T.C.
ISTANBUL ÜNIVERSITESI-CERRAHPAŞA
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

YÜKSEK LİSANS TEZİ

GÜNCELLİK, FREKANS, TUTAR ve SEPET ANALİZİ: HAVAYOLU SADAKAT PROGRAMI UYGULAMA ÖRNEĞİ

RAMAZAN YAŞA

DANIŞMAN
Dr. Öğr. Üyesi TUNCAY ÖZCAN

ENDÜSTRİ MÜHENDİSLİĞİ ANA BİLİM DALI
ENDÜSTRİ MÜHENDİSLİĞİ

İSTANBUL-2019

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

TEZ JÜRİSİ

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

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

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

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

ÖNSÖZ

Tez çalışmasının her aşamasında değerli katkıları ile beni yönlendiren, değerli bilgileriyle yol gösteren, yardımları, destekleri, yorumları ve eleştirileri ile tez çalışmamda sürekli olarak desteğini hissettiğim Dr. Öğr. Üyesi Tuncay ÖZCAN'a teşekkürlerimi sunarım.

Tez çalışmam sürecinde kendisini aksatmama rağmen her daim beni motive eden ve desteklerini sürekli hissettiğim eşim Rabia YAŞA'ya ayrıca teşekkürlerimi sunarım.

Haziran, 2019 **Ramazan YAŞA**

İçindekiler

ŞEKİL LİSTESİ

Sayfa No

TABLO LİSTESİ

Sayfa No

SİMGE VE KISALTMA LİSTESİ

GFT Analizi : Güncellik, Frekans ve Tutar Analizi

GT : Güncellik ve Tutar

G : Güncellik

F : Frekans

T : Tutar

MYD : Müşteri Yaşam Boyu Değeri

BM : Beklenti En Büyüklemesi

MİY : Müşteri İlişkileri Yönetimi

Şehir Kod ve İsimleri :

3'lü Kod	Şehir
ADA	Adana
ANK	Ankara
ASR	Kayseri
AYT	Antalya
BAL	Batman
BJV	Bodrum
DIY	Diyarbakır
DLM	Muğla Dalaman
DNZ	Denizli
ERC	Erzincan
ERZ	Erzurum
EZS	Elazığ
GZT	Gaziantep
HTY	Hatay
IZM	Izmir
KCM	Maraş
KSY	Kars
KYA	Konya
MLX	Malatya
MQM	Mardin

MSR	Muş
MZH	Amasya
NAV	Nevşehir
SZF	Samsun
TZX	Trabzon
VAN	Van
VAS	Sivas
GZP	Antalya
OGU	Ordu

ÖZET

YÜKSEK LİSANS TEZİ

GÜNCELLİK, FREKANS, TUTAR ve SEPET ANALİZİ: HAVAYOLU SADAKAT PROGRAMI UYGULAMA ÖRNEĞİ

Ramazan YAŞA

İstanbul Üniversitesi Cerrahpaşa

Fen Bilimleri Enstitüsü

Endüstri Mühendisliği Anabilim Dalı

Danışman : Dr.Öğr.Üyesi Tuncay ÖZCAN

Havacılık sektöründe artan küresel rekabette mevcut müşterileri elde tutmak ve katlanılacak pazarlama giderlerini doğru müşterilere harcamak önem arz etmektedir. Bu nedenle müşterinin yaşam boyu değerini ölçümlemek hangi müşterilerinin firma açısından daha kârlı olduğunu ortaya çıkarıyor. Bu çalışmamızda ilk aşamada GFT ile müşteri değeri hesaplama yapılmıştır, sonrasında k-ortalamalar, iki aşamalı kümeleme ve SOM (Öz Örgütlemeli Haritalar) yöntemleri ile kümeleme yapılmıştır. Yapılan kümeleme sonuçları Silhoutte indeksi kullanılarak değerlendirilmiş ve en uygun çözüm olarak k-ortalamalar yöntemiyle segmentasyon yapılmıştır. Daha sonrasında her segment için ayrı ayrı Apriori algoritması kullanılarak veri setindeki birliktelik kuralları keşfedilmiştir. Elde edilen sonuçların kampanya öneri sistemine zemin oluşturulması amacıyla kullanılması planlanmıştır.

Anahtar Kelimeler : GFT, k-ortalamalar, Apriori Algoritması, Kümeleme , Segmentasyon

SUMMARY

M.Sc. THESIS

RECENCY, FREQUENCY, MONETARY AND BASKET ANALYSIS: A CASE STUDY AN AIRLINE LOYALTY PROGRAMME

Ramazan YAŞA

İstanbul Üniversity Cerrahpasa

Institute of Graduate Studies in Science and Engineering

Department of Industry Engineering

Supervisor : Dr. Tuncay ÖZCAN

It is important to keep current customers and to spend the marketing expenses to the right customers in the increasing global competition in the civil aviation sector. Therefore, measuring the customer's lifetime value reveals which customers are more profitable for the company. In this study, customer value was calculated with RFM in the first stage and then clustering was performed with k-means, two-step clustering and SOM (Self-Organized Maps) methods. Clustering results were evaluated by using Silhoutte index. It is observed that the optimal solution is the k-means method. After that, the rules of association in the data set were discovered by using the Apriori algorithm for each segment. The results are planned to be used as a basis for the campaign proposal system.

Keywords : RFM, k-means, Apriori Algorithm, Clustering, Segmentation

1. GİRİŞ

Veri; tanım itibariyle, herhangi bir işleme tabi tutulmadan, gözlem veya ölçüm yöntemleri ile ortamdan elde edilen her türlü değerdir (Şeker, 2013). Veri madenciliği ise; genellikle büyük veri setlerinin, veri sahibi için yararlı ve anlaşılır olacak biçimde, umulmadık ilişkiler yakalamak ve özgün bir biçimde özetlemek için analiz edilmesidir (Doğan, 2015).

Birçok sektördeki kurumlar müşteri odaklılığına dayalı yönetim anlayışını benimseyerek değişmektedir. Günümüz dünyasında rekabet artarken, müşterileri daha iyi anlamak, bireysel ihtiyaç ve isteklerine hızlıca cevap vererek her bir müşterinin cüzdan payını artırmak ve daha hedefli promosyonlara odaklanarak maliyetten tasarruf etmek şirketlerin odak noktası olmaktadır (Pritscher ve Feyen, 2001).

2017 yılında 4,07 milyar yolcu taşımacılığı 56 milyon ton kargo taşımacılığı ve 945 milyar dolar[1] lık finansal hacmi olan havayolu sektöründe ise değişen müşteri ihtiyaçlarını karşılama, müşteri değerine göre ayrıcalıklar sunma, sadık müşteriyi elde tutma, yeni müşteriler kazanma, karlılığını en üst düzeye çıkarma ve maliyetleri azaltma hedeflerinini tutturmada sadakat programları önemli bir yer tutmaktadır.

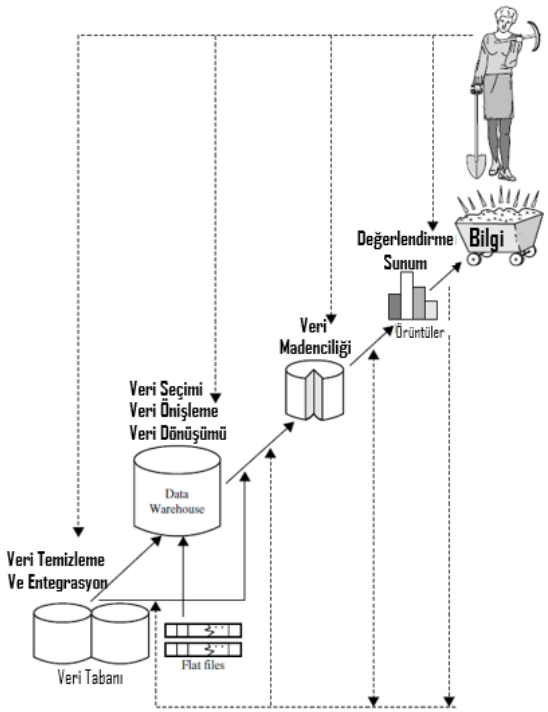
Bu sebeple, bu çalışmanın ikinci bölümünde genel kısımlar başlığı altında veri madenciliği nedir, veri madenciliğinde modeller, kümeleme algoritmaları, birliktelik kuralları, müşteri değeri hesaplama uygulamalarının teorik anlatımına ve literatür araştırmasına yer verilmiştir. Üçüncü bölümde ise uygulama yapılmıştır. Uygulama safhasında; veri seti tanıtılmış ve GFT ile müşteri değeri hesaplama yapılmıştır. Daha sonra, k-ortalamalar, iki aşamalı kümeleme ve SOM (Öz Örgütlemeli Haritalar) yöntemleri ile kümeleme yapılmıştır. Yapılan kümeleme sonuçları Silhoutte indeksi kullanılarak değerlendirilmiş ve en uygun yöntem olarak k-ortalamalar tespit edilmiştir. Dördüncü bölümde ise k-ortalamalar yöntemiyle segmentasyon yapılmıştır. Daha sonrasında her segment için ayrı ayrı Apriori algoritması kullanılarak veri setindeki birliktelik kuralları keşfedilmiştir. Elde edilen sonuçların kampanya öneri sistemine zemin oluşturulması amacıyla kullanılması planlanmıştır. Son bölümde ise çalışmadan elde edilen sonuçlar yorumlanarak, çalışmanın zenginleştirilmesi anlamında neler yapılabileceğine yer verilmiştir.

1 -(ICAO, Presentation of 2017 Air Transport Statistical Results)

2. GENEL KISIMLAR

2.1 VERİ MADENCİLİĞİ NEDİR?

Veri, ham gerçeklerin işlenmemiş halidir. Veriler işlenerek ve düzenlenerek bilgiye dönüşür. Veri işleme, temel bir veriyi organize edebilmek ve bu verilerden örüntüler, kompleks tahminler veya istatistiksel modellerden faydalanılarak sonuçlar ortaya çıkarmaktır. Gerçek anlamda bilgi kaynak gerektirmektedir (Özdemir ve diğ.,2010). Bu noktada bilgi keşfi süreci Şekil 2.1'de sunulmaktadır.



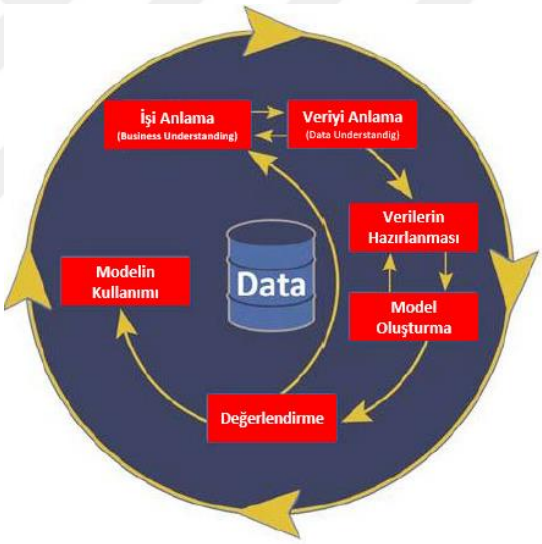
Şekil 2.1: Bilgi Keşfi Süreci (Han ve Kamber, 2012)

Veri madenciliği, ham veriyi istatiksel veya başka analitik teknikler kullanarak içgörü çıkarma, kararlar üretme, iş önerileri ortaya çıkarma ve eğilimler bulmak amacıyla analiz etmektir (Chui ve Tavella, 2008). Kısacası veri madenciliği; veriden bilgi elde etme anlamına gelir (Han ve Kamber, 2012).

Veri madenciliği; büyük miktarda veri içinden, gelecekle ilgili tahmin yapmamızı sağlayacak bağıntı ve kuralların bilgisayar programları kullanılarak aranmasıdır. Veri analizi yapılarak, bir mal için bir sonraki ayın satış tahminleri yapılabilir, müşteriler satın aldıkları

mallara bağlı olarak gruplandırılabilir, yeni bir ürün için potansiyel müşteriler belirlenebilir, müşterilerin zaman içindeki hareketleri incelenerek onların davranışları ile ilgili tahminler yapılabilir (Gündüz, 2018). Veri madenciliği; pazar stratejilerine karar vermede yardımcı olabileceğini ve insanların pazardaki ürünlerlerle ilişkisinin tanımlamasını, çapraz satış yapıp, sıralı satış yapmasını vb. yardım edebileceğini göstermiştir (Liang ve Chen, 2017). Veri madenciliği, veritabanlarında zaten mevcut olan verileri analiz ederek sorunları çözmekle ilgilidir (Witten ve diğ., 2017).

Veri madenciliği projelerinde bir standart yöntem izlenmesi amacıyla günün önde gelen veri depolama ve işleme sağlayıcılarından ikisi olan SPSS ve Teradata, ve ilgili firmaların ürünlerini kullanan anlaşmalı ilk üç şirket (Daimler, NCR ve OHRA), 1996 yılında özel bir ekip oluşturmuştur. Bu ekip günümüzde hala kullanılan veri madenciliği için Çapraz Endüstri Standart Süreç Modelini (CRISP-DM) ortaya koymuştur (Vorhies, 2016).



Şekil 2.2: CRISP-DM Metodolojisi (McCormick ve diğ., 2017)

Veri Madenciliği için Çapraz Endüstri Standart Süreç Modeli (CRISP-DM) 6 aşamadan oluşmaktadır (Larose, 2015):

İşi anlama: Öncelikle, bir bütün olarak iş birimi açısından proje hedeflerini ve gereksinimlerini açıkça belirtilir. Ardından, bu hedefler ve kısıtlamalar bir veri madenciliği problemi tanımının formülasyonuna çevrilir. Son olarak, bu hedeflere ulaşmak için bir ön strateji hazırlanır. İşi anlama aşaması bu şekilde tamamlanır.

Veriyi Anlama: Veriyi anlama aşaması veriyi toplama ile başlar, toplanan veriyi tanımak için veri keşfi yapılır, verilerin kalitesi değerlendirilir ve son olarak da istenirse hangi alt kümelerdeki veriler hangi iş hedeflerine hizmet edebilir bunlar çıkartılır.

Veriyi Hazırlama: Bu aşama emeğin fazla şekilde harcandığı aşamadır. Kirli veriler temizlenir, analiz edilecek olan değişkenler seçilir, gerekiyorsa değişken dönüşümleri gerçekleştirilir ve verilere modele girecek hale getirilir.

Modelleme Aşaması: Uygun model seçilir ve uygulanır, sonuçların optimizasyonu için modellerin kısıtları ve parametreleri belirlenir.

Değerlendirme: Modelleme aşamasında bir veya daha fazla model sunulmuş olabilir. Bu modeller sahada kullanılmak üzere dağıtılmadan önce kalite ve etkinlik açısından değerlendirilir. Modelin ilk aşamada belirlenen hedeflere ulaşıp ulaşmadığı bu aşamada ölçümlenir.

Modelin Kullanımı: Karar verilen modelin çıktıları ilgili departman ve kişilerin kullanımı amacıyla ilgili sistemlere entegre edilir.

2.2 VERİ MADENCİLİĞİNDE MODELLER

En yaygın kullanılan veri madenciliği yöntemlerini altı başlık altında sınıflandırabiliriz (Larose, 2015):

1. Açıklayıcı algoritmalar
2. Kestirimci algoritmalar
3. Tahminleyici algoritmalar
4. Sınıflandırıcı algoritmalar
5. Kümeleme algoritmaları
6. Birliktelik algoritmaları

2.2.1 Açıklayıcı Algoritmalar

Açıklayıcılık veri madenciliği çalışmalarının büyük bir parçasıdır. Açıklayıcı modeller, neler olduğunu gözlemlemek adına raporlara odaklanan modellerdir. Açıklayıcı modeller, algoritmanın kullanıcı yönlendirmesi olmadan ilişkileri tanımladığı denetimsiz öğrenmeye bir örnektir. Bazı hedef değişkenleri tahmin etmezler, fakat veri yapısına, ilişkilere ve bağlantıya ilişkin ipuçları verirler (Olson, 2017).

2.2.2 Kestirimci Algoritmalar

Veri kümesinde parametre tahminleyici modellerde yani kısaca kestirimci algoritmalarda bir sayısal ve/veya kategorik tahmin değişkenleri kümesini kullanarak bir sayısal hedef değişkeninin değerini tahmin ettiğimiz modellerdir. Modeller, tahmin değişkenlerinin yanı sıra, hedef değişkenin değerininin gerçekleşen değerleri de kullanılarak oluşturulur. Daha sonra, yeni gözlemler için, tahmin değişkenlerinin değerlerine dayanarak hedef değişkenin değerinin tahminleri yapılır (Larose, 2015).

2.2.3 Tahminleyici Algoritmalar

Tahmin edici algoritmalarda, sonuçları bilinen verilerden hareket edilerek bir model geliştirilmesi ve kurulan bu modelden yararlanılarak sonuçları bilinmeyen veri kümeleri için sonucun tahmin edilmesi amaçlanmaktadır. Örneğin; pasif ve aktif müşterilerin özelliklerinden oluşan bir veri kümesine sahip isek, bağımlı değişkenimiz müşterilerimizin pasifleşme durumu, bağımsız değişkenlerimiz ise bu müşterilerin daha önce gösterdikleri özellikler olacak, kurulacak model ile sisteme katılan her bir müşteri için firmayı terk edip etmeyeceği tahmin edilebilecektir (Çelik, 2009).

2.2.4 Sınıflandırıcı Algoritmalar

Sınıflandırma, veri sınıflarını veya kavramlarını tanımlayan ve ayırt eden bir model bulma sürecidir. Model, bir dizi eğitim verisinin (yani sınıf etiketlerinin bilindiği veri nesneleri) analizine dayanarak türetilir. Model, sınıf etiketinin bilinmediği nesnelerin sınıf etiketini tahmin etmek için kullanılır (Han ve Kamber, 2012). Sınıflandırma görevi, belirli bir etiketlenmemiş nokta için etiketi veya sınıfı tahmin etmektir. Kategorik bir nitelik olarak sınıflandırma yapılır (Zaki ve diğ., 2014).

2.2.5 Kümeleme Algoritmaları

Kümeleme, bir gruptaki örneklerin birbirine benzer ve diğer gruplardaki örneklerden farklı olacak şekilde verileri gruplara ayırma işlemidir (Tran ve diğ., 2018). Bir küme içindeki noktalar, diğer kümelerdeki noktalardan çok birbirine benzer olmalıdır (Nisbet ve diğ., 2017). Farklılıklar ve benzerlikler, nesneleri tanımlayan özellik değerlerine dayanılarak değerlendirilir ve genellikle mesafe ölçümlerini içerir (Han ve Kamber, 2012).

2.2.6 Birliktelik Kuralları

Pazar sepet analizi olarak da bilinen birliktelik kuralları yöntemleri, özellikler arasındaki ilişkileri ortaya çıkarmaya çalışır; yani, iki veya daha fazla özellik arasındaki ilişkiyi ölçmek için kuralları ortaya çıkarmayı esas almaktadır (Larose, 2015). Birliktelik kurallarını bir sınıflandırma ağacının eğer-sonra kurallarının oluşturularak adalandırılması olarak düşünebiliriz. Temel amaç, alışveriş sepetlerinde yaygın olarak bulunan ürünler arasındaki ilişkilerin belirlenmesidir (King, 2015).

2.3 KÜMELEME ALGORİTMALARI

Kümeleme, en popüler denetimsiz öğrenme tekniklerinden biridir. Bu teknik, verileri analiz etmek ve bu veriler içindeki kümeleri bulmak için kullanılır. Bu kümeleri bulmak için Öklid uzaklığı gibi bir tür benzerlik ölçüsü kullanılır. Bu benzerlik ölçüsü bir kümenin sıkılığını tahmin edebilir. Kümelemenin, verilerimizi, birbirine benzer alt gruplara ayırma süreci olduğunu söylenebilir (Joshi, 2017).

Literatürde birçok kümeleme algoritması vardır. Kümeleme yöntemlerinin net bir şekilde sınıflandırılmasını sağlamak zordur çünkü bu kategoriler üst üste gelebilir, böylece bir yöntem birkaç kategoriden özelliklere sahip olabilir. Genel olarak temel kümeleme yöntemlerini aşağıdaki gibi kategorilere ayırabiliriz (Han ve Kamber, 2012).

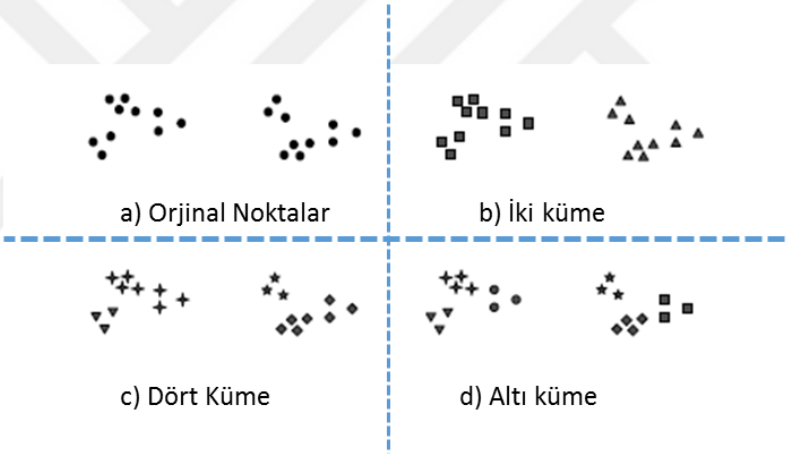
- Bölümlenme Yöntemleri
- Hiyerarşik Yöntemler
- Yoğunluk Tabanlı Yöntemler
- Izgara Tabanlı Yöntemler

İleri kümeleme yöntemlerini ise aşağıdaki gibi gruplayabiliriz.

- Olasılıklı Model Tabanlı Yöntemler
- Yüksek Boyutlu Verilerin Kümelenmesi
- Grafik tabanlı Kümeleme
- Kısıtlarla Kümeleme

2.3.1 Bölümleme Yöntemleri

Bölümleme yöntemleri ile kümeleme, veri nesneleri kümesinin örtüşmeyen alt kümelere bölünmesidir, öyle ki her veri nesnesi tam olarak bir alt kümededir. Şekil 2.3'de yer alan (b-d) küme oluşumu örnekleri bölümleme yöntemlerine örnektir (Tan ve diğ., 2013).



Şekil 2.3: Aynı noktaların farklı kümelenme çeşitleri

2.3.1.1 K-ortalamalar Algoritması

K-ortalamalar en çok bilinen klasik kümeleme tekniğidir. İlk önce, kaç küme arandığını belirlenir: Bu, k parametresidir. Daha sonra, küme merkezleri olarak rastgele k nokta seçilir. Tüm örnekler, normal Öklid uzaklık metriğine göre en yakın küme merkezlerine atanır. Atamalar yapıldıktan sonra yeni kümenin merkezi hesaplanır. Yeni küme merkezlerine göre noktaların oklid uzaklığına göre ataması yapılır. Yeni küme merkezi ve atamaların tekrarı tüm noktaların sabit bir kümede kalmasıyla sonlandırılır. (Witten ve diğ., 2017)

Bir kümeleme yapıldığında, $C = \{C_1, C_2, \cdots, C_k\}$ kalitesini veya iyiliğini değerlendiren bir puanlama fonksiyonuna ihtiyacımız vardır. Bu puanlama fonksiyonu kare hataların toplamı olarak tanımlanır.

$$\text{SSE (C)} = \sum_{i=1}^{k} \sum_{x_j \epsilon C_i} ||X_j - \mu_i||^2 \quad (2.1)$$

Hedef bu ortaya çıkan puanlama fonksiyonunu minimuma indirmektir.

$$C^* = \arg \text{Min}_c \{ \text{SSE (C)} \} \quad (2.2)$$

K-ortalamalar algoritmasının sözde kodu aşağıdaki gibidir (Zaki ve diğ., 2014).

K-ortalamalar (D, k, ǫ):

1. t = 0
2. Rastgele k tane küme merkezi belirle $\mu_1^t, \mu_2^t, \mu_3^t \ldots\ldots, \mu_k^t \ \epsilon R^d$
3. Tekrarla
4. $t \leftarrow t + 1$

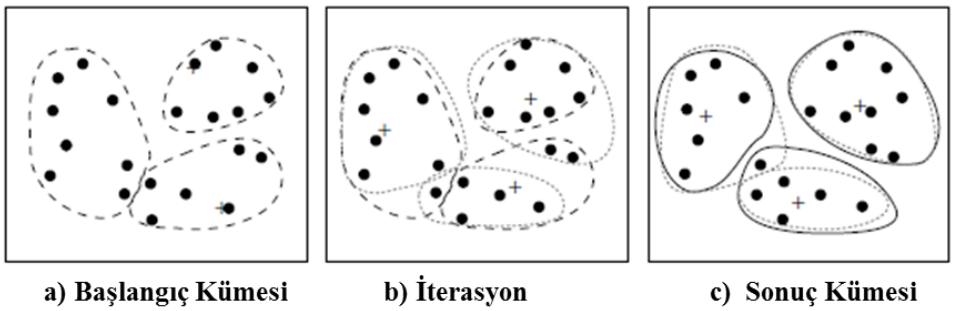
// Küme atama adımı

5. her $x_j \ \epsilon \ D$ için yap
6. $j^* \leftarrow \arg \min_i \{ \ ||X_J - \mu_i||^2 \ \}$ // X_j yi en yakın küme merkezine ata
7. $C_j^* \leftarrow C_j^* \cup \{ X_j \}$

// Küme merkezini güncelle

8. i = 1 den k ya kadar yap
9. $\mu_i^t \leftarrow \frac{1}{|C_i|} \sum_{X_j \ \epsilon \ C_i} X_j$
10. Dur $\sum_{i=1}^{k} ||\mu_i^t - \mu_i^{t-1}|| \leq \epsilon$

Temsili örnek ise aşağıdaki gibidir (Han ve Kamber, 2012) .



Şekil 2.4: K-ortalamalar Kümeleme Örneği

K-ortalamalar algoritmasının avantajlarını sıralayacak olursak; uygulanabilirliğinin kolay olması ve büyük veri kümelerinde hızlı çalışabilmesidir. Büyük veri setlerini işlerken nispeten ölçeklenebilir ve verimlidir. Karmaşıklığı diğer kümeleme yöntemlerine göre azdır. K-ortalamalar algoritması yoğun kitlelerden (bulutlanmış) oluşan veritabanlarında daha iyi uygulanır ve daha verimli sonuçlar ortaya çıkarmaktadır. K-ortalamalar algoritmasının dezavantajlarını sıralayacak olursak; kullanıcının küme sayısını belirleme zorunluluğudur. Kategorik ve metinsel içeren veritabanlarının üzerinde uygulanması gerçekleştirilemez. Gürültülü ve sıra dışı verilere duyarlı bir metottur. Kümelemede kullanılan özelliklerin hangisinin kümelemeye daha fazla katkısı olduğu bilinmemektedir. Kümelemeden sonra bazı kümelerin boş küme olması ihtimali vardır bu yüzden böyle bir durumda bazı optimizasyon işlemlerinin yapılması gerekir (Al-Zand, 2013).

2.3.1.2 K-medoidler Algoritması

K-medoidler algoritması Kaufman ve Rousseeuw (1987) tarafından geliştirilmiştir. K-ortalamalar algoritmasından farklı olarak k-medoidler algoritmasında, küme merkezleri, kümedeki nesnelerin ortalama değerleriyle hesaplanmayıp, küme merkezine en yakın nesne (medoid) bulunarak belirlenmektedir. Böylece ortalamayı çok fazla değiştiren aykırı değerlere karşı k-ortalamalar yöntemine göre daha az hassastır (Nacaroğlu, 2010). K-ortalamalardan diğer bir faklılığı ise; k-ortalamalar yalnızca sayısal gözlemler için kullanılırken, kategorik değişkenler ve diğer farklı değişkenler için k-medoidler kümeleme yöntemi kullanılır (King, 2015). K-medoidler algoritması küme merkezlerinin gerçek veri noktaları olması, önemli olduğunda kullanılmalıdır (Berry ve Linoff, 2011).

K-medoidler kümeleme metodunun temel stratejisi ilk olarak n adet nesnede, merkezi temsili bir medoid olan k adet küme bulmaktır. Geriye kalan nesneler, kendilerine en yakın olan medoide göre k adet kümeye yerleşirler. Bu bölünmelerin ardından kümenin ortasına en yakın olan nesneyi bulmak için medoid, medoid olmayan her nesne ile yer değiştirir. Bu işlem en verimli medoid bulunana kadar devam eder.

K-medoidler algoritması aşağıdaki adımlardan oluşur (Şekerler, 2008).

1. k tane nesne seç (medoid),
2. Nesneleri onlara en yakın medoidlere at,
3. Bu nesne bir medoidmiş gibi ele alınıp toplam performansı hesapla,
4. Eğer daha performanslı sonuç elde ediliyorsa diğeri yerine yeni medoid olarak bu nesneyi ata (yer değiştirilir),
5. Bir değişiklik olmayana dek tekrarla.

K-Medoid algoritmasının dezavantajlarını sıralayacak olursak; veriye uygun k sayısının belirlenmesi için birden fazla denemenin yapılması gerekir. Farklı büyüklüklerde kümelerin tespitinde doğru olmama ihtimali vardır. Karmaşıklık probleminin hassasiyetinin nedeni ile sadece küçük veri setlerinde uygulanabilir durumdadır.

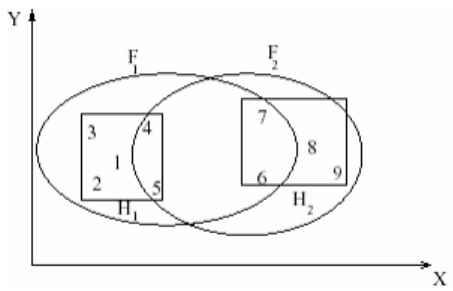
PAM algoritması tanıtılan ilk k-medoid algoritmalarındandır. N tane nesnenin k tane kümeye bölümlenmesini sağlar. K temsili nesnelerin ilk rastgele seçiminden sonra, algoritma art arda küme temsilcileri için daha iyi bir seçim yapmaya çalışır. Mümkün olan her nesne çifti analiz edilir; burada her bir çiftin içindeki bir nesne temsili bir nesne olarak kabul edilir. Elde edilen kümelemenin kalitesi bu kombinasyonların her biri için hesaplanır. Bir nesne, O_j en büyük azalmaya neden olan nesne ile değiştirilir. Bir yineleme içindeki her küme için en iyi nesneler kümesi, bir sonraki yineleme için temsili nesneleri oluşturur. Her yinelemenin karşılığında $O(k(n-k)^2)$ tane kombinasyon ortaya çıkar. Bu sebeple N ve K nın büyük değerleri için uygulaması zordur (Han ve Kamber, 2012).

2.3.1.3 Bulanık C-ortalamalar Algoritması

Bulanık c-ortalamalar (FCM) algoritması, bulanık bölünmeli kümeleme tekniklerinden en iyi bilinen ve yaygın kullanılan yöntemdir. Bulanık c-ortalamalar algoritması 1973 yılında Dunn tarafından ortaya atılmış ve 1981'de Bezdek tarafından geliştirilmiştir Bulanık c-ortalamalar algoritması da amaç fonksiyonu temelli bir metottur (Işık ve Çamurcu, 2007).

Kümeleme analizi kümelere olan üyelik kayıtlarına göre katı, bulanık ve olasılıklı olmak üzere 3 farklı açıdan incelenebilinir. Katı kümelemede kümeye üyelik kaydı ikili değişkendir. Yani gözlemler ya bir kümeye üyedirler ya da değildirler. Bulanık kümelemede, 0 ile 1 arasında değişen bir üyelik kaydı ve verilerin aynı anda birden fazla kümeye üyeliği söz konusu olmaktadır. Olasılıklı kümeleme analizinde ise üyelik yine ikili değişkendir ama bu yöntemde kümelere atanmada bir olasılık dağılımı vardır. Bulanık kümeleme yönteminde üyelik olasılıklarının toplamı daima 1 olmak üzere, bir birimin bir kümede olma olasılığı tüm olası kümeler arasında 0 ile 1 arasında değişir. Bu durum, veri noktalarının aynı anda birden fazla kümeye ait olabileceğini göstermektedir. Kümeye üyelikler bulanık olduğu için veri noktasının hangi kümeye ait olduğunu gösteren tek bir değer yoktur, onun yerine bir değerler kümesi vardır. Birim bir bölümüyle bir kümeye ait iken, bir bölümüyle kümenin dışındadır ve en yüksek olasılığa sahip olduğu kümeye atanır. Dolayısıyla klasik katı kümeleme yöntemlerini, bulanık kümelemenin bir alt durumu olarak ele almak mümkündür (Yılancı, 2010).

Örnek olarak, Şekil 2.5'de H1 ve H2 dikdörtgenleri k-ortalamalar tarafından oluşturulmuş klasik kümeleri, F1 ve F2 elipsleri ise bulanık c-ortalamalar ile oluşturulmuş bulanık kümeleri temsil etmektedir. Nesnelerin bulanık kümelerdeki üyelik değerleri ise Tablo 2.1 'te görülmektedir (Işık, 2006).



Şekil 2.5: Bulanık c-ortalamalar algoritması ile kümeleme (Işık, 2006)

Tablo 2.1: Nesnelerin bulanık kümelerdeki üyelik değerleri (Işık,2006)

Nesne	F1	F2
1	1	0
2	1	0
3	1	0
4	0.8	0.2
5	0.7	0.3
6	0.2	0.8
7	0.2	0.8
8	0	1
9	0	1

Bulanık c-ortalamalar algoritması amaç fonksiyonu temelli bir metodtur. Algoritma, en küçük kareler yönteminin genellemesi olan aşağıdaki amaç fonksiyonunu en küçüklemek için çalışır.

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{c} u_{ij}^{m} \left\| x_i - c_j \right\|^2 \qquad (2.3)$$

X= Veritananında bulunan nesne

m= Üyelik değerinin ağırlığı

c_j=Küme merkezinin prototipi

İlk adımda, "U" üyelik matrisi rastgele atanarak algoritma başlatılır. İkinci adımda ise merkez vektörleri hesaplanır. Hesaplanan küme merkezlerine göre U matrisi yeniden hesaplanır. Eski U matrisi ile yeni U matrisi karşılaştırılır ve belirlenen farkdan küçük olana kadar işlemler devam eder (Chen ve diğ., 2009).

2.3.1.4 CLARA ve CLARANS Algoritmaları

PAM algoritmasını büyük veri setlerini kümelemek için pratik değildir. Bu nedenle özellikle büyük veri setlerinde kullanılmak üzere bir yöntem inşa edilmiştir. k-medoid

yaklaşımını temel alan bu yöntem CLARA (Clustering for Large Applications) olarak adlandırılmıştır. CLARA ile bir dizi nesnenin kümelenmesi iki adımda gerçekleşir. İlk olarak nesne kümesinden bir örneklem kümesi alınır ve k-medoid yöntemi kullanılarak alt kümelerine ayrılır. Daha sonra örnekleme ait olmayan her nesne en yakın k temsili kümelerine atanır. Tüm veri kümesinin kümelenmesi sağlanır. Bu kümelemenin kalitesi veri kümesinin her bir nesnesi ve temsili nesnesi arasındaki ortalama mesafenin hesaplanmasıyla elde edilir. Beş örnek çekilip kümelendikten sonra, en düşük ortalama mesafenin elde edildiği örnek seçilir (Kaufman ve Rousseeuw, 2009). CLARA algoritması için özetlenmiş sözde kod aşağıdaki gibidir (Makhabel, 2015).

CLARA (X,d,k)

bestDissim ← ∞

for t ← 1 $to\ S$

do X' ← $RANDOM - SUBSET\ (X, s)$

D ← $BUILD - DISSIM - MATRIX\ (X', d)$

(C',M) ← $PAM\ (X', D, k)$

C ← $ASSIGN - MEDOIDS\ (X, M, D)$

Dissim ← $TOTAL - DISSIM\ (C, M, D)$

If dissim < bestDissim

Then bestDissim ← $dissim$

C_{best} ← C

M_{best} ← M

Return (C_{best} , M_{best})

CLARANS algoritması ise CLARA'nın sonuçlarını örneklem seçimine bağlı olmaktan kurtarmak amacı ile 1994 yılında bilim dünyasına sunulmuştur. CLARANS örneklem

seçimindeki ön yargıyı gidermek için sabit bir örneklem yerine her aşamada değişen örneklem kavramını ortaya atmıştır. Rastgele seçilen noktalar çevresi dikkate alınarak örneklem oluşturulur (Dinçer, 2006).

CLARANS'ın iki parametresi vardır: İncelenen maksimum komşu sayısı (en yakın) ve elde edilen yerel minimum sayısı. Komşunun değeri ne kadar yüksek olursa, CLARANS PAM'a o kadar yakındır ve herbiri yerel bir minimum araması yapmaktadır (NG ve Han, 2002).

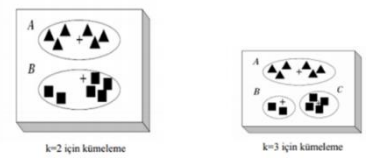
2.3.2 Hiyerarşik Yöntemler

Gruplayıcı ve bölümlemeli olmak üzere iki hiyerarşik yöntem mevcuttur. Gruplayıcı hiyerarşik yöntemde her birim veya her gözlem başlangıçta bir küme olarak kabul edilir. Daha sonra, en yakın iki küme (veya gözlem) yeni bir kümede toplanarak birleştirilir. Böylece her adımda küme sayısı bir azaltılır. Bu süreç dendogram veya ağaç grafiği adı verilen şekille gösterilebilir. Bölümlemeli hiyerarşik yöntemde ise süreç gruplayıcı hiyerarşik yöntemin tam tersidir. Bu yöntemde tüm gözlemlerden oluşan büyük bir küme ile başlanır. Benzer olmayan gözlemler ayıklanarak daha küçük kümeler oluşturulur. Her gözlem tek başına küme oluşturana kadar işleme devam edilir (Çelik Ş., 2013).

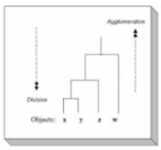
Hiyerarşik kümeleme tekniğinde verilerin normal dağılımlı olması gerektiği varsayımı olmakla birlikte, uygulamalarda uzaklık değerlerinin normalliği yeterli görülmektedir (Doğan ve Başokçu, 2010).

Bölümlemeli ve gruplayıcı kümeleme yöntemlerinin bir örneğini aşağıdaki gibi gösterebiliriz (Han ve Kamber, 2012).

Bölümlemeli ve gruplayıcı kümeleme yöntemleri örneği ;



Şekil 2.6 : Bölümlemeli Yöntem



Şekil 2.7 : Gruplayıcı (Dendogram) Yöntem

Bir gruplayıcı veya bölümlemeli bir metot kullanırken, çekirdek bir ihtiyaç, iki küme arasındaki mesafeyi ölçmektir. Kümeler arasındaki mesafe için yaygın olarak kullanılan dört ölçü, aşağıdaki gibidir (Han ve Kamber, 2012).

Minimum Uzaklık : $\text{dist}_{\text{min}}(C_i, C_j) = min_{p \in C_i, p' \in C_j} \{|p - p'|\}$ (2.4)

Maksimum Uzaklık : $\text{dist}_{\text{max}}(C_i, C_j) = max_{p \in C_i, p' \in C_j} \{|p - p'|\}$ (2.5)

Ortalamaların Uzaklığı : $\text{dist}_{\text{mean}}(C_i, C_j) = |m_i - m_j|$ (2.6)

Ortalama Uzaklık : $\text{dist}_{\text{avg}}(C_i, C_j) = \frac{1}{n_i n_j} \sum_{p \in C_i, p' \in C_j} |p - p'|$ (2.7)

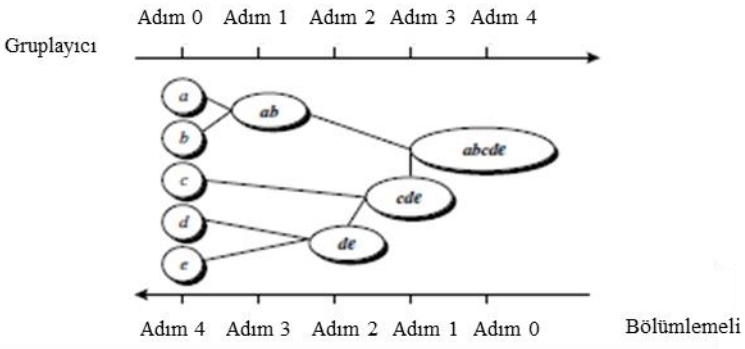
$|p - p'|$: iki nokta arasındaki uzaklık

m_i : Küme Ortalaması

n_i : c_i de bulunan eleman sayısı

c_i : Bağlantılar arası ölçümler

Bölümlemeli ve gruplayıcı yöntemleri şekilsel olarak gösterecek olursak aşağıdaki gibi ifade edebiliriz (Han ve Kamber, 2012).



Şekil 2.8: Bölümlemeli ve Gruplayıcı Yöntemler (Han ve Kamber, 2012)

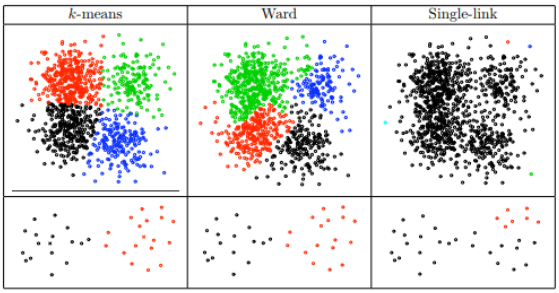
2.3.2.1 Gruplayıcı Hiyerarşik Yöntemler

Gruplayıcı hiyerarşik bağlantı yöntemleri arasında; tek bağlantı, tam bağlantı, ortalama bağlantı, ward yöntemi ve merkezi yöntem yer almaktadır (Kangallı ve diğ., 2014).

Tek Bağlantı Yöntemi: Tek bağlantı yöntemi ile kümeleme, iki küme arasındaki mesafeyi, kümelerde yer alan üyeler arasında birbirine en yakın mesafe olarak tanımlar.

$$d\,(A,B) \equiv min\,||\vec{X} - \vec{Y}\,|| \qquad (\vec{X} \in A, \vec{Y} \in B) \qquad (2.8)$$

Buna "tek bağlantı" denir, çünkü kümelerin tek bir yakın bağlantı noktası varsa bile yakın olduklarını söyler. Bu oldukça karmaşık kümeler için kullanılabilir. Bu algoritma sadece uzaklık ister ve kümeler arası denge ile ilgilenmez. Bu sebeple tek başına bir eleman yalnız başına bir küme oluşturabilir. Bu durum, aşağıdaki şekilde gösterildiği gibi bazen sorunlara yol açabilir (Shalizi, 2009).



Şekil 2.9: Tek Bağlantı Yöntemi Sorunları (Shalizi, 2009)

Tam Bağlantı Yöntemi: Bir önceki metot olan tek bağlantı yönteminin neredeyse aynısıdır. Tek fark kümeler arasındaki uzaklık hesaplanırken gerçekleşir. İki küme arasındaki mesafeyi, kümelerde yer alan üyeler arasında birbirine en uzak mesafe olarak tanımlar. Diğer bir adı olan "en uzak komşu" sözcüğü buradan gelmektedir (Akat, 2007).

$$d\,(A,B) \equiv max\,||\vec{X} - \vec{Y}\,|| \qquad (\vec{X} \in A, \vec{Y} \in B) \qquad (2.9)$$

Ortalama Bağlantı Yöntemi: Tek ve tam bağlantıdan çok farklı bir metot değildir. Algoritmik yapı aynıdır sadece benzerlik matrisinin değerleri farklı bir formülle hesaplanır ki bu formül küme çiftlerinin elemanlarının ortalama uzaklığıdır. Tek bağlantı metotunda iki küme elemanları arasındaki en kısa uzaklık benzerlik değerini verirken, tam bağlantı tekniğinde benzerlik değeri iki küme elemanlarının uzaklıklarının en büyük değeridir, ortalama bağlantı tekniğinde ise benzerlik değeri iki küme elemanları arasındaki uzaklıkların ortalamasıdır (Akat, 2007).

$$d\,(A,B)_{\,w} = \frac{\sum_x \sum_y d_{xy}}{N_{(AB)}\, N_W} \qquad (2.10)$$

formülü ile hesaplaması yapılır.

Ward Yöntemi : Ward'ın yöntemi, her yinelemeli adımda bir hata ölçümü kullanarak, mesafe ölçütlerini veya ilişkilendirme ölçütlerini kullanmak yerine varyans analizini içerir. Bu yöntem yapraklarda başlar ve ilgili dendrogramın gövdesine doğru ilerler. Bu yöntem, sayısal değişkenler için en uygun olanıdır ve ikili değişkenler için uygun değildir (King, 2015).

İki küme arasındaki mesafe, iki küme birleştirildiğinde, hata kareleri toplamındaki (HKT) artış olarak tanımlanır (Zaki, 2014).

Hata kareleri toplamı (HKT) ise aşağıdaki gibi tanımlanır.

$$HKT = \sum_i \sum_j \sum_k | x_{ijk} - \bar{x}_{i.k}|^2 \qquad (2.11)$$

X_{ijk}; k. kümenin j. bireyinin i. gözlem değerini belirtir.

$X_{i.k}$; k kümesindeki j değişkeninin ortalamasıdır.

Toplam Hata Kareleri toplamı (HKT) ise aşağıdaki gibi tanımlanır.

$$HKT = \sum_i \sum_j \sum_k | x_{ijk} - \bar{x}_{..k}|^2 \qquad (2.12)$$

Bu değerler hesaplandıktan sonra bireylerin kümelerde birleştirilmesi aşamalı olarak aşağıdaki şekilde yapılır (King, 2015).

1- İlk başta $HKT_k = 0$ olacak şekilde her birim bir küme kabul edilir.

2- İkinci aşamada HKT'ında en küçük artışı sağlayan (u) ve (v) birleştirilerek (uv) kümesi oluşturulur. HKT'daki bu artış aşağıdaki gibi hesaplanır ve bu suretle n birim (n-1) kümeye ayrılmış olur.

$$\Delta\, HKT_{uv} = HKT_{uv} - HKT_u - HKT_v \qquad (2.13)$$

3- Küme sayısı k=1 oluncaya kadar 2 adım tekrarlanarak tüm bireylerin aşamalı olarak birbirine bağlanmaları sağlanır. Böylece her aşamada HKT'ında ki oluşan minimum artış,

birleştirilen kümelerin (küme ortalamaları) merkez noktaları arasındaki Öklid uzaklığının karesi ile orantılı gerçekleşmiş olur (Kayaalp ve diğ, 2000).

Merkez (Centroid) Yöntem: Birinci kümenin (k_1) merkezi (p elemanlı ortalama vektör) ve ikinci kümenin (k_2) merkezi arasındaki uzaklık hesaplanır (Birant, 2019).

$$U(k_1,k_2) = U\left(\left(\frac{1}{|k_1|}\sum_{x \in k_1}\vec{x}\right),\left(\frac{1}{|k_2|}\sum_{x \in k_2}\vec{x}\right)\right) \quad (2.14)$$

İki küme arasındaki farklılıklar merkezi yöntemler tanımlanacağı zaman öklid yerine öklidin karesi kullanılmalıdır (King, 2015).

İki Aşamalı Kümeleme Yöntemi: İki aşamalı kümeleme (TwoStep Clustering), öncelikle büyük veri kümelerini analiz etmek için tasarlanmış bir algoritmadır. Prosedür olarak gruplayıcı hiyerarşik kümeleme yöntemini kullanır. Klasik küme analizi yöntemleriyle karşılaştırıldığında iki aşamalı kümeleme algoritması, hem sürekli hem de kategorik nitelikleri kümeleyebilir. Ayrıca, bu yöntem optimum küme sayısını otomatik olarak belirler (Schiopu, 2010).

Süreç iki ana adımdan oluşur:

İlk adım, gözlemlerin küçük alt kümelere bölümlenerek ilk kümelenmesinin yapıldığı aşamadır. İkinci adım ise, gözlem değerinin hali hazırda oluşturulmuş kümeye mi yoksa yeni bir kümeye mi katıldığına dair karar verme aşamasıdır, uzaklık kriterleri baz alınarak karar verilir. Alt kümelerin sayısı gözlem sayısından önemli ölçüde daha küçük olduğundan, geleneksel gruplama yöntemlerinin kullanımı kolaydır. Daha fazla alt küme varsa, yöntem daha kesindir (Trpkova ve Tevdovski, 2009).

i ve *j* kümeleri arasındaki mesafe aşağıdaki gibi tanımlanmıştır.

$$d(i,j) = \xi_i + \xi_j + \xi_{j<i,j>} \quad (2.15)$$

$$\xi_i = -N_v \cdot \left(\left(\sum_{k=1}^{K^A}\frac{1}{2}\cdot log(\hat{\sigma}_k^2 + \hat{\sigma}_{v.k}^2)\right) + \left(\sum_{k=1}^{K^B}\hat{E}_{v.k}\right)\right) \quad (2.16)$$

ve

$$\hat{E}_{v.k} = -\sum_{l=1}^{L_k}\left(\frac{N_{v.k.l}}{N_v}.log\left(\frac{N_{v.k.l}}{N_v}\right)\right) \quad (2.17)$$

$d(i, j)$, i ve j kümeleri arasındaki mesafeyi ifade etmektedir.

$< i, j >$, i ve j kümelerinin birleştirilmesiyle oluşan kümeyi temsil eden endeks

K^A, sürekli değişkenlerin sayısıdır

K^B, kategorik değişkenlerin sayısı

L_k, k. kategorik değişken için kategori sayısı

N_k , k kümesindeki toplam veri kaydı sayısıdır

$\hat{\sigma}_k^2$, tüm veriler için k sürekli değişkenin tahmini varyansıdır

$\hat{\sigma}_{v.k}^2$, v kümesindeki k sürekli değişkenin tahmini varyansıdır

$N_{v.k.l}$, v kümesindeki k kategorik değişkeninin l kategorisinden aldığı nesne sayısıdır.

2.3.2.2 Bölümlemeli Hiyerarşik Yöntemler

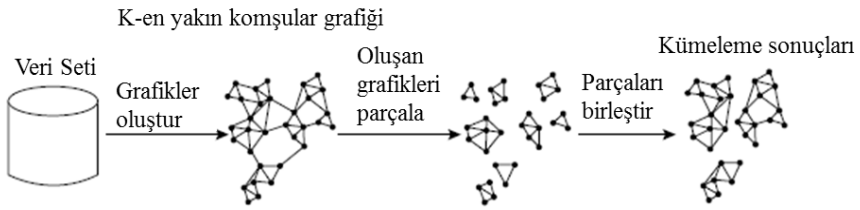
Bölümlemeli hiyerarşik kümeleme yöntemi, yukarıdan aşağıya bir strateji kullanır. Tüm nesneleri, hiyerarşinin kökü olan tek bir kümeye yerleştirerek başlar. Daha sonra kök kümeyi birkaç küçük alt kümeye böler ve bu kümeleri tekrar tekrar küçük bölümlere böler. Bölünme işlemi, en düşük seviyedeki her küme yeterince tutarlı olana kadar devam eder (Han ve Kamber, 2012).

Cure Algoritması: Cure**,** bir diziyi iyi dağınık temsili nokta kullanarak kümeyi temsil eden, yeni bir özellik içeren bir algoritmadır. İki küme arasındaki mesafe, seçilen temsili noktalar arasındaki minimum mesafeye bakılarak hesaplanır. Bu şekilde, CURE hem tek bağlantı hem de grup ortalamalı gruplayıcı hiyerarşik kümeleme yöntemlerinin özelliklerini içerir. Dağınık noktaları seçmek, CURE isteğe bağlı şekil kümelerini yakalamasına yardımcı olur.

Ek olarak, CURE algoritmasında noktaların merkeze doğru bir faktör ile daraldığında büzülme faktörü α kullanır. α daralma, normal noktalara kıyasla aykırı değerlerde daha büyük bir etkiye sahiptir. Bu CURE'i aykırı değerlere karşı daha güçlü kılar. Bu yaklaşıma benzer şekilde, kategorik verileri işlemek için ROCK adlı bir algoritma da önerilmiştir. Bu algoritma

ortak bağlantılar kavramını kullanır ve hiyerarşik kümeleme için aday kümeler arasındaki Jaccard katsayısını belirler (Charu ve Chandan, 2013).

Chameleon Algoritması: Chameleon küme çiftleri arasındaki benzerliği belirlemek için dinamik modellemeyi kullanan hiyerarşik bir kümeleme algoritmasıdır.

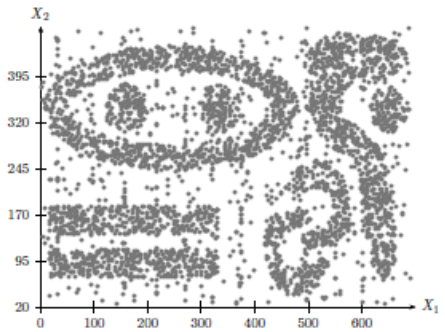


Şekil 2.10: K- En yakın komşular ve dinamik modelleme dayalı hiyerarşik kümeleme (Han ve Kamber, 2012)

Kümeleme işleminde eğer iki küme arasındaki ara bağlantı ve yakınlık degerleri kümelerin kendi içlerindeki ara bağlantı ve yakınlık değerleri ile yüksek oranda ilişkili ise bu kümeler birleştirilir. Bu birleştirme islemi doğal ve homojen kümelerin ortaya çıkarılmasını sağlar ve benzerlik fonksiyonunun tanımlanabildiği her veri tipi için uygulanabilir (Bilgin, 2008).

2.3.3 Yoğunluk Tabanlı Yöntemler

Bölümleme ve hiyerarşik yöntemler, küresel şekilli kümeleri bulmak için tasarlanmıştır. Aşağıdaki şekildeki gibi "S" şekli ve oval kümeler gibi rastgele şekil kümelerini bulmada zorluk çekerler. Bu veriler göz önüne alındığında, gürültünün veya aykırı değerlerin kümelere dahil edildiği dışbükey bölgeleri yanlış bir şekilde belirleyebilirler (Zaki ve diğ., 2014).



Şekil 2.11: Yoğunluk Tabanlı Yöntemler (Zaki ve diğ., 2014)

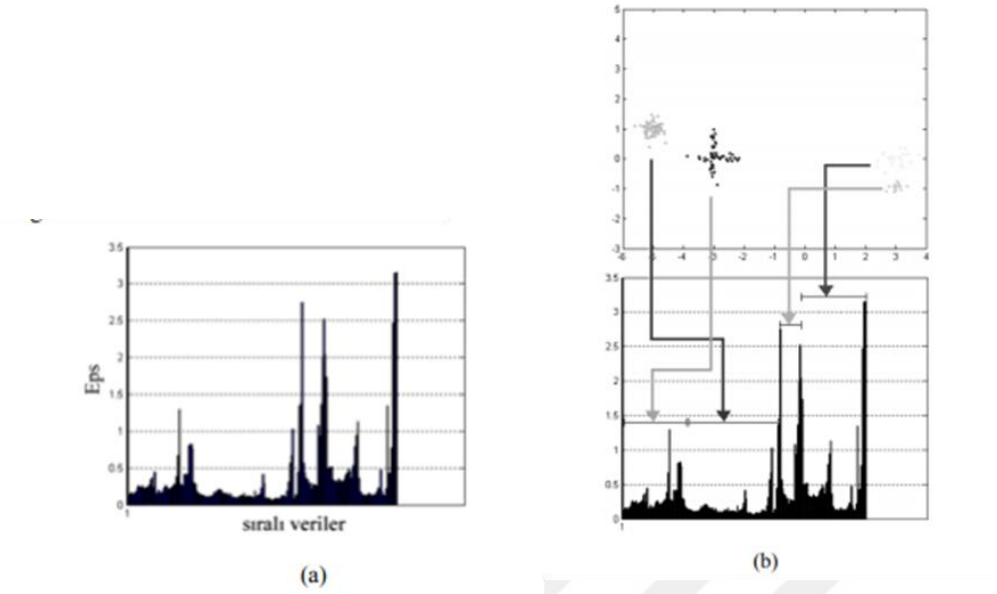
2.3.3.1 DBSCAN Algoritması

DBSCAN, herhangi bir yoğunluk tabanlı kümeleme yaklaşımı için önemli olan bir dizi önemli kavramı gösteren basit ve etkili bir yoğunluk tabanlı kümeleme algoritmasıdır (Tan ve diğ., 2013).

DBSCAN algoritması için çekirdek nesne, Eps, MinPts, doğrudan yoğunluk erişilebilir nokta, yoğunluk erişilebilir nokta, yoğunluk bağlı nokta terimleri temel kavramlardır. Algoritma, Eps ve MinPts değerlerini giriş parametresi olarak alır. Veritabanındaki herhangi bir nesneden başlayarak tüm nesneleri kontrol eder. Eğer kontrol edilen nesne daha önce bir kümeye dahil edilmiş ise işlem yapmadan diğer nesneye geçer. Eğer nesne daha önce kümelenmemiş ise, bir bölgesel sorgu yaparak nesnenin Eps komşuluğundaki komşularını bulur. Komşu sayısı MinPts'den fazla ise, bu nesne ve komşularını yeni bir küme olarak adlandırır. Daha sonra, önceden kümelenmemiş her bir komşu için yeni bölge sorgusu yaparak yeni komşular bulur. Bölge sorgusu yapılan noktaların komşu sayıları MinPts'den fazla ise kümeye dahil eder. Komşuluk bulma işlemi, DBSCAN algoritmasının en fazla işlem gücü gerektiren bölümüdür. Bu bölümde yapılacak performans iyileştirmeleri algoritmanın performansını önemli ölçüde arttırmaktadır. Komşuluk incelemesinde her noktayı incelemek yerine R*-ağaç ya da uzaysal sorgulama gibi çeşitli indeksleme algoritmaları ortaya atılmıştır. Bu algoritmalar ile DBSCAN algoritmasının $O(n*logn)$ olan karmaşıklığını $O(logn)$'e düşürülerek önemli performans artışları sağlanabilmektedir (Bilgin ve Çamurcu, 2005).

2.3.3.2 OPTICS Algoritması

DBSCAN algoritmasının Eps ve MinPts olmak üzere iki adet giriş parametresine bağlı olma dezavantajını gidermek üzere DBSCAN'ı bulan grup tarafından OPTICS algoritması geliştirilmiştir OPTICS algoritması kendi başına bir kümeleme aracı değildir. Algoritma daha çok bir görselleştirme aracı olarak nitelendirilebilir. Veritabanını, değişken Eps değerlerinin dağılımına göre grafik üzerinde göstererek gözlem ya da çeşitli ölçümler ile dolaylı yoldan kümeleri bulmaya olanak sağlar (Bilgin ve Çamurcu, 2005).



Şekil 2.12: (a) MinPts=4 için OPTICS sonucu, (b) OPTICS algoritması sonucunun DBSCAN ile karşılaştırılması (Bilgin ve Çamurcu, 2005)

2.3.3.3 DENCLUE Algoritması

DENCLUE (Density-based Clustering) algoritması, Alexander Hinneburg ve Daniel A. Keim tarafından KDD'98 konferansında sunulmuştur. Bu algoritmada veri tabanında bulunan nesneleri kümelemek için yoğunluk dağılım fonksiyonlarından faydalanılır. DENCLUE algoritmasının DDBSCAN algoritmasına göre 45 kat daha hızlı olduğu deneysel olarak ispatlanmıştır.

DENCLUE algoritması üç aşamadan oluşur (Çelik, 2013).

1. Her bir veri nesnesinin kendi çevresi kapsamında etki fonksiyonu bulunur. Bu fonksiyon parabolik, kare dalga fonksiyonu veya Gauss fonksiyonu olabilir. Daha sonra bu fonksiyon veri tabanındaki bütün nesnelere uygulanır.

2. Hesaplanan etki fonksiyonların toplamı veri tabanının genel yoğunluğunu verir.

3. Matematiksel olarak yoğunluk çekicileri (density-attractors) yardımı ile kümeler tespit edilir.

2.3.4 Izgara Tabanlı Yöntemler

Izgara tabanlı yöntemler nesne uzayını, sonlu sayıda hücrenin bulunduğu bir uzaya indirgeyerek işlem yapar (Çelik, 2009).

2.3.4.1 STING Algoritması

Wang ve arkadaşları uzaysal veritabanlarını kümelemek ve bölgeye yönelik sorguları kolaylaştırmak için bir istatiksel ızgara tabanlı bir kümeleme yöntemi (STING) önermiştir. STING, var olan alanı dikdörtgen hücrelere ayırır ve hücreleri hiyerarşik bir ızgara yapı ağacında saklar. Her hücre (ağaçtaki yapraklar hariç), bir sonraki seviyedeki 4 küçük hücreye ayrılır ve her küçük hücre ana hücrenin çeyreğine karşılık gelir. Bir ana hücre, küçük hücrelerin birliğidir; seviye 1'deki kök hücre, tüm uzaysal alana karşılık gelir. Yaprak seviyeli hücreler, nesnelerin ortalama yoğunluğundan global olarak belirlenen tek tip büyüklüktedir. Her hücre için, istatiksel olarak bağımlı ve bağımsız parametreler korunarak bölümleme işlemi yapılır. Bu parametreler aşağıda gösterilmiştir (Aggarwal ve Reddy, 2014).

Sting algoritmasının parametreleri Tablo 2.2'de sunulmaktadır:

2.3.4.2 CLIQUE Algoritması

Agrawal ve arkadaşları yüksek boyutlu niceliksel verilerin altuzaylarının kümelerini bulabilmek için hibrid yoğunluk ve ızgara tabanlı bir kümeleme algoritması önermişlerdir. Dolayısıyla bu yöntem yoğunluk tabanlı ve ızgara tabanlı kümeleme yöntemlerinin bir entegrasyonu olarak düşünülebilir. CLIQUE algoritması yüksek boyutlu uzayda alt uzayların kümelenmesi için önerilen ilk algoritmadır. CLIQUE algoritması, her bir boyutu ızgara yapıya böler ve hücrelerin ne kadar nokta sayısı içerdiğine bağlı olarak kümelere karar verir.

Tablo 2.2: Sting Algoritmasının Parametreleri (Aggarwal ve Reddy, 2014)

Parametre	Açıklama
N	Hücrede bulunan nokta sayısı
Ort	Hücredeki her boyutun ortalaması
Std	Hücredeki her boyutun standart sapması
Min	Hücredeki her boyutun minimumu
max	Hücredeki her boyutun maksimumu
Dağılım	Hücredeki noktaların dağılımı

Çok boyutlu çok sayıda veri olduğunda algoritma genellikle veri alanı tekdüze dağılmış veri noktalarından oluşmamaktadır. CLIQUE algoritması, uzaydaki seyrek ve yoğun bölgeleri tanımlamaktadır. Böylece veri setinin genel dağılım modeli keşfedilecektir. Toplam veri noktaları, başlangıçta belirlenen model parametresini aşıyorsa, o birimin yoğun olduğu kabul edilir. CLIQUE algoritmasında bir küme, yoğun birimlerin bağlı olduğu maksimum dizi olarak tanımlanır (Pasin, 2015).

2.3.4.3 SOM (Öz Örgütlemeli Haritalar) Algoritması

Öz örgütlemeli haritalar, genellikle iki boyutlu olan küçük alanlar üzerinde projeksiyona izin verdikleri için görselleştirme araçları olarak giderek daha fazla kullanılmaktadır. Kohonen tarafından önerilen temel model, harita adı verilen ayrı bir C hücresi kümesinden oluşur. Bu harita, yönlendirilmemiş grafikle tanımlanan ayrık bir topolojiye sahiptir, genellikle 2 boyutlu normal bir ızgara yöntemidir (Azzag ve Lebbah, 2008). Amaç, bir ızgarada bir dizi nöron düzenlenmesiyle oluşturulan, düşük boyutlu, uzamsal olarak ayrı bir çıktı alanı ile eşleştirmektir. Öz örgütlemeli haritalar, özellik haritası adı verilen doğrusal olmayan bir dönüşüm sağlar. Öz örgütlemeli haritalar algoritmasının aşamaları şu şekilde özetlenebilir (Jassar ve Dhindsa, 2015).

Başlama - İlk ağırlık vektörleri için rasgele değerleri seçin.

Örnekleme - Giriş uzayından örnek olarak eğitim giriş vektörü çizin.

Eşleştirme - Giriş vektörüne en yakın ağırlık vektörü ile kazanan nöron I(x) 'i bulun.

Güncelleme - Ağırlık güncelleme denklemini uygulayın $|W_i - X| \leq |W_k - X|$ ¥**k** (2.18)

Devam - özellik haritası değişmeyi bırakana kadar 2. adıma dönmeye devam edin.

2.3.5 Olasılık Model Tabanlı Algoritmalar

Olasılık modeline dayalı kümeleme yöntemleri görüntü segmentasyonu, el yazı tanıma, doküman kümelemeye kadar birçok uygulamada yaygın bir şekilde kullanılmaktadır. Modele dayalı kümeleme yaklaşımları olasılık yaklaşımlarını kullanarak gözlenen verilerle, bazı matematiksel modeller arasındaki uyumu en iyi duruma getirmektedir. Bazı yöntemler, genellikle verinin ilgili olasılık dağılımlarının karmasından oluşmaktadır. Uygulamada her bir

küme parametrik olasılık dağılımları (Gaussian veya Poisson dağımları) ile matematiksel olarak ifade edilebilir. Dolayısıyla, kümeleme problemi artık parametre tahmini problemine dönüşmektedir. Böylece veri, K karmaşık yapıda bileşen dağılımları ile modellenebilmektedir (Pasin, 2015).

2.3.5.1 Beklenti En Büyüklemesi (EM) Algoritması

EM algoritması bir objenin hangi kümeye ait olduğunu belirlemede kesin mesafe ölçütlerini kullanmak yerine tahminsel ölçütleri kullanmayı tercih eder. Maksimum benzerlik prensibine dayanan beklenti maksimizasyonu (EM) algoritması ilk olarak Dempster, Laird ve Rubin tarafından 1977 yılında ortaya konulmuştur. EM algoritması son yıllarda bir çok alanda kullanılan popüler bir yaklaşım olmuştur (Sezgin ve Çelik, 2013).

Algoritma, bir nesnenin belirli uzaklık ölçütleri kullanmak yerine olasılıksal ölçütler kullanarak mevcut kümelerden birine ait olma olasılığını gösterir. Her yinelemede EM algoritması önce optimum bir alt sınır bulur ve daha sonra gelişmiş bir tahmin elde etmek için bu sınırı en büyükler. Dolayısıyla, algoritma sırasıyla E-adım (beklenti adım) ve M-adımı (en büyükleme adımı) olarak adlandırılan iki adımı içerir (Kiriş ve Tüysüz, 2017).

2.3.5.2 BILCOM Deneysel Bayesian (Kategorik ve Numerik Karışım) Modeli

BILCOM deneysel Bayesian (Kategorik ve Numerik Karışım) modeli iki seviyeden oluşur. Bu kümelemede, kategorik özellikler nesneler hakkında anlamsal bilgileri temsil ederken, numerik bilgiler deneysel sonuçları temsil eder. Bayesian teorisine göre, sayısal verileri kullanmak yerine kategorik özellikleri ilk seviyede kullanmak mantıklıdır. Kategorik bilgilere dayanan benzerlik ilk düzeyde vurgulanır ve numerik özelliklere dayalı benzerlikler ikinci düzeyde vurgulanır. Birinci seviyenin sonucu, ikinci seviyeye girdi olarak verilir ve ikinci seviyenin sonucu BILCOM'un çıktısıdır (Andreopoulos ve diğ., 2006).

2.3.6 Yüksek Boyutlu Verilerin Kümelenmesi

Yüksek boyutlu veri alanı kümelemesinde verimlilik ve kalite adında iki ana sorun ortaya çıkar. Bu veri seti ile sorunsuz başa çıkmak için yeni algoritmalar gerekir. Çözüm olarak iki popüler strateji uygulanmaktadır. Birincisi, kümeyi orijinal veri kümesi alanının alt alanında bulmak için alt uzay kümeleme stratejisidir. Bir diğeri, daha fazla kümeleme için yaratılmış daha düşük boyutlu bir veri alanı olan boyutsallık azaltma stratejisidir (Makhabel, 2015).

2.3.6.1 MAFIA Algoritması

MAFIA, yüksek boyutlu bir veri kümesinin alt alanlarına gömülü kümeleri keşfetmek için her boyutta sonlu aralıkların uyarlamalı hesaplamasını kullanan paralel bir alt uzay kümeleme algoritmasıdır. Aynı zamanda yoğunluk ve ızgara tabanlı kümeleme algoritmasıdır (Gan ve diğ.,2007).

Bir algoritmanın hesaplama maliyeti yoğun hücrelerin belirlenmesine ve bu yoğun hücrelerin yüksek boyuta yayılmasına bağlı olarak değişmektedir. Goil ve arkadaşları 1999 yılında MAFIA kümeleme algoritması, CLIQUE algoritmasının bir uzantısı olarak önermişlerdir. Bu yöntemde boyutlar parçalanırken verinin dağılımına bağlı olan uyarlanabilir aralık genişliği (adaptive interval size) önerilmektedir. Veri tabanı başlangıçta bir kere taranarak histogram oluşturulur ve bu histogram kullanılarak bir boyut için minimum kutu sayısına karar verilir. Benzer histogram değerleri olan bitişik kutular daha büyük kutuları oluşturmak için kombine edilmektedir. Böylece boyut veri dağılımına bağlı olarak bölünmektedir. Uyarlanabilir aralık genişliği kümeleme sonuçlarının kalitesini artırmaktadır. MAFIA kümeleme algoritması sabit genişlikte ızgara yapısı kullanmak yerine kümeleme sonuçlarının etkinliğini artırabilmek için uyarlanabilir ızgara yapılarını kullanmaktadır (Pasin, 2015).

2.3.6.2 Temel Bileşen Analizi (PCA)

Temel bileşen analizi bir veri boyut azaltma tekniğidir. Analiz edilen değişken sayısını azaltır. Ana bileşenler birbiriyle ilişkili olmadığı durumda, orijinal değişkenlerin doğrusal kombinasyonları olan temel bileşenler adı verilen yeni değişkenler yaratır. P orjinal değişkenlerinden türetilen m adet temel bileşeni (TB) olduğu varsayılır. Her ana bileşen, orijinal değişkenlerin doğrusal bir kombinasyonu olarak ifade edilebilir (Chiu ve Tavella, 2008).

2.3.7 Grafik Tabanlı Kümeleme

Grafik tabanlı bir kümeleme algoritması önce bir grafik veya hipergraf oluşturur ve ardından grafiği veya hipergrafiyi bölümlemek için bir kümeleme algoritması uygular (Gan ve diğ., 2007).

2.3.7.1 CACTUS (Özet Verileri Kullanarak Kategorik Verilerin Kümelenmesi)

CACTUS'un ardındaki temel fikir, tüm veri kümesinin bir özetinin, gerçek kümesini belirlemek için doğrulanabilecek bir dizi "aday" kümeyi hesaplamak için yeterli olduğudur. CACTUS üç aşamadan oluşur: Özetleme, kümeleme ve doğrulama. Özetleme aşamasında, veri setinden özet bilgileri hesaplanır. Kümelenme aşamasında, bir dizi aday kümeyi keşfetmek için özet bilgiler kullanılır. Doğrulama aşamasında, aday kümeler grubundan asıl kümeler belirlenir (Ganti ve diğ., 1999).

2.3.7.2 ROCK

ROCK, aşağıdan yukarıya bir hiyerarşik algoritmadır. ROCK, bir ağaç inşa ederek kategorik veri kümelemesini ele alır; her ağaç seviyesinde kümeler, elde edilen küme içi benzerliğin en üst düzeye çıkarıldığı şekilde birleştirilir. Benzerlik, elde edilen küme içindeki benzer nesne çiftlerinin sayısı ile değerlendirilir. ROCK, nesneler arasında özel bir benzerlik ölçüsü varsayar ve benzerliği bir eşiği geçen iki nesne arasında bir "bağlantı" olarak tanımlar. ROCK için motivasyon, nesneler arasındaki bağlantıları göz önünde bulunduran küresel bir kümeleme yaklaşımı geliştirmektir. Bu amaçla, ROCK, bağlantıları tanımlamak için ortak komşular kullanır. A nesnesi, C nesnesi ile komşu ve B nesnesi de, C nesnesi ile komşuysa, A ve B nesneleri birbirine bağlanır (kendileri komşu olmasalar bile). Aynı kümeye ait iki nesnenin çok sayıda ortak komşusu olması gerekirken, farklı kümelere ait nesnelerin birkaç ortak komşusu olması gerekir. Başlangıçta, her nesne ayrı bir kümeye atanır. Daha sonra, kümeler, iki küme arasındaki tüm nesne çiftleri arasındaki bağlantı sayısının toplamı tarafından tanımlanan "yakınlıklarına" göre tekrar tekrar birleştirilir (Aggarwal ve Reddy, 2014).

2.3.8 Kısıtlarla Kümeleme

Kullanıcılar genellikle küme analizine entegre etmek istedikleri temel bilgilere sahiptir. Uygulamaya özel gereksinimler de olabilir. Bu bilgiler kümeleme kısıtlamaları olarak modellenebilir. Kümeler üzerindeki bir kısıtlama, küme analizinde bir çiftin veya bir grup örneğin nasıl gruplandırılması gerektiğini belirtir. İki yaygın kısıtlama türü vardır (Han ve Kamber, 2012).

- Mutlak birliktelik kısıtlaması: İki örneğin aynı kümeye yerleştirilmesi gerektiğini zorunlu kılar.

- Kesinlikle olmamalı kısıtlaması: İki örneğin aynı kümeye yerleştirilmemesi gerektiğini zorunlu kılar.

Kısıtlar belirli durumlara özgü tanımlanabilir. Alternatif olarak, durum değişkenleri veya örneklerin nitelikleri kullanılarak da tanımlanabilir. Aşağıdaki şekilde tanımlanan bir örnek uzaklıklarla ilgili bir kısıta örnek verilebilir (Han ve Kamber, 2012).

Kısıt(x,y) : must-link(x,y) eğer uzaklık(x,y) $\leq \epsilon$

2.3.9 Kümeleme Algoritmalarının Başarısının Değerlendirilmesi

2.3.9.1 Dışsal Kriterler

Dışsal kriterlere dayalı önerilen tüm endeksler, P_1 veya P_2 bölümlerine göre aynı kümeye ait olup olmadıklarına bağlı olarak puan çiftlerini temsil eden bir karışıklık matrisine dayanır. 4 olasılık vardır , aşağıdaki gibi sınıflandırabiliriz (Desgraupes ,2017).

- *YY:* iki nokta hem P_1 hem de P_2'ye göre aynı kümeye ait
- *YN* : iki nokta P_1'e göre aynı kümeye ait ancak P_2'ye değil
- *NY* : iki nokta P_2'ye göre aynı kümeye aittir ancak P_1'e değil
- *NN* : İki nokta, hem P_1 hem de P_2'ye göre aynı kümeye ait değildir.

Rand İndeksi: Kümeler arası benzerliği gösteren bir ölçüttür. Orijinal veritabanındaki U kümesi ile algoritma tarafından üretilen V kümesi için aşağıdaki denklem ile hesaplanır. [0,1] aralığında değerler alır ve en büyüklenmelidir. Doğru etiketlenen nesne sayısının veritabanı boyutuna oranı ile hesaplanır. Bütün elemanların doğru etiketlenmesi durumunda en büyük değeri olan 1 değerini alır (Gemici, 2007).

$$Rand = \frac{a+d}{a+b+c+d} \tag{2.19}$$

a: U ve V ' de aynı kümede olan eleman sayısı

b: U' da aynı V'de farklı kümede olan eleman sayısı

c: U'da farklı V'de aynı kümede olan eleman sayısı

d: U ve V'de farklı kümede olan eleman sayısını ifade etmektedir.

Jaccard İndeksi: Jaccard benzerlik katsayısı (Jaccard indeksi) verilen iki küme arasındaki benzerlik katsayısını istatistiksel olarak değerlendirir. Burada, benzerliği hesaplamak için iki kümedeki kesişen eleman sayısı birleşimin eleman sayısına bölünür. Aşağıda yer alan formülde, nasıl hesaplandığı bulunmaktadır.

$$J\,(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{2.20}$$

Jaccard indeksinin değerinin yüksek çıkması iki kümenin birbirine daha çok benzediği anlamına gelir. İki küme birbirine tamamen eşitse Jaccard indeksi 1 olarak bulunur. İki kümenin hiç ortak özelliği yoksa da değer 0 olarak bulunacaktır. Jaccard indeksi 0 ile 1 arasında değişen bir değer alabilir (Sivri, 2015).

Folkes ve Mallows İndeksi: Girişte tanımladığımız (YY,YN,NY,NN) değerlerine göre formülümüz aşağıdaki gibidir (Desgraupes, 2017).

$$FM = \frac{yy}{\sqrt{(yy + yn)\; x\; (yy+ny)}} \tag{2.21}$$

İndeks değeri 1' e yaklaştıkça benzerlik artar.

Hubert Γ İstatistiği: Hubert'in istatistiği, kümeler aralarındaki korelasyonu sayarak farklı kümeleri bulur. Bunun için aynı büyüklükte kare matris kullanır (Mary ve diğ.,2015).

Hubert İstatistiği [-1,1] aralığında değer alır ve şu şekilde tanımlanır:

$$\Gamma = \frac{M_a - m_1 m_2}{\sqrt{m_1 m_2 (M - m_1)(M - m_2)}} \tag{2.22}$$

Burada $m_1 = a+b$ ve $m_2 = a+c$ olarak tanımlanır. İndeks değerleri 1'e yakınlaştıkça C (Kümelenen) ve P (Gerçekte olan) parçalarının birbirine olan benzerliği, -1'e yaklaştıkça farklılığı artmaktadır (Servi, 2009).

F-ölçütü: F-ölçütü, yaygın olarak kullanılan diğer bir dışsal kalite ölçüm yöntemidir. Kümeleme doğruluğu olarak da bilinir. F-ölçütü için kesinlik ve anma değerlerinin hesaplanması gerekmektedir (Tunalı, 2011).

Kesinlik, C_l kümesinde K_h sınıfına ait nesnelerin payını ifade eder ve şu şekilde tanımlanır:

$$Kesinlik\ (C_l, K_h) = \frac{n_l^{(h)}}{n_l} \quad (2.23)$$

Anma ise C_l kümesinde bulunan K_h sınıfından nesnelerin payıdır ve şu şekilde tanımlanır:

$$Anma\ (C_l, K_h) = \frac{n_l^{(h)}}{|K_h|} \quad (2.24)$$

Burada K_h, h sınıfındaki nesnelerin sayısıdır. C_l kümesinin F değeri, kesinlik ve anma değerlerinin harmonik ortalaması olarak aşağıdaki denklemdeki gibi hesaplanır:

$$F\ (C_l, K_h) = \frac{2\ .\ kesinlik\ (C_l,K_h)\ .\ anma\ (C_l,K_h)}{kesinlik\ (C_l,K_h) + anma\ (C_l,K_h)} \quad (2.25)$$

C_l kümesi için F-Ölçütü ise, tüm sınıflar içerisinde C_l kümesi için elde edilmiş en yüksek F değeridir.

$$F(C_l) = \max_h F\ (C_l, K_h) \quad (2.26)$$

Tüm kümeleme için F-Ölçütü değeri kümelerin F-Ölçütlerinin ağırlıklı ortalamalarının toplamı olarak ifade edilir.

$$F(C) = \sum_{l=1}^{k} \frac{n_l}{n} F(C_l) \quad (2.27)$$

NMI Ölçümü: Normalleştirilmiş karşılıklı bilgi (NMI) olarak adlandırılır. Etiketli iki nesnenin NMI'si şu şekilde ölçülebilir.

$$NMI\ (X,Y) = \frac{I\ (X,Y)}{\sqrt{H(X)H(Y)}} \quad (2.28)$$

Burada, $I(X, Y)$ iki rastgele değişken arasındaki karşılıklı bilgiyi ifade eder. $H(X)$ X'in entropisini belirtir. $H(Y)$ Y'in entropisini belirtir (Rendón ve diğ., 2011).

Entropi: Entropi değeri aşağıdaki formül kullanılarak hesaplanmaktadır.

$$E(C) = -\sum_{i=1}^{m} p_i log_2 p_i \quad (2.29)$$

Formülde yer alan p_i değeri, *i.* kümeye ait olan *m* tane küme için veri nokta olasılıklarıdır (Pasin, 2015).

Saflık: Saflık değeri ise normalleştirilmiş ortak bilgiler kullanılarak, *H(c)* entropi değeri, c_j sınıfına ait olan nesne olasılığı hesaplanarak ve w_i kümesinde sabitlenerek aşağıdaki şekilde hesaplanmaktadır.

$$Saflık\ (v,\ C) = \frac{I\ (v,C)}{[H(v)+\frac{H(C)}{2}]} \tag{2.30}$$

$$\mathrm{I}\ (\mathrm{v}, \mathrm{C}) = \sum_i \sum_j \mathrm{P}(\mathrm{w_i} \cap \mathrm{c_j}) \log \frac{\mathrm{p}(\mathrm{w_i} \cap \mathrm{c_j})}{\mathrm{p}(\mathrm{w_i})\ \mathrm{p}(\mathrm{c_j})} \tag{2.31}$$

$v=\{\omega 1,\omega 2,\ldots,\omega i\}$ kümelerin seti iken $C=\{c1,c2,\ldots,cj\}$ sınıfların setidir (Pasin, 2015).

2.3.9.2 Oransal Kriterler

Kümeleme kalitesi değerlendirmesi için yaygın olarak kullanılan oransal endekslerin en önemlileri aşağıdaki gibi sıralanabilir (Wang ve diğ., 2009):

- Silhouette indeksi
- Davies-Bouldin indeksi
- Calinski-Harabasz indeksi
- RMSSTD indeksi

Silhoutte İndeksi: Silhouette indeksi, Rousseuw tarafından 1987 yılında her bir birimin yer aldığı kümeye uygunluğunu belirlemek amacı ile geliştirilmiştir. $a(x_i)$, i.birimin yer aldığı kümedeki diğer birimlere olan ortalama uzaklıklarını (farklılıklarını; benzememe durumlarını) ve $b(x_i)$ i. birimin diğer kümelerdeki tüm birimlere olan ortalama uzaklıkların minimumunu göstersin. Buna göre, *i.* birim için Silhouette indeksi,

$sil(x_i) = \frac{b(x_i)-a(x_i)}{max(a(x_i),b(x_i))}$ ile elde edilir. sil(x_i), -1 ile 1 arasında değer almaktadır. (2.32)

$sil(x_i) \cong 1\ ise\ i.birim$ doğru sınıflandırılmıştır.

$sil(x_i) \cong 0\ ise\ i.birim$ iki küme arasındadır.

$sil(x_i) \cong -1\ ise\ i.birim$ yanlış sınıflandırılmıştır.

Tüm kümelemenin kalitesi ortalama Silhouette değeri ile ölçülmektedir. Doğal ölçü olarak tüm birimler için ortalama Silhouette değeri,

$sil(C) = \frac{1}{n}\sum_{i=1}^{n} sil(x_i)$ ile hesaplanır. (2.33)

Buna göre, maksimum ortalama Silhouette değerine karşılık gelen küme sayısı uygun küme sayısı olarak alınır. Genel olarak, ortalama Silhouette değeri 0.50'nin üzerinde ise uygun küme sayısı ve dolayısıyla uygun kümelemeye ulaşıldığı kabul edilir (Akgül, 2013).

Davies-Bouldin indeksi: Davies-Bouldin (DB) indeksi, daha iyi bölümleme yeteneği ile bilinir. Dunn indeksine benzer şekilde, Davies-Bouldin endeksi birbirinden uzak ve kompakt olan kümeleri tanımlar. DB endeksi, tüm küme benzerliklerinin ortalaması hesaplanarak elde edilir. Davies Boludin endeksi, her küme ile en benzer olanı arasındaki benzerliğin ortalamasını ölçer. Optimum kümeleme çözümü, en küçük Davies-Bouldin endeks değerine sahiptir (Mary ve diğ., 2015).

Bu endeks, kompakt ve iyi ayrılmış kümeleri belirlemeyi amaçlamaktadır. Davies-Bouldin endeksi aşağıdaki gibi tanımlanır.

$BD = \frac{1}{C}\sum_{\text{İ}=1}^{C} Max_{i\neq j}\{\frac{d(X_i)+d(X_j)}{d(c_i,c_j)}\}$ (2.34)

C küme sayılarını belirtirken, *i,j* küme etiketlerini, $d(X_i)$ ve $d(X_j)$, *i* ve *j* kümelerinin kendi küme merkezlerine ait tüm örnekleridir. $d(c_i, c_j)$ küme merkezleri arasındaki mesafeyi tanımlar. BD nin küçük olması kümelemenin iyi olduğunu gösterir (Rendón ve diğ., 2011).

Calinski-Harabasz indeksi: İyi tanımlanmış kümeleri kümeler arası varyans büyük ve küme içi varyans küçük değere sahip olarak tanımlamıştır. Bu sebeple varyans oranı kriteri (VOK) olarak da adlandırılır. VOK oranı büyüdükçe, veri kümelemesi daha iyi olur. Optimal küme sayısı, Calinski-Harabasz endeks değeri en yüksek olan çözümdür. Calinski-Harabasz kriteri, öklid uzaklığa sahip olan k- ortalamalar kümeleme yöntemi için en uygun yöntemdir (Mary ve diğ., 2015).

Calinski ve Harabasz (1974), k kümeye sahip bir kümelemenin geçerliliğini değerlendirebilmek için,

$CH(k) = \frac{BSS\ (k)\ /\ (k-1)}{WSS(k)/\ (n-k)}$ indeksini önermişlerdir. (2.35)

Burada;

$$WSS(k) = \frac{1}{2}\sum_{i=1}^{k}\sum_{j\in C_i} d(i,j) \quad (2.36)$$

$$BSS(k) = \frac{1}{2}\sum_{i=1}^{k}\sum_{\substack{i\in C_i \\ j\notin C_i}} d(i,j) \quad (2.37)$$

olmak üzere, *WSS(k)* ve *BSS(k)* sırasıyla, kümeler içi ve kümeler arası kareler toplamıdır. Bu kritere göre, CH indeks değerini maksimum yapan küme sayısı, uygun küme sayısı olarak kabul edilir (Hacıoğlu, 2016).

RMSSTD (kök-ortalama-standart sapma karesi) indeksi: RMSSTD indeksi, aşağıdaki denklemde tanımlanmış olup kısaca kümelerin varyansıdır, bu nedenle kümelerin homojenliğini ölçer. Kümelenme işleminin amacı homojen grupları tanımlamaktır, düşük RMSSTD değeri daha iyi kümelenme anlamına gelir (Kovács ve diğ., 2005).

$$RMSSTD = \sqrt{\frac{\sum_{\substack{i=1..nc \\ j=1...d}}\sum_{k=1}^{n_{ij}}(x_k - \overline{x_j})^2}{\sum_{\substack{i=1..nc \\ j=1...d}}(n_{ij}-1)}} \quad (2.38)$$

DUNN indeksi: Yukarıda tanımladığımız indekslere ek olarak en sık kullanılan indekslerden biri de DUNN indeksidir. Dunn indeksi (DI), sıklık teşhisini ve ayrık küme kullanımını önerir. Formülü aşağıdaki gibidir (Murat ve Şekerler, 2009)

$$DI(c) = min_{i\in c}\left\{ min_{k\in c}, i \neq k \left\{ \frac{min_{x\in C_i, y\in C_k}\, d(x,y)}{max_{k\in C}\{ max_{x,y\in C}\, d(x,y)\}} \right\}\right\} \quad (2.39)$$

2.4 BİRLİKTELİK KURALLARI

Pazar sepet analizi ile birliktelik kuralları çıkarımı ilk olarak Agrawal ve diğerleri tarafından 1993 yılında ele alınmıştır. Çalışmada, X ve Y'nin nesne kümesi oluşu X ⇒ Y (X birliktelik Y) şeklinde ifade edilmiş olup, birliktelik kurallarının matematiksel şekli belirlenmiştir. Kuralları oluşturabilmek için destek ve güven değerlerini kullanarak, kullanıcı

tarafından belirlenmiş minimum destek ve minimum güven değerlerinden yaygın birlikteliklerin belirlenmesi amaçlanmıştır (Döşlü, 2008).

Birliktelik kurallarının amacı, büyük veri kümelerindeki kategorik değişkenlerin belirli değerleri arasındaki ilişkileri tespit etmektir. Bu teknik, analistlerin ve araştırmacıların büyük veri setlerinde gizli kalıpları ortaya çıkarmasına izin verir. Erken dönemlerde yapılan bir birliktelik analizinin klasik örneği, biranın çocuk bezi ile satılma eğiliminde olduğunu ortaya çıkarmıştır, pazartesi gecesi futbol izlemenin ve aynı zamanda aile kaygılarını önemsemenin arasında bir ilişki olduğu gözlemlenmiştir (Nisbet ve diğ., 2017).

Birliktelik kuralları bilgisayar bilimleri alanında geliştirilmiş bir analiz türüdür. Birliktelik kural madenciliğinde sıklıkla birlikte bir arada ortaya çıkan ilişkilerin $X \rightarrow Y$ formatında açığa çıkarılması sağlanmaktadır. Bu ifade şu anlama gelmektedir: X'in ortaya çıktığı bir durumda büyük olasılıkla Y de ortaya çıkacaktır. İfadenin sol tarafındakilere (birden fazla olabilir) öncül denilmektedir ve sağ taraftakilere ardıl denilmektedir. Eğer sol taraftaki ifade doğruysa, sağ taraftaki ifade de doğrudur. Bu kuralın önem derecesi iki farklı ölçü ile ölçülebilir. Bunlar destek ve güven parametreleridir.

Destek değeri ile bütün veri setinin yüzde kaçının kuralı kapsadığı ölçülmektedir. Aşağıdaki yardımıyla hesaplanmaktadır:

$$Destek = \frac{n(X \cup Y)}{N} \tag{2.40}$$

Formülde $(X \cup Y)$ ifadesi X ve Y'nin birlikte yer aldığı işlem sayısını, N ise toplam işlem sayısını temsil etmektedir. Bu değerin bir olması, incelenen veri setinde her işlemde X ve Y'nin birlikte gerçekleştiğini, sıfır olması ise veri setinde hiçbir işlemde X ve Y'nin birlikte gerçekleşmediğini ifade etmektedir.

Güven değerinde ise X'i içeren işlemlerin yüzde kaçının Y'yi de içerdiğini ifade etmektedir ve aşağıdaki formül yardımıyla hesaplanmaktadır.

$$Güven = \frac{n(X \cup Y)}{n(X)} \tag{2.41}$$

Formülde destek formülünden farklı olarak paydada toplam X'i içeren gözlem sayısı yer almaktadır. Bu değerin bir çıkması X'i içeren her işlemin Y'yi de içerdiğini, sıfır çıkması ise X'i içeren işlemlerin hiçbirinin Y'yi içermediğini ifade etmektedir (Özçalıcı, 2017).

Destek kriteri, veri içerisinde bulunan nesneler arasındaki bağıntının ne kadar sık olduğunu, güven kriteri ise Y ürününü almış olan bir kişinin hangi olasılıkla X ürününü alacağını göstermektedir. İki ürün arasında elde edilen bağıntının önemli olabilmesi için hem destek değerinin, hem de güven kriterinin olabildiğince yüksek olması gerekmektedir. Ancak bu iki değerin de yüksek olması her zaman önemi yüksek ve ilginç kuralların elde edileceği anlamı taşımamaktadır. Bu nedenle, bir kuralın ne derece ilginç olduğunun tespitine yönelik olarak kaldıraç değeri kullanılmaktadır. Kaldıraç değeri ise aşağıdaki gibi hesaplanmaktadır.

$$kaldıraç\ (\ x \rightarrow y\ ) = \frac{destek\ (x \rightarrow y)}{destek(x).destek(y)} \qquad (2.42)$$

Kaldıraç ölçütünün "1" değerini alması ilginçliğin olmadığını 1'den büyük veya küçük değerler alması ise ilginçliğin arttığını göstermektedir (Yalçın ve Karabatak, 2017).

2.4.1 AIS Algoritması

AIS, tüm olası ürün kombinasyonlarını kontrol eden algoritmadır (Agrawal ve Srikant, 1994). AIS algoritması, Agrawal, Imielinski ve Swami tarafından birliktelik kuralı için önerilen ilk algoritmadır. Karar destek sorgularını işlemek için gerekli işlevsellik ile birlikte veritabanlarının kalitesini iyileştirmeye odaklanır. Bu algoritmada, yalnızca bir öğe sonuç birleştirme kuralı oluşturulur, bu kuralların sonucunun yalnızca bir öğe içerdiği anlamına gelir; örneğin, $X \cap Y \Rightarrow Z$ gibi kurallar üretilebilir ancak $X \Rightarrow Y \cap Z$ gibi kurallar üretilemez (Kumbhare ve Chobe, 2014).

2.4.2 SETM Algoritması

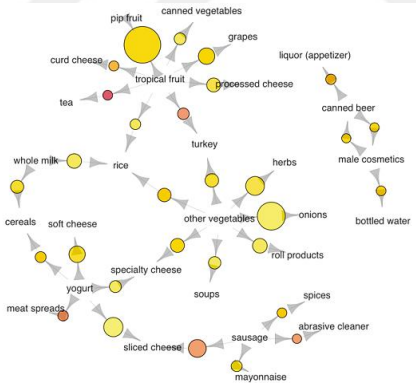
SETM algoritması, büyük öğe setleri hesaplamak için SQL kullanma isteğiyle ortaya çıkmıştır. AIS gibi, SETM algoritmasında aday öğe setleri veritabanı taranırken anında üretilir, ancak geçiş sonunda sayılır. Böylece AIS algoritmasının ürettiği her aday öğe setini üretir ve sayar (Khurana ve Sharma, 2013).

SETM algoritmasında, aday öğe setleri veritabanı taranırken anında üretilir, ancak sonunda sayılır. Ardından, yeni aday öğe setleri AIS algoritmasındaki gibi üretilir, ancak

üretici işlemin işlem tanımlayıcı TID'si aday öğe kümesi ile ardışık bir yapıda kaydedilir. Aday oluşturma sürecini saymaktan ayırır. Geçişin sonunda, aday öğe kümelerinin destek sayısı, sıralı yapının toplanmasıyla belirlenir (Kumbhare ve Chobe, 2014).

2.4.3 Apriori Algoritması

Apriori, minimum destekten büyük tüm öğe kümelerini bulmak için geliştirilmiş bir algoritmadır. Bir öğe için destek, öğe içeren işlem sayısının toplam işlem sayısına oranıdır. Asgari destek kısıtlamasını sağlayan ürün kümelerine sık öğe kümesi denir. Apriori, öğe kümelerinin seviye bazında eksiksiz bir arama algoritması olarak tanımlanır: "Eğer bir öğe kümesi sık değilse, üst kümesi hiçbir zaman sık değildir", buna aşağı doğru kapatma özelliği de denir. Algoritma, veri üzerinden çoklu geçişler yapar (Wu ve Kumar, 2009). Apriori, oldukça verimlidir ve sınırlı sayıda veri okuması olmasına rağmen, bir şekilde ilginç olan bir dizi ilgili kural çıkarabilir. Apriori de ilk önce tek öğeli öğe kümeleri tespit edilir, sonrasında sık kullanılan 2 öğeli öğe kümeleri bulunur, sonrasında 3 öğeli kümeler bulunur ve bu süreç daha fazla öğeli küme bulunmayana kadar devam eder (Azzalini ve Scarpa, 2012). Apriori algoritmasına örnek öğe setleri ise aşağıdaki gibi gösterilebilir (Hatipoğlu, 2018).



Şekil 2.13: Apriori algoritması örnek öğe setleri gösterimi (Hatipoğlu, 2018)

k-öğe (k adet elemana sahip öğe kümesi) kümesi c ile ifade edilirse, öğeleri (ürünler) $c_{[1]}$, $c_{[2]}$, $c_{[3]}$,...,$c_{[k]}$ şeklinde gösterilir ve $c_{[1]} < c_{[2]} < c_{[3]} < \ldots < c_{[k]}$ olacak şekilde küçükten büyüğe doğru sıralıdır. Her öğe kümesine destek ölçütünü tutmak üzere bir sayaç değişkeni eklenmiştir ve sayaç değişkeni öğe kümesi ilk kez oluşturulduğunda sıfırlanır. Sık geçen öğe kümeleri L karakteri ile, aday öğe kümeleri ise C karakteri ile gösterilir. Klasik Apriori algortimasının sözde kodu aşağıdaki gibi ifade edilebilir (Özçakir ve Çamurcu, 2007).

1) **L_1** = { Sık geçen 1-öğe kümesi};

2) **For (k=2; $L_{k-1} \neq \emptyset$; k++) do begin**

3) **C_k = apriori-gen (L_{k-1});** // Yeni adaylar

4) **Forall transactions**-hareketler $\in D$ **do begin**

5) **C_t = subset (C_k , t) ;** // adaylar t içindedir

6) **Forall candidates** – adaylar c $\in C_t$ do

7) **C.count++ ;**

8) **End**

9) **L_k = { c** $\in C_k$ | c.count $\geq$ minimum destek }

10) **End**

11) **Answer $U_k L_k$;**

Apriori algoritması özet kodu incelendiğinde sık geçen öğe kümelerini bulmak için bir çok kez veritabanının tarandığı görülmektedir. İlk aşamadan önce, veri madenciliği uygulanacak veri topluluğunun taranarak öğelerin kaç adet hareket kayıtı içinde yer aldığı tespit edildiği (her öğe için tespit edilen bu değere destek sayacı adı verilir) ve destek sayacı minimum destek değerine eşit veya büyük olan öğelerin L_1 sık geçen 1 – öğe kümesi olarak belirlendiği varsayılarak işleme başlanır.

Kod içinde kurulan döngü yapısı ile ilk aşamada L_1 sık geçen öğe kümesinin öğelerinin ikili kombinasyonuna benzer bir şekilde $(L_1 \infty L_1)$ yeni bir küme oluşur, bu işleme birleştirme adı verilir. Bu işlem ile oluşan kümelere de aday öğe kümeler adı verilir ve *C* harfi ile simgelenir. Oluşan bu aday öğe kümesinin her elemanı iki adet öğeden oluştuğu için C_2 ifadesi ile isimlendirilir. Bu aday küme apriori-gen işlevi ile budama işlemine tabi tutulur ve C_2 kümesinin elemanlarına ait alt kümelerinin L_1 öğe kümesinde olup olmadığına bakılır, alt kümelerden L_1 içinde yer almayan küme elemanları C_2 aday kümesinden silinir.

Apriori algoritması uygulanan veri topluluğu tekrar taranarak budama işleminden geçen C_2 aday kümesi elemanlarının kaç adet hareket kayıtı içinden geçtiği bulunur. Bulunan destek sayaç bilgileri doğrultusunda C_2 aday kümesi elemanlarının destek sayacı minimum destek değerine eşit veya büyük destek değerine sahip olan elemanları L_2 sık geçen öğe kümesini oluşturur. Döngü bir sonraki aşamada L_2 kümesi öğelerinin üçlü kombinasyonu ile yeni bir aday öğe kümesi oluşturur ve bu küme C_3 ifadesi ile simgelenir. İlk aşamada olduğu gibi bu kümede budama işleminden geçer ve budama işleminden sonra minimum destek seviyesinin üstünde kalan elemanları ile L_3 sık geçen öğe kümesi oluşturulur. Döngü her dönüşünde öğe sayısını artırarak devam eder. Bu süreç yeni bir sık geçen öğe kümesi bulunamayana kadar sürer (Özçakir ve Çamurcu, 2007).

Birliktelik kuralları modeli, girdi olarak tanımlanan her kategorik değişken ile hedef değişkenler arasındaki ilişkiyi göstermenin uygun bir yoludur. Birliktelik kuralları, girdi ve hedef değişkenlerin kategorik olmasını gerektirir; SPSS Modeler'de bunlar sözel (nominal), sıralı (ordinal) veya ikili (flag) değişkenleridir (McCormick ve diğ., 2013).

Apriori hangi kuralların tutunacağını belirlemek için çeşitli değerlendirme yöntemleri sunar. Değerler, önceki güven ve ardıl güvene dayalı olarak hesaplanır.

$$C_{önceki} = \frac{c}{N} \tag{2.43}$$

$$C_{sonraki} = \frac{r}{a} \tag{2.44}$$

c: ardılın desteği,

a: öncülün desteği

r: öncülün ve ardılın birleşiminin desteğidir

N: eğitim verilerindeki kayıtların sayısıdır.

Bu formüllere dayalı olarak Apriori'nin ölçüm yöntemleri aşağıdaki gibi sıralanabilir:

Güven Kurali (Rule Confidence): Basitçe kuralın arkasındaki güven değeridir.

$$e = C_{sonraki} = \frac{r}{a} \tag{2.45}$$

Güven Değerleri Farkı (Confidence Difference): Bu ölçüm önceki ve sonraki güven değerlerinin mutlak farkı olarak tanımlanır.

$$e = | C_{sonraki} - C_{önceki} | \quad (2.46)$$

Güven Oranı (Confidence Ratio): Sonraki güven değerinin önceki güvene oranlarından minimum değerli olanının 1 den çıkarılması ile bulunur.

$$e = 1 - min\left(\frac{C_{sonraki}}{C_{önceki}}, \frac{C_{önceki}}{C_{sonraki}}\right) \quad (2.47)$$

Bilgi Kazanımı Farkı (Information Difference) : Bu yöntem, C5.0 ağaçlarının yapımında kullanılana benzer bilgi kazanım kriterine dayanmaktadır. (IBM ,2015, syf. 10-11)

$$e = \frac{r.log\left(\frac{r}{a.c}\right)+(a-r)\,log\left(\frac{a-r}{a.\bar{c}}\right)+(c-r)\,log\left(\frac{c-r}{\bar{a}.c}\right)+(1-a-c+r)log\left(\frac{1-a-c+r}{\bar{a}.\bar{c}}\right)}{log(2)} \quad (2.48)$$

r kural desteği

a öncül destek

c ardıl destek

$\bar{a} = 1 - a$ öncül desteğin tamamlayıcısı

$\bar{c} = 1 - c$ ardıl desteğin tamamlayıcısı

2.4.4 FP-büyüme Algoritması

Çoğu zaman Apriori, adayı üret ve test et yöntemi ile aday setlerinin boyutunu önemli ölçüde düşürerek iyi performans kazancı elde edilmesini sağlar. Bununla birlikte, iki önemsiz maliyetten muzdarip olabilir:

- Hala çok sayıda aday seti üretmesi gerekebilir. Örneğin, 10^4 adet elemandan oluşan bir set varsa, Apriori algoritmasının 10^7'den fazla aday 2 maddeden oluşan bir set üretmesi gerekecektir.

- Tüm veritabanını art arda taramak ve desen eşleşmesine göre çok sayıda adayı kontrol etmek gerekebilir. Aday küme öğelerinin desteğini belirlemek için veritabanındaki her işlemin üstesinden gelinmesi maliyetlidir.

Bu kadar maliyetli bir aday küme üretim süreci olmadan tüm sık kullanılan küme setlerini barındıran bir yöntem tasarlayabilir miyiz?" Bu sorunun cevabı için geliştirilen, basitçe bir böl ve ele geçir stratejisi benimseyen FP-büyüme algoritması bulunmaktadır (Han ve Kamber, 2012).

Bu algoritmanın önceki çoğu algoritmadan daha etkili bir şekilde çalışarak maliyeti azalttığı görülmüştür. Bunun en büyük nedeni, tüm veritabanını daha küçük ve daha yoğun bir veri yapısı, sık örüntü ağacı içinde tutmasıdır. Apriori tabanlı algoritmalardan farklı olarak FP-büyüme içinde tüm veritabanı sadece iki kez taranır. İlki tüm ögelerin destek değerinin hesaplanması için, ikincisi ise ağaç yapısının oluşturulması içindir (Özdemir ve Özdoğan, 2010).

2.4.5 Carma Algoritması

Kullanıcıya her taramadan sonra oluşan kuralları gösterip, minimum destek ve güven seviyelerini değiştirme olanağı veren CARMA algoritması 1999' da Hidber tarafından ortaya atılmıştır (Oktay, 2009).

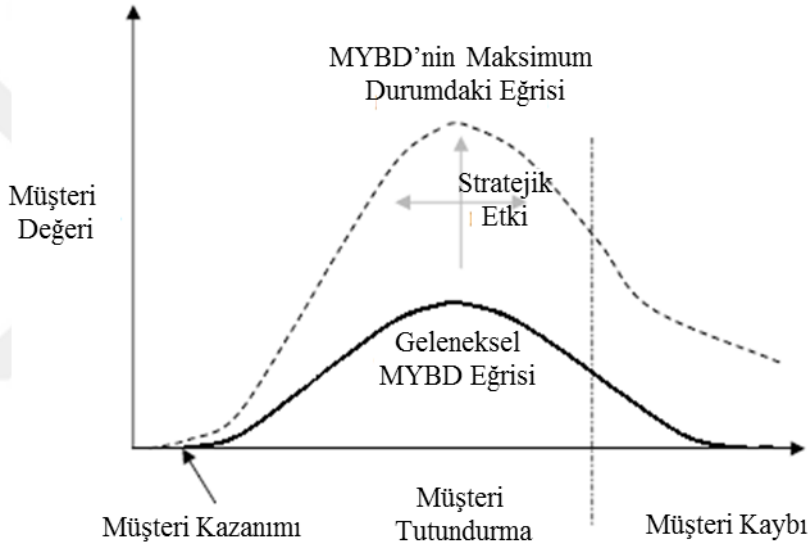
CARMA algoritmasının çalışma mantığı nesne kümelerinin hesaplanması işlemini (online) çevrimiçi olarak gerçekleştirir. Çevrimiçi çalışan CARMA algoritması veritabanı ilk taramasında herhangi bir işleminde kullanıcıya minimum destek ve minimum güven parametrelerini değiştirme imkanı sağlayarak, kullanıcıya online birliktelik kuralları çıkarma imkanı sunar. CARMA algoritması DIC algoritmasına benzer bir mantıkla çalışır. CARMA algoritmasında ilk tarama sırasında nesne kümeleri oluşturulur ve ikinci taramada ise nesne kümelerinin sayılması işlemi tamamlanır. Bu sebeple CARMA algoritması veritabanını en fazla 2 defa taramış olur. CARMA algoritması DIC algoritmasında farklı olarak hareketler üzerinden geçerken nesne kümelerini oluşturur (Yıldırım, 2016).

2.5 MÜŞTERİ DEĞERİ HESAPLAMA

Müşteri yaşam boyu değeri 1974 yılında Kotler tarafından müşterilerin belirli bir zaman diliminde yaptığı işlemlere dayanarak gelecekte beklenen kâr akışının bugünkü değeri olarak tanımlanmıştır (Swinnen ve diğ., 2013). Son yirmi yıl içinde müşteri değeri hesaplama uygulamalarının yaygınlaşmasıyla birlikte pazarlamada ürün odaklı yaklaşımdan, müşteri odaklı yaklaşıma geçiş yaşanmıştır. Bu yaklaşımda müşteriler değer olarak görülmekte, elde

edilmelerine ve elde tutulmalarına odaklanılmaktadır. Ancak müşterilerin hepsinin aynı değere sahip olmaması ve işletme kaynaklarının kısıtlılığı nedeniyle hepsiyle aynı yoğunlukta ilgilenilememesi nedeniyle müsterilerin işletme için değerleri ölçülerek en değerli ve en az kârlı müşterilerin tanımlanması bu doğrultuda pazarlama strateji ve programlarının yapılması gerekmektedir (Yapraklı ve Keser, 2008).

Müşteri yaşam boyu değerini (MYBD) ölçme ve yönetme sürecinin temel bileşenleri, müşteri kazanımı ve müşteriyi elde tutmadır. MYBD'yi hesaplamak için gereken diğer unsurlar, her müşterinin her periyotta sağladığı brüt katkı ve her periyotta her müşteriye yapılan pazarlama maliyetleridir (Kumar ve diğ., 2004).



Şekil 2.14: Müşteri Yaşam Boyu Değeri Yönetme Süreci (Kumar ve diğ., 2004)

Müşteri yaşam boyu değeri (MYBD) basitçe aşağıdaki gibi tanımlanıştır (Chen, 2018).

$$MYBD = \sum_{t=0}^{T} \frac{(P_t - C_t) r_t}{(1+i)^t} - AC \qquad (2.49)$$

P_t Tüketici tarafından t zamanında ödenen fiyat

C_t t zamanında müşteriye hizmet vermenin doğrudan maliyeti

i firma için iskonto oranı veya sermaye maliyeti.

R_t t zamanında müşterinin tekrar satın alma ihtimali

AC Müşteriyi elde etme maliyeti

T MYBD'yi tahmin etmek için gerekli zaman

Müşteri yaşam boyu değeri iki ana kategoriye ayrılabilir. Bunlar olasıksal ve deterministik yaklaşımları içeren modellerdir. İncelenen çalışmalarda birkaç eğilim gözlemlenmiştir. Bunlardan ilki, deterministik modellere göre stokastik modellere artan odaklanmadır. 2005'ten beri, deterministik bağlamda sadece iki tane modele karşı sekiz yeni stokastik model sunulmuştur. Olasılıklı modeller deterministik modellerden önemli ölçüde daha verimlidir. Bu eğilim bu nedenle mantıklı görünmektedir. Deterministik modeller, bir müşterinin durumunu belirleyici bir bağlamda tanımlamak için sıklıkla kullanılır. Bunların en iyi bilinenleri GFT (Güncellik, Frekans, Tutar) segmentasyonudur. Güncellik, bir müşterinin aktif olup olmadığını değerlendirmek için belirleyici faktördür. Müşteriler, az ya da çok geçerli eşikler temelinde bölümlere ayrılmıştır. Geleneksel olarak, kişi G,F,T değerlerine göre 3 seviyeye ayrılır ve bu şekilde müşteriler 27 segmenti temsil eden hale gelir. Bir müşteri ne kadar yakın zamanda bir satın alma işlemi yaptıysa, satın alma sıklığı o kadar yüksek olur ve ortalama sepet ne kadar yüksekse, beklenen potansiyeli o kadar yüksek olur (Castéran ve diğ., 2017).

İki ana kategoriye ayrılmasına ragmen bazı popüler algoritmalar aşağıdaki gibi tanımlanabilir.

Bağımsız Model Tahmini : Sabit bir brüt kâr ve saklama maliyeti ile gelecekteki geliri tahmin etmeye odaklanır. Formül aşağıdaki gibi tanımlanabilir.

$$CLVit = \sum_{t=1}^{T_i} \frac{GC_{it}}{(1+r)^{t/f_i}} - \sum_{l=1}^{n} \frac{\sum_m MC_{i,m,l}}{(1+r)^l} \qquad (2.50)$$

$GC_{i,t}$ = *t* anında *i* müşterisinden satın almadan kaynaklı brüt kâr

$MC_{i,m,l}$ = *i* müşterisinin *l* periyodunda *m* kanalından oluşturduğu pazarlama maliyeti

f_i = *i* müşterisi için satın alma sıklığı

r = indirim oranı

n = tahminlemek için yıl sayısı

T_i = *i* müşterisi tarafından satın alınan ürün miktarı

Model Parametrelerini Aynı Anda Tahmin Etmek: Farklı müşteriler pazarlama aktivitelerine farklı tepki verdiğinde, kâr marjı modelinin, her bir müşteri için regresyon ağırlıklarının farklı olmasına izin vererek yansıtması gerekir. Eşzamanlı olarak satın alma sıklığı, pazarlama maliyeti ve gelecekteki brüt kâr marjının modellenmesi ve müşteri heterojenliğinin hesaplanması bu sorunu çözmekte ve daha doğru sonuçlar vermektedir. Parametreleri aynı anda modelleyen hesaplama yöntemi bu işlemi Bayesian karar teorisini kullanarak yapar. Formülü aşağıdaki gibidir.

Benzerlik fonksiyonu:

$$L = \prod_{i=1}^{n} \prod_{j=1}^{J_i} \sum_{k=1}^{K} \emptyset_{ijk} [f_k (t_{ij} | \alpha_k, \beta_{ijk}, \gamma_k) * p (\Delta Q_{ij} | \delta_{i,k}, \delta_k^*, \sigma_k^2)]^{c_{ij}} S_k(t_{ij} | \alpha_k, \beta_{ijk}, \gamma_k)^{(1-c_{ij})} \quad (2.51)$$

$f(t_{ij}|\alpha_k, \beta_{ijk}, \gamma_k)$ = Genelleştirilmiş gamma dağılımı için yoğunluk fonksiyonu (2.52)

$S(t_{ij}|\alpha_k, \beta_{ijk}, \gamma_k)$ = Genelleştirilmiş gamma dağılımı için yaşam fonksiyonu (2.53)

$p(\Delta Q | \delta_i, \delta^*, \sigma^2)$ = Satın alma miktarı için yoğunluk fonksiyonu (2.54)

c_{ij} = Sansürleme inikatörü

Marka Değiştirme Yaklaşımı: Bu yaklaşım, önceki satın alma vesilesiyle satın alınan marka, farklı markaların satın alınma olasılığı ve müşterilerden belirli harcamalarından puanları hakkında bilgi toplamayı gerektirir. Markov geçiş matrisi sayesinde, müşterilerin bireysel düzeyde hizmet programlarına dayalı olarak bir markadan diğerine geçme olasılığı modellenmiştir. Formülü aşağıdaki gibidir.

$$CLV = \sum_{t=0}^{T_{ij}} \frac{1}{(1+d)^{\frac{t}{f_i}}} V_{ijt} * \pi_{ijt} * B_{ijt} \quad (2.55)$$

T_{ij} = i müşterisinin belirtilen süre içerisinde yaptığı alım sayısı

d_j = Firmanın indirim oranı

f_i = i müşterisinin birim zamanda yaptığı ortalama alım sayısı

V_{ijt} = i müşterisinin j markasını beklenen satın alma hacmi

π_{ijt} = *i* müşterisinin *j* markasının *t* satın alma biriminden beklenen katkı

B_{ijt} = *i* müşterisinin *j* markasının *t* satın alma olasılığı

Monte Carlo Simulasyon Algoritması: Sezgisel ve ampirik modeller, müşteri değeri ile ilgili önemli görüşler ortaya çıkarsa da, simülasyon modellerinin kullanımını da araştırılmıştır. Simülasyon yöntemlerini araştırmanın nedeni, gelecekteki müşteri karlılığını tahmin etmede nispeten başarısız olan modellerden kaynaklanmaktadır. Monte Carlo simülasyon formülü aşağıdaki gibidir.

$$p(\pi_{it}, Pur_{it}, X_{it}) = p(\pi_{it} | Pur_{it} = 1, X_{it}) * p(Pur_{it} = 1 | X_{it}) * p(X_{it}) \quad (2.56)$$

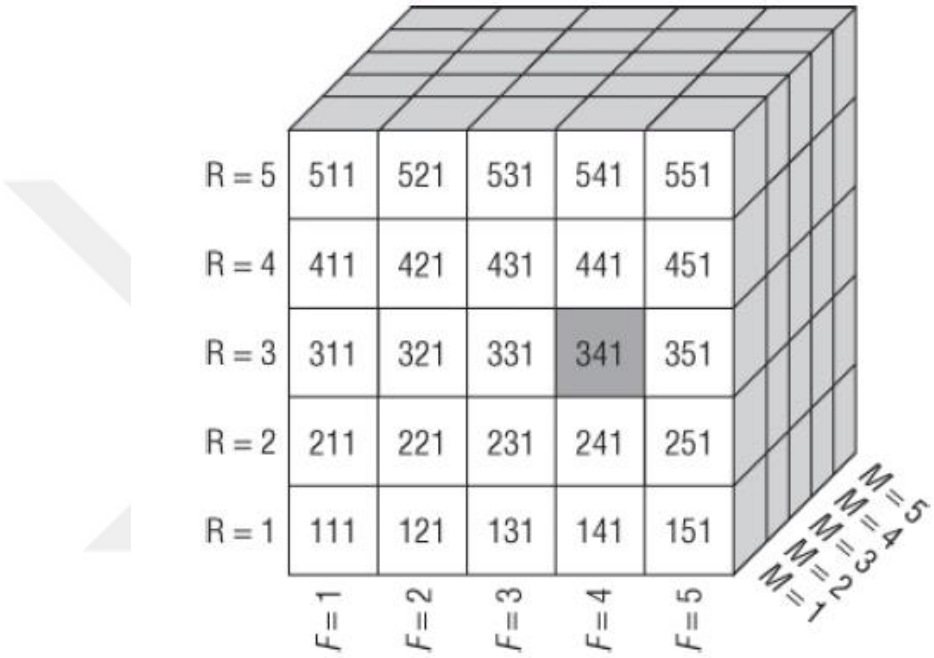
Pur_{it} = eğer i müşterisi t firmasından satın alırsa, satın almanın göstergesidir ve 1'e eşittir, aksi takdirde sıfırdır (Kumar, 2017).

2.5.1 GFT Modeli

GFT modelleri doğrudan pazarlamada 30 yıldan uzun süredir kullanılmaktadır. Sektördeki pazarlama aktivitelerinde düşük yanıt oranları göz önüne alındığında (tipik olarak %2 veya daha az), bu modeller, yanıt oranlarını iyileştirmek amacıyla belirli müşterilere pazarlama programlarını (örneğin; doğrudan posta) hedeflemek için geliştirilmiştir. Bu modellerden önce, şirketler genellikle hedefleme amacıyla müşterilerin demografik profil bilgilerini kullanmışlardır. Bununla birlikte, araştırmalar, geçmiş tüketici alımlarının, gelecekteki satın alma davranışlarının demografik özelliklerinden daha iyi tahmin ediciler olduğunu kuvvetle önermektedir (Gupta ve diğ., 2006).

Doğrudan pazarlama dünyasında, GFT (Güncellik, Frekans, Tutar) müşteri değeri hesaplama modeli için bilinen kısaltmadır. GFT'in arkasındaki mantık basittir. Yakın zamanda bir satın alma işlemi yapan müşterilerin yakın gelecekte bir satın alma işlemi yapma olasılığı daha yüksektir. Geçmişte birçok alım yapmış olan müşterilerin yakın gelecekte bir tane daha alım yapması muhtemeldir ve geçmişte çok para harcayan müşterilerin gelecekte daha fazla para harcaması daha olasıdır. GFT, potansiyel müşterilere uygulanabilecek bir yöntem değildir, çünkü yalnızca mevcut müşterilerin anlamlı bir güncellik, sıklık ve parasal değerleri vardır. GFT, yeni müşterileri çekmek yerine mevcut müşterileri değerlendirmek için kullanılan bir tekniktir.

Güncellik, son satın alma işleminden bu yana geçen gün veya hafta sayısı değerini atamak için kullanılır, genellikle değerleri üç gruba ya da beş gruba ayırarak kullanılır. Sıklık, genellikle önceki alımların toplam sayısı olarak tanımlanır. Müşterilere sıklık miktarına göre sıklık puanı verilir. Son değişken gelir, puanları oluşturmak için kullanılan toplam ömür boyu harcamadır. Aşağıdaki şekil her boyutun beşerli parçalara bölündüğü bir GFT küpünü göstermektedir (Berry ve Linoff, 2011).



Şekil 2.15: Beş Boyutlu GFT Küpü (Berry ve Linoff, 2011)

GFT yaklaşımının popülaritesi, gereken sınırlı bilgi ile müşterileri puanlamak ve bunları gruplara ayırmak için kolay bir yöntem olmasıdır. GFT analizi, sofistike bir yazılım veya analiste gerek duymaksızın kolayca yapılabilir. Bununla birlikte, bu yaklaşımın bazı sakıncaları vardır. Birincisi, GFT, öncelikle her bir müşteri için ayrı bir puan hesaplamak yerine müşterileri bir gruba atayarak bir bölümleme yapmaktadır. İkincisi, yalnızca geçmiş davranışlara odaklandığından, gelecekteki potansiyel veya gelişimsel büyümeyi göz önüne almaz. Ek olarak GFT ile ilgili en büyük sorun, bir müşterinin değerini belirleyen sadece üç değişken olmasıdır, müşterinin ne kadar zamanda, ne sıklıkta ve ne kadar harcadığını varsaymasıdır. Oysaki "en iyi" müşterileri belirleyen çok sayıda alternatif ve/veya tamamlayıcı faktör olabilir (Bejou ve diğ., 2013).

2.6 LİTERATÜR ARAŞTIRMASI

Bu bölümde veri madenciliği, kümeleme analizleri, birliktelik kuralları madenciliği, sadakat programları ile ilgili literatür çalışmaları kronolojik olarak özetlenmektedir.

Shen ve diğ. (2003) müşteri tutundurma faaliyetlerinin başlama zamanı, böyle bir çalışma için gerekli parametreler, kayıp müşteri tanımları gibi havayolu sektöründeki yapılacak çalışmalarda kullanılak girdilerin tanımları yapmışlardır. Bir müşterinin kişisel nedenlerden dolayı arz sahibi kurumu sorgulamaya başladığı an müşteri tutundurma aktivitelerine başlanması gereken zaman olarak tanımlanmıştır. P, Q, X, Y ve Z diye adlandırılan parametreler bir müşteri tutundurma analizi için gerekli olan parametreler olarak tanımlanır.

P: Müşteri sadakati tanımlamak için gerekli olan süre (minimum1 veya 2 yıl olmalı)

Q: Farklı uçuş desenlerini de gözlemlemek için gerekli olan minimum süre (mevsimsellik nedeniyle minimum 2 yıl olmalı)

X, Y : Aylık ortalama uçulan mil (veya sıklık veya gelir) den en az 2 tanesi.

Z: Önceden belirlenmiş iş tecrübesine göre ağırlık katsayısı

Uçulan mesafe, harcanan para, uçuş sıklığı gibi parametrelere göre kıyaslamalarla kayıp müşteri tanımları yapılmıştır. Müşteri değeri hesaplamak için klasik GFT göre daha sofistike bir yaklaşım önermiştir. Bu yaklaşım Metrik Müşteri Değeri Modeli diye adlandırılmıştır. Öncelikle sıklık değeri skorlanmıştır. Müşteriler uçuş sıklıklarına göre büyükten küçüğe sıralanarak eşit sayıda 4 farklı kümeye ayrılmıştır. Bu kümelerde yer alan müşteriler 1-4 aralığında puanlanmıştır. Aynı işlem gelir içinde uygulanmıştır. Gelir, gelir skoru, sıklık, sıklık skoru olmak üzere 4 parametreye göre 100 farklı sınıf yapılmıştır. Öncelikle müşteriler gelir skoru ve sıklık skoruna göre sıralanır. Bu işlemden sonra toplam veri kümesini 100 eşit parçaya bölerek metrik müşteri değeri modeli denilen sınıflandırmayı yapmıştır.

Tsai ve Chiu (2004) aynı kümedeki müşterilerin en yakın satın alma modellerine sahip olmasını sağlamak için bir genetik algoritma yaklaşımı benimsemiştir. Genetik algoritma kullanılarak benzer ürün gruplarını tercih eden müşteriler segmente edilmiştir. Segmentasyon tamamlandıktan sonra, her müşteri kümesinin göreceli karlılığını analiz etmek için Ağırlıklandırılmış bir GFT modeli kullanılmıştır (*G=0.2, F= 0.4, T=0.4*). *G,F,T* değerleri

öncelikle kendi içerisinde normal dağılıma uygun hale getirilmiş, sonrasında müşteri değeri hesaplanmıştır. GFT skoruna göre her müşteri segmentasyonu kendi içinde sıralanarak en karlı müşteriler belirlenmiştir.

Venkatesan ve Kumar (2004) müşteri yaşam boyu değerini satın alma sıklığı ve satın alma zamanlarını incelenerek tahminlemiştir. Kâr, satın alma sıklığı ve değişken maliyet tahminleri göz önüne alınarak müşteri yaşam boyu değeri;

$MYD_i = \sum_{y=1}^{T_i} \frac{CM_{i,y}}{(1+r)^{y/frekans_i}} - \sum_{l=1}^{n} \frac{\sum_m c_{i,m,l} \, x \, x_{i,m,l}}{(1+r)^{l-1}}$, formülasyonu ile hesaplanmıştır.

MYD_i : Müşteri Yaşam Boyu Değeri

$CM\dot{I}_{i,y}$: Müşteriden elde edilmesi ön görülen kar marjı

r : İndirim oranı (çalışmada %15 alınmıştır)

$C_{i,m,l}$: m kanalıyla l yılındaki maliyet

$X_{i,m,l}$: Belirtilen tarih aralığındaki müşteri sayısı

$Frekans_i$: Her müşteri için öngörülen satın alma sıklığı

n: tahminlenecek yıl sayısı

T_i: Planlama yapılan dönemin sonuna kadar satın alınanların tahmini

Malthouse ve Blattberg (2005) çalışmalarında müşterilerin gelecekteki karlılığının ne kadar doğru tahmin edilebileceğine dair detaylı bir değerlendirme sunar. 20–55 ve 80–15 diye adlandırılan kurallar önerilmiştir. En değerli % 20'nin müşterimizin yaklaşık %55'i ileriki dönemlerde değerli olmayacak geri kalan % 80'inin, yaklaşık %15'i ise değerli olacaktır. Müşteri Yaşam değeri, indirim oranını da hesaba katarak $y_i = \sum_{t=1}^{T} c_{it} \, (1+d)^{-t}$ formülü ile hesaplanmıştır. Bu denklem aynı zamanda bir regresyon modelinin $(y_i) = f(x_i) + e_i$, fonksiyonu olarak düşünülmüştür. Çalışmanın temeli, bu şekilde baz alınarak müşteri segmentlerine göre GFT modelinin versiyonları da kullanılmıştır. Her yıl tahminlenen müşteriler ile gerçekleşen arasındaki farklarda gözetilerek bir sonraki yılın tahmini yapılmıştır. 20–55 ve 80–15 diye adlandırdığı kuralların doğruluğu 4 farklı sektörde yapılan araştırmalarla test edilmiştir.

Fader ve diğ. (2005) çalışmalarının temel amacı farklı satın alma geçmişleri olan fakat benzer müşteri değerlerine sahip müşterilerin kümelenmesini sağlayan iso-değer eğrileri

kavramıdır. 23570 müşterinin 78 haftalık verileri incelenmiştir. Veri kümesi 39 haftalık 2 parçaya bölünmüştür. Ilk 39 haftalık veri üzerinden GFT değeri hesaplanarak, daha sonraki satın alma davranışlarına yönelik tahminde bulunulmuştur**.**

Tsay ve Chiang (2005) kümeleme temelli birliktelik kuralları (CBAR) olarak adlandırılan etkin bir algoritma geliştirmiştir. Literatürdeki birliktelik kuralları algoritmalarında performans problemlidir. Çünkü birliktelik kurallarını keşfetme sürecinde veritabanı aday öğe kümesini tüm öğe kümesini tekrar tekrar tarayarak kıyaslar. Veritabanındaki tarama sayısını ve kıyasların sayısını azaltabilecek alternatif bir yöntem performansı iyileştirecektir. CBAR yönteminde veritabanının bir kez taranması yeterlidir, sonrasında kısmi küme tabloları ile kıyas yeterlidir. Bu şekilde sadece veri taramaları için gereken zaman azalmayacak aynı zamanda keşfedilen sonuçların doğruluğu da artacaktır. Makalede önerilen CBAR yöntemi ve Apriori algoritması ile birlikte yapılan bir deneysel çalışma sunulmuş, geliştirilen algoritmanın etkinliği Apriori ile kıyaslanmıştır. Deneyler FoodMart veri seti yapılmıştır ve deney sonuçlarına göre CBAR algoritması Apriori algoritmasından daha iyi sonuç vermiştir. Keşfedilen örüntülerin sayısı ve boyutu arttıkça algoritmalar arasındaki performans farkı daha da aşikâr olmuştur. Geliştirilen algoritmanın geniş veritabanlarından kural keşfinde etkin olduğu görülmüştür.

Gupta ve diğ. (2006) 6 farklı müşteri değeri hesaplama modelinin algoritmasını incelemiştir. Bu 6 model; GFT, olasık temelli modeller, ekonometrik model, kalıcılık modeli, bilgisayar bilimi modelleri, difüzyon/büyüme modelleri olarak adlandırılmaktadır. Aynı zamanda ilerleyen dönemlerde araştırma alanı haline gelebilecek 11 yeni modelden bahsedilmiştir.

Kim ve diğ. (2006) müşteri değerini analiz etmek ve müşterileri değerlerine göre segmentasyon için bir çerçeve önermiştir. Kore'de bulunan bir kablosuz telekomünikasyon şirketinin 6 aylık verisi ile analiz gerçekleştirilmiştir. Veri seti 70-30 eğitim ve test verisi olarak bölümlenerek incelenmiştir. Çalışmada müşteri değeri, potansiyel müşteri değeri, terketme olasılıkları hesaplanmıştır. Müşterilerin demografik ve hesaplanan verilerine göre karar ağacı oluşturularak her segment müşteri için farklı pazarlama stratejileri önerilmiştir.

Wong ve Chung (2007) Tayvan Havayolları'nın iç hat yolcuları üzerinde müşteri segmentasyonu çalışması yapmıştır. GFT çalışmasında yolcuların G,F,T değerleri için her

değerinin modundan küçükler için 0, mod değerinden büyükler için 1 değeri atanmıştır. Mod değerlerini hesaplamak için G,F,T değerleri için belirli kabuller yapılmıştr. Örneğin sıklık için; Haftada 1 den fazla uçanlar için 1, 1 kez uçanlar için 2, 2-3 haftada bir uçanlar için 3 gibi değerler kullanılmıştır. Toplam GFT değeri oluşturulurken ise G,F,T değerleri toplanmıştır. Müşteri değer skorları minimum 0, maksimum 3 olmuştur . Oluşturulan müşteri değeri, müşteri demografik bilgileri, müşteri uçuş davranışları ve tercih sebepleri de ele alınarak C.5 karar ağacı algoritması oluşturularak müşteriler segmente edilmiştir. Yapılan veri madenciliği çalışmasında SPSS Clementine programı kullanılmıştır.

Cavique (2007) Smilis ve Apriori algoritmalarını karşılaştırmışlardır. Veriler öncelikle grafik tabanlı bir yapıya dönüştürülmüştür. Daha sonrasında en sık kullanılan ürün kümelerini bulmak için metasezgisel bir yaklaşım olan Smilis algoritması kullanılarak problem çözülmüştür. Apriori algoritmasıyla Smilis yaklaşımın avantaj ve dezavantajları karşılatırılarak çalışma tamamlanmıştır. Apriori algortiması için SAS Enterprise Miner, Smilis modeli için C++ programı kullanılmıştır.

Maaoluf ve Mansour (2007) havacılık endüstrisinde kümeleme ve birliktelik analizleri üzerine çalışmıştır. Kümeleme aşamasında k-ortalamalar algoritması kullanılarak müşteriler 9 kümeye ayrılmıştır. Son 1 yıllık mil kazanım aktivite sayısı, üyeliği boyunca yapılan mil kazanım aktivite sayısı, üyeliği boyunca kazanılan mil, üyelik yaşı (ay), Mil kazanım/Üyelik yaşı, Yapılan aktivite sayısı/Üyelik yaşı gibi nitelikler kümeleme analizinde kullanılmıştır. Havayolu sadakat programı verisinde havayolu, otel, finansal hizmet ilişkisi Apriori algoritmasıyla incelenmiştir. Daha sonra beraber tercih edilen destinasyonlar Apriori algoritması ile tespit edilmiştir. Bu çalışma yapılırken bütün destinasyonları sepete atıp, sonuçlarını gözlemlemek yerine çıkış noktası sabit tutularak o rotadan yapılan yolculuklar dikkate alınmıştır. Bu çalışmada; Oracle data miner programı kullanılmıştır.

Sohn ve Kim (2008) Kore'de telekomünakasyon sektöründe ek hizmetler üzerine Mart - Mayıs 2001 tarihleri arasındaki 17 bin kayıtlık müşteri verisi kulllanılarak yapılan bir çalışmadır. Faktör analizi, kümeleme ve birliktelik kuralları kullanılmıştır. 115 farklı değişken temel bileşenler analizine göre analiz edildiğinde 5 değişkenin toplam varyansın %75'ini oluşturduğu gözlenmiştir. Bu 5 değişkene göre Newton method kullanılarak yapılan değerlemeye istinaden k-ortalamalar algoritması ile müşteriler 8 kümeye bölünmüştür. Destek değeri %9, güven değeri %80 alınarak 8 kümenin birliktelik kuralları incelenmiştir.

Teichert ve diğ. (2008) 5829 yolcuyla Conjoint anket değerlendirme yöntemiyle değerlendirilen ankette müşterilerin uçuş nedenlerinin iş veya eğlence için gerçekleştirildiğinin ayrımının yapılmadığı sonucuna varmıştır. Bu sebeple tüketici tercihlerine uygun teklifler üretilmediği dile getirilmiştir. 5 farklı yolcu segmentine göre ücret, tarife, sadakat programı, ikram, yer hizmetleri, zamanında kalkış değerlemesi yapılmıştır. Aynı zamanda yolcuların iş ve eğlence amaçlı uçuşlarında hangisinin önem arz ettiğine dair bir çıkarım yapılmıştır. 40 yaş üstü business kabinde seyahat eden yolcular için iş uçuşlarında en önemli etken dakiklik olurken, eğlence amaçlı uçuşlarda ücret/performans önem arzetmektedir.

Sun ve Bai (2008) birliktelik kuralları sonucunda ortaya çıkan destek değerinin ağırlıklandırılması üzerine çalışma yapmıştır. Birliktelik kurallarına göre aynı destek değerine sahip ürünler, değerli ürünlerle bulunma olasıklarına göre ağırlık almaktalar ve bu şekilde değerlendirilmektedirler. Gerçek hayat uygulamasında ise Netflix'de 480.189 müşterinin 17.770 film üzerinde yaptıkları değerlendirmelerin yer aldığı yaklaşık 100 milyonluk bir veri seti kullanılmıştır. İlk önce en popüler 10 film seçilmiştir. Ağırlıklandırma bu filmlerle birlikte tercih edilme olasılığına göre yapılmıştır. Sonuç olarak incelendiğinde ise ilk 19 birliktelik kuralından 15 i ortak olmasına ragmen 4 farklı kural daha keşfedilmiştir.

Wu ve diğ. (2008) veri madenciliğinde en etkili 10 modelin (C4.5, k-Ortalamalar, SVM, Apriori, EM, PageRank, AdaBoost, kNN, Naive Bayes, and CART) tanımlanması, etkisi ve mevcut ve ileri araştırma seviyeleri incelenmiştir. Dezavantajlarına rağmen, k-ortalamalar en yaygın kullanılan kümeleme algoritmasıdır. Algoritma basit, kolay anlaşılabilir ve kolayca başka problemlere uygulanabilir niteliktedir. Veri madencilerinin yapmaya çalıştığı ilk şey Apriori benzeri algoritma ile deneme yapmaktır. Apriori'nin geliştirilmesi amacıyla algoritmayı tüm veri setine uygulamak yerine tüm verinin rastgele örneklenmiş küçük bir alt kümesini inceleyerek her farklı küme için farklı algoritmalar geliştirmişlerdir. Apriori'de en göze çarpan gelişme, aday jenerasyonun ortadan kaldırılmasında başarılı olan FP-Büyüme (sık rastlanan büyüme) ve FP-ağaç (Sık desen ağacı) adı verilen yöntemlerin ortaya çıkmasıdır. Bu yöntemler öncelikle tüm veri setini tarayarak ilişkisel kuralları bulmaktadır. İkinci taramada ise bu kuralları oluşturan alt kümeleri bulmaya yönelik çalışmadır. Bu sayede alt kümelerde daha anlamlı analizler keşfedilmiştir.

Baesens ve diğ. (2009) veri madenciliği ve yöneylem araştırmasının birbirleri arasındaki ilişkiyi ve bunların bir dizi eğilim ve zorluklarını incelemiştir. Veri madenciliğinin 1941 yılında yöneylem araştırmasından doğduğu ifade edilmektedir. Markov zincirleri, Monte Carlo simulasyonu gibi yöneylem araştırması tekniklerinin veri madenciliğinde de kullanıldığı, yöneylem temelli tahminleme metodojilerinin veri madenciliğindeki kullanımlarından (lineer regresyon ve zaman serileri analizi) bahsedilmiştir. Bu çalışmada; veri kalitesi, veri madenciliği modellerinin yorumlanabilirliği, alan bilgisinin veri madenciliği sürecine dahil edilmesi, test etme, ağ temelli öğrenme ile ilgili birçok önemli konu ve mevcut eğilimler ve bir dizi yeni uygulama alanı tespit edilmiştir. Bunların çoğu için, veri madenciliğinin, yöneylem araştırması alanından tekniklerin kullanılmasından yararlanmaya devam edebileceğinden bahsedilmiştir.

Ngai ve diğ. (2009) müşteri ilişkileri yönetiminin ana faktörleri olan müşteri tanımlama, müşteri gelişimi, müşteriye yönelik aksiyonlar, müşteri tutundurmanın altında yapılan 87 akademik çalışmayı sınıflandırmıştır. Müşteri tanımlama, müşteri segmentasyonu ve hedef grup analizlerini, müşteri gelişimi müşteri yaşam değeri, birliktelik kuralları ve üst/çarpraz satış analizlerini; müşteriye yönelik aksiyonlar direk pazarlamayı; müşteri tutundurmada ise sadakat programı, geribildirim yönetimi ve birebir pazarlama faaliyetlerini alt başlık altında ele almıştır. Kullanılan veri madenciliği tekniklerini ise birliktelik analizleri, sınıflandırma, kümeleme, tahminleme, ardıl desen keşfi, görselleştirme ve regresyon olarak 7 başlık altında tanımlamıştır. MİY ana faktörleri, MİY alt faktörleri, veri madenciliği fonksiyonları ve kullanılan teknikler olarak 4 başlık altında ilgili makaleler sınıflandırılmıştır.

Cheng ve Chen (2009) GFT özelliklerinin kantitatif değerine ve k-ortalamalar algoritmasına kaba küme teorisini dahil ederek, anlamlı kuralları çıkarmak için yeni bir prosedür önermektedir ve bu şekilde mevcut algoritmaların dezavantajlarının etkin bir şekilde iyileştirebileceğini ifade etmektedir. GFT hesaplanırken G,F,T değerlerine göre sırayla veriler büyükten küçüğe göre sıralanır ve 5 eşit parçaya bölünür ve 1-5 aralığında puanlar atanır. GFT değerleri Maksimum 125, minimum 1 değerini almaktadır. K-ortalamalar algoritmasında k=5 değeri alınmıştır. Bu adımlardan sonra LEM2 algoritması kullanılarak kurallar oluşturulmuştur. Çalışmada, önerilen yöntemin karar agaçları ve yapay sinir ağlarına göre yüksek başarı oranları elde ettigi sonucuna varılmıştır.

Jiale ve Huiying (2010) Çinli bir havayolu verileri üzerinden analiz yapmıştır. Mevcut müşteri değeri toplam gelirden, toplam maliyetin çıkarılması yoluyla elde edilen değere göre

hesaplanmaktadır. Frekans değeri 5 farklı, en son işlem yaptığı tarih 8 farklı kategoriye bölümlendirilerek puanlanmıştır. GFT değerleri analitik hiyerarşi süreci ile birleştirilerek müşterilerin potansiyel müşteri değerleri hesaplanmıştır.

Trnka (2010) sepet analizi sonuçlarının 6 Sigma aşamalarından birine uygulanmasıyla, sonuçları iyileştirilmesi ve sürecin Sigma performans seviyesini değiştirilmesi üzerine bir çalışma önermiştir. 6 sigmanın tanımla, ölç, analiz et, geliştir, kontrol döngüsünde "geliştir" aşamasında sepet analiziyle iyileştirme hedeflenmiştir. Pazar analizi sonuçlarına göre müşteri davranışlarını öngörerek üretim sürecinin ana ürünlerini ve müşteri gruplarını belirlemede kullanılmıştır.

Forgas ve diğ. (2010) havayolu kullanıcı sadakatinin öncüllerini tanımlamaktadır. Barselona-Londra uçuşlarını gerçekleştiren üç havayolu şirketinin kullanıcı anket sonuçlarına göre, temel sadakat öncülünün duygusal sadakat olduğu ortaya konulmuştur. Duygusal sadakatin ana önceliklerinin ise memnuniyet ve güven olduğu ortaya çıkmıştır. Duygusal sadakatin aynı zamanda havayolu şirketi ile kullanıcıları arasındaki ilişkilerin başarısını garanti altına almak ve uzun vadeli satın alma davranışını anlamak için anahtar değişken olduğu ortaya çıkmıştır.

Birant (2011) 15 kayıtlık örnek bir veri seti üzerinden segmentasyon için yeni bir yöntem ortaya koymuştur. Değerler G,F,T değerlerine göre büyükten küçüğe sıralanarak 5 eşit parçaya bölünmüştür. Her bir değer için En yüksek değerli müşteriler 5, en düşük değerliler 1 değerini almıştır. En iyi müşteri 555, en kötü müşteri ise 111 değerini almıştır. Kümeleme yapılırken G,F,T değerlerinin 5 üzerinden ortalaması alınmıştır. Her küme de ortalamadan büyük değerler yukarı, ortalamadan küçük değerler aşağı yönlü oklarla 8 farklı müşteri kümesi tanımlanmıştır. Geliştirilen bu yöntem 2666 müşteriye ait veri setine uygulanmıştır.

Miguéis ve diğ. (2011) çalışmalarında homojen müşteri gruplarına yönelik promosyon tasarımını incelemişlerdir. Bu kapsamda ilk adımda müşterileri frekans ve gelir odaklı bir GFT modeli kullanılarak müşteri değeri hesaplamışlardır. Daha sonra, k-ortalamalar metodu ile kümeleme gerçekleştirmişlerdir. Optimum küme sayısı, Davies-Bouldin indeksi ve Elbow eğrisi yöntemleri kullanılarak 5 olarak belirlenmiştir. Her segmentin alışveriş davranışlarıyla ilgili olarak birliktelik analizi yapılmıştır.

Dolnicar ve diğ. (2011) havayolu sadakat programlarının kilit öncül noktalarını bulmak için 687 yolcunun vermiş olduğu cevaplara dayalı bir araştırma gerçekleştirmiştir. Sonuçlar, havayolu sadakat programı üyeliğinin, bilet fiyatının, ulusal bir taşıyıcı olma statüsünün ve havayolu şirketinin çevrede algılandığı itibarının, havayolu şirketine sadık olan ve olmayanlar arasında en iyi ayrımcı olan değişkenler olduğunu göstermektedir. Hesaplanan modellerin hiçbiri için memnuniyet, havayolu sadakatinin ana itici gücü olarak ortaya çıkmamıştır.

Chiang (2011) GFT modeline dayalı yeni bir müşteri değeri hesaplama yöntemi ortaya koymuştur. Bu yöntem; GFT modeline indirim oranı ve maliyetin geri dönüş parametresi de eklenerek oluşturulan bir modeldir. G,F,T,D,R değerleri 1-5 skalasında değerlendirilmiştir. Daha sonra, Apriori algoritması kullanılarak da hangi skorlara sahip müşterilerin hangi pazarı tercih ettiğine yönelik analizler gerçekleştirilmiştir.

Khajvand ve diğ. (2011) bir sağlık ve güzellik firmasının müşteri segmentasyonunda müşteri yaşam değerini kullamıştır. Bu çalışmada; iki yaklaşım kullanmıştır: İlk yaklaşımda, müşterilerin segmentasyonu için GFT analizi, ikinci yaklaşımda ise genişletilmiş GFT analizi yöntemi kullanılmıştır. GFT analiziyle birlikte k-ortalamalar metoduyla kümeleme yapılmıştır. En iyi küme sayısının belirlenmesi için Dunn indeksi kullanılmıştır. Bu indekse göre k=4 bulunmuştur. GFT için G,F,T değerleri 5 kategoriye ayrılmıştır. (çok düşük, düşük, orta, yüksek, çok yüksek). Kategorilerin puanlarına göre 4 farklı müşteri segmenti oluşturulmuştur.

Khajvand ve Tarokh (2011) İran menşeili bir bankanın 2008 ilkbahar- 2009 yaz ayları arasındaki 5000 satırlık bir veri setini kullanarak, GFT ve k-ortalamalar kümeleme metodunu birlikte kullanarak bir kümeleme çalışması yapmıştır. Küme sayısını belirlemede Dunn indeksi kullanılarak optimum k sayısı 4 olarak bulunmuştur. G,F,T değerleri bulunduktan sonra doğrusal olarak dönüştürülerek toplam müşteri değeri hesaplanmıştır. Müşterilerin ait oldukları küme belirlendikten sonra geçmişe yönelik 6 dönem boyunca ilgili kümedeki müşterilerin yaşam boyu değeri hesaplanmıştır. Bu sayede trend gözlemlenerek, gelecek yılın potansiyel müşteri değerleri tamninlenmiştir.

Kristiani ve diğ. (2013) Endonazya'da gelirlerinin %75'i, %20'lik yolcu portfoyünden gelen bir havayolunda GFT modeli kullanarak müşteri değerini hesaplamış ve hesaplanan

değerle müşterinin sosyo-demografik profili arasındaki ilişkiyi analiz etmiştir. Çalışmada; GFT ve k-ortalamalar algoritmaları birlikte kullanılmıştır. GFT modeli kullanılırken temel bileşenler analizinden de yararlanılmıştır. Temel bileşenler analizi çalışması G, F, T değerlerinin hangi ağırlıkta modele katılması hususunda yardımcı olmuştur. *Müşteri değeri = a*(R) +b*(F)+ c*(M)* formülasyonuyla hesaplanmıştır. Güncelliği her 45 günlük periyodun değerlendirilmesine göre puanlamışlardır. Son 45 günde uçanlar en yüksek puanı almıştır. K-ortalamalar algoritmasındaki küme sayısına karar vermek içinse R^2 analizi yapılmıştır. Bu analize göre; 7 farklı küme olmasına karar verilmiştir. Yaş, üyelik seviyesi, çalışma pozisyonu ile GFT skoru karşılaştırması yapılarak çalışma tamamlanmıştır.

Nikumanesh ve Albadvi (2014) İran'da bir bankaya ait 25 değişkenli 12.359 satırlık veri setini incelemiştir. Çalışmada ağırlıklı GFT yöntemi kullanılarak müşteri değeri hesaplaması yapılmıştır. Çalışmanın detaylarında, ağırlıkları saptamak için 30 tane üst düzey yöneticiye anket yapılmıştır, anket sonuçlarına göre G,F,T değerlerinin ağırlıkları hesaplanmıştır. Ortalamaya göre G, F, T değerleri karşılaştırılarak 8 farklı küme oluşturulmuştur.

Coussement ve diğ. (2014) veri kalitesinin önemli bir boyutu olan veri doğruluğu sorunlarının etkisini, doğrudan pazarlama açısından öne çıkan üç önemli segmentasyon tekniği (GFT, lojistik regresyon ve karar ağaçları) ile incelemiştir. Optimum veri doğruluğu kısıtı altında CHAID tekniğinin GFT ve lojistik regresyona göre kullanılması önerilmektedir. GFT'in, düşük yanıt verme oranına sahip veri setinde lojistik regresyona göre daha iyi performans sergilemesine karşın orta ve yüksek dosya derinliği olan veri setlerinde lojistik regresyon tekniği tercih edilmektedir. Genel olarak, bu sonuçlar, üç segmentasyon tekniğinin de, veri doğruluğundaki sorunlara duyarlı olduğunu göstermektedir.

Hu ve Yeh (2014) müşteri kimlik bilgileri olmadan GFT müşteri modellerini bulmayı amaçlamıştır. Ürün özelinde kurulan GFT modelleri ile ürünleri şirket açısından katma değeri ortaya konulmuştur.

Sukri ve diğ. (2014) Malezya'daki Malezya Hava Yolları (tam servis havayolu) ve AirAsia (düşük maliyetli havayolu) arasındaki müşteri memnuniyeti ve müşteri sadakati arasındaki farkları incelenmiştir. Kuala Lumpur'daki iki büyük havayolu terminalinde 152 anket formu üzerinden analiz yapılmıştır. Sonuç olarak Malezya'daki Malezya Hava Yolları

(tam hizmet havayolu) ve AirAsia (düşük maliyetli havayolu) arasındaki müşteri memnuniyeti ve müşteri sadakati ile ilgili ücret duyarlılığı, hizmet kalitesi ve verilen hizmetler olmak üzere üç boyutun kritik olduğu ortaya konulmuştur. Düşük maliyetli havayolunun yolcuları ücretten memnunken, tam servis havayolu yolcuları hizmetin kalitesini önemsemektedir.

Akamavi ve diğ. (2015) servis aksaklığı yaşamış 286 yolcuyla anket yaparak, düşük maliyetli havayollarında kilit öncüllerin yolcu sadakati üzerindeki etkisini incelemektedir. Bu çalışmanın sonucunda, havayolunda müşteriyle temas eden çalışanların kalitesinin, yalnızca müşteri geri kazanımını ve fiyatını olumlu yönde etkilemekle kalmayıp aynı zamanda şirketin yolcu gözündeki güvenini de arttırdığını göstermektedir. Çalışanın yeterliliği, kötü hizmet deneyimlerini azaltır ve yolcu memnuniyetini arttırmaktadır. Bu çalışmada, yolcu sadakatini yükseltmenin en üst faktörünün fiyat değil yolcu memnuniyeti olduğu ortaya konulmuştur.

Dursun ve Caber (2016) GFT tekniği kullanılarak Antalya'da faaliyet gösteren 5 yıldızlı 3 otele ait müşterileri gruplandırmıştır. Çalışma yapılırken öncelikle otellerdeki veriler bir merkezi müşteri veritabanı projesi gibi birleştirilerek bir müşterinin bütün otellerde yaptığı konaklamalar ele alınmıştır. 6000 kişilik bir büyüklüğün %95 güvenilirlik ± 0,05 hata payı ile örneklem büyüklüğü (362) hesaplanmıştır. Çalışma bu örneklem üzerinden yapılmıştır. Çalışmada; yolcuların Nisan 2014-Nisan 2015 yılları arasındaki aktiviteleri baz alınarak GFT skorları hesaplanmıştır. Skorlara göre, k-ortalamalar metoduyla yapılan gruplandırmada 8 adet grup ortaya çıkmıştır. Daha sonra her grubun kendi içerisindeki ortalama *G,F,T* değerlerinin ortalaması, örneklemdeki bütün verilerin *G,F,T* ortalamasına göre kıyaslanarak *G,F,T* değerlerinin ortalamadan büyük, küçük olduğuna karar verilmiştir. *G,F,T* deki büyük küçüklüğe göre de 8 grup isimlendirilmiştir. *G+,F+,T+* ise sadık müşteriler, *G-,F+,T+* sadık sezon müşterileri, *G-, F+,T-* toplu alım müşterileri, *G+,F+,T-* kış sezon müşterileri, *G-,F-,T-* kayıp müşteri, *G-,F-,T+* yüksek potansiyelli müşteriler, *G+F-,T-* yeni müşteriler, *G+,F-,T+* kış sezonu potansiyel müşteriler olarak kümelenmiştir.

Insani ve Soemitro (2016) bir telekomünakasyon şirketinin 2015 yılı son çeyrek verilerini kullanılarak GFT modeli ile müşteri değerini hesaplamıştır. Hesaplanan GFT değerine göre k-ortalamalar ve kohonen ağlarını kullanılarak segmentasyon yapılmıştır. Bu iki algoritma karşılaştırılmıştır. K-ortalamalar algoritmasında k değeri 4 alınmıştır. Kümelerin kalitesi karşılaştırılırken Silhouette indeksi kullanılmıştır. Kullanılan veri seti için k-

ortalamalar algoritmasının daha doğru sonuçlar verdiği görülmüştür. Çalışmada aynı zamanda müşterilerin telefon, internet ve kablo TV servisinin kullanım durumuna yönelik sepet analizi de yapılmıştır.

Çalışır ve diğ. (2016) havayollarında kurumsal görüntü, memnuniyet, fiyat ve hizmet kalitesi gibi faktörlerin tam hizmet sağlayıcı taşıyıcılara ve düşük maliyetli taşıyıcılara yönelik yolcu sadakati üzerindeki etkilerini analiz etmeyi amaçlamaktadır. Yolcu sadakatini önemli derecede etkileyen faktörleri belirlemek için yapısal eşitlik modellemesi yaklaşımı uygulanmıştır. LISREL kullanılarak, önerilen modeli test etmek için Frankfurt'tan İstanbul'a seyahat eden ve Atatürk Havalimanı'na gelen bagaj bekleme alanında yüz yüze görüşmelerde bulunan 237 yolcu ile gerçekleştirilmiştir. Bu 237 anketten yüz yetmiş beşi geçerli olarak kabul edilmiştir. Sonuç olarak, yolcu sadakatinin % 71'inin kurumsal görünüm ile açıklandığını göstermektedir. Ayrıca, hizmet kalitesi ve fiyatın memnuniyet üzerinde olumlu etkisi olduğu tespit edilmiştir. Fiyatla karşılaştırıldığında, hizmet kalitesinin daha güçlü bir memnuniyet belirleyicisi olduğu tespit edilmiştir.

Sarvari ve diğ. (2016) GFT analizini kullanarak kapsamlı bir müşteri kümeleme modelini oluşturmak için ağırlıklandırma faktörleri ve demografik değişkenler gibi diğer bazı özellikleri de değerlendirerek müşteriler farklı gruplarda farklı GFT skorlarına göre kümelemiştir. GFT Skoru ve demografik özelliklerin birlikteliklerine göre 42 farklı senaryo bulunmaktadır. Bu çalışmada; k-ortalamalar ve SOM (Self Organizing Map) kümeleme yöntemleri kullanılarak karşılaştırılmıştır. Dunn indeksine göre k-ortalamalar hem küme kalitesi hem zaman olarak en iyi çözümü sunduğu için bu algoritma kullanılmıştır. Bu aşamadan sonra birliktelik kuralları analizi yapılmıştır. 3 farklı (Apriori, Eclat, FP-Büyüme) algoritmaları incelenerek en iyi birliktelik kuralı modeli seçilmiştir.

Chen ve diğ. (2016) temel bileşenler analizini kullanarak ağırlıklandırılmış müşteri değeri hesaplama konusunu incelemişlerdir. 12 farklı değişken temel bileşenler analizi yöntemine göre analiz edildiğinde 9 tane değişkenin modele eklenmesine karar verilmiştir. Bu 9 değişken verinin %98'ini yansıtmaktadır. Müşteriler GFT değerine göre 8 segmente ayrılmıştır. Buradaki modele GFT'deki değerlere ek olarak ortalama indirim faktörü denilen bir değişken daha eklenmiştir.

Solnet ve diğ. (2016) konaklama sektöründe kısıtlı bir kullanım alanı olan sepet analizi misafirlerin oda tercihinden çok hangi ilişkisel ürün gruplarını tercih ettikleri üzerine bir analiz yapmıştır. 2009 ile 2014 yılları arasında 5 yıldızlı bir otelin 5 yıllık toplamda 119.244 kayıtlık verisi incelenmiştir. Her bir müşteri için otelde satın alınan yiyecekler, içeçekler, tercih edilen oda ve diğer giderler (kuru temizleme, internet, park) olmak üzere 4 farklı kategorideki satın almaları incelenmiştir. SAS programı kullanılarak modeller kurulmuştur.

Weng (2017) ağırlıklandırılmış F,T değerleri kullanılarak müşteri değeri hesabı ve müşteri gruplarının satın alma alışkanlıklarını Apriori algoritmasını kullanarak incelemiştir. Ağırlıklandırma işlemi F ve T değerlerini 0-1 aralığında bir değere getirerek çarpılması sonucunda bulunmaktadır. En küçük sıklık veya yenilik değerlerine sahip müşteriler sıfır değerini almaktadır. Sonrasında, Apriori algoritması kullanılarak satın alınan ürünler arasındaki ilişkiler incelenmiştir.

Chiang (2017) GFT ve FSLC (frekans, yüksek sezon, ikametgah, seyahat zamanı, seyahatleri iptal zamanı) yöntemlerine dayalı veri madenciliği tekniklerini kullanarak müşteri segmentasyonu ve bu segmentlere yönelik ilişkilendirme kuralları üzerine çalışmıştır. Her bir parametre 1-5 skalasında değerlendirilmiştir. Müşterilerin demografik ve FSLC skorlarına yönelik tercih ettikleri pazarlar birliktelik kurallarına göre analiz edilerek çalışma tamamlanmıştır.

Özçalıcı (2017) Apriori algoritmasını kullanarak 2016 yılının Temmuz ve Ağustos aylarından 2 aylık bir veri seti üzerine çalışmıştır. Araçların donanımsal, multimedya, performans, fiyat özellikleri 0-1 veri seti haline dönüştürülmüştür. Çıkan sonuçlara göre araçların belli fiyat aralıklarında en çok hangi özelliklere sahip olduğu hangi ikili özelliğin birlikte daha çok tercih edildiği gibi bilgiler elde edilmiştir.

Song ve diğ. (2017) GFT analizi ile zaman serileri analizini birlikte öneren bir çalışma gerçekleştirmiştir. Telekom sektöründe 1 milyonluk müşterinin 481 Milyonluk arama verisi üzerine çalışma yapılmıştır. Çalışmada program olarak SPARK kullanılmıştır. Öncelikle GFT analizinde uç değerlerin fazlalığından kurtulmak ve buna sebebiyet vermemek için çok ölçütlü analiz (MCA) yapılmıştır. Çalışmada GFT değerleri 8 farklı zaman aralığında değerlendirilerek ele alınmıştır. Sonuç itibariyle GFT nin her bir değerinin zaman serileri yöntemiyle bölümlenmesi üzerine istatistik tabanlı bir yaklaşım önerilmiştir.

Meyer ve diğ. (2017) olasıklı ve deterministik müşteri yaşam boyu değeri hesaplama modellerini anlatmıştır. Deterministik modellerde Müşteri Tutundurma oranı t-1 zamanı ile t zamanı kohort analizine göre alınan satın alımların oranıdır. Kayıp müşteri oranı ise bu değerin 1 den çıkartılmasıyla bulunur. Deterministik modellerde müşteri değeri hesaplama $MYD = \sum_{t=1}^{n} \pi\,(t) \frac{\rho^t}{(1+d)^t}$ formülü ile hsaplanmaktadır.

$\pi\,(t)$: t periyodunda elde edilen kar

ρ : Müşteri tutundurma oranı

d : İndirim Oranı

Olasıksal modeller ise parametrik ve yarı parametrik modeller olmak üzere iki farklı grupta ele alınmıştır.

Pakyürek ve diğ. (2018) k-ortalamalar ve GFT modelini kullanarak kümeleme yapmışlardır. GFT'de yer alan özniteliklerin dağılımına bakılarak harcanan para 4 kademede, sıklık 4 kademede ve güncellik ise 3 kademede (aktif, potansiyel ve kayıp) değerlendirilmiştir. Herbir factor 1-4 arasında puanlanarak çarpımı sonucunda GFT değeri hesaplanmıştır. Silhoutte katsayısı kullanılarak en uygun küme sayısı bulunmuştur. Çıkan sonuçlar k-ortalamalar ve GKM (Gaussian Karısımı Modeli) yöntemleri ile kümelenmiş ve sonuçları kıyaslanmıştır. Ayrıca k-ortalamanın, GKM'den daha iyi performans gösterdigi gözlemlenmiştir.

Ferhatosmanoglu (2018) havayolu endüstrisinde müşteri yaşam değeri değerini tahmin etmek için müşterilerin sosyal ağları ve uçuş bilgileri entegre edilerek oluşturulan yeni bir model geliştirilmiştir.İlk aşamada müşteri yaşam değerini hesaplamak için bir regresyon modeli uygulanmıştır, daha sonra ise bu temel modeli müşterilerin yaptıkları dolaylı katkıları da içerecek şekilde müşterinin sosyal çevresiyle ilgili girdilerle zenginleştirilmiştir. Önerilen yöntem sadece uçuş verisiyle yapılan analizle karşılaştırılarak doğruluğun ve güvenilirliğinin daha fazla olduğu gözlemlenmiştir. Müşteri değeri hesaplamak için çoklu doğrusal regresyon yöntemi kullanılmıştır.

Chen (2018) Çin'de bir havayolu şirketi için müşteri yaşam boyu değeri modeli geliştirmiş ve bu modelin TravelSky skorları ile uyumunu incelemiştir. Uçuş sıklığı, indirim

seviyesi, ödeme seviyesi, toplam mil miktarı ve uluslar arası uçuşları gibi ölçütlerle TravelSky yolcuların 0-100 arasında bir değerlendirmesini aylık olarak yapmaktadır. Çalışmada geliştirilen yolcu değerlendirme modeli skoru ile TravelSky skorları eşleştirilerek havayolunun müşteri değerini zamanında ve maliyet etkin olarak tespiti ve uygun müşteri segmentasyon ve pazarlama hizmetlerinin sunumunun sağlanması amaçlanmıştır.

Qadadeh ve Abdallah (2018) 2 farklı veri madenciliği tekniği üzerinden kümeleme yapmıştır. K-ortalamalar kümeleme tekniği yapay sinir ağları temelli SOM (self-organizing map) tekniği ile kombine edilerek bir kümeleme yapılmıştır. 9822 müşteri kayıtlı veri tabanında 86 farklı özelliği ele alınmıştır. 86 özelliğin 43 tanesi profil ve demografik bilgilerden oluşurken 43 tanesi aktiviteleri içeren özelliklerdir. K-ortalamalar algoritmasında en iyi küme sayısını bulmak için Elbow metodu kullanılmıştır ve veri seti 5 kümeye ayrılmıştır.

Christy ve diğ. (2018) GFT analiziyle birlikte k-ortalamalar, bulanık c-ortalamalar ve mevcut k-ortalamalar algoritmalarını ufak değişiklikler yaparak, yeni bir model GT k-ortalamalar algoritmaları kullanmış ve kümeleme yapmıştır. Önerilen yeni algoritma daha az yineleme ve daha az zamanla bir kümelemeyi önermektedir. GFT hesabı yapılırken G, F, T değerleri 5 farklı kategoriye ayrılmıştır. En iyiden en kötüye olmak üzere 5'den 1'e puanlanmıştır. Daha sonra, hesaplanan GFT değerleri 1-125 arasında değerler almıştır. Kümeleme kısmında ise 10 kümeye bölünmüştür. 3 metodla yapılan kümeleme sonuçları karşılaştırılmıştır.

Khalili-Damghani ve diğ. (2018) k-ortalamalar yöntemini kullanarak ilk aşamada bir kümeleme yapmışlardır. İkinci aşama olarak ise filtrelemeler yapılarak hibrit özellik seçim yöntemi ve TOPSIS çok özellikli karar verme yöntemini incelenmişlerdir. Yönteme dayalı olarak eğer-sonra kuralları oluşturulmuştur. Oluşturulan kurallar sigorta ve telekomünikasyon sektörlerinde uygulanmıştır. K-ortalamalar algoritmasında en iyi küme sayısı için Davies–Bouldin indeksi, aykırı değerler ve boş değerler için LOF (Lokal uçdeğer faktörü) metodu ve özellik seçimi için Shannon entropi metodu kullanılmıştır. Bu metot birbirleri arasındaki korelasyonu yüksek değişkenlerin, en düşük seçim ağırlığına sahip olduğuna dayanmaktadır. Bu çalışmadaki bütün analizler RapidMiner programı kullanılarak yapılmıştır.

Ilham ve diğ. (2018) FP-Büyüme algoritması ile pazar sepet analizi yapmışlardır. Berkah Mart, Pekanbaru şehrinde perakende sektöründe faaliyet gösteren, ürünlerin yerleşimi

geleneksel yöntemlerle yapan müşteri bakış açısı dikkate alınmayan bir perakedencidir. Bu çalışmada, Fp-Büyüme algoritması ile ürünlerin yerleşimi ve ürünlerin ulaşılabilir olmasının planlamasını sağlayacak çok sayıda birliktelik kuralı çıkarılmıştır. Çalışmada hem apriori algoritması hem de FP-Büyüme algoritması ile elde edilen sonuçlar kıyaslanmıştır. Yapılan 7 deney için FP-Büyüme algoritmasının daha etkin olduğu görülmüştür. Çıkarılan kurallar ile müşteriler için özel promosyonlar tanımlanması mümkündür. Deneysel sonuçlar kullanılan algoritmanın hızlı ve etkin bir şekilde tüketicilerin alışveriş örüntülerini tespit ettiği ve bu sayede marketin karlılığının artırılabileceği görülmüştür.

Ponyiam ve Arch-int (2018) hedeflenen müşterilerin satın alma davranışını analiz ederek ürün tercih alışkanlıklarına ilişkin yüksek boyutlu verilerde uygulanabilir gelişmiş bir yaklaşım önermektedir. Yöntem 3 aşamaya ayrılmıştır. İlk aşamada k-ortalamalar kullanarak müşterilerin kümelere ayrılmıştır. Bu çalışmada, müşterilerin kümelere ayrılmasında temel parametreler toptan, yiyecek alımı ve endüstriyel ürün satın alma alışkanlıklarıdır. En iyi küme sayısı Elbow yöntemi ile 5 olarak belirlenmiştir. İkinci aşamada ise Apriori algoritması kullanılarak birlikte satın alınan ürün grupları belirlenmiştir. Üçüncü aşamada ise ürün satın almalarının uzmanlar tarafından değerlendirildiği aşamadır.

Sagin ve Ayvaz (2018) perakende sektöründe faaliyet gösteren büyük bir donanım şirketinin beş buçuk yıllık verilerine ilişkin pazar sepeti analizi yapmış ve birbiriyle ilgili ürün kategorileri belirlenmiştir. İlişkilendirme kurallarının belirlenmesinde, hem Apriori hem de FP-büyüme algoritmaları ayrı ayrı çalıştırılarak faydaları karşılaştırılmıştır. Ek olarak, veri seti 2 parçaya bölünmüştür, böylece kuralların tutarlılığı ardışık zamanlanmış olan ikinci veri setinden türetilen kurallarla türetilen ilk veri setinden elde edilen kuralların doğruluğu karşılaştırılmıştır. 81.384 fatura verisi incelenerek 3.081 farklı ürün için çalışma yapılmıştır. Çalışmada Weka programı kullanılmıştır. Sonuç olarak, 1.veri setinde FP-Büyüme algoritması Apriori algoritmasına göre 14 kat daha hızlı sonuç vermesine ragmen ortaya çıkan kurallar anlamında zayıf kalmıştır. 2. veri setinde ise benzer sonuçlara ragmen Apriori algoritması 41 kat daha yavaş sonuç vermiştir.

Chiang (2018) bulanık C-ortalamalar ve Apriori kullanılarak farklı pazarlar özelinde hangi algoritmaların daha iyi sonuç verdiği üzerine çalışmıştır. Bu araştırmanın sonuçları ve önerilen veri madenciliği yaklaşımı; bir pazar için bulanık c-ortalamalar yöntemiyle, geriye kalan 3 pazar için birliktelik kuralları kullanılması gerektiğini ortaya koymuştur.

3. MALZEME VE YÖNTEM

3.1 UYGULAMANIN AMACI

Bu tez çalışmasında bir havayolu şirketinin örnek müşteri verileri üzerinden ilk olarak GFT metodolojisi ile müşteri değeri hesaplanmaştır. Daha sonra, GFT sonuçlarına göre k-ortalamalar, SOM (Öz Örgütlemeli Haritalar) ve iki aşamalı kümeleme yöntemleri uygulanarak segmentasyon yapılmıştır. Kümeleme performansları Silhoutte indeksi kullanılarak kıyaslanmıştır. En iyi kümeleme yöntemi ve küme sayısına göre belirlenen her segment için ayrı ayrı yurtiçi uçuşlar özelinde Apriori algoritması kullanılarak birliktelik kuralları ortaya çıkarılmıştır. Çalışmada IBM SPSS Modeler 17.0 programı kullanılmıştır.

Elde edilen sonuçların kampanya yönetim sistemine girdi olarak kullanılması hedeflenmiştir. Örneğin; GFT modeline göre yapılan segmentasyonda küme-4 de yer alan bir müşterinin uçuşlarında Gaziantep (GZT) ve Antalya'ya (AYT) uçuş yaptığı fakat İzmir'e (IZM) uçuş yapmadığını gözlemlenmiştir. Geçmiş verilerden elde ettiğimiz sonuçlara göre Gaziantep ve Antalya'ya birlikte uçan müşterilerin tamamının İzmir'e de uçtuğu gözlemlenmiştir. O yüzden uçak doluluk durumlarına göre uygun bir zamanda bu müşteriye İzmir uçuşu belirli bir indirim veya ayrıcalıkla teklif edilebilir. Bu sayede yapılacak kampanyalarda kampanya geri dönüş oranı yükseltilecektir.

3.2 VERİLERİN TANITILMASI VE ÖZELLİKLERİ

Kullanılıcak veri setimiz müşterilerin profil ve uçuş bilgilerini içeren iki ayrı tablodan oluşmaktadır. Profil tablosunda müşterilerin müşteri numarası, cinsiyet, doğum tarihi ve üyelik tarihi yer almaktadır. Müşterilerin ortak özelliği olarak ise tüm müşteriler İstanbul ikamet adresine sahiptir Analiz toplamda 2726 müşteri verisi ile gerçekleştirilecektir. Profil tablomuzun örnek görüntüsü aşağıda gösterilmiştir.

1	ID	Cinsiyet	Dogum_Tarihi	Üyelik_Tarihi
2	5050605101	F	4/22/2003	6/23/2012
3	5050611401	F	6/5/1983	6/7/2012
4	5060108973	F	8/1/1990	6/8/2012
5	5060129749	F	4/5/1988	6/9/2012
6	5060166611	M	9/1/1981	5/30/2012
7	5060179841	F	9/5/1978	5/28/2012
8	5072297184	F	4/13/1949	5/22/2012
9	5072482250	F	5/21/1985	6/7/2012
10	5072669913	M	9/15/1977	6/27/2012
11	5072704241	F	8/12/1972	6/16/2012
12	5072744008	M	5/15/1981	6/20/2012
13	5072842134	F	6/24/1945	5/18/2012
14	5072842365	F	6/28/1984	5/23/2012

Şekil 3.1: Müşteri Profil Bilgileri Tablosu

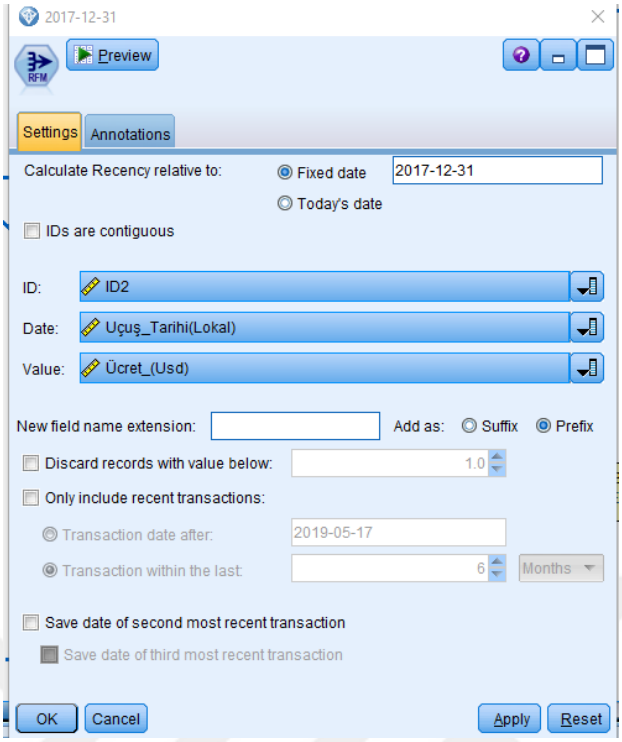
Yolculara ait uçuş bilgisinde ise müşteri numarası, yolculuğun iç hat veya dış hat olma bilgisi, yolculuğun başlangıç noktası, varış noktası, uçuş tarihi ve uçuşun ücreti bilgileri yer almaktadır. örnek görüntüsü aşağıdaki gibidir. Veri setimiz 1 Mayıs 2012 ile 31 Aralık 2017 tarihleri arasındaki uçuşları içeren toplam 100.801 kayıttan oluşmaktadır.

ID2	DOM/INT	Çıkış Noktası(Şehir)	Varış Noktası(Şehir)	Ücret_(Usd)	Uçuş_Tarihi (Lokal)
5050605101	DOM	IST	ADA	17.6	12/27/2016
5050611401	DOM	IST	ADA	19.8	12/17/2017
5060129749	DOM	IST	AYT	71.44	5/18/2012
5060166611	DOM	IST	AYT	15.68	5/4/2015
5072297184	DOM	IST	MLX	63.44	12/12/2017
5072482250	DOM	IST	MLX	58.35	6/24/2014
5072848280	DOM	IST	MLX	73.41	6/26/2012
5072916229	DOM	IST	VAS	38.61	12/6/2017
5078327299	DOM	IST	VAS	16.77	12/12/2017
5078330617	DOM	IST	XHQ	95.4	6/24/2014
5078366758	DOM	IST	XHQ	91.5	6/22/2014

Şekil 3.2: Müşteri Uçuş Bilgileri Tablosu

3.3 GFT İLE MÜŞTERİ DEĞERİ HESAPLANMASI

Veri setimiz yukarıda bahsettiğimiz üzere 31 Aralık 2017 tarihine kadar olan verileri içermesi sebebiyle müşteri değeri hesaplanırlen bu tarih baz alınarak hesaplanacaktır. Müşteri değeri hesaplanırken IBM SPSS Modeler 17.0 programında yer alan "RFM Aggregate" işlemcisi kullanılmıştır. Bu işlemcinin ayarlarını içeren ekran görüntüsü aşağıda verilmiştir.



Şekil 3.3: SPSS Modeler RFM Aggregate Ayarlar Görüntüsü

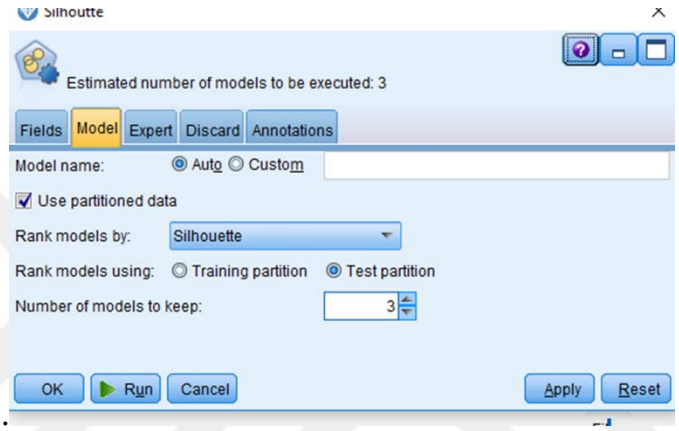
"RFMAggregate" kullanılarak yapılan müşteri değeri hesaplamasında her bir müşteri için güncellik, sıklık ve gelir değerleri hesaplanmıştır. Müşteriler için hesaplanan değerlerin örnek bir kesiti Şekil 3.4'de sunulmaktadır.

	ID2	Recency	Frequency	Monetary
1	2888792154	51	35	5044.350
2	3311631067	294	64	9312.410
3	6144926840	9	133	25179.780
4	3198203578	55	93	36558.740
5	5061237534	9	63	12042.860
6	5783203982	29	91	22780.030
7	5965031355	79	58	7114.590
8	3279799533	42	16	3672.510
9	3309303602	78	18	2049.160
10	3236271727	102	20	2489.290

Şekil 3.4: SPSS Modeler RFM Aggregate Çıktı Görüntüsü

3.4 EN İYİ KÜMELEME YÖNTEMİ SEÇİMİ VE KÜME SAYISININ BELİRLENMESİ

ID2, Recency, Frequency, Monetary adlı kolonları içeren yeni veri setimiz programda yer alan "Auto Clustering" işlemcisi ile analiz edilerek veri setine en uygun model ve küme sayısı belirlenmesi amaçlanmıştır. "Auto Clustering" işlemcisinin ayarlarını içeren ekran görüntüsü aşağıda verilmiştir. Bu aşamada; k-ortalamalar, iki aşamalı kümeleme ve öz örgütlenmeli kümeleme yöntemleri kullanılmıştır



Şekil 3.5: SPSS Modeler Auto Clustering Ayarlar Görüntüsü

İlgili işlemcinin çalıştırılması sonucunda elde edilen görüntü ise aşağıdadır.

Use?	Graph	Model	Build Time (mins)	Silhouette	Number of Clusters	Smallest Cluster (N)	Smallest Cluster (%)	Largest Cluster (N)	Largest Cluster (%)	Smallest/ Largest	Importance
☑		K-m...	< 1	0.619	5	70	2	1556	57	0.045	0.0
☐		Two...	< 1	0.444	4	132	4	1677	61	0.079	0.0
☐		Koh...	< 1	0.41	12	24	0	624	22	0.038	0.0

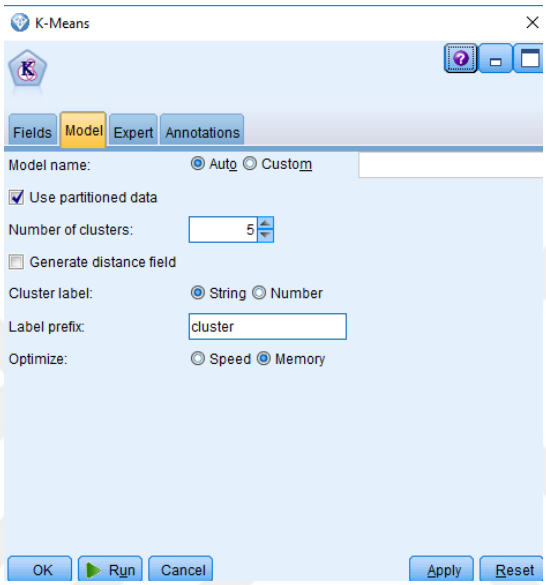
Şekil 3.6: SPSS Modeler Auto Clustering Çıktı Görüntüsü

Veri setimizin k-ortalamalar, iki aşamalı kümeleme ve öz örgütlenmeli kümeleme yöntemleri ile kümelenmesi durumunda Silhoutte indeksi değerleri sırasıyla "0.619 , 0.444, 0.41" olarak elde edilmiştir. Elde edilen sonuçlara göre silhouette değeri en yüksek olması sebebiyle en iyi kümeleme yöntemi k - ortalamalar ve optimum küme sayısı ise 5 olarak belirlenmiştir.

4. BULGULAR

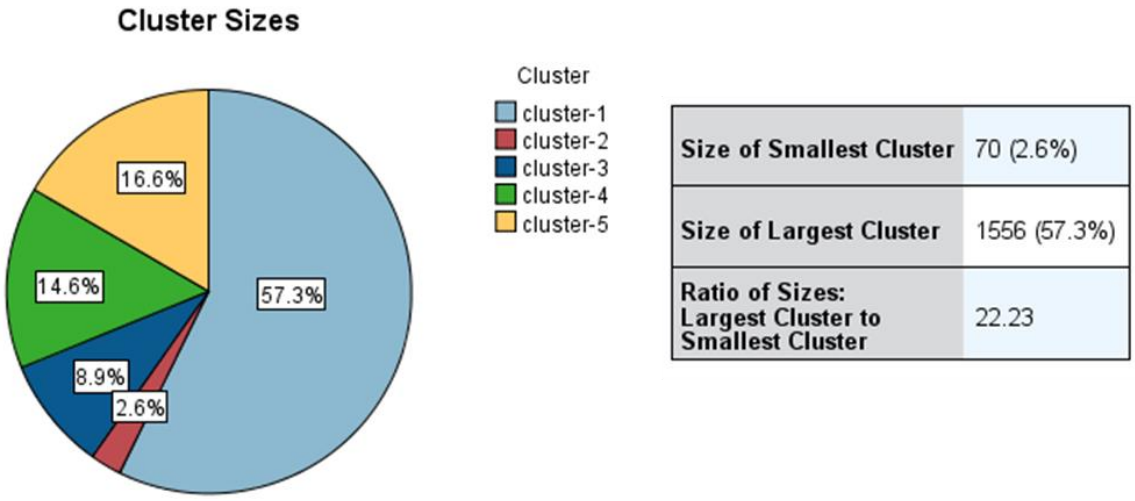
4.1 K-ORTALAMALAR METODUYLA KÜMELEME

Optimum küme sayısı 5 olmak üzere programda yer alan "k-ortalamalar" işlemcisinin ayarlarını içeren ekran görüntüsü aşağıdadır.



Şekil 3.7: SPSS Modeler k-ortalamalar Ayarlar Görüntüsü

Model çalıştırıldıktan sonra çıkan sonuca göre; en küçüğü 70 öğe, en büyüğü 1556 öğe içeren 5 küme oluşmuştur.



Size of Smallest Cluster	70 (2.6%)
Size of Largest Cluster	1556 (57.3%)
Ratio of Sizes: Largest Cluster to Smallest Cluster	22.23

Şekil 3.8: SPSS Modeler k-ortalamalar Çıktı Görüntüsü

Küme merkezleri ve eleman sayıları aşağıdaki gibidir:

	Güncellik (G)	Frekans (F)	Tutar (T)	Kümedeki Eleman Sayısı
Küme - 1	108	26	2.753	1.556
Küme - 2	36	211	26.005	70
Küme - 3	1.613	9	901	243
Küme - 4	49	86	9.963	396
Küme - 5	606	20	2.187	452

Kümelerin herbir nitelik için minimum, maksimum, standart sapma, medyan. varyans, 1.çeyrek, 3. çeyrek değerleri aşağıdaki tabloda sunulmaktadır.

Tablo 4.1: K-ortalamalar yöntemiyle ayrılan kümelerin istatistiki bilgiler tablosu

	Küme-1	Küme-2	Küme-3	Küme-4	Küme-5
Güncellik_Min	0,0	0,0	1.112,0	0,0	353,0
Güncellik_Maks	363,0	704,0	2.052,0	436,0	1.127,0
Güncellik_Standart Sapma	90,0	95,6	235,3	64,9	180,4
Güncellik_Medyan	89	10,5	1.594,0	26,5	556,0
Güncellik_Varyans	8.104,5	9.133,6	55.358,4	4.199,2	32.546,7
Güncellik_1.Çeyreklik	33,0	3,0	1.418,0	10,0	463,5
Güncellik_3.Çeyreklik	160,0	25,0	1.795,0	60,5	727,5
Frekans_Min	3,0	100,0	2,0	32,0	3,0
Frekans_Maks	63,0	426,0	55,0	161,0	149,0
Frekans_Standart Sapma	13,5	59,0	7,8	27,4	17,6
Frekans_Medyan	24,0	201,0	6,0	77,0	14,0
Frekans_Varyans	181,2	3.486,3	60,6	752,7	310,1
Frekans_1.Çeyreklik	16,0	176,0	4,0	64,0	9,0
Frekans_3.Çeyreklik	36,0	245,0	11,0	105,0	25,0
Tutar_Min	53,0	8.762,1	43,9	1.865,9	52,5
Tutar_Maks	14.843,5	90.991,5	5.429,9	36.558,7	20.926,2
Tutar_Standart Sapma	2.149,0	16.741,9	1.019,5	5.763,0	2.758,9
Tutar_Medyan	2.211,1	21.761,4	510,2	8.309,5	1.344,3
Tutar_Varyans	4.618.357,4	280.292.026,6	1.039.305,2	33.212.043,3	7.611.738,4
Tutar_1.Çeyreklik	1.162,7	15.098,3	263,5	6.162,8	560,0
Tutar_3.Çeyreklik	3.845,3	29.837,0	1.113,6	12.052,5	2.674,9
Kümedeki_Eleman_Sayısı	1556	70	243	396	452

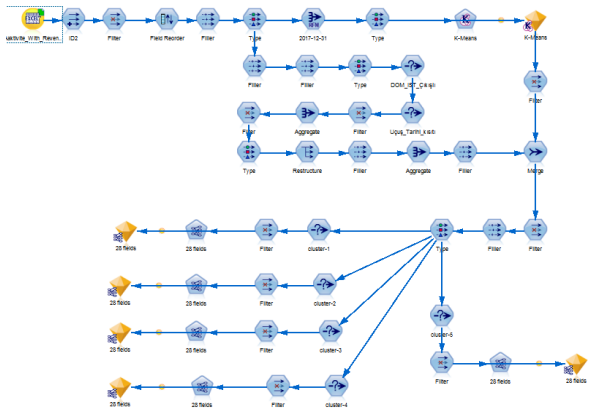
4.2 APRIORI ALGORİTMASI İLE BİRLİKTELİK KURALLARI KEŞFİ

Apriori algoritması oluşturulurken kabullerimiz Shen ve arkadaşlarının 2003 yılında US Patent Ensitüsünden aldıkları patentte yer alan kıstaslara göre oluşturulmuştur. Bu kıstaslar

analiz edilecek müşterinin en az 1-2 yıllık bir müşteri olması gerektiğini, analiz edilecek uçuş verisinin sezonsallıktan dolayı minimum 2 yıllık olması gerektiğini ve veri setinin parçalanmadan bütünsel yıllar alınarak gözlemlenmesi gerektiğini aktarmıştır. Bu bilgiler ışığında kabullerimizi sıralayacak olursak;

- Müşterilerimiz İstanbul ikametli oldukları için çıkış noktası İstanbul olan uçuşları analiz edilmiştir.
- Minimum üyelik yaşı 2 olan üyelerin 2 yaşından sonraki uçuşları yani 1 Temmuz 2014 sonrasındaki uçuşlar analiz edilecektir.
- Sezonsallık etkisini kaçırmamak için 1 Temmuz 2017 yılına kadar olan uçuşlar analiz edilecektir.
- Algoritmamızda destek değeri minimum %7, güven değeri minimum %40 olacak şekilde alınmıştır.

Veri setini model uygulamaya uygun hale getirmek için aşağıdaki gibi bir veri madenciliği süreci uygulanmıştır.



Şekil 3.9: SPSS Modeler Apriori Algoritması Model Görüntüsü

Veri hazırlama sürecinin sonunda veri setimiz Apriori algoritmasında kullanılmak üzere hazır hale gelmiştir. Kolonlarda ilgili varış noktalarının 3'lü il kodları yer almaktadır.

	ID2	$KM-K-Means	ADA	ANK	ASR	AYT	BAL	BJV	DIY	DLM	DNZ	ERC	ERZ	EZS	GZT	HTY	IZM	KCM	KSY	KYA	MLX	MQM	MSR	MZH	NAV	OGU	SZF	TZX	VAN	VAS
	1414413710	cluster-5	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	1414415432	cluster-1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	1414422915	cluster-4	1	1	1	1	0	0	0	0	1	0	0	0	0	0	1	0	0	1	1	0	0	0	0	1	1	1	1	1
	1414437328	cluster-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
	1414457257	cluster-5	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
	1414480952	cluster-1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
	1414483563	cluster-1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	1	0	0
	1414490801	cluster-1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
	1414490955	cluster-1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
	1414491739	cluster-1	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Şekil 3.10: SPSS Modeler Apriori Algoritması Veri Görüntüsü

Küme-1 için Apriori algoritmasını minimum destek değeri %7, minimum güven değeri %40 olacak şekilde düzenleyip çalıştırdığımızda çıkan en güçlü kurallar aşağıda gösterilmektedir.

Tablo 4.2: Küme-1 için Apriori Sonuçları

Ardıl	Öncül	Destek	Güven	Lift
IZM	AYT ve ANK	9,8	46,6	1,3
ANK	KYA	7,6	43,5	1,4
ANK	AYT ve IZM	10,5	43,4	1,4
IZM	GZT	11,0	41,9	1,2
IZM	ASR	9,9	41,3	1,2
IZM	ANK	31,2	41,0	1,1
IZM	BJV ve AYT	7,8	40,7	1,1

İlk çıkan sonuca göre yorumlayacak olursak ;

- Antalya (AYT) ve Ankara(ANK) uçan yolcuların %46,6 sinin İzmir (IZM) yolculuğu yaptığı da gözlemlenmiştir.
- Antalya (AYT) , Ankara(ANK) ve İzmir (IZM) uçuşlarının birlikte bulunma eğilimi, tüm uçuş davranış desenlerinden % 9,8 inde beraber görülmektedir.

Küme-2 için Apriori algoritmasını minimum destek değeri %7, minimum güven değeri %40 olacak şekilde düzenleyip çalıştırdığımızda çıkan en güçlü kurallar aşağıda gösterilmektedir. Bu kurallar güven değeri 100 olan sonuçlardan destek değerine göre en değerli 10 kuraldır.

Tablo 4.3: Küme-2 için Apriori Sonuçları

Ardıl	Öncül	Destek	Güven	Lift
IZM	AYT and ADA	42,9	100,0	1,3
IZM	AYT and ADA and ANK	41,4	100,0	1,3
IZM	TZX	40,0	100,0	1,3
IZM	ASR and ADA	38,6	100,0	1,3
IZM	KYA and ADA	37,1	100,0	1,3
IZM	SZF and ADA	37,1	100,0	1,3
IZM	TZX and ADA	35,7	100,0	1,3
IZM	TZX and ANK	35,7	100,0	1,3
IZM	GZT and AYT	35,7	100,0	1,3
IZM	ASR and ADA and ANK	35,7	100,0	1,3

İlk çıkan sonuca göre yorumlayacak olursak ;

- Antalya (AYT) ve Adana (ADA) uçan yolcuların tamamının İzmir (IZM) yolculuğu yaptığı da gözlemlenmiştir.

- Antalya (AYT) , Adana(ADA) ve İzmir (IZM) uçuşlarının birlikte bulunma eğilimi, tüm uçuş davranış desenlerinden % 42,9 sında beraber görülmektedir.

Küme-3 için Apriori algoritmasını minimum destek değeri %7, minimum güven değeri %40 olacak şekilde düzenleyip çalıştırdığımızda çıkan en güçlü kurallar aşağıda gösterilmektedir.

Tablo 4.4: Küme-3 için Apriori Sonuçları

Ardıl	Öncül	Destek	Güven	Lift
DLM	TZX	25,0	100,0	2,7
TZX	DLM	37,5	66,7	2,7
ANK	IZM	25,0	50,0	2,0
IZM	ANK	25,0	50,0	2,0
AYT	ANK	25,0	50,0	1,3

İlk çıkan sonuca göre yorumlayacak olursak ;

- Dalaman (DLM) uçan yolcuların tamamının ve Trabzon (TZX) yolculuğu yaptığı da gözlemlenmiştir.
- Dalaman (DLM) ve Trabzon (TZX) uçuşlarının birlikte bulunma eğilimi, tüm uçuş davranış desenlerinden %25 inde beraber görülmektedir.

Apriori algoritmasını minimum destek değeri %7, minimum güven değeri %40 olacak şekilde düzenleyip çalıştırdığımızda güven değerine göre en değerli 10 kural aşağıda sunulmaktadır.

Tablo 4.5: Küme-4 için Apriori Sonuçları

Ardıl	Öncül	Destek	Güven	Lift
IZM	BJV ve ADA ve AYT	7,3	96,6	1,5
IZM	KYA ve ADA ve AYT	8,8	94,3	1,5
IZM	KYA ve GZT	8,3	93,9	1,5
IZM	GZT ve ADA ve AYT ve ANK	7,3	93,1	1,4
IZM	TZX ve ADA ve AYT ve ANK	7,1	92,9	1,4

IZM	GZT ve ADA ve AYT	9,3	91,9	1,4
IZM	GZT ve AYT ve ANK	9,1	91,7	1,4
ANK	TZX ve ADA ve AYT	7,8	90,3	1,6
IZM	ASR va ADA ve AYT ve ANK	7,8	90,3	1,4
IZM	ASR va ADA ve AYT	10,3	90,2	1,4
ADA	KYA ve AYT ve ANK ve IZM	7,3	89,7	2,3

İlk çıkan sonuca göre yorumlayacak olursak ;

- Bodrum(BJV), Antalya (AYT) ve Adana(ADA) uçan yolcuların %96,6 i İzmir (IZM) yolculuğu yaptığı da gözlemlenmiştir.
- Bodrum (BJV), Antalya (AYT), Adana(ADA) ve İzmir (IZM) uçuşlarının birlikte bulunma eğilimi, tüm uçuş davranış desenlerinden % 7,3 sında beraber görülmektedir.

Küme-5 için Apriori algoritmasını minimum destek değeri %7, minimum güven değeri %40 olacak şekilde düzenleyip çalıştırdığımızda çıkan kurallar aşağıda gösterilmektedir.

Tablo 4.6: Küme-5 için Apriori Sonuçları

Ardıl	Öncül	Destek	Güven	Lift
IZM	KYA	7,1	46,9	1,4
IZM	ASR	9,6	44,2	1,3

İlk çıkan sonuca göre yorumlayacak olursak ;

- Konya (KYA) uçan yolcuların %46,9‘u İzmir (IZM) yolculuğu yaptığı da gözlemlenmiştir.
- Konya (KYA) ve İzmir (IZM) uçuşlarının birlikte bulunma eğilimi, tüm uçuş davranış desenlerinden % 7,1 inde beraber görülmektedir.

5. TARTIŞMA VE SONUÇ

Ülkemizin dünya genelinde havacılık sektörü hizmet ihracatı sıralamasında 11.3 milyar Usd ciro ile Dünya 3.sü konumda bulunması[1], ülkemizin coğrafik avantajları ve küresel prestij anlamında havacılık sektörünün cazibesi bu alandaki büyük boyutlu rekabetide beraberinde getirmektedir. Günümüz dünyasında müşteri istek ve beklentilerinin değişmesi, kişiselleştirilmiş deneyimler isteği ve firmaların müşterilerini tanıyarak hangi müşteriye ne kadarlık yatırım yapması gerektiğini bilmek istemesi müşteri segmentasyonunun gerekliliğini ortaya koymuştur.

Bu nedenle, çalışmamızın ilk bölümünde çalışmanın amacından ve yapılan işlemlerden bahsedilmiştir. İkinci bölümde genel kısımlar başlığı altında veri madenciliği, veri madenciliği modelleri, kümeleme algoritmaları, birliktelik kuralları, müşteri değeri hesaplama uygulamalarının teorik anlatımına ve literatür araştırmasına yer verilmiştir. Üçüncü bölümde ise uygulama yapılmışbir havayolu şirketinden alınan verilerle kümeleme analizi ve birliktelik kuralları madenciliği uygulaması yapılmıştır. Uygulama safhasında; veri seti tanıtılarak GFT modeli ile müşteri değeri hesaplanmıştır. Daha sonra, k-ortalamalar, iki aşamalı kümeleme ve SOM (Öz Örgütlemeli Haritalar) yöntemleri ile kümeleme yapılmıştır. Kümeleme algoritmaları Silhoutte indeksi kullanılarak değerlendirilmiş ve en uygun küme sayısı 5 ve en uygun kümeleme yöntemi k-ortalamalar olarak belirlenmiştir. Dördüncü bölümde ise k-ortalamalar yöntemiyle müşteri segmentasyonu yapılmıştır. Herbir küme için müşterilerin uçuş davranışlarını gözlemlemek amacıyla Apriori algoritması kullanarak müşterilerin yurtiçi uçuşlarında nasıl bir desen ortaya çıkardığını keşfedilmiştir.

Çalışmanın zenginleştirilmesi anlamında ise müşterilerin hobileri, cinsiyetleri, yaşları gibi farklı demografik özelliklerini de göz önünde bulundurarak GFT tabanlı bir kümeleme analizi yapılabilir. Uçuş davranışlarını ise daha anlamlı hale getirmek amacıyla uçuşların iş amaçlı mı yoksa gezi amaçlı mı yapıldığını gösteren değişkenler modele eklenebilir. Bu tür yöntemlerin uygulanması halinde destek ve güven değerlerinde düşüş olmakla birlikte daha anlamlı sonuçlar elde edilecektir.

1 - SHGM 2018 Faaliyet Raporu

KAYNAKLAR

1. Aggarwal, C. C., & Reddy, C. K.,2014, *Data Clustering Algorithms and Applications*, CRC Press, Florida, ISBN -13 : 978-1-4665-5822-9 (eBook) .

2. Agrawal, R., & Srikant, R. , 1995, Fast algorithms for mining association rules, *Proceedings of the 20th International Conference on Very Large Data Bases*, 12-15 September 1994 Santiago, California, Morgan Kaufmann Publishers, 1215, 487-499.

3. Akamavi, R. K., Mohamed, E., Pellmann, K., & Xu, Y., 2015, Key determinants of passenger loyalty in the low-cost airline business, *Tourism management*, 46, 528-545.

4. Akat, Y., 2007, *Ülkelerin Askeri Benzerliklerine Göre Kümeleme Analizi Yardımıyla Sınıflandırılması*, Doktora Tezi, İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü.

5. Akgül, F. G., 2013, Bankaların 2008-2012 Yılları Arasında Aktif Büyüklüklerini Etkileyen Kriterler Bakımından Hiyerarşik Kümeleme ve PAM Algoritması ile Sınıflandırılması, *Bankacılık ve Sigortacılık Araştırmaları Dergisi*, 1, 5-6.

6. Al- Zand, H. R. A., & Karacan, H., 2013, Bölümleyici kümeleme algoritmalarının farklı veri yoğunluklarında karşılaştırılması, *Erciyes Üniversitesi Fen Bilimleri Enstitüsü Fen Bilimleri Dergisi*, 30(1), 56-62.

7. Andreopoulos, B., An, A., & Wang, X., 2006, Bi-level clustering of mixed categorical and numerical biomedical data, *International Journal of Data Mining and Bioinformatics*, 1(1), 19-56.
8. Anon., 2015, *IBM SPSS Modeler 17 Algorithms Guide*, , IBM Corporation 1994.

9. Azzag, H., & Lebbah, M., 2008, Clustering of Self-Organizing Map, *26th European Symposium on Artificial Neural Networks*, 23-25 April 2018 Bruges, d-side publi., ISBN:2-930307-08-0, 209-214.

10. Azzalini, A., & Scarpa, B., 2012, *Data analysis and data mining: An introduction*, Oxford University Press, New York, ISBN:973-0-19-976710-6.

11. Baesens, B., Mues, C., Martens, D., & Vanthienen, J., 2009, 50 years of data mining and OR: upcoming trends and challenges, *Journal of the Operational Research Society*, 60, 16–23.

12. Bejou, D., Keningham, T. L., & Aksoy, L., 2013, *Customer lifetime value: Reshaping the way we manage to maximize profits*, Routledge, Binghamton, ISBN-13: 978-0-7890-3435-9.

13. Berry, M. J., & Linoff, G. S., 2011, *Data mining techniques: for marketing, sales, and customer relationship management*, John Wiley & Sons, Indianapolis, ISBN: 978-1-118-08745-9 (ebook).

14. Bilgin, A., 2008, *Merkez Tabanlı Kümeleme Algoritmalarının Karşılaştırılması,* Yüksek Lisans Tezi, Kocaeli Üniversitesi Fen Bilimleri Enstitüsü.

15. Bilgin, T. T., & Çamurcu, Y., 2005, DBSCAN, OPTICS ve K-Ortalamalar Kümeleme Algoritmalarının Uygulamalı Karşılaştırılması, *Politeknik Dergisi,* 8(2).

16. Birant, D., (2011), *Data mining using RFM analysis,* In Knowledge-oriented applications in data mining, In: Funatsu, K. (ed.), Chapter 2, InTech, Rijeka, Croatia, 91-108.

17. Birant, D., 2019, Farklı Bağlantı Yöntemleri ile Hiyerarşik Kümeleme Topluluğu, *Selçuk Üniversitesi Mühendislik, Bilim ve Teknoloji Dergisi*, 7(1), 154-164.

18. Çalışır, N., Basak, E., & Çalışır, F,. 2016, Key drivers of passenger loyalty: A case of Frankfurt–Istanbul flights. *Journal of Air Transport Management*, 53, 211-217.

19. Castéran, H., Meyer-Waarden, L., & Reinartz, W., 2017, *Modeling customer lifetime value, retention, and churn*, Handbook of Market Research, In: C. Homburg et al. (ed.),Chapter 7, Springer International Publishing, E-book, ISBN : 978-3-319-05542-8, 1-33.

20. Cavique, L., 2007, A scalable algorithm for the market basket analysis, Journal of Retailing and Consumer Services, 14(6), 400-407.

21. Çelik, G., 2013, *Meslek yüksekokulu öğrencilerinin başarı durumlarını etkileyen faktörlerin veri madenciliği kümeleme teknikleri kullanılarak analizi: Ağrı Meslek Yüksekokulu örneği,* Yüksek Lisans Tezi, Atatürk Üniversitesi Bilgisayar Mühendisliği Ana Bilim Dalı.

22. Çelik, M., 2009, *Veri Madenciliğinde Kullanılan Sınıflandırma Yöntemleri ve Bir Uygulama,* Yüksek Lisans Tezi, İstanbul Üniversitesi Sosyal Bilimler Enstitüsü.

23. Çelik, Ş., 2013, Kümeleme analizi ile sağlık göstergelerine göre Türkiye'deki illerin sınıflandırılması, *Doğuş Üniversitesi Dergisi*, 14, 175-194.

24. Charu, C. A., & Chandan, K. R, 2013, *Data clustering: algorithms and applications,* CRC Press, Florida, ISBN : 978-1-4665-5822-9 (eBook - PDF).

25. Chen, S., 2018, *Estimating Customer Lifetime Value Using Machine Learning Techniques,* Data Mining, In: Thomes C. (ed.), Chapter 2, IntechOpen, London, ISBN: 978-1-78923-597-5, 18-32.

26. Chen, S., Huang, Y., & Huang, W., 2016, *Big Data Applications in Business Analysis,* In Big Data Concepts, Theories, and Applications, In: Yu S., Guo S. (ed.), Chapter 12, Springer International Publishing, Cham, ISBN: 978-3-319-27763-9, 413-437.

27. Chen, Y.F., Kuo, I.T., Ku, H.C., (2009), An Intelligent Market Segmentation System Using K-Ortalamalar And Particle Swarm Optimization, *Expert Systems with Applications,* 36, 4558–4565.

28. Cheng, C. H., & Chen, Y. S., 2009, Classifying the segmentation of customer value via RFM model and RS theory, *Expert systems with applications,* 36(3), 4176-4184.

29. Chiang, W. Y., 2011, To mine association rules of customer values via a data mining procedure with improved model: An empirical case study, *Expert Systems with Applications,* 38(3), 1716-1722.

30. Chiang, W. Y., 2017, Discovering customer value for marketing systems: an empirical case study, *International Journal of Production Research,* 55(17), 5157-5167.

31. Chiang, W. Y., 2018, Applying data mining for online CRM marketing strategy: An empirical case of coffee shop industry in Taiwan, *British Food Journal*, 120(3), 665-675.

32. Chiu, S., & Tavella, D., 2008, *Data mining and market intelligence for optimal marketing returns,* Elsevier, Burlington, ISBN: 978-3-319-27763-9.

33. Christy, A. J., Umamakeswari, A., Priyatharsini, L., & Neyaa, A., 2018, RFM ranking–An effective approach to customer segmentation, *Journal of King Saud University-Computer and Information Sciences,*1-10.

34. Coussement, K., Van den Bossche, F. A., & De Bock, K. W., 2014, Data accuracy's impact on segmentation performance: Benchmarking RFM analysis, logistic regression, and decision trees, *Journal of Business Research*, 67(1), 2751-2758.

35. Davidson, I., & Ravi, S. S., 2005, Clustering with constraints: Feasibility issues and the k-ortalamalar algorithm. *In Proceedings of the 2005 SIAM international conference on data mining*, 21-23 April 2005 California, Philadelphia, Society for Industrial and Applied Mathematics, ISBN: 978-0-89871-593-4, 138-149.

36. Desgraupes, B., 2013, Clustering indices, *University of Paris Ouest-Lab Modal'X,* 1, 34.

37. Dinçer, E., 2006, *Veri Madenciliğinde K-Ortalamalar Algoritması ve Tıp Alanında Uygulanması,* Yüksek Lisans Tezi, Kocaeli Üniversitesi Fen Bilimleri Enstitüsü.

38. Doğan, O., 2015, Bir E-Ticaret Sitesi Kullanıcı Hesaplarında Şifre Yapılarının Birliktelik Kuralları ile İncelenmesi, *İnternet Uygulamaları ve Yönetimi Dergisi,* 6(2), 49-61.

39. Doğan, N., & Başokçu, T. O., 2010, İstatistik tutum ölçeği için uygulanan faktör analizi ve aşamalı kümeleme analizi sonuçlarının karşılaştırılması, *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi,* 1(2), 65-71.

40. Dolnicar, S., Grabler, K., Grün, B., & Kulnig, A., 2011, Key drivers of airline loyalty, *Tourism Management*, 32(5), 1020-1026.

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

42. Dursun, A., & Caber, M., 2016, Using data mining techniques for profiling profitable hotel customers: An application of RFM analysis, *Tourism management perspectives*, 18, 153-160.

43. Fader, P. S., Hardie, B. G., & Lee, K. L., 2005, RFM and CLV: Using iso-value curves for customer base analysis, *Journal of marketing research*, 42(4), 415-430.

44. Ferhatosmanoglu, N., 2018, Airline customer lifetime value estimation using data analytics supported by social network information, *Journal of Air Transport Management*, 67(C), 19-33.

45. Forgas, S., Moliner, M. A., Sánchez, J., & Palau, R., 2010, Antecedents of airline passenger loyalty: Low-cost versus traditional airlines, *Journal of Air Transport Management*, 16(4), 229-233.

46. Gan, G., Ma, C., & Wu, J., 2007, *Data clustering: theory, algorithms, and applications (Vol. 20)*, ASA-SIAM, Philadelphia.

47. Ganti, V., Gehrke, J., & Ramakrishnan, R. ,1999, CACTUS-clustering categorical data using summaries. *In KDD* ,Vol. 99,. 73-83.

48. Gemici, Ö., 2007, *Demetleme Problemi İçin Paralel Karınca Yaklaşımı,* Yüksek Lisans Tezi, İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü.

49. Gündüz, E.B., 2018, *Veri Madenciliğinde Kümeleme Algoritmaları ile Mağaza Segmentasyonu,* Lisans Tezi, İstanbul Üniversitesi Endüstri Mühendisliği.

50. Gupta, S., Hanssens, D., Hardie, B., Kahn, W., Kumar, V., Lin, N., & Sriram, S., 2006, Modeling customer lifetime value, *Journal of service research*, 9(2), 139-155.

51. Hacıoğlu, H.,K., 2016, *Kümeleme Analizinde Kullanılan Bazı Benzerlik İndekslerinin Karşılaştırılması,* Yüksek Lisans Tezi, Gazi Üniversitesi Fen Bilimleri Enstitüsü

52. Han, J., Pei, J., & Kamber, M., 2012, *Data mining: concepts and techniques*, Elsevier, Waltham, ISBN:978-0-12-381479-1.

53. Hatipoğlu.E., 2018, Machine Learning—Association Rule Mining (Birliktelik Kural Çıkarımı)—Apriori Algorithm —Eclat Algorithm—Part 14, https://medium.com/@ekrem.hatipoglu/machine-learning-association-rule-mining-birliktelik-kural-%C3%A7%C4%B1kar%C4%B1m%C4%B1-apriori-algorithm-4326b8f224c3, [Ziyaret Tarihi : 17 Nisan 2019].

54. Hu, Y. H., & Yeh, T. W., 2014, Discovering valuable frequent patterns based on RFM analysis without customer identification information, *Knowledge-Based Systems*, 61, 76-88.

55. Ilham, A., GS, A. D., Laumal, F. E., Kurniasih, N., Iskandar, A., Manulangga, G., & Rahim, R., 2018, Market Basket Analysis Using Apriori and FP-Growth for Analysis Consumer Expenditure Patterns at Berkah Mart in Pekanbaru Riau, *In Journal of Physics: Conference Series*, 1114 (1), 012131.

56. Insani, R., & Soemitro, H. L., 2016, Data mining for marketing in telecommunication industry, *2016 IEEE Region 10 Symposium (TENSYMP)*, 9-11 May 2016 Bali,New Jersey, IEEE, ISBN -13 : 978-1-5090-0931-2, 179-183.

57. Işık, M., 2006, *Bölünmeli Kümeleme Yöntemleri İle Veri Madenciliği Uygulamaları*, Yüksek Lisans Tezi, Fen Bilimleri Enstitüsü.

58. Işık, M., & Çamurcu, A. Y. ,2007, K-ortalamalar, K-medoids ve bulanık C-ortalamalar algoritmalarının uygulamalı olarak performanslarının tespiti, *İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi*, 6(11), 31-45.

59. Jassar, K. K., & Dhindsa, K. S., 2015, Comparative study and performance analysis of clustering algorithms, *International Journal of Computer Applications*, 975, 8887.

60. Jiale, L., & Huiying, D., 2010, Study on airline customer value evaluation based on RFM model, *2010 International Conference On Computer Design and Applications* (ICCDA), 25-27 June 2010 Qinhuangdao, New Jersey, IEEE, ISBN: 978-1-4244-7164-5, 4, 278-281.

61. Joshi, P., 2017, *Artificial intelligence with python*, Packt Publishing Ltd, Birmingham, ISBN: 978-1-78646-439-2.

62. Kangallı, S. G., Uyar, U., & Buyrukoğlu, S., 2014, OECD Ülkelerinde Ekonomik Özgürlük: Bir Kümeleme Analizi, *Journal of Alanya Faculty of Business/Alanya Isletme Fakültesi Dergisi*, 6(3).

63. Kaufman, L., & Rousseeuw, P. J., 2009, *Finding groups in data: an introduction to cluster analysis*, John Wiley & Sons, New Jersey, ISBN: 0-471-73578-7.

64. Kayaalp G., Yazgan E., Şahinler S.,2000, Kümeleme Analiz Yöntemlerinin Karşılaştırmalı Olarak İncelenmesi , *Türk Matematik Derneği-Devlet İstatistik Enstitüsü Araştirma 2000 Sempozyumu*., 15- 18 Ekim 2000 Ankara, Ankara, Türk Matematik Derneği Yayınevi,15-18 Ekim 2000, ss.24-34

65. Khajvand, M., & Tarokh, M. J., 2011, Estimating customer future value of different customer segments based on adapted RFM model in retail banking context, *Procedia Computer Science*, 3, 1327-1332.

66. Khajvand, M., Zolfaghar, K., Ashoori, S., & Alizadeh, S., 2011, Estimating customer lifetime value based on RFM analysis of customer purchase behavior: Case study, *Procedia Computer Science*, 3, 57-63.

67. Khalili-Damghani, K., Abdi, F., & Abolmakarem, S., 2018, Hybrid soft computing approach based on clustering, rule mining, and decision tree analysis for customer segmentation problem: Real case of customer-centric industries, *Applied Soft Computing*, 73, 816-828.

68. Khurana, K., & Sharma, S., 2013, A comparative analysis of association rule mining algorithms, *International Journal of Scientific and Research Publications*, 3(5), 0.

69. Kim, S. Y., Jung, T. S., Suh, E. H., & Hwang, H. S., 2006, Customer segmentation and strategy development based on customer lifetime value: A case study, *Expert systems with applications*, 31(1), 101-107.

70. King, R. S., 2015, *Cluster Analysis and Data Mining: An Introduction*, Mercury Learning & Information, Amerika Birleşik Devletleri, ISBN: 978-1-938549-38-0.

71. Kiriş, S. B., & Tüysüz, F., 2017, İmalat hücresi oluşturulması için farklı kümeleme yöntemlerinin performans karşılaştırması, *Sakarya University Journal of Science*, 21(5), 1031-1044.

72. Kovács, F., Legány, C., & Babos, A., 2005, Cluster validity measurement techniques, *In 6th International symposium of hungarian researchers on computational intelligence*, 18-19 November 2005 Budapest, Budapest, Budapest, Teknoloji ve Ekonomi Üniversitesi, ISBN: 963-7154-43-4.

73. Kristiani, E., Sumarwan, U., Yulianti, L. N., & Saefuddin, A., 2013, Customer Loyalty and Profitability: Empirical Evidence of Frequent Flyer Program, *International Journal of Marketing Studies*, 5(6), 62.

74. Kumar, V., Ramani, G., & Bohling, T., 2004, Customer lifetime value approaches and best practice applications, *Journal of Interactive marketing*, 18(3), 60-72.

75. Kumar V.,2017, *CLV Models*, http://www.vkclv.com/about-clv/clv-models/, [Ziyaret Tarihi: 5 Mayıs 2019].

76. Kumbhare, T. A., & Chobe, S. V., 2014, An overview of association rule mining algorithms, *International Journal of Computer Science and Information Technologies*, 5(1), 927-930.

77. Larose, D. T., 2015, *Data mining and predictive analytics*, John Wiley & Sons, Kanada, ISBN: 978-1-118-11619-7.

78. Liang, D. L., & Chen, H. B., 2017, An Online Mall CRM Model Based on Data Mining, *Quantitative Logic and Soft Computing 2016*, 14-17 October 2017 Hangzhou, Cham, Springer, ISBN: 978-3-319-46206-6.

79. Liu, Y., Li, Z., Xiong, H., Gao, X., Wu, J., & Wu, S., 2013, Understanding and enhancement of internal clustering validation measures, *IEEE transactions on cybernetics*, 43(3), 982-994.

80. Maalouf L., Mansour N., 2007, Mining Airline Data for CRM Strategies, *Proceedings of the 7th WSEAS International Conference on Simulation, Modelling and*

Optimization, 15-17 September 2017 Beijing, Wisconsin, World Scientific and Engineering Academy and Society (WSEAS), 345-350.

81. Makhabel, B., 2015, *Learning data mining with R*. Packt Publishing Ltd, Birmingham, ISBN: 978-1-78398-210-3.

82. Malthouse, E. C., & Blattberg, R. C., 2005, Can we predict customer lifetime value?, *Journal of interactive marketing*, 19(1), 2-16.

83. Mary, S. A. L., Sivagami, A. N., & Rani, M. U., 2015, Cluster validity measures dynamic clustering algorithms, *ARPN Journal of Engineering and Applied Sciences*, 10(9).

84. McCormick, K., Abbott, D., Brown, M. S., Khabaza, T., & Mutchler, S. R., 2013, *IBM SPSS modeler cookbook*, Packt Publishing, Birmingham, ISBN 978-1-84968-546-7.

85. Miguéis, V. L., Camanho, A. S., & e Cunha, J. F., 2011, Mining customer loyalty card programs: The improvement of service levels enabled by innovative segmentation and promotions design, *Exploring Services Science*, 16-18 February Geneva, Berlin, Springer, ISBN: 978-3-642-21547-6, 83-97.

86. Murat Y.S., Şekerler A., 2009, Trafik Kaza Verilerinin Kümelenme Analizi Yöntemi ile Modellenmesi, *İMO Teknik Dergi,*1(1), 4759-4777.

87. Nacaroğlu, E., 2010, *Deprem etkisiyle oluşan boru hasarlarının coğrafi bilgi sistemleri (CBS) ve kümeleme analizi ile değerlendirilmesi*, Yüksek Lisans Tezi, Pamukkale Üniversitesi Fen Bilimleri Enstitüsü.

88. Ng, R. T., & Han, J., 2002, CLARANS: A method for clustering objects for spatial data mining, *IEEE Transactions on Knowledge & Data Engineering*, (5), 1003-1016.

89. Ngai, E. W., Xiu, L., & Chau, D. C., 2009, Application of data mining techniques in customer relationship management: A literature review and classification, *Expert systems with applications*, 36(2), 2592-2602.

90. Nikumanesh, E., & Albadvi, A., 2014, Customer's life–time value using the RFM model in the banking industry: a case study, *International Journal of Electronic Customer Relationship Management*, 8(1-3), 15-30.

91. Nisbet, R., Elder, J., & Miner, G., 2017, *Handbook of statistical analysis and data mining applications,* Elseiver, Burlington, ISBN: 978-0-12-374765-5.

92. Oktay, H. M., 2009, *Web Kullanım Madenciliğinde Birliktelik Kurallarının Uygulanması,* Yüksek Lisans Tezi, Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı.

93. Olson, D. L., 2017, Descriptive data mining, Springer, Singapore, ISBN: 978-981-10-3340-7 (eBook).

94. Özçakır, F.C. ve Çamurcu, A.Y., 2007, Birliktelik Kuralı Yöntemi İçin Bir Veri Madenciliği Yazılımı Tasarımı ve Uygulaması, *İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi*, 6(12), 21-37.

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

96. Özdemir Özdoğan, G., 2010, *Öbek bilgisayarlarda paralel FP-growth gerçekleştirimi*, Yüksek Lisans Tezi, TOBB Ekonomi ve Teknoloji Üniversitesi-Fen Bilimleri Enstitüsü-Bilgisayar Mühendisliği Anabilim Dalı.

97. Özdemir, A., Aslay, F. Y., & Çam, H., 2010, Verin tabanında bilgi keşfi süreci: Gümüşhane Devlet Hastanesi uygulaması, *Selçuk Üniversitesi İktisadi ve İdari Bilimler Fakültesi Sosyal ve Ekonomik Araştırmalar Dergisi,* 10(20), 347-365.

98. Pakyürek, M., Sezgin, M. S., Kestepe, S., Bora, B., Düzağaç, R., & Yıldız, O. T., 2018, Customer clustering using RFM analysis, *2018 26th Signal Processing and Communications Applications Conference (SIU),* 2-5 May 2018 İzmir,New Jersey, IEEE, 1-4.

99. Pasin, Ö., 2015, *Sağlık Alanında Yapılan Araştırmalarda Kümeleme Algoritmalarının Kullanımı: Bir Uygulama*, Yüksek Lisans Tezi, Düzce Üniversitesi Biyoistatistik ve Tıbbi Bilişim Anabilim Dalı.

100. Ponyiam, P., & Arch-int, S., 2018, Customer Behavior Analysis Using Data Mining Techniques, *2018 International Seminar on Application for Technology of Information and Communication*, 21-22 September 2018 Semarang, New Jersey, IEEE, 549-554.

101. Pritscher, L., & Feyen, H. ,2001, Data mining and strategic marketing in the airline industry. *Data Mining for Marketing Applications*, 39.

102. Qadadeh, W., & Abdallah, S., 2018, Customers Segmentation in the Insurance Company (TIC) Dataset, *Procedia computer science*, 144, 277-290.

103. Rendón, E., Abundez, I. M., Gutierrez, C., Zagal, S. D., Arizmendi, A., Quiroz, E. M., & Arzate, H. E., 2011, A comparison of internal and external cluster validation indexes, *In Proceedings of the 5th WSEAS International Conference on Computer Engineering and Applications*, 158-163.

104. Sagin, A. N., & Ayvaz, B., 2018, Determination of Association Rules with Market Basket Analysis: Application in the Retail Sector, *Southeast Europe Journal of Soft Computing*, 7(1).

105. Sarvari, P. A., Ustundag, A., & Takci, H., 2016, Performance evaluation of different customer segmentation approaches based on RFM and demographics analysis, *Kybernetes*, 45(7), 1129-1157.

106. Schiopu, D., 2010, Applying TwoStep cluster analysis for identifying bank customers' profile, *Buletinul*, 62, 66-75.

107. Şeker, Şadi Evren., 2013, *İş Zekası ve Veri Madenciliği*, Cinius, İstanbul.

108. Şekerler, A., 2008, *Trafik kaza verilerinin kümeleme analizi yöntemi ile incelenmesi*, Yüksek Lisans Tezi, Pamukkale Üniversitesi Fen Bilimleri Enstitüsü.

109. Sezgin, E., & Çelik, Y., 2013, Veri madenciliğinde kayıp veriler için kullanılan yöntemlerin karşılaştırılması. *Akademik Bilişim 2013 Konferansı,* 23-25 Ocak 2013 Antalya, Online, İnternet Teknolojileri Derneği , 23-25.

Sezgin E., "Veri Madenciliğinde Kayıp Veriler İçin Kullanılan Yöntemlerin Karşılaştırılması", Akademik Bilişim 2013, ANTALYA, TÜRKIYE, 23-25 Ocak 2013, ss.194-198

110. Shalizi, C., 2009, Distances between Clustering Hierarchical Clustering, *Lectures notes.*

111. Shen, X., Burnett, D., Vishwanath, N., & Lee, Y. S., 2003, *Retention Modeling Methodology for Airlines. U.S. Patent Application No. 10/046,226.*

112. Sivri, E. Ş., 2015, Veri madenciliği: *e-ticaret için ürün tavsiye sistemi geliştirilmesi*, Yüksek Lisans Tezi, İstanbul Ticaret Üniversitesi.

113. Sohn, S. Y., & Kim, Y., 2008, Searching customer patterns of mobile service using clustering and quantitative association rule, *Expert systems with Applications,* 34(2), 1070-1077.

114. Solnet, D., Boztug, Y., & Dolnicar, S., 2016, An untapped gold mine? Exploring the potential of market basket analysis to grow hotel revenue, *International Journal of Hospitality Management,* 56, 119-125.

115. Song, M., Zhao, X., Haihong, E., & Ou, Z., 2017, Statistics-based CRM approach via time series segmenting RFM on large scale data, *Knowledge-Based Systems,* 132, 21-29.

116. Sukri, S., Abdullah, F., & Waemustafa, W., 2014, Customer satisfaction and loyalty in the airline industry: A case study of Malaysia Airlines (MAS) and Air Asia, *International Case Study Conference*, 18-19 August 2014 Putra World Trade Centre, Malaysia, online, 43-69.

117. Sun, K., & Bai, F, 2008, Mining weighted association rules without preassigned weights, *IEEE transactions on knowledge and data engineering*, 20(4), 489-495.

118. Swinnen, G., Estrella-Ramon, A. M., Sanchez-Perez, M., & VanHoof, K., 2013, A marketing view of the customer value: Customer lifetime value and customer equity, *South African Journal of Business Management*, 44(4), 47-64.

119. Tan, P. N., Steinbach, M., & Kumar, V., 2013, *Data mining cluster analysis: basic concepts and algorithms*, Springer, Heidelberg.

120. Teichert, T., Shehu, E., & von Wartburg, I., 2008, Customer segmentation revisited: The case of the airline industry, *Transportation Research Part A: Policy and Practice*, 42(1), 227-242.

121. Tran, C. T., Zhang, M., Andreae, P., Xue, B., & Bui, L. T., 2018, Improving performance of classification on incomplete data using feature selection and clustering, *Applied Soft Computing*, 73, 848-861.

122. Trnka, A., 2010, Market basket analysis with data mining methods, *2010 International Conference on Networking and Information Technology*, 11-12 June 2010 Manila, New Jersey, IEEE, ISBN: 978-1-4244-7578-0, 446-450.

123. Trpkova, M., & Tevdoski, D., 2009, Twostep cluster analysis: Segmentation of largest companies in Macedonia, *Challenges for Analysis of the Economy, the Businesses, and Social Progress*, pp. 302–320.

124. Tsai, C. Y., & Chiu, C. C., 2004, A purchase-based market segmentation methodology, *Expert Systems with Applications*, 27(2), 265-276.

125. Tsay, Y. J., & Chiang, J. Y., 2005, CBAR: an efficient method for mining association rules, *Knowledge-Based Systems*, 18(2-3), 99-105.

126. Tunalı, V., 2011, *Metin Madenciliği İçin İyileştirilmiş Bir Kümeleme Yapısının Tasarımı Ve Uygulaması*, Doktora Tezi, Marmara Üniversitesi Fen Bilimleri Enstitüsü.

127. Venkatesan, R., & Kumar, V., 2004, A customer lifetime value framework for customer selection and resource allocation strategy, *Journal of marketing*, 68(4), 106-125.

128. Vorhies W., 2016, *CRISP-DM – a Standard Methodology to Ensure a Good Outcome,* https://www.datasciencecentral.com/profiles/blogs/crisp-dm-a-standard-methodology-to-ensure-a-good-outcome, [Ziyaret Tarihi : 4 Nisan 2019].

129. Wang, K., Wang, B., & Peng, L., 2009, CVAP: validation for cluster analyses, *Data Science Journal*, 8, 88-93.

130. Weng, C. H., 2017, Revenue prediction by mining frequent itemsets with customer analysis, *Engineering Applications of Artificial Intelligence,* 63, 85-97.

131. Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J., 2017, *Data Mining: Practical machine learning tools and techniques*, Morgan Kaufmann, Cambridge, ISBN: 978-0-12-804291-5.

132. Wong, J. Y., & Chung, P. H., 2007, Managing valuable Taiwanese airline passengers using knowledge discovery in database techniques, *Journal of Air Transport Management*, 13(6), 362-370.

133. Wu, X., & Kumar, V. (Eds.)., 2009, The top ten algorithms in data mining, CRC press, New York, ISBN : 13: 978-1-4200-8964-6.

134. Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., & Zhou, Z. H., 2008, Top 10 algorithms in data mining, *Knowledge and information systems*, 14(1), 1-37.

135. Yapraklı, T. Ş., & Keser, E., 2008, Müşteri Yaşam Boyu Değerinin Analizi : Bir Saha Araştırması, *Atatürk Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, 12(2), 483-503.

136. Yılancı, A. G. V., 2010, Bulanık Kümeleme Analizi İle Türkiye'deki İllerin Sosyoekonomik Açıdan Sınıflandırılması, *Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi,* 15(3), 453-470.

137. Yıldırım, M., 2016, *İldeki Kurumlar Arası Çalışma Performanslarının Arttırılmasında Veri Madenciliği Tekniklerinin Kullanılması*, Yüksek Lisan Tezi, Fırat Üniversitesi Fen Bilimleri Enstitüsü.

138. Zaki, M. J., Meira Jr, W., & Meira, W., 2014, Data mining and analysis: fundamental concepts and algorithms, Cambridge University Press, New York, ISBN: 978-0-521-76633-3.

ÖZGEÇMİŞ

Kişisel Bilgiler	
Adı Soyadı	Ramazan YAŞA
Doğum Yeri	Malatya
Doğum Tarihi	01.01.1992
Uyruğu	☑ T.C. ❒ Diğer:
Telefon	0(506) 525 67 78
E-Posta Adresi	rmznyasa@hotmail.com
Web Adresi	

Eğitim Bilgileri	
Lisans	
Üniversite	Gazi Üniversitesi
Fakülte	Mühendislik Fakültesi
Bölümü	Endüstri Mühendisliği
Mezuniyet Yılı	2013

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