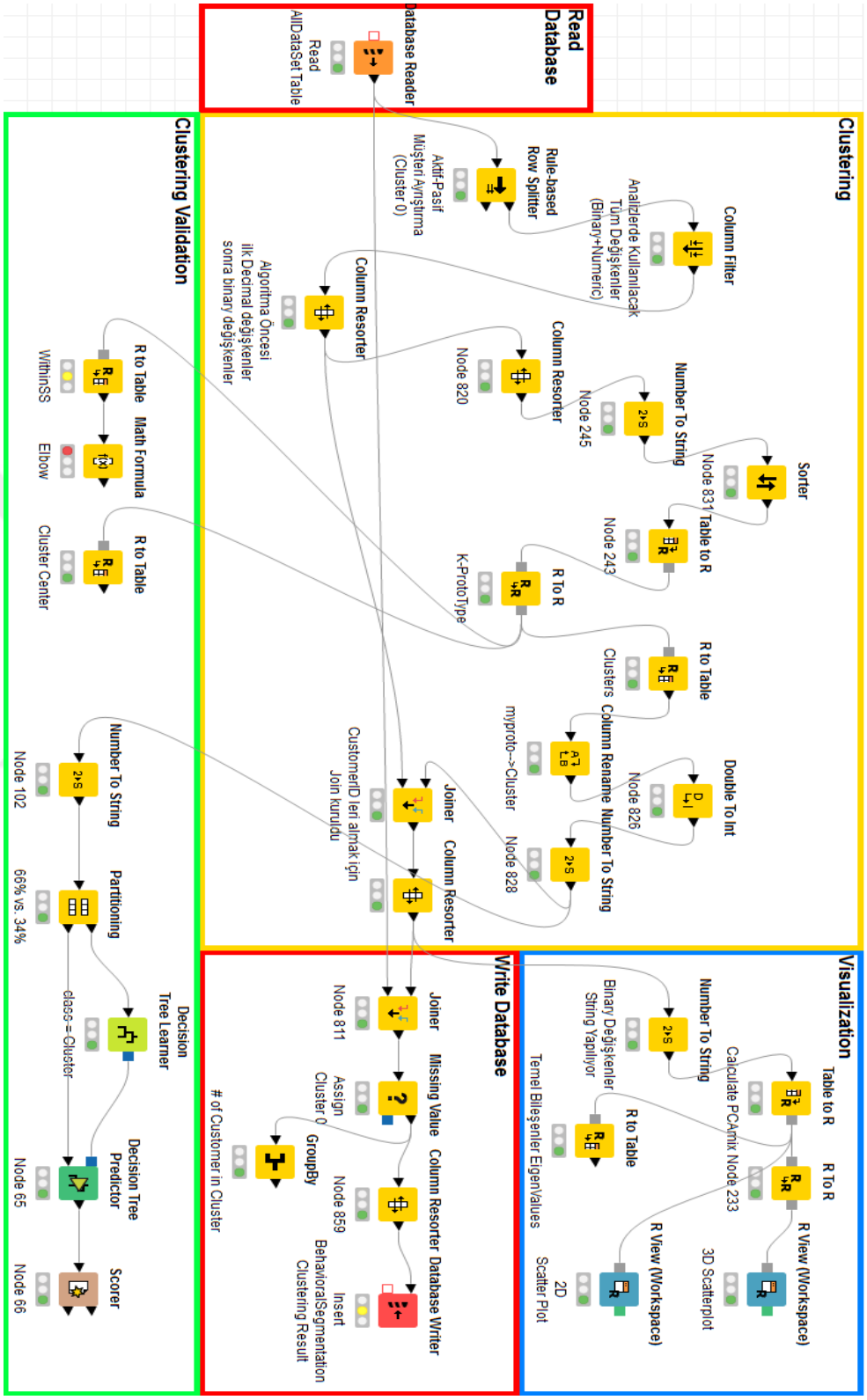
- **Algoritma çalışma süresi:** seçilecek algoritma büyük ve çok boyutlu veri kümesini işleyebilecek bir çalışma süresinde olması gerekmektedir. Bu tez çalışmasında, 2.261.173 nesneli 35 değişkenli büyük ve çok boyutlu bir veri kümesi kullanılmıştır.
- **Nesnelerin kümelere atanma biçimi:** kümeleme analizi sonucunda her nesne yalnızca bir kümenin üyesi olmalıdır. Bu nedenle katı kümeleme yapan bir algoritma seçilmelidir.

Bölüm 3.4'te açıklanmış olan yöntemler karışık veri türlerini işleyebilen kümeleme algoritmalarıdır. Bu algoritmalar içerisinde yer alan Gower Metrik Uzaklığı ve İki Aşamalı Kümeleme algoritmaları benzerlik/uzaklık matrisini kullanarak çalışan algoritmalardır. Benzerlik/Uzaklık matrisi nesnelerin birbirleri arasında olan uzaklık/yakınlıklarını hesapladığından veri kümesi 2.261.173 x 2.261.173 boyutlu bir matris oluşturmaktadır. Böyle bir matris üzerinde kümeleme analizi, çalışma süresinin uzunluğu nedeniyle tercih edilmemiştir.

Karışım modelleri kullanılarak karışık veri tiplerinde olan veri setine kümeleme analizi uygulanabilir. Ancak karışım modelleri sonucunda veri nesnesinin her kümeye ait olma olasılığı vardır. Katı kümeleme yapılması zorunluluğundan dolayı bu algoritmada problemin çözümü olarak tercih edilmemiştir.

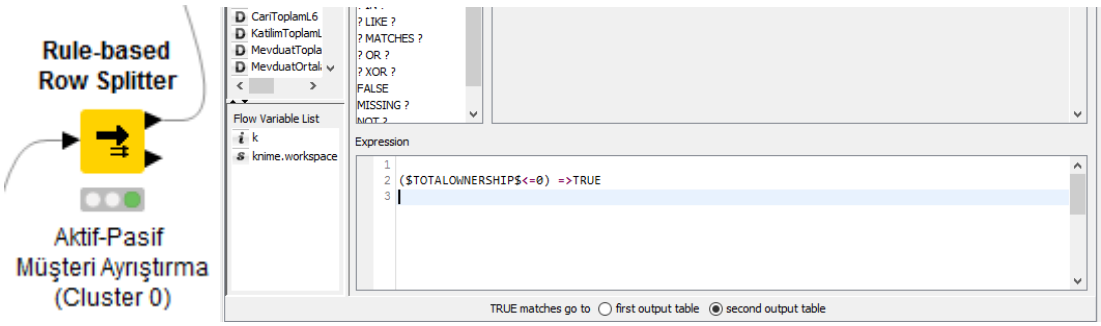
Karışık veri tipi işleyebilme, kısa çalışma süresi, katı kümeleme yapabilmesi açısında en uygun olan k-prototip algoritması uygulamada kullanılacak kümeleme algoritması olarak tercih edilmiştir.

Uygulamanın kümeleme analizi çalışması Workflow Süreci bölümünde yer almaktadır. Uygulamanın KNIME veri analiz yazılımında oluşturulan akış diyagramı Şekil 21'de gösterilmektedir.



Şekil 21. KNIME üzerinde kümeleme analizi iş akışı.

Workflow sürecinde iş akışı KNIME düğümlerinin birbirine bağlanması ile oluşturulur. Her düğüm kendisine ait bir görev yerine getirmektedir. Akış, ETL sürecinin çıktısı olan "AllDataSetTable" tablosu okunarak başlar. Bankalar müşteri veri tabanlarında, banka faaliyet göstermeye başladığı tarihten itibaren tüm müşterilerinin bilgisini tutmaktadır. Banka ile yıllar öncesinde çalışmış ancak uzun süredir çalışmayan bir müşterinin kayıtları hâlâ banka müşteri veri tabanında saklıdır. Bu tip müşterilere pasif müşteri denmektedir. Tez çalışmasında kullanılan veri kümesinde pasif müşteriler de yer almakta olup, veri kümesinin büyük bir çoğunluğunu bu pasif müşteriler teşkil etmektedir. Pasif müşterilerin veri kümesindeki değişken değerleri sıfırdır. Pasif müşterilerin veri kümesinden ayırılması ve bir kümeye atanması hem kümeleme analizinin sağlıklı bir sonuç üretmesi hem de bu müşterilerin tespit edilmiş olması açısından önemlidir. Çalışmada, pasif müşteriler kümeleme analizine sokulmadan önce ayrıştırılmış ve bu müşteriler "Cluster 0" kümesine atanmıştır. Müşterilerin banka ile çalışıp, çalışmadığı TOTEL_OWNERSHIP değişkeni ile tespit edilmektedir. Bu değişkenin değeri 0 "sıfır" olan müşteriler pasif müşteri olarak atanmıştır. Pasif müşterileri ayrıştırmayı sağlayan "Rule-based Row Splitter" düğümü Şekil 22'de gösterilmiştir.

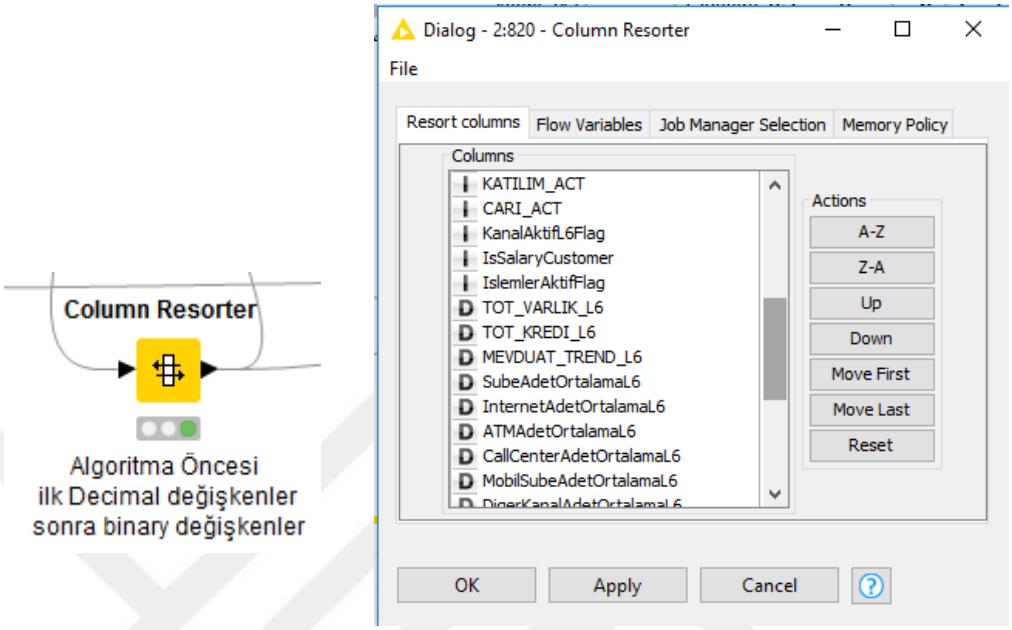


Şekil 22. Pasif Müşterilerin Tespiti.

Böylelikle ilk davranışsal segmentasyon kümesi "Cluster 0 - Pasifler" olarak oluşturulmuştur.

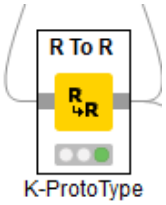
Akışın kümeleme analizi kısmına geçilmeden önce değişkenler veri türlerine göre sıraya dizilmiştir. Değişkenler, İkili değişkenler önce, sayısal değişkenler sonra olacak

şekilde “Column Resorter” düğümü ile sıralanmıştır. Sıralama işlemi Şekil 23’de gösterilmiştir.



Şekil 23. Değişkenlerin Sıralanması.

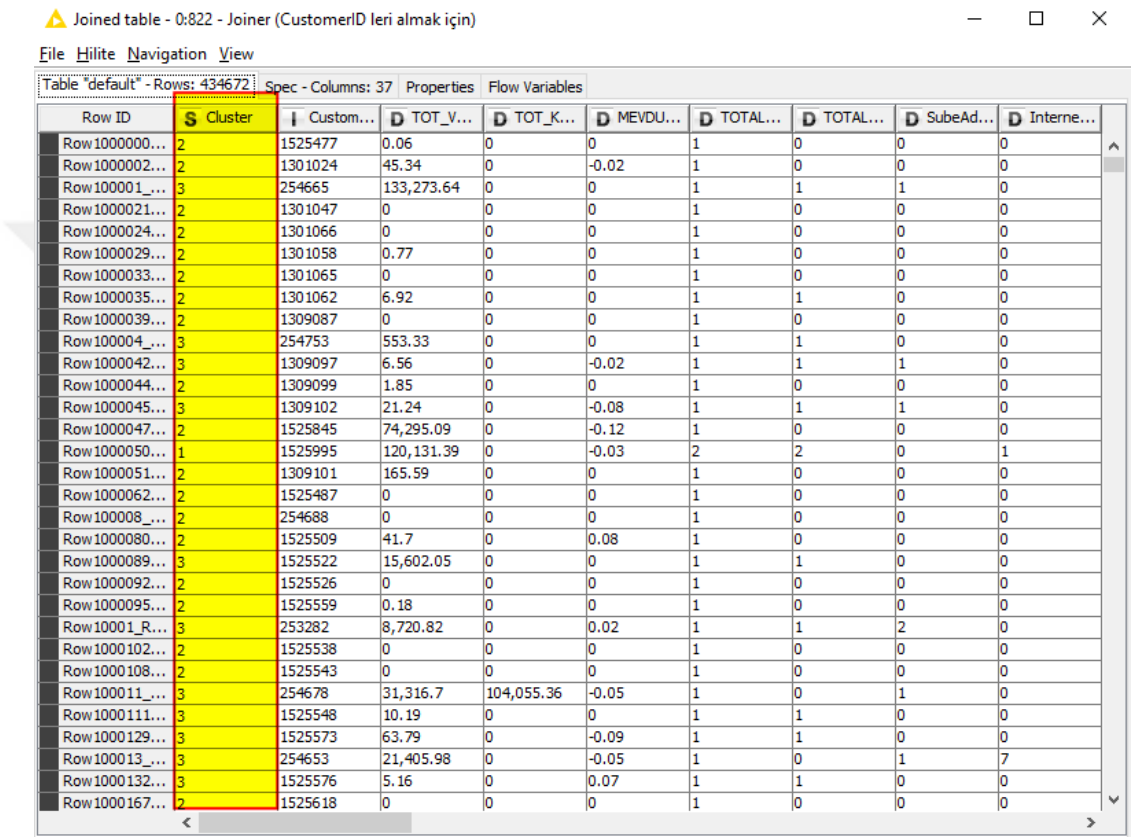
Değişkenlerin türlerine göre sıralanmasının nedeni k-prototip algoritmasının sayısal değişkenler üzerinde farklı işlemleri yapabilmesi içindir. Veri kümesine kümeleme analizi, “R To R” düğümü içerisine yazılan R kodları ile gerçekleştirilmiştir. Algoritmanın çalıştığı düğüm Şekil 24. Kümeleme Analizi Düğümünde gösterilmiştir.



Şekil 24. Kümeleme Analizi Düğümü.

Bu düğüm içerisinde yazılan R kodu; k-prototip algoritmasına girdi olarak küme sayısı (k) ve kümeleme analizi yapılacak veri kümesini girdi parametresi olarak alır.
Algoritma çıktı olarak veri kümesinde yer alan her nesnenin (banka müşterisi) atandığı küme indeks değerini döndürür. “R To R” düğümünde yazılmış olan k-prototip algoritmasının R kodu “EK-B K-Prototip Algoritması R Kodu bölümünde verilmiştir.

Uygulanan k-prototip algoritmasının çalışması sonrasında müşterilerin atanmış olduğu küme ve değişkenlerin değerleri Şekil 25'te yer almaktadır. Şekil 25'te yer alan kırmızı çerçevede müşterilerin atandığı davranışsal küme bilgisi vurgulanmıştır.


Joined table - 0:822 - Joiner (CustomerID leri almak için)

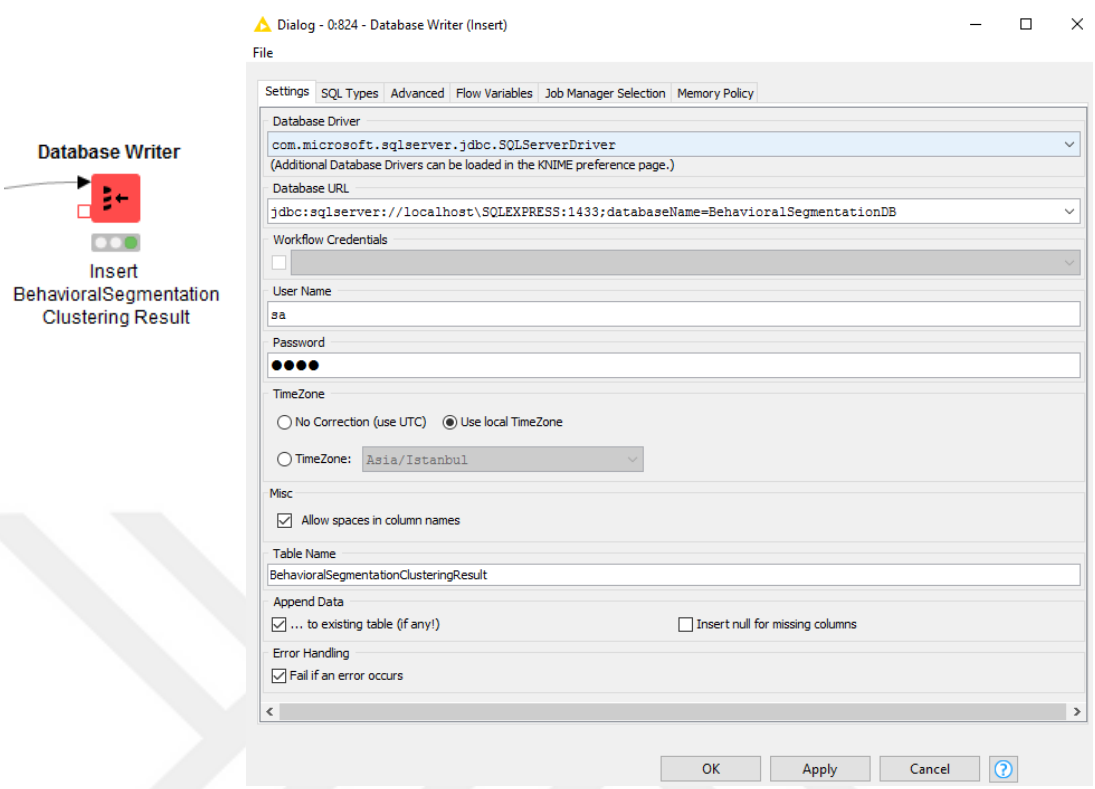
File Hilite Navigation View

Table "default" - Rows: 434672 | Spec - Columns: 37 | Properties | Flow Variables

Row ID	Cluster	Custom...	TOT_V...	TOT_K...	MEVDU...	TOTAL...	TOTAL...	SubeAd...	Interne...
Row1000000...	2	1525477	0.06	0	0	1	0	0	0
Row1000002...	2	1301024	45.34	0	-0.02	1	0	0	0
Row100001_...	3	254665	133,273.64	0	0	1	1	1	0
Row1000021...	2	1301047	0	0	0	1	0	0	0
Row1000024...	2	1301066	0	0	0	1	0	0	0
Row1000029...	2	1301058	0.77	0	0	1	0	0	0
Row1000033...	2	1301065	0	0	0	1	0	0	0
Row1000035...	2	1301062	6.92	0	0	1	1	0	0
Row1000039...	2	1309087	0	0	0	1	0	0	0
Row100004_...	3	254753	553.33	0	0	1	1	0	0
Row1000042...	3	1309097	6.56	0	-0.02	1	1	1	0
Row1000044...	2	1309099	1.85	0	0	1	0	0	0
Row1000045...	3	1309102	21.24	0	-0.08	1	1	1	0
Row1000047...	2	1525845	74,295.09	0	-0.12	1	0	0	0
Row1000050...	1	1525995	120,131.39	0	-0.03	2	2	0	1
Row1000051...	2	1309101	165.59	0	0	1	0	0	0
Row1000062...	2	1525487	0	0	0	1	0	0	0
Row100008_...	2	254688	0	0	0	1	0	0	0
Row1000080...	2	1525509	41.7	0	0.08	1	0	0	0
Row1000089...	3	1525522	15,602.05	0	0	1	1	0	0
Row1000092...	2	1525526	0	0	0	1	0	0	0
Row1000095...	2	1525559	0.18	0	0	1	0	0	0
Row10001_R...	3	253282	8,720.82	0	0.02	1	1	2	0
Row1000102...	2	1525538	0	0	0	1	0	0	0
Row1000108...	2	1525543	0	0	0	1	0	0	0
Row100011_...	3	254678	31,316.7	104,055.36	-0.05	1	0	1	0
Row1000111...	3	1525548	10.19	0	0	1	1	0	0
Row1000129...	3	1525573	63.79	0	-0.09	1	1	0	0
Row100013_...	3	254653	21,405.98	0	-0.05	1	0	1	7
Row1000132...	3	1525576	5.16	0	0.07	1	1	0	0
Row1000167...	2	1525618	0	0	0	1	0	0	0

Şekil 25. Kümeleme Analizi Sonucu.

Kümeleme analizi sonucu davranışsal kümeleri belirlenmiş müşteriler, bankanın diğer uygulamaları tarafından kullanılabilmesi için Şekil 26'da gösterilen KNIME düğümü ile veri tabanına kaydedilir.



Şekil 26. Veri Tabanına Yaz Düğümü.

5.8. Kümeleme Geçerliliği

Kümeleme analizinde geçerlilik sınaması aşağıdaki yöntemlerle sağlanabilmektedir:

1. İç ve dış indeks kullanımı
2. Örnekleme ile model seçimi
3. Görselleştirme

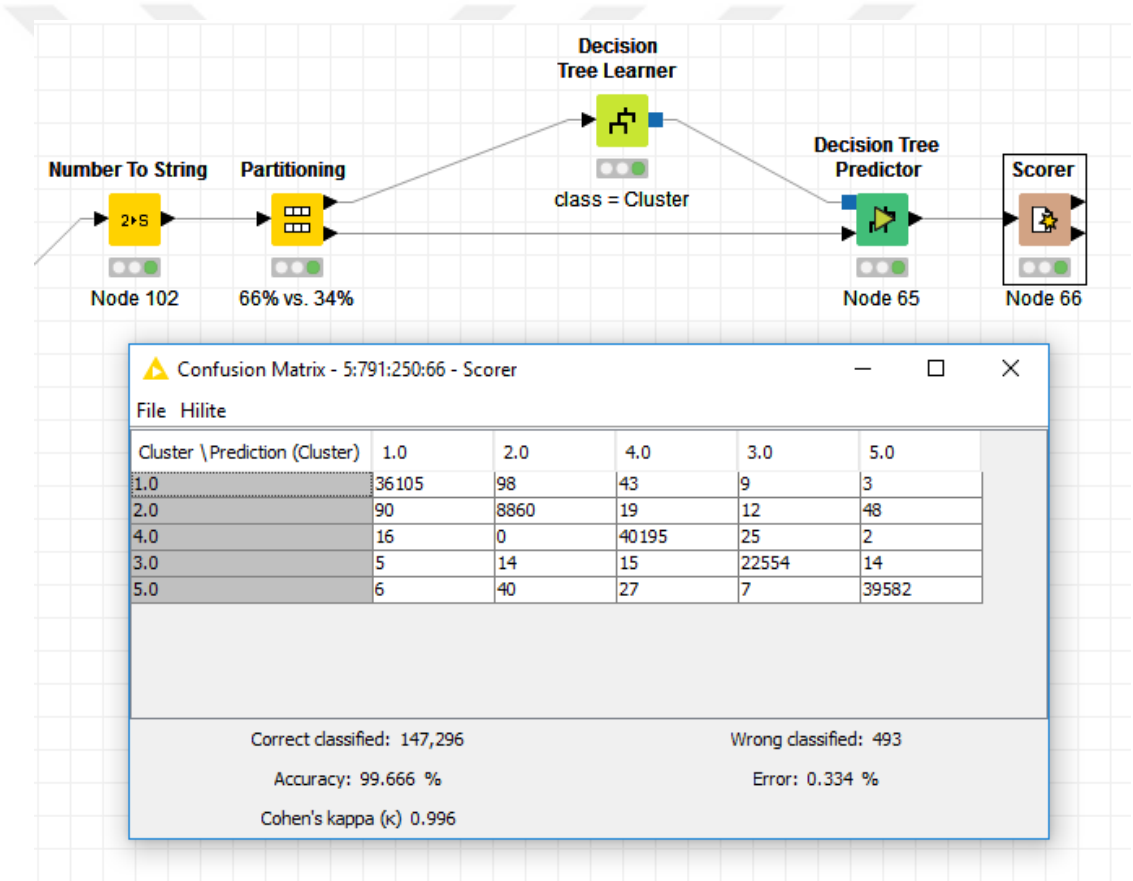
İç indeksler, oluşan küme yapısı ve kümenin oluşmasını sağlayan veri arasındaki ilişkinin ölçümüdür. İndeks değeri ne kadar iyi ise kümeleme analizi de o kadar güvenilirdir.

Dış indeksler ise uzman tarafından sunulan veya başka bir kümeleme algoritması tarafından bulunan kümeleme analizi ile çalışmada oluşan kümeleme analizinin karşılaştırılmasını sağlayan indekslerdir [62].

Örnekleme ile geçerlilik sınaması, veri kümesinden örneklem (resampling) alındıktan sonra verinin bir modele tabi tutulup kümeleme sonuçlarının güvenirliğinin test edilmesidir.

Bu çalışmada, kümeleme analizi sonrasında oluşan "Cluster" bilgisi, karar verme değişkeni olacak şekilde bir karar ağacı algoritması uygulanmış ve bu sayede kümeleme analizinin doğruluğu test edilmiştir. Uygulamada kümeleme geçerliliği için kullanılan karar ağacı workflow akışı ve karar ağacı skorlaması Tablo 1. Karar Ağacı sonrası Karışıklık Matrisinde yer almaktadır.

Tablo 1. Karar Ağacı sonrası Karışıklık Matrisi.



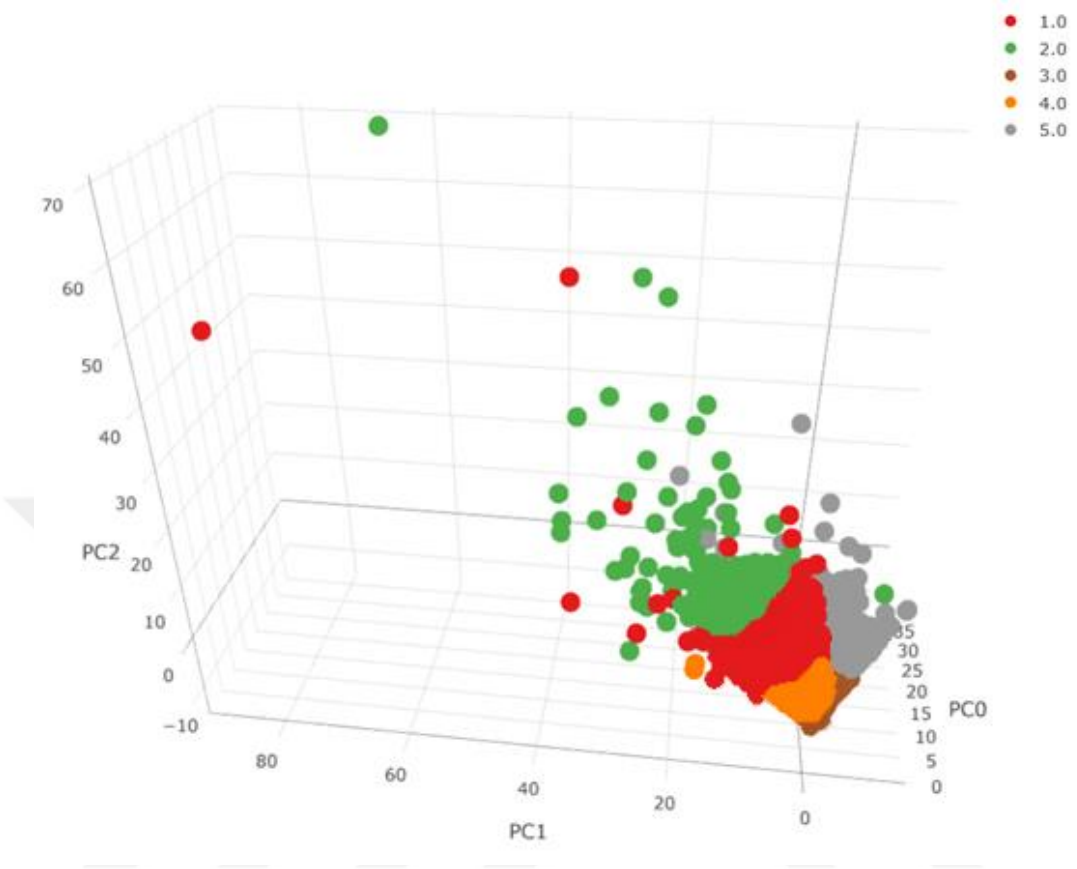
Cluster \ Prediction (Cluster)	1.0	2.0	4.0	3.0	5.0
1.0	36105	98	43	9	3
2.0	90	8860	19	12	48
4.0	16	0	40195	25	2
3.0	5	14	15	22554	14
5.0	6	40	27	7	39582

Kümeleme analizi sonuçları 2 veya 3 boyutlu uzayda görselleştirme araçları ile izlenebilmektedir. Görselleştirmede, küme içi üyelerin birbirlerine olan uzaklığının küçük (compactness), kümeler arası uzaklığın yani ayrışmanın ise belirgin olması (well-separeted) kümeleme analizinin geçerliliğinin sınanmasını sağlar.

Bu tez çalışmasının veri kümesi, çok boyutlu ve karışık veri türlerinden oluştuğundan boyut indirgeme yöntemi olarak karışık veri türleri için kullanılabilen Temel Bileşenler Analizi uygulanarak veri kümesi 3 boyuta indirgenmiş ve bu 3 boyutun görselleştirilmesi yapılmıştır. Temel bileşenler analizi için R programlama ortamında "PCAmixdata" kütüphanesi kullanılmıştır. Karışık tip veriler için uygulanan temel bileşenler analizine ait R koduna EK-C Karışık Tip Veriler için Temel Bileşenler Analizi R Kodu bölümünde yer verilmiştir.

3 boyuta indirgenmiş kümeleme sonuç verilerinin görselleştirilmesi için R programlama ortamında "plotly" kütüphanesi kullanılmıştır. Görselleştirmeyi sağlayan R koduna EK-D Kümeleme Analizi Sonucunu Görselleştiren R Kodu bölümünde yer almaktadır.

Görselleştirmeye ait çıktının görüntüsüne Şekil 27'de yer verilmiştir. Plotly Kütüphanesi ile 3 boyutlu uzayda yapılan görselleştirmede 5 kümenin küme içi uzaklıklarının yakın ve kümeler arası ayrışmanın da belirgin olduğu gözlemlenmektedir.

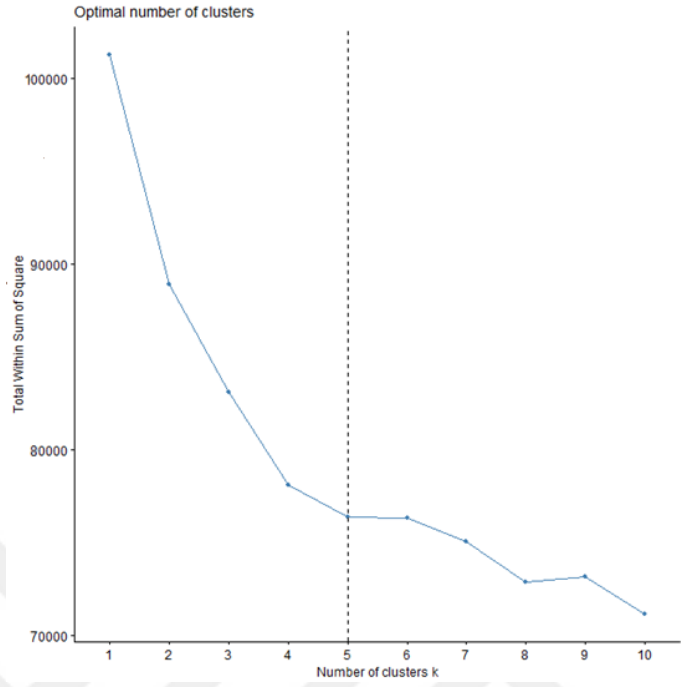


Şekil 27. Kümeleme Analizinin 3 Boyutlu Görselleştirilmesi.

k-means, k-modes, k-prototype gibi k türü bölümleme algoritmalarında kümeleme geçerliliği için önemli bir geçerlilik unsuru da doğru küme sayısının seçilmesidir. Bu çalışmada küme sayısı tespiti için küme içi hata karelerinin toplamına dayanan "Elbow Metodu" kullanılmıştır. Elbow metodu,

1. Kümeleme analizi farklı küme sayılarına göre çalıştırılır.
2. Analiz sonucu ortaya çıkan küme prototipleri ve bu kümelere ait olan küme üyelerinin arasındaki uzaklığın karesi alınır ve bu kareler toplanarak hesaplanır.

Farklı küme sayıları için oluşturulan elbow grafiği ile optimum küme sayısı 5 olarak tespit edilmiştir. Veri kümesine hiyerarşik kümeleme algoritması uygulayarak da küme sayısı tespit edilebilmektedir. Veri kümesinin çok boyutlu olması nedeniyle küme sayısı tespitinde hiyerarşik kümeleme yerine elbow metodu tercih edilmiştir. Elbow metodu sonucu oluşan grafik Şekil 28'de gösterilmiştir.



Şekil 28. Elbow Grafiği.

Küme sayısına göre oluşan küme içi uzaklık kareleri toplam değerlerine de Tablo 2. Küme İçi Uzaklık Kareleri Toplamı tablosunda yer verilmiştir.

Tablo 2. Küme İçi Uzaklık Kareleri Toplamı

Küme Sayısı (k)	**Küme İçi Uzaklık Kareleri Toplamı**
1	10.626.735.364.625.824
2	87.959.379.877.360.384
3	84.782.233.669.023.632
4	78.431.349.757.750.416
5	77.409.961.333.918.320
6	77.923.154.573.313.648
7	75.921.871.298.366.200
8	72.757.922.994.259.992
9	72.475.918.326.256.816
10	70.021.326.522.712.956

5.9. Kümelerin Profillemesi

Profillemede amaç 5 kümeye göre kümelere ayrıştırılması yapılan müşterilerin ortak özelliklerinin ortaya konmasıdır. Kümeleme algoritması tamamlandığında, her nesnenin atandığı küme tespit edildiği gibi her kümenin nihai küme merkezleri olan prototiplerde sonuç olarak elde edilmiş olacaktır.

Küme merkezleri, kümeye ait nesnelerin ortak özelliklerini en net şekilde temsil eder. Profilleme çalışması, küme merkezleri olan prototiplerin değişken değerleri yorumlanarak oluşturulur. 5 küme için uygulanan kümeleme analizi sonrasında oluşan prototiplerin aldığı değişken değerleri Tablo 3'te yer almaktadır.

Tablo 3. Küme Prototiplerinin (Merkezlerinin) Değişken Değerleri.

Değişken	Küme 1	Küme 2	Küme 3	Küme 4	Küme 5
GNAKDI_KREDI_OWN	0	1	0	0	0
NAKDI_KREDI_OWN	0	0	0	0	0
BIREYSEL_KREDI_OWN	1	1	0	1	1
KK_OWN_by_KK_HARCAMA	1	1	0	1	1
KATILIM_OWN	0	1	1	0	1
CARI_OWN	1	1	1	1	1
GNAKDI_KREDI_ACT	0	1	0	0	0
NAKDI_ACT	0	0	0		0
BIREYSEL_KREDI_ACT	1	1	0	1	1
KK_ACT_by_KK_HARCAMA	1	1	0	0	1
KATILIM_ACT	0	1	0	0	1
CARI_ACT	1	1	1	1	1
KanalAktifL6Flag	1	1	0	0	1
IsSalaryCustomer	0	0	0	0	1

Değişken	Küme 1	Küme 2	Küme 3	Küme 4	Küme 5
IslemlerAktifFlag	1	1	0	0	1
GNKrediOrtalamaL6	16,155	1.251,90	53,76	8,65	68,7
TOT_VARLIK_L6	4.454,97	151.484,75	4.819,02	863,02	9.967,82
TOT_KREDI_L6	4.461,21	36.709,31	2.980,00	1.378,85	7.650,15
MEVDUAT_TREND_L6	-0.004	0.022	-0.028	-0.033	0,073
SubeAdetOrtalamaL6	0,825	1,835	2,028	2,021	0,654
InternetAdetOrtalamaL6	0,432	0,918	0,002	0,002	0,392
ATMAdetOrtalamaL6	0,662	0,503	0,001	3,023	0,858
CallCenterAdetOrtalamaL6	0,007	0,004	0	0	0,003
MobilSubeAdetOrtalamaL6	0,572	0,715	0,001	0,002	0,437
DigerKanalAdetOrtalamaL6	0,026	0,461	0	0	0,264
SalaryCustomerHacimL6	47,03	38,91	11,42	2,28	2.599,73
TOTALOWNERSHIP	3,09	5,13	2,06	3,04	4,5
TOTALACTIVITY	3,06	4,87	0,85	2,09	4,425
HavaleTurleriAdetOrtalamaL6	0,08	5,36	0	0	0,12
EFTTurleriAdetOrtalamaL6	0,23	7,98	0	0	0,33
LikitTurleriAdetOrtalamaL6	0,38	6,99	0	0	3,75
ToplamIslemlerHacimOrtalamaL6	2.986,88	80.289,50	6,88	15,45	7.363,95
CreditCardLimit	4.788,37	5.079,04	48,16	2.181,13	816,34
CreditCardExpense_L6	36.029,82	20.797,53	7,29	403,15	4.629,59

Küme prototiplerine ait değişkenlerin değerleri bankacılık konusunda uzmanlarla beraber değerlendirilmiştir. Bu değerlendirme sonucunda her kümeye bir profil ismi verilmiştir. Müşteri profilleri;

- Ürün Kullanım,
- Kanal Kullanımı,
- İşlemler

başlıkları altında gruplanmış ve her grubun özellikleri ortaya çıkartılmıştır. Oluşan müşteri profilleri Tablo 4. Müşteri Profilleri Tablosunda gösterilmiştir.

Tablo 4. Müşteri Profilleri Tablosu.

Küme ve Profil İsmi	**Müşteri Profili**
Küme 1 (Harcama Severler)	**Ürün Kullanımı** Kredi Kartı olan ve aktif harcama yapan Kredi Kartı Harcama Hacmi En Yüksek müşteri kümesi Cari Mevduatında varlığı olan Bireysel Kredi Riski Bulunan Mevduat Bakiyesi Azalan Trend te olan **Kanal Kullanımı** İşlemlerini Şube, Mobil, İnternet ve ATM kanallarından yapmayı tercih eden **İşlemler** Düşük Ölçekte EFT ve Likit (Fatura, SSK, Vergi vb.) işlemleri yapan
Küme 2 (Varlıklılar)	**Ürün Kullanımı** Kredi Kartı Olan ve aktif olarak harcama yapan Cari, Vadeli Mevduat Hesaplarında varlığı bulunan Bireysel Kredisi riski olan Gayri Nakdi (Çek) Riski olan Mevduat Bakiyesi Artan Trend te olan **Kanal Kullanımı** İşlemlerini Şube, Mobil, İnternet ve ATM kanallarından yapmayı tercih eder. **İşlemler** En Yüksek ölçekte EFT, Havale ve Likit (Fatura, SSK, Vergi vb.) işlemleri yapan müşteri kümesi

Küme ve Profil İsmi	Müşteri Profili
Küme 3 (Kaybedilecekler)	**Ürün Kullanımı** Cari, Vadeli Mevduat Sahibi fakat vadeli hesaplarını kapatmış Mevduatı azalan eğilimde **Kanal Kullanımı** İşlemlerini Şube Kanalı kullanarak yapan **İşlemler** EFT, Havale para transferleri ve Likit İşlemleri yapmayan
Küme 4 (Krediciler)	**Ürün Kullanımı** Bireysel Kredi Riski bulunan Kredi Kartı Sahip fakat kullanmayan Düşük Mevduat sahip olan ve mevduatı azalma eğiliminde olan **Kanal Kullanımı** Şube ve ATM Kanallarını kullanan **İşlemler** EFT, Havale para transferleri ve Likit İşlemleri yapmayan yalnızca kredi ödemesi yapan
Küme 5 (Teknolojikler)	**Ürün Kullanımı** Cari ve Vadeli Mevduat varlığı olan Mevduatı artan eğilimde olan Bireysel Kredi Riski bulunan Maaş Hesabı bulunan Kredi Kartı sahibi olan ve aktif harcama yapan **Kanal Kullanımı** İnternet, ATM ve Mobil Kanallarda işlem yapmayı tercih eden

Küme ve Profil İsmi	Müşteri Profili
	İşlemler EFT, Havale para transferleri işlemlerini yapmayan Likit (Fatura, SSK, Vergi vb.) işlemleri yapan

Profilleme sonucunda banka veri tabanında yer alan 2.261.173 müşteri Tablo 5'te belirtilen şekilde kümelere dağılmıştır. Tablo, oluşan müşteri profillerini ve bu profillerdeki müşteri sayılarını göstermektedir.

Tablo 5. Küme Büyüklükleri.

Küme	Müşteri Sayısı
Küme 0 (Pasifler)	1.826.501
Küme 1 (Harcama Severler)	106.641
Küme 2 (Varlıklılar)	26.555
Küme 3 (Kaybedilecekler)	66.477
Küme 4 (Krediciler)	118.346
Küme 5 (Teknolojikler)	116.653
Toplam Müşteri Sayısı	2.261.173

6. SONUÇLAR VE ÖNERİLER

Bu çalışmada, kümeleme analizi sürecinin tüm aşamaları bir banka müşteri veri tabanında yer alan tüm müşteri kayıtlarına uygulanmış ve karışık veri türleri içeren bir veri kümesine kümeleme analizi gerçekleştirilmiştir.

Başlangıçta 528 adet olan değişken, veri ön işleme aşamasında 35 adet değişkene indirgenmiştir. Bu işlem sırasında toplama, ortalama alınarak değişken sayısı indirgenmiş ve yeni değişkenler türetilmiştir. Yürütülen çalışmada amaç, bir Türk bankasının veri tabanında yer alan tüm müşterilerin ürün kullanım, kanal kullanım ve yaptığı işlemlerine dayalı olarak müşteri davranışlarını tespit etmektir. Tüm müşterilerin en az ve en çok bir kümeye ait olması amaçlandığından veri kümesindeki 2.261.173 adet nesneden (müşteri) hiçbiri kümeleme işlemi dışında tutulmamıştır.

Denemeler sırasında kullanılan bazı kümeleme algoritmaları esnek kümeleme yaptığından müşterileri birden fazla kümeye atamıştır. Bazı kümeleme algoritmaları sadece sayısal ya da sadece kategorik verileri işleyebilmiştir. Bazı kümeleme algoritmaları ise uzaklık matrisi kullanmaları nedeniyle çalışma süresi ya uzamış ya da herhangi bir sonuç üretememiştir.

Çok sayıda değişken ve nesne sayısına sahip olan veri kümesi kullanılması, nesnelerin sadece bir kümeye ait olması zorunluluğu ve veri kümesini teşkil eden değişkenlerin karışık veri türlerini ihtiva etmesi sebebiyle bu teze ait problemi çözmeye en uygun kümeleme algoritmasının k-prototip olduğu sonucuna varılmıştır.

Uygun küme sayısı seçiminde küme merkezlerine uzaklıkların kareleri toplamı hesaplanarak en uygun kareler toplamını veren küme sayısı 5 olarak tespit edilmiştir. Kümeleme analizi sonrasında ortaya çıkan sonuçların geçerliliği görselleştirme ve karar ağacı sınıflandırılması ile yapılmıştır.

Kümeleme analizi sonrasında ortaya çıkan kümelerin, özel Türk bankasında kullanılabilmesi adına konu uzmanlarının görüşü ile kümelerin profillemesi yapılarak 5 kümenin kendilerine özgü davranışsal özellikleri tespit edilmiştir.

Çalışmada ortaya çıkan veri madenciliği uygulaması özel Türk bankasının belirleyeceği sıklıkla çalıştırılabilir halde tasarlanmıştır. Banka istediği takvimde uygulamayı çalıştırarak tüm müşterilerinin davranışsal segmentasyonunu tespit edebilecektir.

Gerçekleştirilen davranışsal segmentasyon çalışmasında en zorlu kısım verinin hazırlaması, boyut indirgenmesi ve karışık veri türlerine ait etkin bir kümeleme algoritmasının seçilmesi olmuştur. Uygulamanın çıktı olarak sunduğu, müşterilerin davranış segment değerleri özel Türk bankasının veri tabanına yazılarak, bankanın başka uygulamaları tarafından kullanılabilecek bir halde sunulmuştur. 5 adet davranış segmentine göre, banka farklı pazarlama stratejileri geliştirme imkânına kavuşmuştur. Banka, her davranış segmentine özel olacak şekilde kampanya yönetimi, fiyat yönetimi ve müşteri ilişkileri yönetimi uygulayabilecektir.

Bu çalışma kapsamında incelenen kümeleme analizi yöntemlerinin sonucunda:

1. Büyük nesne sayılı ve değişkenli veri kümeleri için k-means ve türevi olan k-modes, k-prototip, PAM, CLARA ve alt uzay kümeleme yöntemlerinin kullanılması önerilmektedir.
2. Nesne ve değişken sayısının az olduğu veri kümeleri için küme sayısı gibi başlangıç parametreleri belirlemek zorunda kalınmadığından hiyerarşik kümeleme yöntemleri olan AGNES, BIRCH yöntemleri tavsiye edilir.
3. Veri kümesi değişkenlerinin hepsi sayısal veri türünde ve küme ayrışması doğrusal değil ise yoğunluk tabanlı kümeleme analizlerinin kullanması tavsiye edilir.
4. Esnek kümeleme türünde bir kümeleme analizi yapılacak ise SOM, Karışım Modelleri ve Fuzzy Kümeleme analizlerinin üzerinde çalışılmalıdır.

Birçok kümeleme analizi metodu bulunmasına karşın her problemi çözebilen tek bir kümeleme analizi metodu yoktur. Bu alanda çalışacak araştırmacılara problemlerine, verilerinin türüne ve veri büyüklüğüne en uygun kümeleme analizleri seçmeleri tavsiye edilmektedir.

KAYNAKLAR

[1] Tsiptsis, K., & Chorianopoulos, A. (2011). Data Mining Techniques in CRM : Inside Customer Segmentation. John Wiley & Sons Ltd.

[2] Sheth, J. N., Parvatiyar, A., & Shainesh, G. (2001). Conceptual Framework of Customer Relationship Management. Customer Relationship Management-Emeging Concepts, Tools and Applications (s. 3-25). içinde Tata McGraw-Hill Publishing Company Ltd.

[3] Ryals, L., & Payne, A. (2001). Customer Relationship Management in Financial Services : towards information-enabled relationship marketing. Journal of Strategic Marketing, 3-27.

[4] Payne, A., & Frow, P. (2005). A Strategic Framework for Customer Relationship Management. Journal of Marketing, 167-176.

[5] Wang, J. (2009). Encyclopedia of Data Warehousing and Mining. Information Science Reference.

[6] Chotianopoulos, A. (2016). Effective CRM Using Predictive Analytics. John Wiley & Sons, Ltd.

[7] Shili, S. (2009). An Analysis on the Conditions and Methods of Market Segmentation. International Journal of Business and Management, 63-69.

[8] Aras, Ü. (2008). Finansal Veri Madenciliği. İstanbul: Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü.

[9] Bounsaythip, C., & Runsala, E. R. (2001). Overview of Data Mining for Customer. Vtt Information technology Research Report.

[10] Sexton, D. E., & Kathryn, B. (1980). A Behavioral Segmentation of The Arts Market in NA - Advances in Consumer Research Volume 07.

[11] Vardar, T. (2010). Bankaların Tüzel Müşteri Segmentasyonunun, Niteliksel ve Niceliksel Kümeleme Analizi. İstanbul.

[12] Akarsu, E. (2010). Customer Retention via Hybrid Modeling for Banking Industry. İstanbul.

[13] Charles W. Lamb, J. F. (2003). Marketing. Beijing: Peking University Press.

[14] Mehdi B., M. J. (2010). Behavioral rules of bank's point-of-sale for segments description and scoring prediction. International Journal of Industrial Engineering Computations.

[15] Zhou, Z., Miao, X., & Guangcan, L. (2009). Customer Segmentation Algorithm of Wireless Content Service Based on Ant K-Means. International Forum on Computer Science-Technology and Applications.

[16] Garima, G., & Pooja, C. (2014). Eco-tourists and Environment Protection: A Pro-Environment Behavioral Segmentation Approach. Amity Global Business Review, 90-95.

[17] Birkhead, B. (2000). Behavioural Segmentation Systems : A perspective. Journal of Database Marketing, 105-112.

[18] Kasturi, R., Moriarty, R. T., & Swartz, G. S. (1992). Segmenting Customers in Mature Industrial Markets. Journal of Marketing, Vol. 56, No. 4, 72-82.

[19] Van Raaij, W. F., & Verhallen, T. M. (1991). Domain-Specific Market Segmentation. IAREP/SASE. Stockholm: Stockholm School of Economics.

[20] Valters, K., & Roberts, Š. (2011). Paradigm Shift in Consumer Segmentation to Gain Competitive Advantages in Post-Crisis FMCG Markets : Life Style or Social Values. EKONOMIKA IR VADYBA, 1266-1277.

[21] Nakıpoğlu, B. (2007). Tüketimin Çevreci Boyutu : Çevreci Tutum ve Davranışlara göre Pazar Bölünlenmesi. Ç.Ü. Sosyal Bilimler Enstitüsü Dergisi, Cilt 16, Sayı 2, 423-438.

[22] Kotler, P. (1994). Marketing Management Analysis, Planning, Implementation and Control. NJ: Prentice-Hall Inc.

[23] Hosseini, Z. Z., & Mohammadzadeh, M. (2016). Knowledge Discovery From Patients' behavior via Clustering-Classification Algorithms based on Weighted eRFM and CLV Model : An empirical study in public health care services. Iranian Journal of Pharmaceutical Research, 355-367.

[24] Braun, E., Geurten, B., & Egelhaaf, M. (2010). Identifying Prototypical Components in Behaviour Using Clustering Algorithms. PLoS ONE.

[25] Juković, S., Pejić Bach, M., Dumičić, K., & Šarlija, N. (2012). Segmentation in Banking using Self-Organizing Maps: A Case Study of Business Customer. International Conference of the Faculty of Economics, (s. 767-777). Sarajevo.

[26] Watson, H. W. (2007). The Current State of Business Intelligence. Computer Volume: 40 Issue: 9, 767-777.

[27] Sinha, K., & Uniyal Prasad, D. (2005). Using Observational Research for Behavioural Segmentation of Shoppers. Journal of Retailing and Consumer Services, 35-48.

[28] Thomas, A., & Pickering, G. (2003). Behavioural Segmentation : A New Zealand Wine Market Application. Journal of Wine Research, 127-138.

[29] Akpınar, H. (2000). Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği. İstanbul Üniversitesi İşletme Fakültesi Dergisi, 1-22.

[30] Kaufman, L., & Rousseeuw, P. J. (1990). Finding Groups in Data An Introduction to Cluster Analysis. Wiley Interscience.

[31] Özdamar, K. (2004). Paket Programlar ile İstatistiksel Veri Analizi (Çok Değişkenli Analizler). Eskişehir: Kaan Kitabevi.

[32] Aggarwal, C. C., & Chandan, R. K. (2014). Data Clustering Algorithms and Applications. Florida: CRC Press.

[33] Anderberg, M. (1973). Cluster analysis for applications. New York: NY: Academic Press.

[34] Jain, A. K., & Dubes, R. C. (1988). Algorithms for clustering data . NJ: Prentice-Hall.

[35] Han, J., Kamber, M., & Pei, J. (2012). Data Mining Concepts and Techniques. Morgan Kaufmann.

[36] Rokach, L., & Maimon, O. (2005). Clustering Methods. Data Mining and Knowledge Discovery Handbook (s. 321-352). içinde Boston: Springer.

[37] Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data Clustering: a review. ACM computing surveys Volume 31 Issue 3, 264-323.

[38] Guha, S., Rastogi, R., & Shim, K. (1998). CURE: An Efficient Clustering Algorithm for Large Databases. ACM SIGMOD. Seattle,USA.

[39] Guha, S., Rastogi, R., & Shim, K. (1999). ROCK : A robust clustering algorithm for categorical attributes. Proceedings 15th International Conference on Data Engineering (s. 512-521). IEEE.

[40] Karypis, G., Han, E. H., & Kumar, V. (1999). Chameleon: hierarchical clustering using dynamic modeling. Computer(Volume: 32, Issue: 8), 68-75.

[41] Zhang, T., Ramakrishnan, R., & Linvy, M. (1996). BIRCH: An Efficient Data Clustering Method for Very Large Databases. SIGMOD.

[42] MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, (s. 281-297).

[43] Kaufman, L., & Rousseeuw, P. J. (1987). Clustering by Means of Medoids. Computational Statistics & Data Analysis Volume 5 Issue 4, 405-416.

[44] Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. Data mining and knowledge discovery, 283-304.

[45] Huang, Z. (1997). Clustering large data sets with mixed numeric and categorical values. Proceedings of the 1st pacific-asia conference on knowledge discovery and data mining (s. 21-34). World Scientific.

[46] Ng, R. T., & Han, J. (1994). Efficient and Effective clustering methods for spatial data mining. In Proceeding of VLDB, 144-155.

[47] Wang, W., Yang, J., & Muntz, R. (1997). STING: A Statistical Information Grid Approach to Spatial Data Mining. International Conference of Very Large Databases, (s. 186-195). Athens.

[48] Gholamhosein, S., Surojit, C., & Aidong, Z. (1998). Wavecluster: A Multi-Resolution Clustering Approach for Very Large Spatial Databases. Proc. 24th International Conference on Very Large Databases, (s. 428-439). New York.

[49] Aggrawal, R., Gehke, J., Gunopulos, D., & Raghavan, P. (1998). Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications. Proc. ACM SIGMOD International Conference on Management of Data, (s. 94-105). Seattle.

[50] Wolfe, J. H. (1970). Pattern clustering by multivariate mixture analysis. Multivariate Behavioral Research, 329-350.

[51] Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. Biological Cybernetics, 59-69.

[52] Fisher, D. H. (1987, September). Knowledge acquisition via incremental conceptual clustering. Machine Learning, 139-172.

[53] Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. 2nd International Conference on Knowledge Discovery and Data Mining, 226-231.

[54] Hinneburg, A., & Keim, D. A. (1998). An efficient approach to clustering in large multimedia databases with noise. Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining . New York: AAAI Press.

[55] Ankerst, M., Breunig, M. M., Kriegel, H. P., & Sander, J. (1999). OPTICS: Ordering Points To Identify the Clustering Structure. ACM SIGMOD international conference on Management of data (s. 49-60). Philadelphia: ACM.

[56] Roy, D. K., & Sharma, L. K. (2010). Genetic k-means Clustering Algorithm For Mixed Numeric and Categorical Data sets. International Journal of Artificial Intelligence & Apllications 1(2), 23-28.

[57] Gower, J. C. (1971). A general coefficient of similarity and some of its properties. Biometrics Vol. 27 No. 4, 857-871.

[58] Sautot, L., Faivre, B., Journaux, L., & Molin, P. (2015). The hierarchical agglomerative clustering with Gower index: A methodology for automatic design of OLAP cube in ecological data processing context. Ecological Informatics, 217-230.

[59] Zhong, X., Yu, T., & Xia, H. (2017). A New Partition-based Clustering Algorithm For Mixed Data. Proceedings of the International MultiConference of Engineers and Computer Scientists. Hong Kong: IMECS.

[60] Arı, A. (2013). Gauss Karışım Modeli Kullanılarak Hareketli Nesnelerin Tespit. T.C. Fırat Üniversitesi Fen Bilimleri Enstitüsü Elektronik-Bilgisayar Eğitimi Anabilim Dalı Yüksek Lisans Tezi, (s. 28). Elazığ.

[61] Shih, M. Y., Jheng, J. W., & Lai, L. F. (2010). A Two-Step Method for Clustering Mixed Categroical and Numeric Data. Tamkang Journal of Science and Engineering 13(1), 11-19.

[62] Mirkin, B. (2005). Clustering for Data Mining A Data Recovery Approach. Boca Raton: Chapman & Hall/CRC.

[63] Xu, R., & Wunsch, D. C. (2009). Clustering. New Jersey: John Wiley & Sons, Inc.

[64] Figueiredo, M. A., & Anil, J. K. (2002, March). Unsupervised Learning of Finite Mixture Models. IEEE Transactions on Pattern Analysis and Machine Intelligence Vol.24,No.3, s. 381-396.

[65] Tunalı, V., Bilgin, T. T., & Çamurcu, Y. A. (2016). An Improved Clustering Algorithm for Text Mining: Multi-cluster Spherical K-means. The International Arab Journal of Information Technology (IAJIT) 13(1), 12-19.

[66] Agis, G. (2017, 11 27). Guillaume Agis's blog: http://blog.guillaumeagis.eu/k-means-clustering-apache-mahout/ adresinden alındı.

ÖZGEÇMİŞ

Kamil Balıkçı, 1975 yılı Ankara doğumludur. İlköğrenimini, ortaokulu ve liseyi Ankara' da tamamladı. 1999 yılında Hacettepe Üniversitesi İstatistik Bölümünde mezun oldu. Kısa bir dönem ticaretle uğraştıktan sonra 2003 yılında özel bir banka da Sistem Analisti olarak görev aldı. Yine aynı bankada 4 yıl Proje Yöneticiliği görevini üstlendi. 3 yıldır Uygulama Geliştirme Müdürlüğünde yöneticilik görevini sürdürmektedir. 2015 yılında başladığı Maltepe Üniversitesi Bilgisayar Mühendisliği Yüksek Lisans programına tez aşamasında devam etmektedir.

EKLER

EK-A Kümeleme Analizinde Kullanılan Değişkenler

Değişken İsimleri	Değişken Türleri	Değişken Açıklamaları
CustomerIdMap	Sınıflayıcı	Müşterinin banka iç sisteminde kullanılan benzersiz numarasıdır.
CustomerTenure	Oran Ölçekli	Müşterinin bankada müşteri olma süresinin ay türünde gösterimidir.
TOT_VARLIK_L6	Oran Ölçekli	Eğer CustomerTenure (Ay türünden müşteri olma süresi) 6 aydan küçük ise - (OrtalamaToplamHesapBakiyesi_M1+ OrtalamaToplamHesapBakiyesi_M2 + OrtalamaToplamHesapBakiyesi_M3 + OrtalamaToplamHesapBakiyesi_M4 + OrtalamaToplamHesapBakiyesi_M5 + OrtalamaToplamHesapBakiyesi_M6) / CustomerTenure Büyük ise - (OrtalamaToplamHesapBakiyesi_M1+ OrtalamaToplamHesapBakiyesi_M2 + OrtalamaToplamHesapBakiyesi_M3 + OrtalamaToplamHesapBakiyesi_M4 + OrtalamaToplamHesapBakiyesi_M5 + OrtalamaToplamHesapBakiyesi_M6) / 6
TOT_KREDI_L6	Oran Ölçekli	TicarKrediOrtalamaL6 + BireyselKrediOrtalamaL6 değişkenlerinin toplamıdır.
MEVDUAT_TREND_L6	Aralık Ölçekli	-1 ve 1 aralığında değer alır. Denklem 8.Tek Değişkenli Doğrusal Regresyon Modeli. Den elde edilir. Mevduat Eğilimini veren değişkendir.

Değişken İsimleri	Değişken Türleri	Değişken Açıklamaları
TOTALOWNERSHIP	Oran Ölçekli	Müşterinin sahip olduğu farklı ürün adedini gösterir. GNKredi_OWN + CekKarne_OWN>0 + HappyKart_OWN + TFKart_OWN + BusinessKart_OWN + TicariKart_OWN + EkKrediKart_OWN + EkTicariKart_OWN + KurumsalFinansmanDestegi_OWN + FinansalKiralama_OWN + KarZararOrtakligi_OWN + 0_faizli kredi_OWN + VergiTahsilati_OWN + SSK_OWN + OtomatikOdemeTalimat_OWN + TüketiciTasitKredisi_OWN + TicariTasitKredisi_OWN + TüketiciKonutKredisi_OWN + TaksitliTicariIsyeriKredisi_OWN + TuketiciIhtiyacKredisi_OWN + TaksitliTicari IhtiyacKredisi_OWN + VadeliT_OWN + VadeliY_OWN + VadeliZ_OWN + VadeliI_OWN + VadeliA_OWN + CariHesapTP_OWN + CariHesapYP_OWN + Emeklilik_OWN + DebitKart_OWN + Uyeisyeri_OWN + KiralikKasa_OWN+VadeliDövizi_OWN

Değişken İsimleri	Değişken Türleri	Değişken Açıklamaları
TOTALACTIVITY	Oran Ölçekli	Müşterinin aktif kullandığı farklı tür ürün sayısını gösterir. Müşterinin sahip olduğu farklı ürün adedini gösterir. GNKredi_ACT + CekKarne_ACT>0 + HappyKart_ACT + TFKart_ACT + BusinessKart_ACT + TicariKart_ACT + EkKrediKart_ACT + EkTicariKart_ACT + KurumsalFinansmanDestegi_ACT + FinansalKiralama_ACT + KarZararOrtakligi_ACT + 0_faizli kredi_ACT + VergiTahsilati_ACT + SSK_ACT + OtomatikOdemeTalimat_ACT + TüketiciTasitKredisi_ACT + TicariTasitKredisi_ACT + TüketiciKonutKredisi_ACT + TaksitliTicariIsyeriKredisi_ACT + TuketiciIhtiyacKredisi_ACT + TaksitliTicari IhtiyacKredisi_ACT + VadeliT_ACT + VadeliY_ACT + VadeliZ_ACT + VadeliI_ACT + VadeliA_ACT + CariHesapTP_ACT + CariHesapYP_ACT + Emeklilik_ACT + DebitKart_ACT + Uyeisyeri_ACT + KiralikKasa_ACT + VadeliDövizi_ACT
SubeAdetOrtalamaL6	Oran Ölçekli	Müşterinin son 6 ay içinde Şube ye gelerek yaptığı işlem sayısının ortalamasıdır. Eğer CustomerTenure (Ay türünden müşteri olma süresi) 6 aydan küçük ise - (SubeKullanimAdetAy1 + SubeKullanimAdetAy2 + SubeKullanimAdetAy3 + SubeKullanimAdetAy4 +

Değişken İsimleri	Değişken Türleri	Değişken Açıklamaları
		SubeKullanimAdetAy5 + SubeKullanimAdetAy6) / CustomerTenure Büyük ise - (SubeKullanimAdetAy1 + SubeKullanimAdetAy2 + SubeKullanimAdetAy3 +SubeKullanimAdetAy4+ SubeKullanimAdetAy5 + SubeKullanimAdetAy6) / 6
InternetAdetOrtalamaL6	Oran Ölçekli	Müşterinin son 6 ay içinde İnternet Bankacılığını kullanarak yaptığı işlem sayısının ortalamasıdır. Eğer CustomerTenure (Ay türünden müşteri olma süresi) 6 aydan küçük ise - (InternetToplamAdetL6) / CustomerTenure Büyük ise - (InternetToplamAdetL6) / 6
ATMAdetOrtalamaL6	Oran Ölçekli	Müşterinin son 6 ay içinde ATM cihazlarını kullanarak yaptığı işlem sayısının ortalamasıdır. Eğer CustomerTenure (Ay türünden müşteri olma süresi) 6 aydan küçük ise -ATMKullanimAdetAy1+ATMKullanimAdetAy2 + ATMKullanimAdetAy3+TMKullanimAdetAy4 + ATMKullanimAdetAy5 + ATMKullanimAdetAy6) / CustomerTenure Büyük ise - (ATMKullanimAdetAy1 + ATMKullanimAdetAy2 + ATMKullanimAdetAy3 + ATMKullanimAdetAy4 + ATMKullanimAdetAy5 + ATMKullanimAdetAy6) / 6

Değişken İsimleri	Değişken Türleri	Değişken Açıklamaları
CallCenterAdetOrtalamaL6	Oran Ölçekli	Müşterinin son 6 ay içinde Çağrı Merkezini arayarak yaptığı işlem sayısının ortalamasıdır. Eğer CustomerTenure (Ay türünden müşteri olma süresi) 6 aydan küçük ise - (CallCenterKullanimAdetAy1 + CallCenterKullanimAdetAy2 + CallCenterKullanimAdetAy3 + CallCenterKullanimAdetAy4 + CallCenterKullanimAdetAy5 + CallCenterKullanimAdetAy6) / CustomerTenure Büyük ise - (CallCenterKullanimAdetAy1 + CallCenterKullanimAdetAy2 + CallCenterKullanimAdetAy3 + CallCenterKullanimAdetAy4 + CallCenterKullanimAdetAy5 + CallCenterKullanimAdetAy6) / 6
MobilSubeAdetOrtalamaL6	Oran Ölçekli	Müşterinin son 6 ay içinde Mobil Telefon üzerinde Mobil Şube üzerinden yaptığı işlem sayısının ortalamasıdır. Eğer CustomerTenure (Ay türünden müşteri olma süresi) 6 aydan küçük ise - (MobilSubeKullanimAdetAy1 + MobilSubeKullanimAdetAy2+MobilSubeKullanimAdetAy3+ MobilSubeKullanimAdetAy4 + MobilSubeKullanimAdetAy5 + MobilSubeKullanimAdetAy6) / CustomerTenure Büyük ise - (MobilSubeKullanimAdetAy1+ MobilSubeKullanimAdetAy2 +

Değişken İsimleri	Değişken Türleri	Değişken Açıklamaları
		MobilSubeKullanimAdetAy3 + MobilSubeKullanimAdetAy4 + MobilSubeKullanimAdetAy5 + MobilSubeKullanimAdetAy6) / 6
DigerKanalAdetOrtalamaL6	Oran Ölçekli	Müşterinin son 6 ay içinde PTT, Ödeme Merkezleri vb. kanallardan yaptığı işlem sayısının ortalamasıdır. Eğer CustomerTenure (Ay türünden müşteri olma süresi) 6 aydan küçük ise - (DigerKanalKullanimAdetAy1+DigerKanalKullanimAdetAy2+DigerKanalKullanimAdetAy3+DigerKanalKullanimAdetAy4+DigerKanalKullanimAdetAy5 + DigerKanalKullanimAdetAy6) / CustomerTenure Büyük ise - (DigerKanalKullanimAdetAy1+DigerKanalKullanimAdetAy2 + DigerKanalKullanimAdetAy3+ DigerKanalKullanimAdetAy4+DigerKanalKullanimAdetAy5 + DigerKanalKullanimAdetAy6) / 6
HavaleTurleriAdetOrtalamaL6	Oran Ölçekli	GelenHavaleAdetOrtalamaL6 + GidenHavaleAdetOrtalamaL6 + GelenYDHAdetOrtalamaL6 + GidenYDHAdetOrtalamaL6

Değişken İsimleri	Değişken Türleri	Değişken Açıklamaları
EFTTurleriAdetOrtalamaL6	Oran Ölçekli	GelenEFTAdetOrtalamaL6+GidenEFTAdetOrtalamaL6 + KKEFTAdetOrtalamaL6
LikitTurleriAdetOrtalamaL6	Oran Ölçekli	ParaYatırmaAdetOrtalamaL6 + ParaÇekmeAdetOrtalamaL6
ToplamIslemlerHacimOrtalamaL6	Oran Ölçekli	HavaleTurleriHacimOrtalamaL6 + EFTTurleriHacimOrtalamaL6 + LikitTurleriHacimOrtalamaL6
SalaryCustomerHacimL6	Oran Ölçekli	(EmployeeSalaryAmountAy1 + EmployeeSalaryAmountAy2 + EmployeeSalaryAmountAy3 + EmployeeSalaryAmountAy4 + EmployeeSalaryAmountAy5 + EmployeeSalaryAmountAy6)
CreditCardLimit	Oran Ölçekli	Banka tarafından müşteriye tahsis edilen geçerli müşteri kart limitidir.
CreditCardExpense_L6	Oran Ölçekli	(CreditCardExpenseM1+CreditCardExpenseM2+ CreditCardExpenseM3+CreditCardExpenseM4+CreditCardExpenseM5+CreditCardExpenseM6)
GNAKDI_KREDI_OWN	İkili	Müşterinin Son 12 ay içinde kullandığı Gayri Nakdi Kredisi var ise 1 aksi halde 0 değerini alır.
NAKDI_KREDI_OWN	İkili	Müşterinin Son 12 ay içinde kullandığı Nakdi Kredisi var ise 1 aksi halde 0 değerini alır.
BIREYSEL_KREDI_OWN	İkili	Müşterinin Son 12 ay içinde kullandığı Bireysel Kredisi var ise 1 aksi halde 0 değerini alır.
KK_OWN_by_HARCAMA	İkili	Müşterinin Son 12 ay içinde Kredi Kartı harcaması sıfırdan büyükse 1 aksi halde 0 değerini alır. 12 aylık kredi harcaması (CreditCardExpenseM1+CreditCardExpenseM2+ CreditCardExpenseM3+CreditCardExpenseM4+CreditCardExpenseM5+CreditCardExpenseM6+Cr

Değişken İsimleri	Değişken Türleri	Değişken Açıklamaları
		editCardExpenseM7+CreditCardExpenseM8+CreditCardExpenseM9+CreditCardExpenseM10+CreditCardExpenseM11+CreditCardExpenseM12) değişkenlerinin toplanması ile elde edilir.
KATILIM_OWN	İkili	Müşterinin Son 12 ay içinde açık olan ve bakiyesi bulunan Vadeli Hesabı var ise 1 aksi halde 0 değerini alır.
CARI_OWN	İkili	Müşterinin Son 12 ay içinde açık olan ve bakiyesi bulunan Vadesiz Hesabı var ise 1 aksi halde 0 değerini alır.
GNAKDI_KREDI_ACT	İkili	Müşterinin Son 6 ay içinde kullandığı Gayri Nakdi Kredisi var ise 1 aksi halde 0 değerini alır.
NAKDI_KREDI_ACT	İkili	Müşterinin Son 6 ay içinde kullandığı Nakdi Kredisi var ise 1 aksi halde 0 değerini alır.
BIREYSEL_KREDI_ACT	İkili	Müşterinin Son 6 ay içinde kullandığı Bireysel Kredisi var ise 1 aksi halde 0 değerini alır.
KK_ACT_by_HARCAMA	İkili	Müşterinin Son 6 ay içinde Kredi Kartı harcaması sıfırdan büyükse 1 aksi halde 0 değerini alır. 6 aylık kredi harcaması (CreditCardExpenseM1+CreditCardExpenseM2+CreditCardExpenseM3+CreditCardExpenseM4+CreditCardExpenseM5+CreditCardExpenseM6) değişkenlerinin toplanması ile elde edilir.
KATILIM_ACT	İkili	Müşterinin Son 6 ay içinde açık olan ve bakiyesi bulunan Vadeli Hesabı var ise 1 aksi halde 0 değerini alır.

Değişken İsimleri	Değişken Türleri	Değişken Açıklamaları
CARI_ACT	İkili	Müşterinin Son 6 ay içinde açık olan ve bakiyesi bulunan Vadesiz Hesabı var ise 1 aksi halde 0 değerini alır.
IsSalaryCustomer	İkili	Müşterinin bir maaş müşterisi olup, olmadığını gösteren değişkendir. (EmployeeSalaryAmountAy1+EmployeeSalaryAmountAy2+EmployeeSalaryAmountAy3+EmployeeSalaryAmountAy4+EmployeeSalaryAmountAy5+EmployeeSalaryAmountAy6) değişkenlerinin toplamı 0'dan büyükse 1 aksi halde 0 değerini alır.
KanalAktifL6Flag	İkili	Son 6 ayda herhangi bir kanaldan işlem yapılıp, yapılmadığını gösteren değişkendir. (SubeToplamAdetL6 + InternetToplamAdetL6 + ATMToplamAdetL6 + CallCenterToplamAdetL6 + MobilToplamAdetL6 + DigerKanalToplamAdetL6)> 0 ise 1 diğerini aksi halde 0 değerini alır.
IslemlerAktifFlag	İkili	Son 6 ayda Havale, EFT, Fatura Ödemesi, SSK Ödemesi, Döviz Alım Satım işlemleri vb. bankacılık işlemi yapılıp, yapılmadığını gösteren değişkendir. ToplamIslemlerAdetL6 değişkeni 0 ise işlem yapılmamış 0 eğer 0'dan büyükse 1 değerini alır.

EK-B K-Prototip Algoritması R Kodu

```
#küme sayısını al.
k<-knime.flow.in[["k"]]

#CustomerIdMap kolonunu çıkar.
mydataset<-mydataset[,-match(c("CustomerIdMap"),names(mydataset))]

#k-prototip algoritması
kprototype<-function(mydataset,k,lambda=NULL,iter.max=100,nstart=1,keep.data=TRUE)
{
  if(!is.data.frame(mydataset)) stop("Veri setim bir data frame olmalıdır.")
  if(ncol(mydataset) < 2) stop("Kümeleme Analizi için en az 2 değişken olması
gerekir.")
  if(iter.max < 1 | nstart < 1) stop("İterasyon sayısı ve başlangıç noktası 1 den
büyük olmalıdır.")
  if(!is.null(lambda))
  {
    if(any(lambda < 0)) stop("Lambda 0 veya daha büyük bir değer olmalıdır.")
    if(!any(lambda > 0)) stop("En az 1 değişken 0 dan büyük olmalıdır.")
  }
  numvars <- sapply(mydataset, is.numeric)  #numeric değişkenler
  anynum <- any(numvars)            #Hiç numeric değişken var mı ?
  catvars <- sapply(mydataset, is.factor)   #kategorik değişkenler
  anyfact <- any(catvars)           #Hiç kategorik değişken var mı ?
  if(!anynum) cat("\n Veri setimde hiç numeric değişken bulunmuyor.\n\n")
  if(!anyfact) cat("\n Veri setimde hiç kategorik değişken bulunmuyor.\n\n")

  #Lambda değerinin otomatik hesaplanması
  if(length(lambda) > 1)
  {
    if(length(lambda) != sum(c(numvars,catvars))) stop("Lambda bir vektör ise,
Vektörün uzunluğu veri setindeki sayısal ve kategorik değişkenlerin toplam sayısına
eşit olmalıdır.")
  }
  if(is.null(lambda))
  {
    if(anynum & anyfact) #hem kategorik hem numeric değişken varsa
    {
```

```
      #her numeric değişkenin varyansı bulunur. Bu varyansların ortalaması alınır.
      vnum <- mean(sapply(mydataset[,numvars, drop = FALSE], var))
      #her kategorik değişkenin sıklığı/kayıt sayısının kareleri toplamı alınarak
      varyansı bulunur.
      vcat <- mean(sapply(mydataset[,catvars, drop = FALSE], function(z) return(1-
      sum((table(z)/length(z))^2))))
      if (vnum == 0)
      {
        warning("Tüm numeric değişkenler 0 varyansa sahip")
        anynum <- FALSE
      }
      if (vcat == 0)
      {
        warning("Tüm kategorik değişkenler 0 varyansa sahip.")
        anycat <- FALSE
      }
      if(anynum & anyfact)
      {
        #lambda değeri numeric değişkenlerin toplam varyansı bölü kategorik
       değişkenlerin toplam varyansı şeklinde bulunur.
        lambda <- vnum/vcat; cat("Tahmini Lambda:", lambda, "\n\n")
      }
      else lambda <- 1

    }
  }
  #Başlangıç Prototipleri
  if (length(k) == 1)
  {
    selectedids <- sample(nrow(mydataset), k) #veri setimden rasgele 5 satır
seçmeyi sağlar. nrow fonksiyonu kullanılmaz ise 5 sütün seçer.
    protos <- mydataset[selectedids,]
  }
  #Kümelemeye Başla
  kumeler <- numeric(nrow(mydataset))
  tot.dists <- NULL
  moved <- NULL
```

```
keep.protos <- rep(TRUE,k)
for(l in 1:(k-1))
{
  for(m in (l+1):k)
  {
     # Öklid Uzaklığı
     d1 <- sum((protos[l,numvars, drop = FALSE]-protos[m,numvars, drop =
     FALSE])^2)
     # kategorik değişkenlerde eşleşmeyenlerin sayısı simple matching
     d2 <- sum(protos[l,catvars, drop = FALSE] != protos[m,catvars, drop =
     FALSE])
     if((d1+d2) == 0) keep.protos[m] <- FALSE
  }
}
if(!all(keep.protos))
{
  protos <- protos[keep.protos,]
  k <- sum(keep.protos)
  cat("Eşit olan prototipler birleştirildi.")
}
#Kümeleme iterasyonuna başla
iterasyon<-1
while(iterasyon < iter.max)
{
  nrows <- nrow(mydataset)
  #nesneler ile prototipler arasındaki uzaklık
  dists <- matrix(NA, nrow=nrows, ncol = k)
  for(i in 1:k)
  {
    #numeric değişken değerlerinin prototipteki numeric değişkenlerle arasındaki
     farkın karesini al. Çıktı n*p matris

     #numeric değişkenin prototipteki numeric değişkenlere uzaklığın karesi
     d1 <- (mydataset[,numvars, drop = FALSE] -
     matrix(rep(as.numeric(protos[i,numvars, drop = FALSE]), nrows),
     nrow=nrows, byrow=T))^2
```

```
    #Satırlar toplamı
    if(length(lambda) == 1) d1 <- rowSums(d1)
    if(length(lambda) > 1) d1 <- d1 %*% lambda[numvars]
    #lambda değeri ile distance ların matris çarpımı yapılıyor. lambda tek
    olduğu için numeric değişkenler aslında lambda değeri ile çarpılmıyor.

    d2 <- sapply(which(catvars), function(j) return(mydataset[,j] !=
    rep(protos[i,j], nrows)))
    if(length(lambda) == 1) d2 <- lambda * rowSums(d2)
    #satırlar toplamının lambda değeri ile çarpılmıştır.
    if(length(lambda) > 1) d2 <- d2 %*% lambda[catvars]
    dists[,i] <- d1 + d2
  }
  #Küme ataması yap.
  eski.kumeler <- kumeler
  kumeler <- apply(dists, 1, function(z) {a <- which.min(z); if (length(a)>1)
  a <- sample(a,1); return(a)})
  size <- table(kumeler)
  min.dists <- apply(cbind(kumeler, dists), 1, function(z) z[z[1]+1])
  within <- as.numeric(by(min.dists, kumeler, sum))
  tot.within <- sum(within)

  if (length(size) < k)
  {
    k <- length(size)
    protos <- protos[1:length(size),]
    cat("Boş kümeler oluştu. Küme sayısı azaltıldı :", k, "\n\n")
  }
  tot.dists <- c(tot.dists, sum(tot.within))
  moved <- c(moved, sum(kumeler != eski.kumeler))
  remids <- as.integer(names(size))

  for(i in remids)
  {
    protos[which(remids == i), numvars] <- sapply(mydataset[kumeler==i,
    numvars, drop = FALSE], mean)
```

```
    protos[which(remids == i), catvars] <- sapply(mydataset[kumeler==i,
    catvars, drop = FALSE], function(z) levels(z)[which.max(table(z))])
  }
  keep.protos <- rep(TRUE,k)
  for(l in 1:(k-1))
  {
    for(m in (l+1):k)
    {
     # Öklid Uzaklığı
      d1 <- sum((protos[l,numvars, drop = FALSE]-protos[m,numvars, drop =
     FALSE])^2)
     # kategorik değişkenlerde eşleşmeyenlerin sayısı simple matching
     d2 <- sum(protos[l,catvars, drop = FALSE] != protos[m,catvars, drop
     = FALSE])
     if((d1+d2) == 0) keep.protos[m] <- FALSE
    }
  }
  if(!all(keep.protos))
  {
    protos <- protos[keep.protos,]
    k <- sum(keep.protos)
    cat("Eşit olan prototipler birleştirildi.")
  }
  # Döngü durma kuralı
  if(moved[length(moved)] == 0) break
  if(k == 1) break
  iterasyon <- iterasyon + 1
}
#Sonuçları oluştur.
res <- list(cluster = kumeler,       #Küme vektörü.
            centers = protos,        #Küme Prototipleri
            lambda = lambda,         #Lambda parametresi.
            size = size,             #Küme büyüklükleri vektörü.
           #Küme bazında küme içi hata kareleri toplamı.
            withinss = within,
           #Küme içi hata kareleri toplamı.
```

```
                tot.withinss = tot.within,
                #Tüm nesneler ile küme prototipleri arasındaki uzaklık matrisi.
                dists = dists,
                #Maksimum iterasyon sayısı.
                iter = iterasyon,
                #Tüm İterasyon boyunca izleme matrisi
                trace = list(tot.dists = tot.dists, moved = moved))

  # loop: if nstart > 1:
  if(nstart > 1)
    for(j in 2:nstart)
      {
                res.new <- kprototype(mydataset=mydataset, k=k, lambda = lambda,
                iter.max = iter.max, nstart=1)
                if(res.new$tot.withinss < res$tot.withinss)
                res <- res.new
      }

  if(keep.data) res$data = mydataset
  class(res) <- "kprototype"
  return(res)
}
#k-prototip fonksiyonu çağırılıyor.
mykprotomodel<-kprototype(mydataset,k,lambda=NULL,iter.max=100,nstart=1,
                keep.data=TRUE)
```

EK-C Karışık Tip Veriler için Temel Bileşenler Analizi R Kodu

```
library(PCAmixdata)
datasetPCAmix<-knime.in
split<-splitmix(datasetPCAmix[,-match(c("Cluster"),names(datasetPCAmix))])
X1<-split$X.quanti
X2<-split$X.quali

PCAmixObj<-PCAmix(X.quanti=X1,X.quali=X2,ndim=3,rename.level=TRUE,graph = FALSE)

PC0<-data.frame(PCAmixObj$ind$coord[,1])
PC1<-data.frame(PCAmixObj$ind$coord[,2])
PC2<-data.frame(PCAmixObj$ind$coord[,3])
Clust<-data.frame(datasetPCAmix[,match(c("Cluster"),names(datasetPCAmix))])

df<-c(PC0,PC1,PC2,Clust)
names(df)[1]<-"PC0"
names(df)[2]<-"PC1"
names(df)[3]<-"PC2"
names(df)[4]<-"Cluster"
```

EK-D Kümeleme Analizi Sonucunu Görselleştiren R Kodu

```
library(plotly)
df<-as.data.frame(df)
plot_ly(df, x = ~PC0, y = ~PC1, z = ~PC2, color = ~Cluster, colors = "Set1") %>%
  add_markers() %>%
  layout(scene = list(xaxis = list(title = 'PC0'),
                      yaxis = list(title = 'PC1'),
                      zaxis = list(title = 'PC2')))
```