T.C.
BAHÇEŞEHİR ÜNİVERSİTESİ

OTEL REZERVASYON DATALARINA GÖRE MÜŞTERİ PROFİLİ BELİRLEME

Yüksek Lisans Tezi

MURAT KAYA

İSTANBUL, 2019

T.C.

BAHÇEŞEHİR ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

BİLGİ TEKNOLOJİLERİ

OTEL REZERVASYON DATALARINA GÖRE MÜŞTERİ PROFİLİ BELİRLEME

Yüksek Lisans Tezi

MURAT KAYA

Tez Danışmanı: DR. ÖĞR. ÜYESI MUSTAFA EREN YILDIRIM

İSTANBUL, 2019

T.C.
BAHÇEŞEHİR ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİ TEKNOLOJİLERİ

Tezin Adı: Otel Rezervasyon Datalarına Göre Müşteri Profili Belirleme
Öğrencinin Adı Soyadı: Murat KAYA
Tez Savunma Tarihi: 21.08.2019

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

Dr. Öğr. Üyesi Yücel Batu SALMAN
Enstitü Müdürü

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

Prof. Dr. Mehmet Alper TUNGA
Program Koordinatörü

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

Jüri Üyeleri	İmzalar
Tez Danışmanı Dr. Öğr. Üyesi M. Eren YILDIRIM	---
Üye Dr. Öğr. Üyesi Yücel Batu SALMAN	---
Üye Dr. Öğr. Üyesi Atınç YILMAZ	---

ÖZET

OTEL REZERVASYON DATALARINA GÖRE MÜŞTERİ PROFİLİ BELİRLEME

Murat KAYA

Bilgi Teknolojileri Yüksek Lisans Programı

Tez Danışmanı: Dr. Öğr. Üyesi Mustafa Eren YILDIRIM

Ağustos 2019, 34 sayfa

Veri madenciliği, büyük kapasiteli veriler arasından yararlı bilgilere ulaşma işidir. Büyük veri ambarları içerisinden gelecekle ilgili ipuçları bulunabilmemizi sağlayabilecek bağıntıların bilgisayar algoritmaları kullanarak aranması olarak da belirtilebilir. Veri madenciliği tekniği, rekabet içerisinde bulunan farklı sektörlerdeki kurumlar; "nokta atış" diye tabir edebileceğimiz saptamalar yapabilmeleri ve ürünleri veya verdikleri hizmetleri, ilgili müşterilerine ulaştırabilmeleri için çok büyük bir öneme sahiptir. Bu çalışma, elde bulunan büyük ölçekteki müşteri verileri üzerinden, veri madenciliği metotları ile müşteri profili çıkarmayı amaçlamaktadır. Bu çalışmada, turizm sektöründe halen aktif olarak faaliyet gösteren bir turizm firmasına ait, kişisel verilerin korunması kanunu gereğince isimsiz bir veri tabanı kullanılmıştır. Bu data üzerinde k-means kümeleme algoritması kullanılarak müşteriler, demografik niteliklerine göre gruplara ayrıştırılmıştır.

Anahtar Kelimeler: Müşteri Profili Belirleme, Veri Madenciliği

ABSTRACT

CUSTOMER PROFILE DETERMINATION BY HOTEL RESERVATION DATAS

Murat KAYA

Information Technology

Thesis Supervisor: Asisist. Prof. Dr. Mustafa Eren YILDIRIM

August 2019, 34 pages

Data mining is the task of accessing useful information from large capacity of data. It can also be referred to as searching for correlations that can provide clues about the future in large data warehouses by using computer algorithms. Data mining technique, institutions in different sectors competing; It is of great importance for them to make the determinations that we can call “point shot” and to deliver the products or services they provide to the relevant customers. This study aims to create a customer profile through data mining methods based on large scale customer data available. In this study, an anonymous database was used in accordance with the law on protection of personal data of a tourism company which is still active in tourism sector. By using the k-means clustering algorithm on this data, customers were divided into groups according to their demographic characteristics.

Keywords: Customer Profiling, Data Mining

İÇİNDEKİLER

TABLOLAR

ŞEKİLLER

1. GİRİŞ

Firmalar için geçmişte veriyi muhafaza etmek her ne kadar önemli olsa da firma menfaatlerini yükseltmek ve geleceğe yönelik stratejilerin maksimum verimlilik ile geliştirilebilmeleri için günümüzde, geleceğe yönelik bu ipuçlarını içeren verileri işlemek ve yönetmek daha büyük önem arz etmektedir. Gelişen bilgisayar teknolojileri, çok daha fazla veririn geçmişe oranla, çok daha hızlı bir şekilde toplanması, depolanması ve işlenmesine olanak sağlamaktadır. Depolanan verinin işlenmesi kısmında veri madenciliği bilimi devreye girmektedir. Veri madenciliği; önceden bilinmeyen, çok net olmayan bilginin, veri tabanlarından dinamik bir süreç ile elde edilmesi olarak tanımlanabilir. Veri madenciliği, istatistiksel bir yöntemler dizisidir. Ancak veri madenciliği metotları, geleneksel istatistik metotlarından farklılıklar göstermektedir. Veri madenciliğinde amaç, verinin, kolaylıkla ifade edilebilecek mantıksal kurallara ya da görsel ifadelere çevrilebilecek nitelikteki modellerinin çıkarılmasıdır.

Turizm sektörü, ülke ekonomimizin büyük pasta payına sahip lokomotiflerinden biridir. Türkiye istatistik kurumu verilerine göre, Turizm geliri 2018 yılında bir önceki yıla göre yüzde12,3 artarak 29 milyar 512 milyon 926 bin $ oldu. 2018 yılı gelirinin 22 milyar 546 milyon 616 bin $'ını kişisel harcamalar, 6 milyar 966 milyon 311 bin $'ını ise paket tur harcamaları oluşturdu. Ziyaretçi sayılarına bakacak olursak; Yabancı ziyaretçi 39.488.401, Yurtdışı ikâmetli Türkiye Cumhuriyeti Vatandaşı 6.624.191 olmak üzere toplam 46.112.592. İç turizm rakamlarına bakacak olursak; 2018 yılı toplam seyahat sayısı bir önceki yıla göre yüzde1,7 artarak 78 milyon 523 bin olarak gerçekleşti. 2018 yılı seyahate çıkanların yaptıkları toplam geceleme sayısı 633 milyon 721 bin. Ortalama geceleme sayısı 2018 yılında 8,1 gece, seyahat başına ortalama harcama ise 513 TL oldu. Bu yılda, yurt içindeki seyahatlerde yapılan toplam seyahat harcamaları geçen yıla göre yüzde14 artarak 40 milyar 266 milyon 153 bin TL oldu. Türkiye seyahat acenteleri birliği(TÜRSAB) verilerine göre; Turizm gelirleri ülke ihracatının yüzde17.5 ini oluşturmaktadır.

Tablo 1.1'de görüldüğü üzere;

Tablo 1.1: Turizmin 2018 Yılı İhracat Raporu

TURİZM GELİRLERİNİN İHRACAT GELİRLERİ VE TURİZM GİDERLERİNİN İTHALAT GİDERLERİ ORANI (2018) (1 000 000 $)						
YILLAR		TURİZM	TURİZM GELİRLERİNİN		TURİZM	TURİZM GİDERLERİNİN
	IHRACAT	GELİRLERİ	İHRACAT GELİRLERİNE ORANI (%)	ITHALAT	GİDERLERİ	İTHALAT GİDERLERİNE ORANI (%)
2018	167 967,2	29 512,9	17.5	223 046,4	4 896,3	2.2

Bütün turizm firmaları bu pazarda kendilerine düşen pasta payını büyütmek için rekabet halindedirler. Bu rekabet ortamında üst sıraları hedefleyebilmek için sahip oldukları verileri en iyi şekilde yönetmeli ve işleyebilmelidirler. Gelişen teknoloji, günümüzde veriyi çok efektif bir şekilde işlemeye olanak sağlamaktadır. Bu teknolojik gelişmeler ışığında firmalar, kendileri için en doğru müşteri profillerini belirleyebilirler ise, hem reklam için harcadıkları maliyetlerini düşürebilirler hem de kendileri için uygun kitleyle doğrudan iletişime geçerek hızlı geri dönüşler sağlayabilirler. Veri madenciliği yöntemleri ile ürününüzün çoğunlukla genç nüfus tarafından tercih edildiğini tespit ettiğinizi varsayalım. Gençlere hitap eden bu ürününüzün reklamları için neden gsm operatörlerine bütün yaş grupları için para ödeyesiniz ki? Veri madenciliği ile mevcut müşteriler daha iyi tanınarak firmaların müşteri ilişkileri departmanlarında düzenleme ve geliştirmeler yapılabilir. Müşterinin iyi tanınması, kişiselleştirilmiş ürün ve hizmetlerin tanıtılmasını mümkün kılacaktır. Bu sayede firmaların "müşteri gibi düşünme" empatisi de geliştirilebilir.

Bu çalışmadaki amacımız, sahip olduğumuz turizm sektörüne ait bu büyük ölçekli veriyi, veri madenciliği metotları ile analiz ederek müşteri profillerini belirlemek. Veri madenciliği işlemleri sonrasında oluşan modeller, bize bu kümeler içerisindeki müşterilerin ortak özellikleri hakkında bilgiler verecek ve onların gelecekte yapacakları tercihleri tahmin etme konusunda bizlere olanak sağlayacaktır. Çalışmada turizm sektöründe halen aktif rol alan bir firmanın dataları kullanılmıştır. Datalar en yaygın kümeleme tekniklerinden biri olan k-means tekniği ile farklı kümelere ayrıştırılmıştır. Bu dataların nasıl toplandığı, hangi boyutlara göre analiz edildiği ve elde edilen bulguları ilerleyen bölümlerde anlatılacaktır.

1.1 TEKNOLOJİNİN TURİZM İŞLETMELERİ ÜZERİNDEKİ ETKİSİ

Son zamanlarda teknoloji dünyasında meydana gelen gelişmeler seyahat sever insanların tercihlerinde değişimlere yol açtığı gibi, onlara bu seyahat seçeneklerini sunmak isteyen turizm firmalarının da faaliyetlerinde değişime yol açmıştır. Bir turizm firmasının sektörde tutunabilmesi, varlığını sürdürmeye devam edebilmesi, o firmanın güncel teknolojileri kullanması ile doğru orantılıdır. Turizm firmaları teknoloji sayesinde müşterilerinin medeni durumu, yaşı, gelir düzeyi, çocukları olup olmadığı, eğitim durumu gibi demografik verilerle, boş vakitlerinde ne yapmaktan hoşlandıkları, nerelere seyahat ettikleri veya etmek istedikleri, seyahat ederken genellikle yılın hangi dönemini kullandıkları vb. pek çok bilginin yer aldığı bir veri tabanı oluşturarak ve bu verileri iyi yöneterek pazarda yer alan rakiplerine karşı çok büyük avantaj elde edebilmektedirler. Turizm firmalarının, müşterilerini sunduğu hizmetler, satış anında kontrol edilebilir veya denenebilir ürünler değildir. Ürünlerin çoğu farklı bir coğrafyada hatta farklı bir iklimde bulunan bir otel, bir tur gezisi veya bir ibadet ziyareti olabilir. Bu ürünleri müşterilere satabilmek için, Turizm firmaları bilgi teknolojilerini kullanarak müşterilerine ürünlerini en iyi şekilde tanıtmalıdırlar. Müşterilerin beklentilerinin karşılanması için, Gerekli bütün bilgi altyapısı bu teknolojik enstrümanlarla sağlanmalıdır. Turizm firmaları, web siteleri, mobil aplikasyonlar, rezervasyon sistemleri ve elektronik satış noktaları gibi rezervasyon ve bilgi yönetim sistemleri kullanmaya, global dağıtım sistemlerinin içinde yer almaya başlamışlardır. Bunu başarabilmek için interneti efektif kullanmak, turizm firmaları için çok önemli bir hale gelmiştir. Yeni teknolojileri sayesinde firmalar pazar gereksinimlerini en iyi şekilde yorumlayabilmekte ve rekabet stratejilerini günden güne geliştirebilmektedirler. Günümüzde internet ve bilgi teknolojilerini kullanmak giderek ucuzlamakta bununla birlikte daha fazla yaygınlık kazanarak uluslararası alandaki değişim sürecinde globalleşmeyi hızlandırmaktadır. Bilgi ve iletişim teknolojileri sayesinde bu teknolojileri kullanan donanımların; bilgisayar, cep telefonları, faks makineleri, uydu yayınları gibi yeni ürünlerin üretim sıklığı ve adedinin de artması, bu donanımları kullanan bireyler için de yeni iş olanaklarına imkan sağlamaktadır. Internet kanalı ile yapılan e-ticaret, dijital pazarlama etkinlikleri, yeni pazarlama teknikleri, yeni organizasyon ve yönetim yöntemleri ile ekonomik etkinlik artmaktadır. Gelişmekte olan ülkelerin kalkınmasında turizmin önemi sürekli vurgulanmakta, turizmin yeni ekonomik

sürecin bir parçası olduğu açık bir şekilde görülmektedir. Dünya seyahat ve turizm endüstrisi büyük değişimlere maruz kalmaktadır ve bu alanda varlık gösteren turizm firmalarının rekabet edebilme becerilerini korumalarının tek yolu, meydana gelen bu değişim ve gelişimlere mümkün olduğu kadar hızlı bir şekilde adapte olabilmeleridir. Aynı şekilde global anlamda turizm pastasından büyük pay almak isteyen ülkeler, turizm politikalarını oluştururken, bu değişim ve gelişimleri en iyi şekilde yorumlamalı ve turizm politikalarını bu yönde oluşturmalıdırlar. Global dağıtım sistemleri, turizm endüstrisine gerekli tüm bilgileri sağlamak için turizm endüstrisi koşullarınca oluşturulmuş bir sistemlerdir. Bu sistemlerin ortam özelliği hizmeti satın alacak müşteri için turizm firmalarının sunduğu hizmetleri(otel rezervasyonu, gezi veya kültür turları gibi) bir araya getirmesidir. Başlarda yalnızca seyahat acentalarının kullandığı bu sistemler, internetin hızlanması ve yaygınlık kazanması ile artık bütün müşterilerin kullanabileceği dijital platformlar haline gelmişlerdir. Sabre, Amadeus, Galileo ve Worldspan global dağıtım sistemlerinin en büyük ve en önemlileri olarak gösterilebilir. Bu şirketlerin sağladığı veriler ve teknik destekler ile turizm firmaları, müşterilerine, talepleri doğrultusundaki en iyi seçenekleri profesyonel anlamda sunabilmektedirler. Internet üzerinden son kullanıcılara da ulaşabilen bu şirketler, turizm firmasının sunduğu bütün hizmetleri bir araya getirerek müşterilerin çeşitli alternatifler arasından kendileri için en uygun olan seçeneği bulmalarında yardımcı olmaktadır. Global dağıtım sistemleri 90'lı yılların başından itibaren sürekli bir gelişim ve değişim eğilimi içerisindedirler. Firmalara sağladığı faydalar, sebep oldukları maliyetler ile karşılaştırıldığında firmalara büyük avantaj sağladığı ve geleneksel yöntemlere göre daha uygulanabilir olduğu söylenebilir. Teknoloji ve sehayat endüstrisi değişmeye devam ettikçe, bunların global tedarikçileri olan Sabre, Amadeus, Galileo ve Worldspan‘da gelişmeye devam edecek ve günün koşullarına göre sürekli kendilerini güncelleyeceklerdir. Teknolojinin ilerlemesi, turizm firmalarını ve tedarikçileri, varlıklarını sürdürebilmeleri için gelişmeye mecbur kılmaktadır.[19] Erdem, 2018

2. LİTERATÜR İNCELEMESİ

Müşteri profili oluşturma çalışmaları birçok sektörde yaygın olarak yapılmaktadır. Örnek olarak sağlık sektöründe [1] Frederick A. Barber, lazerle göz ameliyatı için müşteri profili çıkarmak ve bu ameliyat için en uygun kişileri belirlemek amacıyla lojistik regresyon analizi gerçekleştirilmiştir. (2001) Bir diğer çalışmayı [2] T. K. Das, geçmişteki satın alma hareketlerini inceleyerek, hangi ürünleri hangi müşteri profilinin almasının daha yatkın olacağına yönelik çalışmasında naive Bayes, k en yakın komşu, destek vektör makinesi algoritmalarını kullanmıştır. (2015). [3] Özer, online müzik endüstrisini ele almış ve bulanık küme analizi uygulamıştır. Analiz sonucunda online müzik sunucularını kullanan potansiyel kullanıcılar arasında homojen grupların bulunduğu tespit edilmiştir. Elde edilen her grup için farklı stratejiler geliştirilmiştir. (2001). Perakende sektöründe yapılan çalışmalara bakacak olursak [4] Mehpare TİMOR, sektörde faaliyet gösteren bir firmanın alışveriş kayıtları ve bu alışverişleri gerçekleştiren müşterileri ele alarak, Birliktelik kuralları analizi kullarını kullanarak müşterilerin alışveriş alışkanlıklarını belirlemeye çalışmıştır. Kümeleme algoritmaları ile müşteriler demografik özelliklerine göre ayrıştırılmıştır (2011). Bu sektörde bir diğer çalışmada veri madenciliği kümeleme yaklaşımlarından kohonen ağlarını kullanan [5] Gül Gökay EMEL, ele aldığı işletmede, pazar bölümlendirmesi, hedef pazar seçimi gibi firma stratejilerinin geliştirilmesinde yardımcı olabilmek için, önceden bilinmeyen kritik müşterilerin ve firma için ne kadar değerli olduklarının anlaşılmasında gerekli öngörüyü sağlamaya çalışmıştır. kohonen ağlarını tercih etme sebebi ise, büyük ölçekli veriler üzerinde çalışabilme yetisi ve kümeleme analizinde en önemli karar olan küme sayısını, bu yöntem ile en uygun olarak belirlenebilmesidir.(2010) Turizm sektöründe gerçekleştirilen Hedef kitle analizi çalışmasında [6]Eyüp Erkan Özbek ve arkadaşları, uygun ürünü uygun kitle ile buluşturmak için 2 sınıflı bir sınıflandırma tekniği kullanmışlardır. Sınıflar, seçilen bir kampanya için hedef olan ve hedef olmayan kişilerin özelliklerini içermektedir. Böylece sınıflandırma yöntemleri kullanılarak ulaşılmak istenen müşteriler belirleneceklerdir (2018). Sigortacılık sektöründe [7]Buket DOĞAN, Türkiye'de aktif olarak varlık gösteren, sektöründe öncü bir sigorta şirketinin müşterilerine ait verileri, veri madenciliğinin en çok tercih edilen birliktelik kuralı algoritmalarından Apriori algoritması ile analiz ederek, Bu analiz neticesinde müşterilerin sıklıkla hangi ürün

gruplarını bir arada satın almayı tercih ettiği tespit etmektedir. Müşteri ilişkileri yönetimi bakımından, birliktelik kuralı analiz sonuçlarından yararlanılarak daha olumlu sonuçlar verecek satış kampanyası ve pazarlama stratejisi geliştirmeyi amaçlamıştır. (2014). Telekomünikasyon sektöründeki bir diğer çalışmada [8]Umman Tuğba Şimşek Gürsoy, Türkiye'de telekomünikasyon sektöründe aktif faaliyet gösteren büyük bir şirketin, ayrılma şansı yüksek olan abonelerini belirleyerek; bu abonelere özel pazarlama stratejileri geliştirilmesini hedeflemiştir. Veri madenciliğinin yoğunlukla kullanıldığı uygulama alanlarından biri, ayrılma eğilimi gösteren abonelerin tahmin edilmesidir. Churn adı verilen bu analiz, firmaların kaybetme potansiyeli olan abonelerine özel pazarlama kampanyalarını geliştirmelerine yardımcı olmaya yöneliktir. Ayrılacak abone profilini belirlemek için veri madenciliği tekniklerinden Lojistik Regresyon Analizi ve sınıflandırma tekniklerinden Karar Ağaçları kullanılmıştır (2010).

Benzer bir çalışma da [9] Emel Kızılkaya Aydoğan; yine ayrılma eğilimi olan müşterileri tespit etmeye yönelik gerçekleştirdiği çalışmasında, Ayrılma eğilimi gösteren müşteri kesitini belirleyerek, şirketlerin bu müşterilere özel pazarlama kampanyalarını geliştirmelerine yardım etmeyi hedeflemektedir. Ayrıca müşteri bölümlenmesi yaparak benzer özellikler gösteren müşterileri gruplandırarak; gruplara özel, pazarlama programlarının geliştirilmesine yardımcı olmayı amaçlamaktadır. Bir kozmetik firmasının verilerini kullanarak gerçekleştirdiği çalışmasında bölümleme için kümeleme teknikleri, ayrılacak müşterileri belirlemek için ise sınıflama tekniklerini kullanmıştır. Bankacılık sektöründe, kaybedilen müşteri profilini tespit etmek için yapılan çalışmada [10] Tuğba Tosun, Yapı Kredi bankasına ait 30.000 müşteri datası üzerinde veri madenciliği yöntemlerinden karar ağaçlarını kullanarak, kaybedilmiş bir müşterinin profilini çıkartılmaya çalışmıştır. Karar ağaçları sonucu ortaya çıkarılan kurallar incelenerek, müşteri kayıplarının sebepleri ve ne zaman gerçekleştiği bilgisine ulaşılmaya çalışılmıştır. Uyguladıkları algoritmanın hızlı ve güvenilir olması amacıyla C programlama dili seçilmiştir. Algoritmayı gelişime açık bir halde yazarak daha sonra faklı müşteri dataları ile de işlem yapılabilir hale getirmeye çalışmışlardır (2006). [21] Songül Şekeroğlu, Kuyumculuk sektörü üzerinde yaptığı çalışmasında, bu sektörün tarihsel yapısı ve günümüzdeki ,pozisyonunu da dikkate alarak veri madenciliği teknikleri ile bu sektöre ait müşterileri segmentlerine göre ayrıştırmaya çalışmıştır. Müşteriler segmentlere ayrılırken şirketin stratejileri ve

kuyumculuk sektörünün koşulları dikkate alınmıştır. Analiz kriterleri bu iki parametreye göre belirlenmiştir. Analiz sonucunda elde edilen müşteri segmentlerine yönelik izlenecek stratejiler ve kampanya önerileri oluşturulmuştur (2010). Cumhuriyet Üniversitesi Sosyal Bilimler Enstitüsü ait veriler kullanılarak gerçekleştirilen bir diğer çalışmada [22] Mehmet Ali Alan, Lisansüstü öğrencilerin verilerini kullanarak, hem bu verileri en başarılı şekilde analiz eden algoritmayı tespit etmeye hem de öğrencilerin bölümü, cinsiyeti, hangi ilden geldikleri, bulunduğu kadronun araştırma görevlisi olup olmaması ve ders döneminin farklı olmasının notlarını etkileyip etkilemediğini incelemeyen bir uygulama geliştirmeye çalışmıştır. Çalışma neticesinde SimpleCART algoritmasının diğer algoritmalara göre sınıflandırma oranının daha yüksek olduğu, doktora programında öğrenim gören öğrencilerin, yüksek lisans programında öğrenim gören öğrencilere göre daha yüksek notlara sahip oldukları cinsiyetin, hangi ilden olduklarının, ders dönemlerinin ve kadrolarından araştırma görevlisi olup olmamasının başarılarını etkilemediği tespit edilmiştir (2012).

3. VERİ VE YÖNTEM

Veri madenciliği, veri ambarlarında bulunan çok çeşitli verileri kullanarak, daha önce keşfedilmemiş verileri elde etmek ve bu verileri karar verme ve gerçekleştirmek için kullanmaktır. Bu tanıma göre, veri madenciliğini bir istatistiksel süreç olduğunu görmekteyiz. Karar verme işinde alınan kararın doğruluğu her ne kadar kararı alan kişinin bilgisine, becerisine ve tecrübesine bağlı olsa da, bu kararı almasında sahip olduğu veri kümesinin rolü de büyüktür. Diğer bir anlamda kararın doğrulunda, verilerin doğru sınıflandırılması, doğru depolanması, doğru işlenmesi ve doğru yorumlanması çok önemlidir. Karar süreçlerinin karışıklığı, daha fazla veri üzerinden karar verme gereksinimi ortaya çıkartmış, bununla birlikte bu verileri manuel olarak yönetme ve yorumlama işlerini imkansız hale getirmiştir. Örneğin bir turizm firması bir takım keşfedilmemiş verileri son yıldaki satışlarına bakarak çıkarabilir fakat son 10 yıla bakarak çıkartacağı sonuçlar ve bu sonuçlar sonunda alacağı kararlar, sadece bir yılı incelemiş ve alınmış kararlara göre daha sağlıklı olacaktır. Yanlış karar verme riskinden kaçınmak için mümkün olduğunca daha fazla veri dikkate alınarak karar verilmelidir. Bu durum karar vericiyi daha fazla veriyi depolamaya zorlamaktadır. İnternetin globalleşmesi, müşteri memnuniyetinin giderek zorlaşması, firmalar arası rekabetlerin artması da veri madenciliğini kullanmayı mutlak kılmaktadır. Firmalar doğru veriyi depolama ve doğru veriye ulaşma adına büyük yatırımlar yapmakta, veriye sahip olmak kadar veriye ulaşmak da büyük rol oynamaktadır.

Veri madenciliğinin yararlarını sıralayacak olursak;

- Mevcut müşteri profillerinin karar vericiler tarafından daha iyi tanınmalarını sağlar.
- Müşterileri statülerine göre ayırarak, hedef kitle analizi ile şirketlerin satış politikaları oluşturmalarına yardım sağlayabilir.
- Müşteri memnuniyetinin azaldığı ve rekabetin arttığı günümüzde şirketlerin kendileri için hızlı ve en doğru kararı almaları sağlayabilir.
- Şirketlerin sahip olduğu ürünler üzerinden müşterilerine çarpraz satış kapasitesini arttırıcı politikalar üretebilir.
- Müşterilerin davranışları ve alışkanlıkları tespit edilerek, reklam politikalarında en pozitif sonuçlar anılabilir.

- En iyi ve sadık müşterileri tespit edilebilir ve bu müşterilere yönelik yeni pazarlama stratejileri geliştirilebilir.

Veriyi enformasyonel ve operasyonel veri olmak üzere iki çeşide ayırabiliriz. Enformasyonel, kişiye yönelik, bütünleşmiş, birleştirilmiş veriler olarak tanımlanabilir. Operasyonel, uygulamaya yönelik, dağınık, kısa zamanda elde edilmiş ve tekrar edebilen veriler olarak adlandırabiliriz. Veri ambarları, tarihsel bir derinliğe sahip, aynı amaç için bir araya getirilmiş verileri barındıran, bu veriler üzerinde analiz yapılabilmesi için özel olarak modellenmiş, veri depolama sistematikleridir. Veri ambarları enformasyonel verilerle kurulur. Bir veri ambarı az önce de bahsettiğimiz üzere ilgili karara özeldir ve o karar sürecine özel olarak modellenir.

Verinin kaliteli olması veri madenciliği için en önemli husustur. Veri madenciliğinde, sonuca güvenilirliğin artması için, elde bulunan veri ön işlemlerden geçirilmelidir. Aksi takdirde hatalı girdiler bizi yanlış sonuçlara götürebilir. Veri ön işleme çoğu zaman yarı otomatik olan ve zaman isteyen bir veri madenciliği aşamasıdır. En sağlıklı sonuca ulaşabilmek için mümkün olduğunca çok veri ile çalışma gerekliliği, bu verileri otomatik olarak işleyecek teknikleri önemli kılmaktadır. Verilerin ön işlemlerden geçirilebilmeleri için aşağıdaki sebepler gösterilebilir,

a. Verilerin içerisinde herhangi bir analiz türünün uygulanmasını engelleyecek türden veri problemleri çözülmeli
b. Verilerin doğası anlaşılmalı ve verilerin anlamlı bir şekilde analiz edilmesi
c. Verilen bir veri kümesinden daha anlamı bir bilginin üretilmesi

Çok sayıda veri ön işleme tekniği bulunmaktadır. Bu sebeple elde bulunan verinin hangi ön işleme tekniğine tabii tutulacağını tespit etmek, doğru sonuca ulaşma açısından önemli rol oynamaktadır. ([18] Famili). Veri ön işleme teknikleri sırasıyla;

i. Veri Temizleme (data cleaning)
ii. Veri Birleştirme (data integration)
iii. Veri Dönüştürme (data transformation)
iv. Veri İndirgeme (data reduction)

Eksik, tutarsız, yanlış veriler bütün veri tabanlarında bulunabilir. Bu tarz verilerin oluşmasında insan ve bilgisayar hataları, test çalışmaları, database ilk oluşturulduğunda bulunmayan bir özelliğin sonradan dikkate alınıp geç loglanmaya başlanması, verilerin kasıtlı veya kasıtsız silinmesi/kayıp edilmesi, verileri tablolar arası taşırken veya transfer

ederken değişmesi ve bozulması etkenleri rol oynayabilir. Bu tarz deformasyonlara uğramış veriler "kirli veri" olarak adlandırılır. Doğru sonuca ulaşmak için bu "kirli veri" olarak adlandırdığımız veriler yukarıda sıraladığımız veri ön işleme teknikleri ile tespit edilip düzeltilmeli veya veri tabanımızdan çıkartılmalıdır. Bundan sonraki kısımlarda bizi kirli verilerden kurtaracak olan ön işleme tekniklerini açıklanmaya çalışılacaktır. [20]Ayşe Oğuzlar, 2003

3.1 VERİ TEMİZLEME

Veri temizleme, eksik verilerin tamamlanması, aykırı değerlerin tespit edilmesi, verilerdeki tutarsızlıkların ortadan kaldırılması gibi aşamalar gerektirmektedir. Bu bölümde veri temizleme için temel yöntemlere kısaca değinilecek, bu problemin çözümü olarak seçilmiş yöntem olan kümeleme analizi detaylı bir şekilde açıklanacaktır. Herhangi bir eksik verinin doldurulması hususunda izlenebilecek farklı yollar vardır. Bunlardan bazılarını aşağıda açıklanmaktadır;

a. Eksik değer içeren veri kayıtlardan çıkartılabilir.
b. Eksik değerin yerine bu değerin ortalaması kullanılabilir.
c. Aynı sınıfta bulunan ve eksik değer içeren veriler için bu değerin ortalaması o değeri doldurmak için kullanılabilir. Örneğin aynı "Son Dakika Rezervasyon Müşterisi" sınıfına giren müşteriler için ortalama rezervasyon bedeli eksik değerler yerine kullanılabilir.
d. Diğer veriler analiz edilerek(regresyon veya karar ağacı gibi teknikler kullanılarak) eksik olan değer için en uygun değer bulunup kullanılabilir. Örneğin yaşı x, bütçesi y olan bir müşterinin tatil amaçlı hangi bölgeyi seçeceği az önce bahsettiğimiz teknikler kullanılarak tespit edilebilir ve eksik değerler için bu bulunan değer kullanılabilir. Veri temizleme tekniğinin kullanılmasını gerektiren bir diğer veri türü ise gürültülü verilerdir. Gürültü, veri giriş veya toplanması esnasında oluşan sistem dışı hatalara denir. Gürültülü veriyi değişken varyans ya da rassal hata olarak adlandırılabilir. Aşağıda veri temizleme tekniklerinden bazıları açıklanmıştır.

3.1.1 Binning Yöntemi

Binning yöntemi küçükten büyüğe veya büyükten küçüğe sıralanmış olan verileri düzenlemek için kullanılır. Bu yöntem öncelikleri verileri eşit büyüklüklerdeki binlere ayırır. Sıralanmış datadaki minimum ve maksimum değerler tespit edilir. Minimumlar sol tarafa, maksimumlar sağ tarafa yerleştirilir. Bin sınırındaki orta değerler, daha az mesafeli en yakın komşu değerlerine taşınır ve data bin sınırları yardımı ile düzeltilmiş olur.

3.1.2 Kümeleme Yöntemi

Kümeleme yöntemi ile benzer değerler aynı kümelere ayrıştırılarak, küme dışarısında kalan aykırı değerlerin tespit edilmesi hedeflenmektedir. Küme dışında kalan veriler analizden çıkartılmalıdır.

3.1.3 Regresyon Yöntemi

Regresyon, eksik olan dataları telafi etmek amacıyla kullanılan bir veri temizleme yöntemidir. Bu yöntemde iki veya daha fazla değişkenin, kendilerine uygun bir regresyon fonksiyonu yardımıyla daha önce bilinen datalardan yola çıkarak eksik olan dataların tahmin edilmesi hedeflenmektedir. Regresyon aynı zamanda yüksek gürültülü verilerinde düzeltilmesinde kullanılmaktadır.

3.2 VERİ BİRLEŞTİRME

Veri madenciliği yapabilmek için genellikle farklı veri tabanlarında bulunan datalar birleştirilmelidir. Bu veri tabanları birleştirilerek veri ambarları oluşturulur. Genellikle veri madenciliği metotları bu offline daha üzerinde yapılır. Çünkü veri tabanlarında anlık veri girişi, veri kaybı veya manipülasyon olabilir. Bir veri madenciliği fonksiyonunu çalıştırdığımızda, üzerinde çalıştırdığımız datanın izole olması, sonucun doğruluğu açısından önemlidir. Bu veri ambarının oluşması esnasında farklı veri tabanlarında aynı değişken için farklı kolon isimleri kullanılmış olabilir. Örneğin bir tabloda “Customer-Id” olan bir alan diğer tabloda “MusteriNo” olarak tutuluyor olabilir. Bu tip şema birleşmelerinde meydana gelebilecek hatalardan kaçınmak için meta data kullanılır. Meta

data o tabloda bulunan dataya ilişkin bilgidir. Meta dataları uyuşan veriler birleştirilirken alanlardan birinin adı diğeriyle güncellenebildiği gibi her iki alan yeni bir alan adı altında veri ambarındaki yerini alabilir. Örneğin tablo 1'deki "Customer-Id" alanı ile tablo 2'deki "MusteriNo" alanları veri ambarında "Musteri-Id" olarak saklanabilir.

Veri birleştirmelerde önemli bir diğer konu ise veri indirgemeleridir. Bir değişken sadece bir tabloya özel türetilmiş ise fazlalık veri olabilir. Bu fazlalık veri sonuçta elde edilen veri ambarı için fazlalık olabilir. Örnek olarak yukarıda bahsedilen "Customer-Id" ve "MusteriNo" değişkenleri için korelasyon katsayısı bulunabilir. Eğer korelasyon yüksek bulundu ise değişkenlerden biri veri tabanından çıkartılmalı ve indirgeme yapılmalıdır.

3.3 VERİ DÖNÜŞTÜRME

Veri dönüştürme işlemi ile veri seti, veri madenciliği için uygun formata dönüştürülmüş olur. Veri dönüştürme, birleştirme, düzeltme, normalleştirme ve genelleştirme gibi değişik işlemlerden birini veya bir kaçını içerebilir. Veri normalleştirme bu yöntemlerden en sık kullanılanıdır. Veri normalleştirme ile veriler kendileri için tayin edilen yeni aralığa doğrusal dönüşüm uygulanarak dönüştürülürler. Bu veriler için tayin edilen veri aralığı genellikle 0-1 aralığıdır. Örneğin bir rezervasyon sistemi içerisinde bir rezervasyon için en yüksek ödenen para 1 olarak dönüştürülebilir. Daha sonra gelecek olan değerler 0.9… şeklinde olacaktır. Ödenen değer düştükçe yeni tayin edildiği dönüşüm değeri 0'a yaklaşacaktır. Minimum değer 0, maksimum değer ise 1 olarak normalleştirilmiş olur. Diğer bir normalleştirme yöntemi ise z skor yöntemidir. Z skor normalleştirme, bir önceki normalleştirmeden farklı olarak ortalama değerleri ve standart sapmayı baz almaktadır. Standart sapma standart skor olarak adlandırılır. Örneğin bir veri setinin maksimum noktası 29, minimum noktası 5 ve bu veri setinin ortalama değeri (bütün verilerin aritmetik ortalaması alınarak bulunan ortalama değeri) 17 olsun. 5 in 17'ye olan uzaklığı ile 29'un 17 olan uzaklığı eşittir. Bu değerlerin normalleştirilmiş halleri aynı değere sahip olacaktır fakat yönleri farklıdır. 5, 17'den 12 birim geride olduğun için yönü –(negatif) olurken 29, 17'den 12 birim ileride olduğu için yönü +(pozitif) olacaktır. Z-skor normalleştirmesinde sayılar ortalama değere olan uzaklıklarına göre

normalleştirilmektedir. Ayrıca standart sapmaya bölünerek, veriler arasındaki değişim hızı ortalamaya olan uzaklığı normalize etmektedir.

3.4 VERİ İNDİRGEME

Veri indirgeme teknikleri, daha küçük hacimli ve tutarsız veri içermeyen indirgenmiş bir küme elde etmek amacı ile uygulanmaktadır. Bu indirgenmiş kümede uygulanacak veri madenciliği teknikleri ile daha etkin sonuçlar elde etmek mümkün olacaktır.

3.5 KÜMELEME ANALİZİ

Kümeleme, adından da anlaşılacağı üzere benzer öğeleri bir arada gruplamayı amaçlayan çok değişkenli istatistik tekniklerinden biridir. Bu işlem sonucunda elde edilen küme içerisindeki öğeler, önceden belirlenmiş bir özellik bakımından benzerlik göstermektedirler. Kümeleme tekniğinin esas amacı, dağınık halde bulunan datayı, birbirine benzer gruplar haline getirerek, işlenebilirliğini sağlamaktır.

Kümeleme modellerinde amaç, içerik olarak birbirlerine çok benzeyen elemanlardan oluşan, özellikleri olarak kıyaslandıklarında da birbirlerinden olabildiğince farklı olan kümeler yaratmak ve veri tabanındaki datayı bu kümelere bölüştürmektir.

Bu şekilde küme içi homojenlik ve kümeler arası heterojenlik maksimum seviyede tutulmuş olacaktır. Sonuç itibari ile bir kümenin elemanı bulunduğu kümedeki elemanlar ile benzeşirken başka kümenin elemanları ile benzeşmeyecektir.

Kümeleme Analizi genel olarak üç aşamadan oluşmaktadır. İlk aşamada veri matrisi hazırlanır. Veriler kümelemeye uygun şekilde girilir ve uzaklıklar matrisi elde edilir. İkinci aşamada hangi kümeleme yöntemi kullanılacak ise belirlenir ve uygulanır. Üçüncü ve son aşamada ise bulgular elde edilir ve değerlendirilir.

En yaygın olarak kullanılan kümeleme tekniği k-ortalamalar (k-means) yöntemidir.

3.5.1 K-Means Kümeleme Algoritması

1967 yılında J. B. MacQueen tarafından geliştirilmiştir. K-means algoritması, data mining alanında en çok kullanılan kümeleme algoritmalarının başında gelmektedir. Bu algoritma en basit tabir ile büyük veri toplulukları, birbirine benzer veriler aynı kümelerde bulunacak şekilde birbirinden farklı özelliklere sahip kümelere ayrıştırmaktadır. Bir eleman yalnızca bir kümeye ait olabilir. Kümenin merkezi, kümeyi temsil eden değerdir. k-means adındaki k harfi küme sayısını belirtmektedir. Bu algoritma temel olarak 4 adımdan oluşur;

a. Küme merkezleri belirlenir.

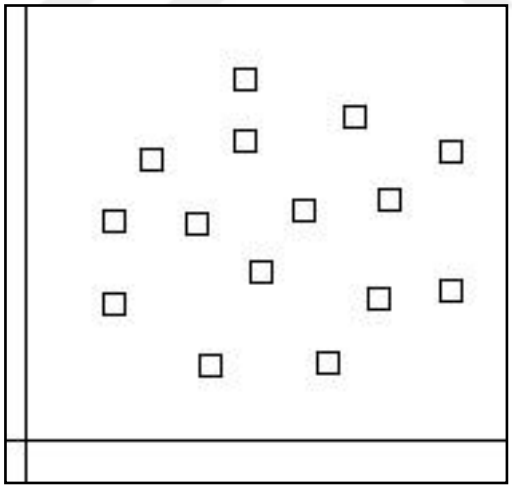
b. Merkezlere yakın olan elemanlar sınıflandırılır.

c. Yapılan sınıflandırmaya göre yeni merkezler belirlenir (eski merkezler, sınıflandırmaya göre kaydırılarak modelin ağırlık merkezindeki yerini alır)

d. Merkezler kararlı hale gelene kadar sınıflandırma ve kaydırma işlemi devam eder.

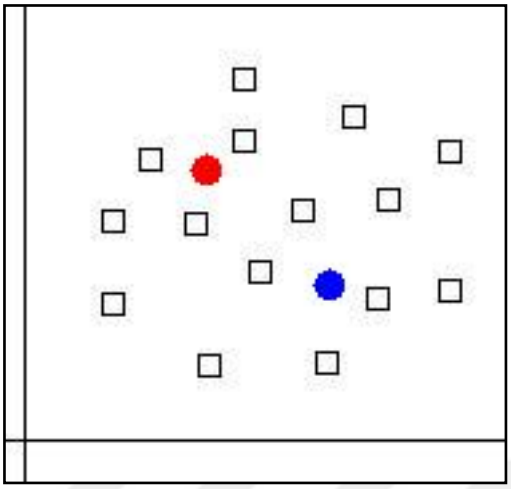
Çalışmanın daha net anlaşılabilmesi için;

Şekil 3.1: Örnek uzay



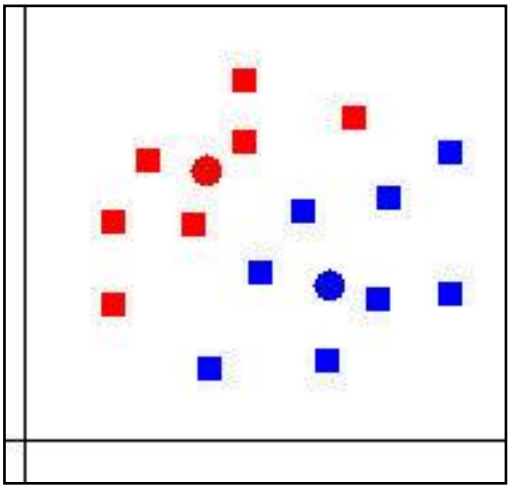
Yukarıda verilen düzlemdeki elemanları, iki farklı kümede gruplamaya çalışalım;

Şekil 3.2: Hedef Kümelerin seçilmesi



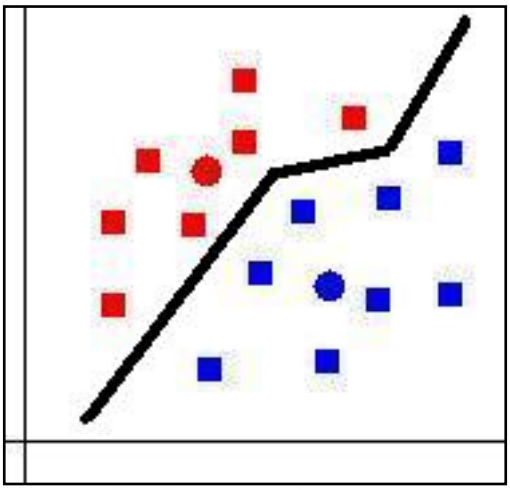
elemanları gruplayabilmek için öncelikle iki hedef küme tanımlıyoruz ve bu elemanları bu kümelere olan uzaklıklarına göre sınıflandırıyoruz. (hangi renge daha yakın mesafede ise)

Şekil 3.3: Hedef Kümelere göre diğer noktaların sınıflandırılması



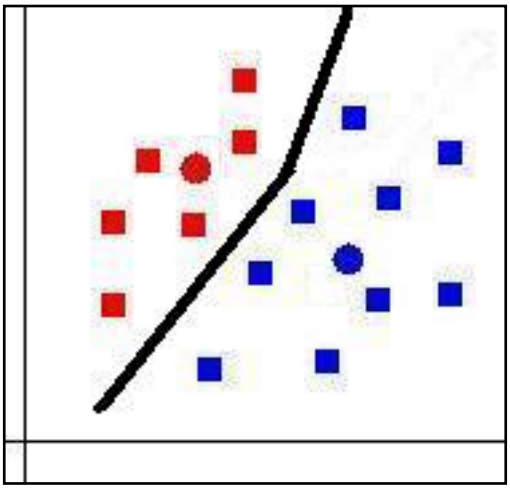
Sınıflandırmayı ayıracak bir hat, aşağıdaki şekildeki gibi çekilebilir;

Şekil 3.4: Noktaların ayrıştırılması



Daha önceden sınıflandırdığımız bu elemanlarımızın yeni merkezini buluyor ve eski merkezleri yenilerine doğru kaydırıyoruz.

Şekil 3.5: Ayrışma sonrası yeni ağırlık merkezinin seçilmesi ve noktaların yeniden sınıflandırılmaları



Merkezleri hareket ettirdikten sonra, merkezlerin yeni pozisyonlarına göre elemanları yeniden sınıflandırıyoruz. (Yeni merkezlere göre, eskiden maviye daha yakın olan bir

nokta var fakat şu anki güncellemeden sonra eğer bu nokta kırmızı merkeze daha yakın ise, noktanın rengini kırmızı olarak güncelliyoruz.)
Noktalarda herhangi bir renk değişimi kalmayana kadar merkez kaydırma ve kaydırma sonucu oluşacak renk değişimlerini gerçekleştirmeye devam ediyoruz. Sonuç olarak kararlı (stabil) merkezlere sahip kümeler elde ediyoruz. [12](Şadi Evren Şeker, 2008)
Küme elemanlarını tekrar tekrar sınıflandırma ve sonrasında yeni merkezlerin belirlenmesi işlemlerini belirlemek için k-means noktalarının ağırlık merkezini bulma fonksiyonu kullanılır;

$$minimize[J = \sum_{AllRedPoints} distance(C1, Red_Point) + \sum_{AllBluePoints} distance(C2, Blue_Point)]$$

Burada C1 eski kırmızı merkezimiz, C2 ise eksi mavi merkezimiz olmak üzere
K tane döngünün de formüle eklenmiş hali; k-means noktalarının ağırlık merkezini bulma fonksiyonu;

$$argmin_s \sum_{i=1}^{k} \sum_{x_j e S_i} ||x_j - u_i||^2$$

Bu formülde amaç, her bir merkez çifti için sınıflandırmadaki tüm elemanların bu merkezlere olan uzaklıkları toplanır. Merkez çiftleri kaydırıla kaydırıla bu işlem için en uygun merkez çiftleri bulunur. [13](Mustafa Akça)

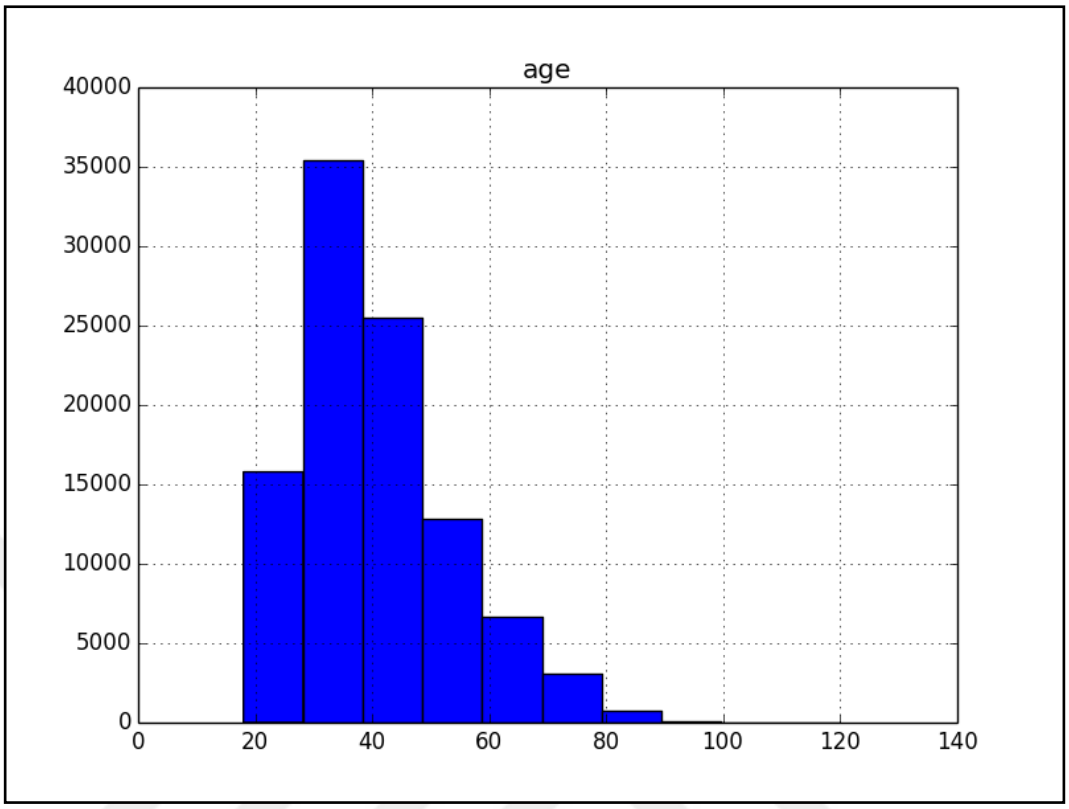
3.6 VERİ ANALİZ İŞLEMLERİ

Analizde kullanılan veri, bir turizm firmasının son altı yıllık müşteri ve rezervasyon bilgilerinin bir araya getirilmesiyle ilişkisel veri tabanı tablosu yapısında oluşturulmuştur. Rezervasyon bilgisi müşteri bazında gruplanarak müşteri bilgisine eklenmiştir. (Veri, kişisel verilerin korunması kanunu (KVKK) gereğince isim, soy isim, cep telefonu, email gibi kişiyle doğrudan ilişkilendirilebilecek özniteliklerden arındırılmıştır.) Veri kaynaktan alınırken bir ETL ve veri işleme yazılımı olan ve big data cluster'ı üzerinde

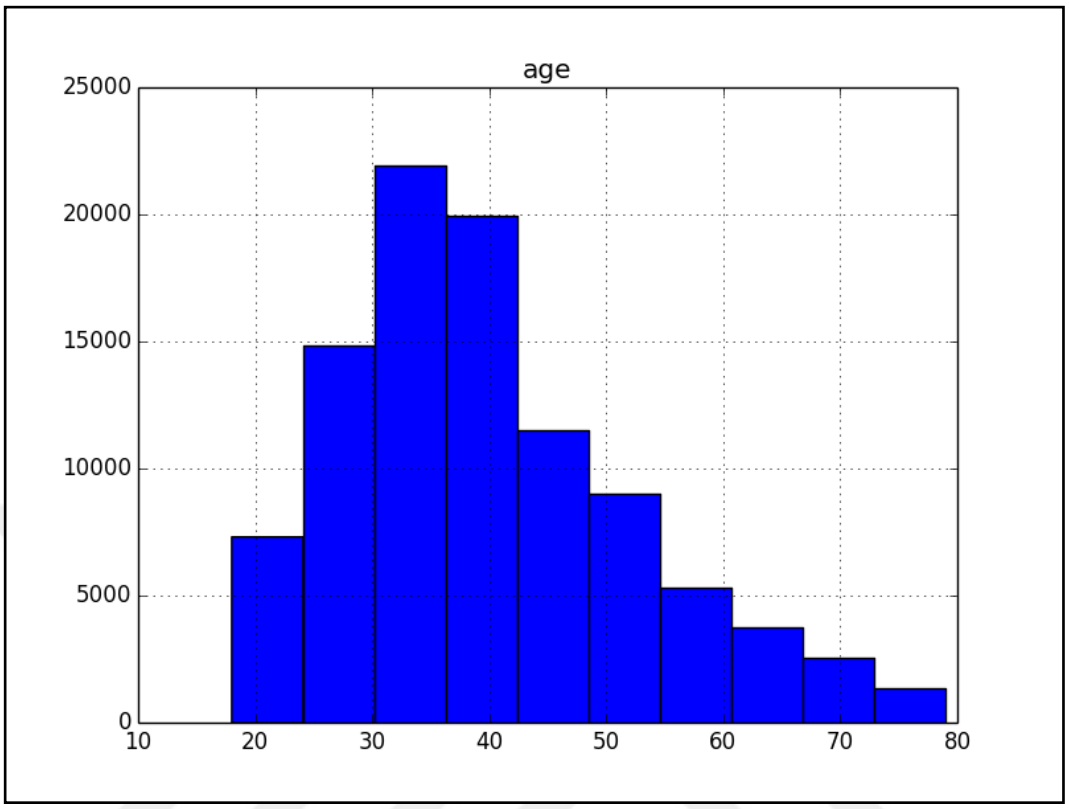
çalışan Datameer kullanılmıştır. Turizmde bulunan rezervasyon verisi telekomünikasyon ya da banka verileriyle karşılaştırıldığında daha kirli ve eksik halde bulunmaktadır. Bunun sebebi, müşterilerinin genel olarak sadece doldurmak zorunda oldukları alanlarla ilgili bilgi vermesi ve bu bilgilerin doruluğunun bahsedilen diğer sektörlerdeki kadar kritik olmamasıdır. Veri önişleme adımının yapılması analizler üzerinde olumlu yönde, önemli ölçüde etki etmektedir. Önişleme sonucunda verinin temizlenmesiyle, yapılacak analiz daha tutarlı ve anlamlı hale gelmektedir. Bu yüzden veri analizde kullanılmadan önce bazı önişleme adımlarından geçirilmiştir. Özniteliklerdeki boş değerler hesaplanmış, %70'nin üzerinde boş olan kolonlar atılmış. Bu ortalamanın altındaki veriler kendi kolonundaki diğer değerlerin ortalaması ile doldurulmuştur. Kategorik veri içeren ürün tiplerinin bulunduğu kolon üzerinde one-hot encoding yapılarak değerler farklı kolonlarda {0, 1} değerlerini alacak şekilde ifade edilmiştir [13].Aykırı değer, uzayda gözlemlenen bir noktanın, gözlemlenen diğer noktalardan uzak olması durumunda ortaya çıkar ve özellikleriyle kümenin diğer elemanlarından ayrılır. Aykırı değerler oluşmasının en önemli sebebi, veri çeşitliliği ile beraber, veriler kaydedilirken ya da toplanırken hata yapılmasıdır. Aykırı değerlerin fazla olması veri setinin normal dağılımdan sapmasına ve yapılacak analizlerin olumsuz yönde etkilenmesine sebep olabilir. Bu adımda önemli olan hata sonucu oluşmuş değerlerin ihmal edilmesidir. Diğer değerlerden farklı olduğu için aykırı olarak belirlenmiş olan değerlerin, gerçekten aykırı olup olmadığı kontrol edilmelidir. Z-score sonucunda veri setinin dağılımı normal dağılıma daha çok benzetilerek sonucun daha güvenilir ve tutarlı olması sağlanmıştır.

Çalışmada aykırı değerlerin tespitinde z-score yöntemi kullanılmıştır. Z-score yönteminde, verilerin, veri gruplarıyla standart sapma ve ortalama açısında ilişkisi incelenir. Z-score sonucunda değerler aynı ölçek, yani 0 ortalama, 1 standart sapma üzerine gelir.Şekil 3.6'da "age" kolonu için, veri üzerinde Z-score hesaplanması ve aykırı olan veriler atılmasından önceki, Şekil 3.7'de ise sonraki dağılım görülmektedir.

Şekil 3.6: Aykırı değerler atılmadan önce yaş verisi



Şekil 3.7: Aykırı değerler atıldıktan sonra yaş verisi



[14] Sonrasında da kolonlar üzerinde min-max normalizasyon işlemi yapılarak tüm değerler [0, 1] arasına alınmıştır. Ayrıca, öznitelikler arasındaki korelasyon hesaplanmış, korelasyonu yüksek çıkan özniteliklerden sadece bir tanesi kalacak şekilde diğerleri elenmiştir.

Önişleme adımlarından sonra, 1000000 satırlık ve 29 kolonluk bir veri, analizde kullanılmak üzere oluşturulmuştur. Analizde kullanılan öznitelikler ve dağılımları Tablo 2'de görülmektedir. Tablo 3'de ise özniteliklerle ilgili sayısal verilere yer verilmiştir.

Tablo 3.1: Analizde Kullanılan Öznitelikler

Öznitelik	Açıklama
age	Müşterinin yaşı
total_res_count	Müşterinin toplam yaptığı rezervasyon sayısı
months_since_last_res	Müşterinin yaptığın son rezervasyondan bugüne kadar geçen ay ortalaması
avg_month_between_vocation_arrival_dates	Tatiller arasındaki ay ortalaması
total_amount	Tüm tatillere harcanan toplam para
total_accomodation	Tüm tatillerde kalınan gece sayısı
unit_price	Bir kişi için bir gece ödenen para miktarı
total_days_between_res_and_arrival	Rezervasyon yapılmasıyla tatile gidilmesi arasında geçen toplam gün sayısı
average_days_between_res_and_arrival	Rezervasyon yapılmasıyla tatile gidilmesi arasında geçen ay ortalaması
min_days_between_res_and_arrival	Rezervasyon ile tatile gidiş arasında geçen en küçük gün sayısı
max_days_between_res_and_arrival	Rezervasyon ile tatile gidiş arasında geçen en büyük gün sayısı
average_res_amount	Rezervasyon başına harcanan ortalama para
average_res_accomodation	Rezervasyonlarda kalınan ortalama gece sayısı
total_pax	Rezervasyonlarda toplam para ödenen kişi sayısı
total_adult	Rezervasyonlarda toplam yetişkin sayısı

total_child	Rezervasyonlarda toplam çocuk sayısı
_2014_res_count	Müşterinin 2014 yılında yaptığı toplam rezervasyon sayısı
_2015_res_count	Müşterinin 2015 yılında yaptığı toplam rezervasyon sayısı
_2016_res_count	Müşterinin 2016 yılında yaptığı toplam rezervasyon sayısı
_2017_res_count	Müşterinin 2017 yılında yaptığı toplam rezervasyon sayısı
_2018_res_count	Müşterinin 2018 yılında yaptığı toplam rezervasyon sayısı
_2019_res_count	Müşterinin 2019 yılında yaptığı toplam rezervasyon sayısı
q1	Müşterinin yılın ilk çeyreğinde yaptığı toplam rezervasyon sayısı
q2	Müşterinin yılın ikinci çeyreğinde yaptığı toplam rezervasyon sayısı
q3	Müşterinin yılın üçüncü çeyreğinde yaptığı toplam rezervasyon sayısı
q4	Müşterinin yılın dördüncü çeyreğinde yaptığı toplam rezervasyon sayısı
active_year_count	Müşterinin rezervasyon yaptığı toplam yıl sayısı
tkoy_res_count	Tatil köyü rezervasyon sayısı
kult_res_count	Kültür turu rezervasyon sayısı
ydis_res_count	Yurtdışı rezervasyon sayısı

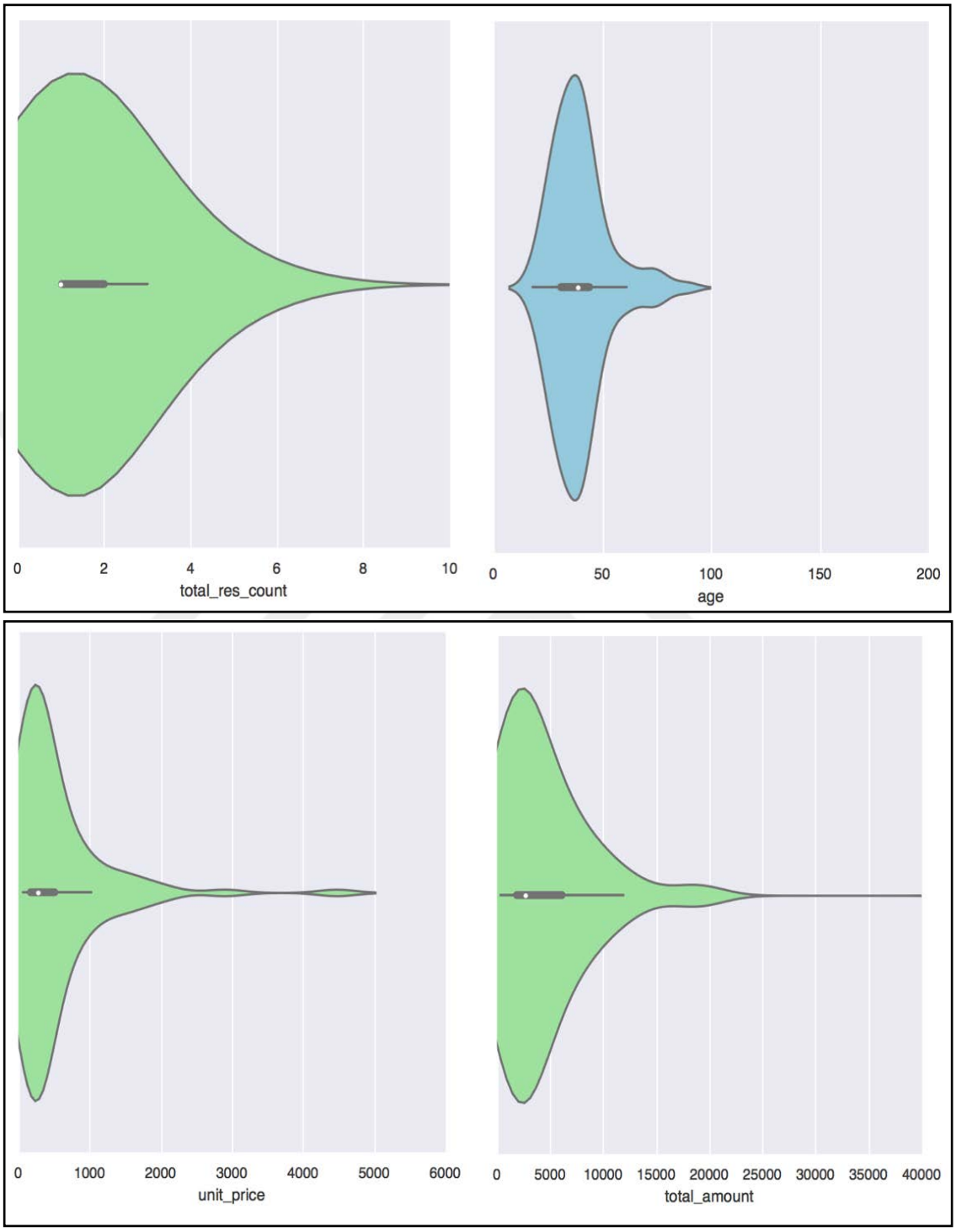
Tablo 3.2: Özniteliklerin Dağılımı

Öznitelik	Ortalama	Standart Sapma	En Küçük Değer	En Büyük değer
Age	40,309	13,002	18,000	80,000
total_res_count	1,874	2,132	1,000	190,000
months_since_last_res	33,216	24,862	0,000	92,200
avg_month_between_vocation_arrival_dates	16,440	8,364	0,010	89,960
total_amount	6189,682	10261,250	0,000	494154,000
total_accomodation	7,843	9,308	0,000	661,000
unit_price	557,302	845,441	0,000	50887,000
total_days_between_res_and_arrival	98,179	177,286	0,000	4926,000
average_days_between_res_and_arrival	49,565	59,048	0,000	344,000
min_days_between_res_and_arrival	38,654	56,746	0,000	344,000
max_days_between_res_and_arrival	63,160	72,697	0,000	356,000
average_res_amount	3088,389	2759,260	0,000	245706,000
average_res_accomodation	4,212	2,040	0,000	35,000
total_pax	4,920	5,735	0,000	261,000
total_adult	4,030	4,507	0,000	260,000
total_child	0,891	1,880	0,000	53,000
_2014_res_count	0,135	0,413	0,000	21,000
_2015_res_count	0,168	0,476	0,000	25,000
_2016_res_count	0,172	0,478	0,000	33,000
_2017_res_count	0,233	0,654	0,000	94,000

_2018_res_count	0,310	0,621	0,000	32,000
_2019_res_count	0,328	0,632	0,000	25,000
q1	0,154	0,540	0,000	48,000
q2	0,499	0,938	0,000	67,000
q3	1,052	1,342	0,000	94,000
q4	0,169	0,539	0,000	59,000
active_year_count	1,130	0,994	0,000	6,000
tkoy_res_count	1,370	1,803	0,000	181,000
kult_res_count	0,044	0,252	0,000	15,000
ydis_res_count	0,072	0,363	0,000	15,000

Şekil 3.8’de toplam rezervasyon sayısı, yaş, gecelik ortalama ödeme miktarı ve toplam ödeme miktarı özniteliklerinin dağılım diyagramları görülmektedir.

Şekil 3.8: Özniteliklerin Dağılımları



3.4 ELBOW YÖNTEMİ

K-means, bir veri kümesini kullanıcı tarafından belirlenen sayıda (k) kümede gruplayan basit bir veri madenciliği algoritmasıdır. Algoritma için verilen k değeri, doğru küme sayısı olmasa bile, verileri k kümelerine böler. Bu nedenle, k-means kümelemesini kullanırken, kullanıcılar doğru sayıda küme kullanıp kullanmadıklarını tespit etmek için bir yola ihtiyaç duyarlar.

Küme sayısını doğrulamak için bir yöntem Elbow yöntemidir. Elbow yönteminin fikri; k-means kümesinde bir dizi k değeri için kümeleme yapmak ve her k değeri için küme içi noktaların merkeze olan uzaklıklarının karesi alınmış ortalamasını hesaplamaktır (WCSS). Bunun gibi:

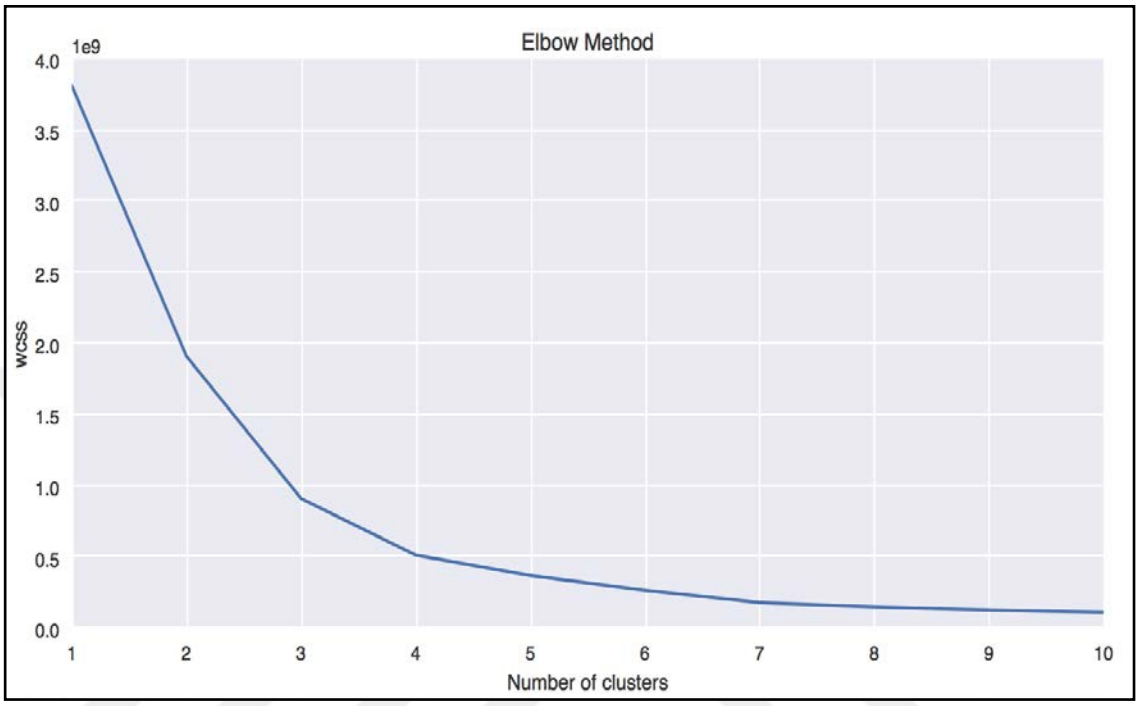
```
var sse = {};
for (var k = 1; k <= maxK; ++k) {
   sse[k] = 0;
   clusters = kmeans(dataset, k);
   clusters.forEach(function(cluster) {
      mean = clusterMean(cluster);
      cluster.forEach(function(datapoint) {
         sse[k] += Math.pow(datapoint - mean, 2);
      });
   });
}
```

Ardından, k'nin her değeri için WCSS bir çizgi grafiğini çizilir. Çizgi grafiği bir kol gibi görünüyorsa, koldaki "dirsek" en iyi k değeridir. Fikir, küçük bir WCSS istediğimizdir, ancak WCSS, k'yi artırdıkça 0'a düşme eğilimindedir (k, veri kümesindeki veri noktalarının sayısına eşit olduğunda WCSS 0'dır, çünkü her veri noktası kendi küme ve kümesinin merkezi arasında uzaklık farkı yoktur). Bu yüzden hedefimiz hala düşük WCSS'ye sahip küçük bir k değeri seçmek. Dirsek genellikle k değerini artırarak azalan geri dönüşlere başladığımız yeri temsil ediyor. Elbow yöntemi her zaman iyi çalışmıyor; özellikle veriler çok kümelenmemişse. Bu sebeple en iyi sonucu elde edebilmek için verilerin iyi şekilde kümelenmiş olmaları gerekmektedir.

Çalışmada müşteriler k-means yöntemiyle öbeklenmiştir. Yöntemde k sayısı 3 olarak seçilmiştir. [15] Bunun seçiminde elbow yöntemi kullanılmıştır.

Şekil 3.9'dan görüleceği üzere, en fazla değişim öbek sayısı 3 seçildiğinde görülmüştür.

Şekil 3.9: Elbow Yöntemi (Küme içi noktaların merkeze uzaklığının kareler ortalamasının küme sayısına etki grafiği)



Analiz, python programla dilinde sklearn kütüphanesi kullanılarak yapılmıştır. Kullanılan KMeans algoritmasındaki parametre değerleri şu şekildedir; *n_clusters=3, max_iter=300, init='k-means++'*

3.5 GEÇERLİK VE GÜVENİRLİK

Kümeleme analizinde genel amaç birbirine benzer olan bireylerin aynı gruplarda toplanmasıdır. Kümelerin başarısını değerlendirmek için sınıflara atanan veriler arasındaki benzerliklere ve kümeler arasındaki farklılığa bakılmaktadır. Bu bakımdan kümeler arasındaki benzerliği ölçmede hangi ölçünün kullanılacağı kümeleme analizinin en önemli sorunlarından birini teşkil eder. Kümeleme yapan bir model ortaya konduktan sonra bu modelin başarısının ölçülmesi gerekmektedir. Silhouette score yapılan kümeleme işleminin başarısının ölçülmesinde kullanabilecek bir yöntemdir. Bu yöntemde her veri için iki uzaklığı baz alır. Bu uzaklıklardan ilki verinin bulunduğu

kümeye ait diğer verilere olan uzaklıkların ortalamasıdır. Silhouette score, kümeleme işlemi sonucunda öbekler arasındaki mesafenin incelenmesinde kullanılmaktadır. Bir öbekteki noktaların komşu bir öbekteki noktalara ne kadar yakın olduğu hesaplanır. Silhouette scoru'nun değeri [-1, 1] arasında çıkmaktadır. Değerin 1 olması, bir öbekteki noktanın, komşu öbekten en uzak noktada olduğunu gösterir. Benzer şekilde, değerin 0 olması ise örnek noktanın komşu öbekle mevcut öbeğin sınırında olduğunu gösterir. Silhouette score'nun negatif olması ise değer sahibi noktaların yanlış öbekte olduğunu göstermektedir. [16] Yapılan öbekleme analizinde, öbek sayısının başarısı için Tablo 3.3'de silhouette score hesaplanmıştır. Tablo değerlerinde de görüleceği üzere, öbek sayısının 3 olarak seçildiği durumda en yüksek silhouette score elde edilmiştir.

Tablo 3.3: Silhouette Score Hesaplanışı

Öbek Sayısı	Silhouette Score
3	0,324222114436
4	0,307782351272
5	0,260950552594
6	0,256316559071
7	0,253742155711
8	0,257216372113
9	0,264428504859
10	0,265572065843

4. BULGULAR

Analizde özniteklerin öbeklenmedeki önem sırası (feature importance) Tablo 4.1'de görülmektedir.

Tablo 4.1: Öznitelikerin Önem Sırası

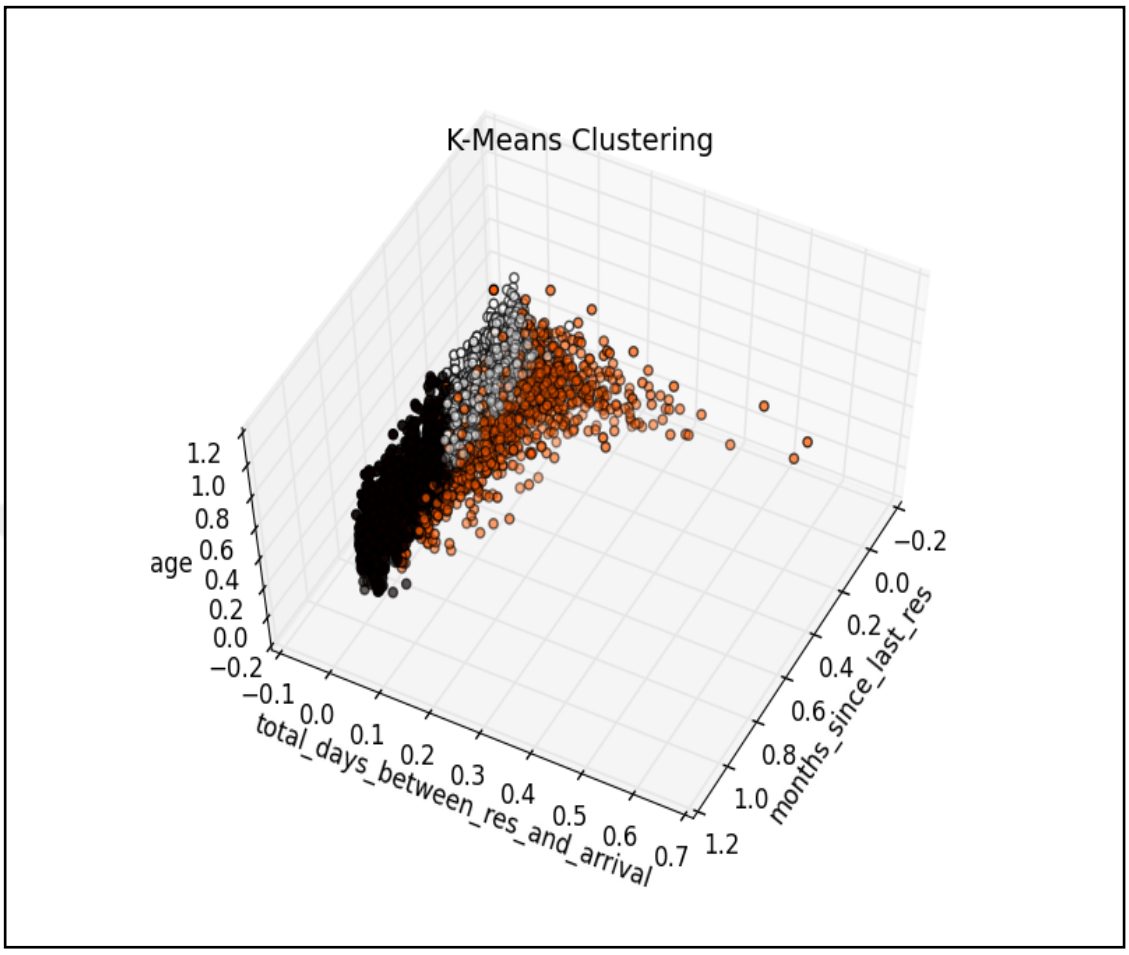
Öznitelik	Önem Katsayısı
months_since_last_res	0.3
total_days_between_res_and_arrival	0.26
average_days_between_res_and_arrival	0.25
max_days_between_res_and_arrival	0.25
active_year_count	0.22
total_res_count	0.19
min_days_between_res_and_arrival	0.19
age	0.19
agency_count	0.15
total_adult	0.14
total_pax	0.14
tkoy_res_count	0.13
total_accomodation	0.13
_2019_res_count	0.11
avg_month_between_vocation_arrival_dates	0.11
q3	0.1
total_amount	0.1
_2017_res_count	0.1
_2018_res_count	0.09
_2015_res_count	0.09
_2016_res_count	0.08
unit_price	0.08

_2014_res_count	0.07
total_child	0.06
average_res_accomodation	0.05
q2	0.03
average_res_amount	0.03
q1	0.02
q4	0.01
kult_res_count	0.01
ydis_res_count	0.01

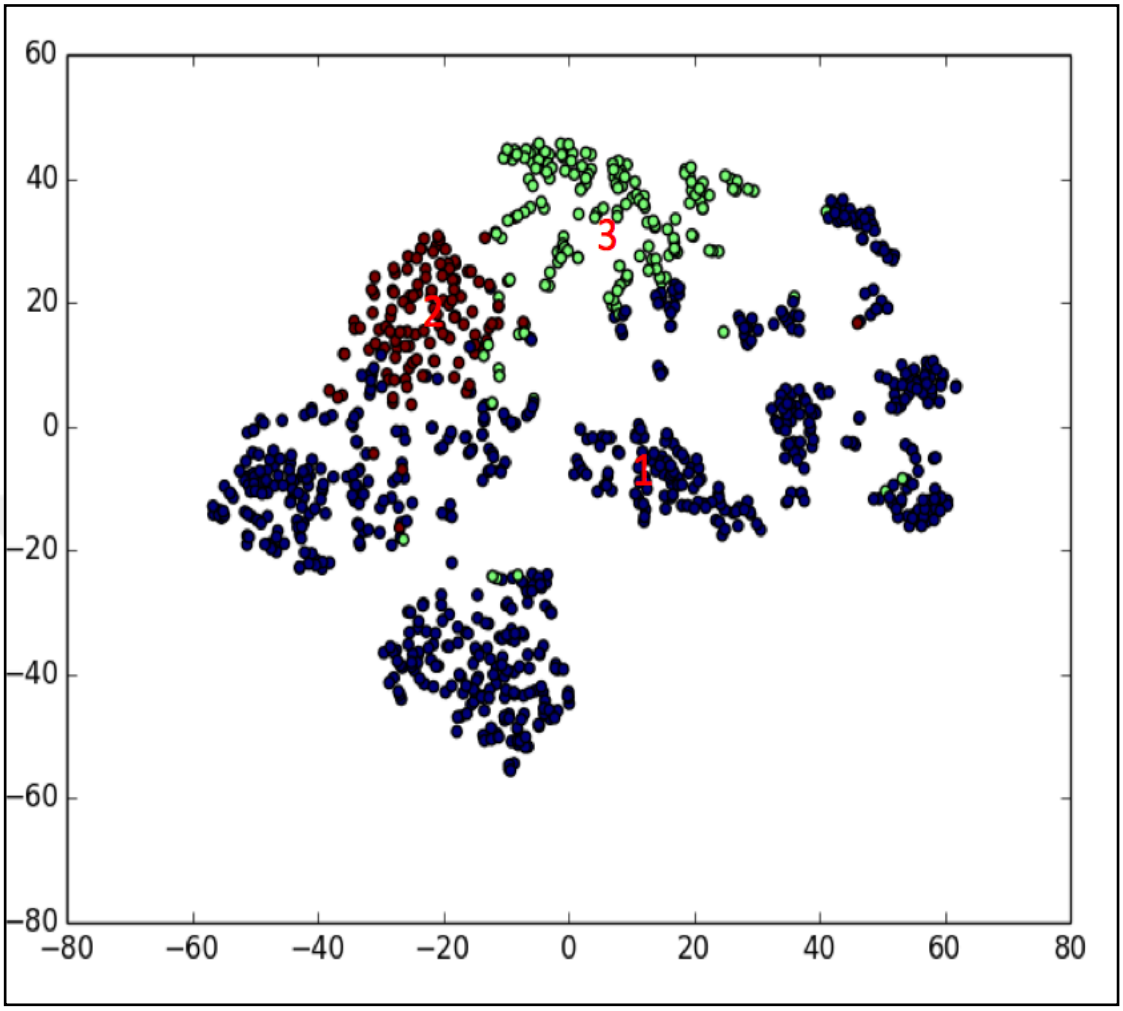
Tablodaki sonuçlar incelendiğinde, verideki en önemli kırılımın rezervasyon yapma sıklığıyla alakalı olduğu görülmektedir. Ayrıca erken rezervasyondaki indirimlerden dolayı, yapılan rezervasyon ile tatile gitme arasındaki sürenin de önemli bir parametre olduğu görülmüştür.

Analizdeki önemli özniteliklerden olan "months_since_last_res", "total_days_between_res_and_arrival" ve "age" için öbekler Figure 10'de görülmektedir. Veri, k-means yöntemiyle 3 öbeğe ayrılırken 29 farklı öznitelik kullanıldığından, öbeklerin 3 boyutlu olarak ifade edilebilmesi için bu önemli olan bu 3 öznitelik seçilmiştir [17].

Şekil 4.1: 3 öznitelik için öbek dağılımı



Şekil 4.2: t-SNE çıktısı



t-SNE lineer olmayan bir boyut düşürmek algoritmasıdır. Bu algoritmada, noktaların benzerliklerine göre bulunmuş kümeler incelenerek veride şablonlar bulunur. Çok boyutlu uzayı daha düşük boyutlu bir uzaya dönüştürmesi sebebiyle, t-SNE çıktısında baştaki örnek kümeleri birebir temsil etmez. Bununla beraber, algoritma çıktısının görselleştirilmesiyle, öbeklerin birbirinden ne kadar iyi ayrıldığı daha düşük ve basit bir uzayda gözlemlenebilir.

Şekil 4.2'de 3 öbek için t-SNE (t-Distributed Stochastic Neighbor Embedding) çıktısı görülmektedir. Bu 3 öbekte öne çıkan özellikler Tablo 4.2'de verilmiştir.

Tablo 4.2: Öbeklerdeki Müşteri Özellikleri

Öbek numarası	Öne Çıkan Özellik	Müşteri Sayısı
1	Erken rezervasyoncular & çocukla tatile çıkanlar	143866
2	Tatilden bir hafta önce rezervasyon yaptıranlar & rezervasyon sayısı 3'ten az olanlar	637419
3	Yaşı 30'dan büyük olanlar & rezervasyon başına 1000TL'den fazla harcama yapanlar	218715

Bu sonuca göre, müşterilerin yüzde14.3'ünün konaklama yapmadan 3 ile 6 ay öncesinde rezervasyon yaptıran (erken rezervasyoncu) ve çocuklarıyla tatile gittiği görülmüştür. Müşterilerin yüzde63.7'si konaklamadan bir hafta önce rezervasyon yaptırıp, toplam rezervasyon sayısı 3'ten azdır. Diğer öbekteki müşteriler, toplam müşterilerin yüzde21.8'ini oluşturmakta ve bu gruptaki müşterilerin yaşı 30'dan büyük olup rezervasyon başına yaptıkları harcama 1000TL'den fazladır.

5. SONUÇLAR

Yapılan çalışmada, turizm sektöründe faaliyet gösteren bir firmanın müşteri verileri içerisinden normal dağılıma göre rasgele seçilmiş 1000000 adet örnek veri kullanılmıştır. İlk olarak verideki özniteliklerin dağılımına bakılmış ve yüzde70'in üzerinde boş olan kolonlar atılmıştır. Diğer kolonlardaki boş değerlere, kendi kolon ortalamaları atanmıştır. Ürün tiplerinin kategorik olarak bulunduğu kolon üzerinde one-hot encoding yapılarak veri farklı kolonlara değerleri {0, 1} olacak şekilde dağıtılmıştır.

Sonraki adımda, Z-score ile aykırı değerler atılmış, kalan veriler için kolonlardaki değerlere min-max normalizasyonu uygulanarak değerler [0, 1] arasına alınmıştır.

Elbow method ile en uygun öbek sayısı olan 3 bulunduktan sonra python programla dilinde sklearn kütüphanesi kullanılarak, önişlemeden geçirilmiş veri üzerinde K-means algoritması, sırasıyla şu parametre edeğerleri için çalıştırıştır; n_clusters=3, max_iter=300, init='k-means++' çalıştırılmıştır. Analiz başarısı için silhouette score hesaplanmış ve en uygun öbek sayısının 3 olduğu doğrulanmıştır. Öbeklerin dağılımını görsel olarak ifade edip inceleyebilmek için t-SNE çıktısı oluşturulmuştur.

Öbeklerin kırılımında öne çıkan özniteliklerin anlaşılması için özniteliklerin önem katsıları çıkarılmıştır. Buna göre, verinin öbeklenmesinde rezervasyon sıklığı, sayısı ve sürelerinin analiz sonucunda ön plana çıktığı görülmüştür.

Çalışma sonucunda, müşteriler dahil oldukları öbek numarası ile işaretlenmiştir. Yapılması muhtemel ve belirli gruplara hitap edecek kampanyalarda, bu 3 öbeğin özelliklerinin kullanılması, kampanyaların doğru kişilere ulaşmasını ve ulaşacağı kitlenin büyüklüğünü öngörmek açısından faydalı olacaktır.

KAYNAKÇA

Kitaplar

[2] T. K Das, “A customer classification prediction model based on machine learning techniques” 2015 International Conference on Applied and Theoretical Computing and Communication Technology 2015.

[13] Alguliyev, Rasim & Aliguliyev, Ramiz & Sukhostat, Lyudmila. (2017). Anomaly Detection in Big Data based on Clustering. Statistics, Optimization & Information Computing. 5. 10.19139/soic.v5i4.365.

[14] Patro, S Gopal & Sahu, Kishore Kumar. (2015). Normalization: A Preprocessing Stage. IARJSET. 10.17148/IARJSET.2015.2305.

[8] Umman Tuğba Şimşek Gürsoy, “Telekominikasyon sektöründe müşteri ayrılma Analizi” İstanbul Üniversitesi İşletme Fakültesi dergisi Cilt/Vol 39 Sayı/No:1 2010, 35-49

[9] Emel Kızılkaya Aydoğan, Cevriye Gencer, Sinem Akbulut, “Veri Madenciliği teknikleri ile bir kozmetik markanın ayrılan müşteri analizi ve müşteri bölümlenmesi” Mühendislik ve Fen Bilimleri Dergisi Cilt/vol 26 Sayı/No:1 2008

[7] Buket DOĞAN, Bahar Erol, Ali Buldu, “Sigortacılık Sektöründe Müşteri ilişkileri yönetimi için birliktelik kuralı kullanılması” Marmara Fen Bilimleri Dergisi 2014, 3: 105-114

[12] Mustafa Akça, İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi Yıl :6 Sayı: 11 , 2007

[16] Şenol, Ali & Karacan, Hacer. (2018). Akan Veri Kümeleme Teknikleri Üzerine Bir Derleme. European Journal of Science and Technology. 17-30. 10.31590/ejosat.446019.

[17] Maaten, Laurens van der and Geoffrey E. Hinton. “Visualizing Data using t-SNE.” (2008).

[18] FAMILI A., SHEN W, WEBER R. and E. SIMOUDIS (1997), ‘Data Preprocessing and Intelligent Data Analysis’, Intelligent Data Analysis, 1,USA, pp.3-23.

[20] Ayşe Oğuzlar, Erciyes Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, Sayı : 21 , 2003 ss:67-76

[21] Songül Şekeroğlu, İstanbul Teknik Üniversitesi, Yüksek Lisans Tezi, 2010

[22] Mehmet Ali Alan, Dumlupınar Üniversitesi Sosyal Bilimler Dergisi,Sayı : 33, 2012

Süreli Yayınlar

[1] E.F. Ayetiran, A. V Adeyemo, “A Data Mining-Based Response Model for Target Selection in Direct Marketing” International Journal of Information Technology and Computer Sience, vol 4, no. 1 pp 9-18, March 2012.

[15] Kodinariya, Trupti & Dan Makwana, P.R.. (2013). Review on Determining of Cluster in K-means Clustering. International Journal of Advance Research in Computer Science and Management Studies. 1. 90-95.

Türkiye İSTATIK kurumu Haber Bülteni;
Sayılar: 27612, 27613, 27614, 27615, 27616, 27617, 27618, 27619, 30599

Diğer Yayınlar

[3] Özer M. “User segmentation of online music services using fuzzy clustering” 2001

[4] Mehpare TiMOR / Ayşegül EZERÇE / U. Tuğba GÜRSOY, “Müşteri Profili ve Alışveriş davranışlarını belirlemede kümeleme ve birliktelik kuralları Analizi” 2001

[5] Arş.Gör.Dr. Çağatay TAŞKIN / Yrd.Doç.Dr. Gül Gökay Emel “Veri madenciliğinde kümeleme yaklaşımları ve kohonen ağları ile perakendecilik sektöründe bir uygulama” 2015 vol. 15 No. 3 pp. 395-409.

[6] Sinan Keçeci, Eyüp Erkan Özbek, Mustafa Sertaç Türkel, Remzi Düzağaç “Doğrudan pazarlama amaçlı Hedef Kitle Analizi” 2018

[10] Tuğba Tosun / Atadan Sabih, “Veri Madenciliği Teknikleriyle Kredi Kartlarında Müşteri Kaybetme Analizi” İstanbul Teknik Üniversitesi, Fen Bilimleri Enstitüsü, 2006

[11] Şadi Evren Şeker, bilgisayarkavramlari.sadievrenseker.com , 2008

[19] Erdem Uzun www.uzakrota.com/teknolojinin-turizm-isletmeleri-uzerine-etkisi/ ,2018

ÖZGEÇMİŞ

Adı Soyadı: Murat KAYA

Sürekli Adresi: Osmangazi Mah. Atayolu Cad. Seçkin Teletaş Sit. D blok Kat:4 No: 38 Sancaktepe İstanbul

Doğum Yeri ve Yılı: İzmir / 10.02.1987

Yabancı Dili: İngilizce

İlk Öğretim: Hamdi Helvacıoğlu ilköğretim okulu / 2001

Orta Öğretim: Cumhuriyet Lisesi / 2005

Lisans: Akdeniz Üniversitesi / 2011

Enstitü Adı: Fen Bilimleri Enstitüsü

Program Adı: Bilgi Teknolojileri

Çalışma Hayatı:

Ets Ersoy Turistik Servisleri / 01.08.2019 –

Altar Yüksek Bilişim Teknolojileri / 03.05.2013 – 31.07.2019