Müşteri Segmentasyonu için K-Means, Bulanık C-Means ve Bulanık Mantık Tabanlı Kümeleme Algoritmalarının Karşılaştırılması

Muhammed Ali BAYRAM[1*], Cemil KÖZKURT[2]

[1*] *Bandırma Onyedi Eylül Üniversitesi, Bandırma, Türkiye, muhammedbayram@ogr.bandirma.edu.tr, ORCID: 0000-0001-7784-2992*

[2] *Bandırma Onyedi Eylül Üniversitesi, Bandırma, Türkiye, ckozkurt@bandirma.edu.tr, ORCID: 0000-0003-1407-9867*

Müşteri segmentasyonu, işletmelerin pazarlama stratejilerini optimize etmeleri ve müşteri sadakatini artırmaları için kritik bir araçtır. Bu çalışma, K-Means, Bulanık C-Means (FCM) ve Bulanık Mantık tabanlı kümeleme algoritmalarını kullanarak, müşterilerin satın alma davranışlarına dayalı olarak segmentlere ayrılmasını amaçlamaktadır. K-Means algoritması, verileri belirli sayıda kümeye ayırarak her küme için bir merkez belirlerken, FCM algoritması, veri noktalarının birden fazla kümeye değişen üyelik dereceleri ile ait olmasına izin vermektedir. Bulanık Mantık tabanlı kümeleme ise, her özellik için tanımlanan bulanık üyelik fonksiyonları ve kurallar sistemi kullanarak daha nüanslı bir segmentasyon sağlar. Çalışmada, Online Retail veri seti kullanılarak her üç algoritmanın performansı Silhouette skoru, Calinski-Harabasz skoru ve Davies-Bouldin skoru ile değerlendirilmiştir. Sonuçlar, K-Means'in genellikle daha yüksek Silhouette skorları ve Calinski-Harabasz indeksi değerleri sunduğunu, FCM'nin kümeler arası benzerlikleri yönetme esnekliği sağladığını ve Bulanık Mantık tabanlı yaklaşımın özellikler arasındaki karmaşık ilişkileri daha iyi yakaladığını göstermektedir. Bulgular, müşteri segmentasyonu ve pazarlama stratejilerinin optimizasyonu için değerli içgörüler sunmaktadır.

Keywords: *Bulanık mantık tabanlı kümeleme, K-Means Algoritması, Bulanık C-Means (FCM), Kümeleme kıyaslama metrikleri*

Comparison of K-Means, Fuzzy C-Means and Fuzzy Logic Based Clustering Algorithms for Customer Segmentation

Customer segmentation is a critical tool for businesses to optimize their marketing strategies and increase customer loyalty. This study aims to segment customers based on their buying behavior using K-Means, Fuzzy C-Means (FCM) and Fuzzy Logic based clustering algorithms. While the K-Means algorithm divides the data into a certain number of clusters and determines a center for each cluster, the FCM algorithm allows data points to belong to more than one cluster with varying degrees of membership. Fuzzy Logic based clustering provides a more nuanced segmentation by using a system of fuzzy membership functions and rules defined for each feature. In this study, the performance of all three algorithms is evaluated with Silhouette score, Calinski-Harabasz score and Davies-Bouldin score using Online Retail dataset. The results show that K-Means generally delivers higher Silhouette scores and Calinski-Harabasz index values, FCM provides flexibility to manage similarities across clusters, and the Fuzzy Logic-based approach better captures complex relationships between features. The findings provide valuable insights for customer segmentation and optimization of marketing strategies.

Keywords: *Fuzzy logic based clustering, K-Means Algorithm, Fuzzy C-Means (FCM), Clustering benchmark metrics*

1. Giriş

Müşteri segmentasyonu, işletmelerin pazarlama stratejilerini optimize etmeleri ve müşteri sadakatini artırmaları açısından kritik bir öneme sahiptir. Bu çalışma, müşterilerin satın alma davranışlarına dayalı olarak segmentlere ayırmak için çeşitli kümeleme metodolojilerini kullanmakta ve özellikle K-Means, Bulanık C-Means (Fuzzy C-Means) ve Bulanık Mantık tabanlı kümeleme algoritmalarına odaklanmaktadır.

Anitha ve Patil, müşteri satın alma davranışlarını analiz etmek için K-Ortalamalar algoritmasını kullanan bir RFM (Recency, Frequency, Monetary) modelini incelemişlerdir. Çalışmada, satış geçmişi ve tüketici davranışlarını analiz ederek perakende endüstrisinde potansiyel müşterilerin belirlenmesi hedeflenmiştir [1]. Elde edilen veriler, K-Ortalamalar algoritması kullanılarak farklı kümelere ayrılmış ve Silhouette Katsayısı kullanılarak bu kümeler değerlendirilmiştir. Sonuçlar, satış işlemleriyle ilgili çeşitli parametrelerle karşılaştırılmıştır.

Hu ve Yeh, müşteri segmentasyonunun önemli bir bileşeni olan RFM analizini kullanarak değerli ve sık alışveriş desenlerini keşfetmişlerdir [2]. Bu çalışmada, müşteri tanımlama bilgileri olmaksızın RFM analizi gerçekleştirilmiştir. Aynı şekilde, Khajvand ve arkadaşları müşteri satın alma davranışlarını analiz etmek için RFM analizini kullanmış ve müşteri yaşam boyu değerini tahmin etmişlerdir [3].

Khalili-Damghani ve arkadaşları, müşteri segmentasyon problemi için kümeleme, kural çıkarımı ve karar ağacı analizi gibi yöntemleri birleştiren hibrit bir yumuşak hesaplama yaklaşımı önermişlerdir [4]. Bu yöntem, posta hizmetleri gibi müşteri odaklı endüstrilerde gelecekteki işlemleri tahmin etmek için kullanılmıştır. Ayrıca, Griva ve arkadaşları, perakende iş analitiği için müşteri ziyaret segmentasyonunu incelemiş ve market sepeti verilerini kullanarak müşteri gruplarını belirlemişlerdir [5].

Qadadeh ve Abdallah, RFM modelini kullanarak sigorta şirketi veritabanında müşteri segmentasyonu yapmışlardır [6]. Bu çalışma, müşteri ihtiyaçlarını ve özelliklerini belirlemede etkili olmuştur. Aynı şekilde, Arunachalam ve Kumar, fayda tabanlı tüketici segmentasyonu ve kümeleme yaklaşımlarının performans değerlendirmesini yapmışlardır [7].

Yoo ve Jang, elektronik ticaret araştırmalarını hizmet ilişkileri, iş modelleri ve teknolojiler açısından inceleyerek üç aşamalı bir çerçeve önermişlerdir [8]. Bu çerçeve, uygulayıcılar tarafından bildirilen sorunların çözüm önerilerini içermektedir. Zerbino ve arkadaşları, büyük verinin CRM üzerindeki etkisini incelemiş ve müşteri ilişkileri yönetimi için bütüncül bir yaklaşım önermişlerdir [9].

Han ve arkadaşları, perakende verimliliğini artırmak ve müşteri segmentasyonunu geliştirmek için veri madenciliği tekniklerini kullanmışlardır [10]. Brito ve arkadaşları, moda endüstrisinde müşteri segmentasyonunu incelemiş ve müşteri tercihlerini anlamak için veri madenciliği tekniklerini kullanmışlardır [11].

Bahari ve Elayidom, müşteri davranışlarını tahmin etmek için veri madenciliği ve CRM çerçevesi geliştirmişlerdir[12]. Sheng ve arkadaşları, büyük veri ve yönetim araştırmalarının entegrasyonunu inceleyerek kapsamlı bir çerçeve sunmuşlardır[13]. Nguyen ve arkadaşları, çevrimiçi perakendecilikte müşteri davranışlarını ve sipariş karşılama süreçlerini incelemişlerdir[14].

Patak ve arkadaşları, eczane müşterilerinin segmentasyonunu yaparak müşteri sadakatini artırmayı hedeflemişlerdir[15]. Carnein ve Trautmann, akış kümelemesi yöntemini kullanarak perakende verilerini analiz etmişlerdir [16]. Kaur ve arkadaşları, veri kümeleme için çiçek tozlaşması ve kara delik algoritmaları gibi sürü tabanlı algoritmaları karşılaştırmıştır [17].

Fu ve arkadaşları, çevrimiçi oyun müşterilerinin segmentasyonunu yaparak müşteri bağlılığını artırmayı hedeflemişlerdir [18]. Holy ve arkadaşları, müşteri davranışlarına dayalı olarak perakende ürünlerini kümelemişlerdir [19].

Bu çalışmalar, müşteri segmentasyonunun ve veri analitiğinin işletmelerin pazarlama stratejilerini optimize etmeleri ve müşteri sadakatini artırmaları açısından önemli olduğunu göstermektedir.

Bu çalışmalar, müşteri davranışlarının analiz edilmesi ve segmentasyonun optimize edilmesi konusundaki çeşitli yaklaşımları ve metodolojileri ele almaktadır. Literatürdeki bu bulgular, müşteri segmentasyonu ve kümeleme yöntemlerinin işletme performansını artırmada ve müşteri memnuniyetini sağlamada nasıl kullanılabileceğine dair önemli bilgiler sunmaktadır.

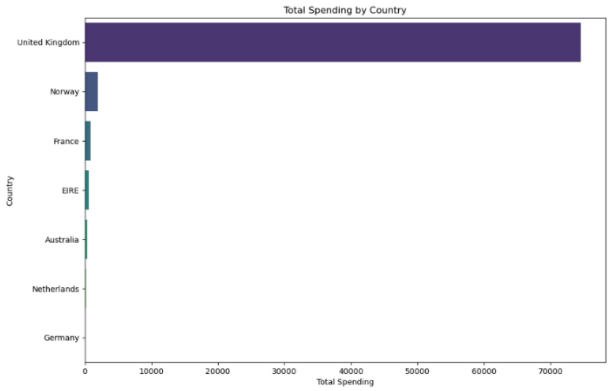
Bu çalışmada kullanılan veri seti, bir perakende mağazasının işlem verilerinden elde edilmiş olup, faturalar, ürün kodları, açıklamalar, miktarlar, fatura tarihleri, birim fiyatlar, müşteri kimlikleri ve satın alma yapılan ülkeler hakkında ayrıntılı bilgiler içermektedir. Bu kapsamlı veri seti, müşteri davranışlarının çok boyutlu bir analizini mümkün kılmaktadır.

Verilerin ön işlenmesi, toplam harcamanın her işlem için toplam fiyat ve miktarın çarpımıyla hesaplanmasını içermektedir. Daha sonra, toplam harcama, satın alma sıklığı, ortalama sepet büyüklüğü ve ürün çeşitliliği gibi temel

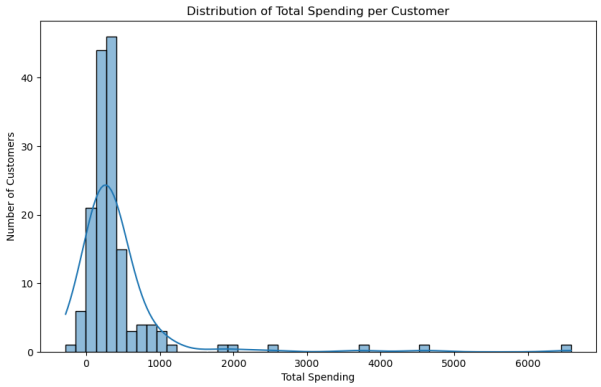
müşteri özellikleri toplanmıştır. Veriler, yalnızca ilgili işlemlerin dikkate alınmasını sağlamak için negatif ve sıfır değerlerin çıkarılmasıyla temizlenmiştir.
K-Means algoritmasının uygulanması, verilerin MinMaxScaler kullanılarak normalize edilmesini ve ardından benzer satın alma kalıplarına sahip müşteri kümelerini tanımlamak için K-Means modelinin uygulanmasını içermektedir. K-Means, her veri noktasını yinelemeli olarak en yakın küme merkezine atayan ve küme merkezlerini optimize eden yaygın bir kümeleme yöntemidir. Küme değerlendirmesi, kümelerin optimal sayısını ve segmentasyonun kalitesini belirlemek için Silhouette skorları, Calinski-Harabasz indeksi ve Davies-Bouldin skorları kullanılarak gerçekleştirilmiştir.
Benzer şekilde, FCM algoritması da kullanılmış olup, bu yöntem K-Means'ten farklı olarak veri noktalarının birden fazla kümeye değişen üyelik dereceleri ile ait olmasına izin vermektedir. Bu yöntem, kümeler arasındaki sınırların net olmadığı durumlarda özellikle kullanışlıdır. FCM algoritmasının performansı da aynı değerlendirme metrikleri kullanılarak değerlendirilmiş ve kümeleme esnekliği ve uyarlanabilirliği hakkında içgörüler sağlanmıştır.
Ek olarak, Bulanık Mantık tabanlı bir kümeleme yaklaşımı uygulanmış olup, her özellik için bulanık üyelik fonksiyonlarının tanımlanmasını içermektedir. Bu fonksiyonlar, birden fazla özellik arasında üyelik derecesini dikkate alan kurallara dayanarak veri noktalarını kümelere ayırır. Bulanık kurallar, özellikler arasındaki ilişkileri yakalamak için tasarlanmış olup, örtüşen kümeleri yönetebilen nüanslı bir segmentasyon sağlar. Bu yöntem, her kuralın girdi özelliklerinin özelliklerine dayalı olarak formüle edildiği bir dizi kural kullanarak küme atamasını belirler. Kontrol sistemi, bu kuralları simüle ederek veri noktalarını kümelere atar.
Üç modelin karşılaştırılması, K-Means'in genellikle daha yüksek Silhouette skorları ve Calinski-Harabasz indeksi değerleri sunduğunu, bunun da daha iyi tanımlanmış kümeleri gösterdiğini ortaya koymuştur. Ancak, FCM algoritması, Davies-Bouldin skorlarıyla yansıtıldığı gibi, örtüşen kümeleri yönetme esnekliği göstermiştir. Bulanık Mantık tabanlı yaklaşım ise, veri sınırlarını bulanık şekilde yönetme konusunda sofistike bir yöntem sunarak, özellikler arasındaki karmaşık ilişkileri yakalamada sağlamlık göstermiştir.
Bu kümeleme metodolojilerini entegre ederek, çalışma müşteri segmentasyonu hakkında değerli içgörüler sağlamaktadır ve daha hedefli pazarlama stratejileri ve geliştirilmiş müşteri ilişkileri yönetimine olanak tanımaktadır. Bulgular, veri setinin spesifik özelliklerine ve analizden elde edilmek istenen sonuçlara dayalı olarak uygun kümeleme algoritmalarının seçilmesinin önemini vurgulamaktadır. K-Means, FCM ve Bulanık Mantık tabanlı kümelemenin birleşimi, müşteri davranış kalıplarını anlamak ve kullanmak için kapsamlı bir araç seti sunarak, nihayetinde daha etkili iş stratejilerine katkıda bulunmaktadır.

2. Veri Seti

Online Retail veri seti, bir çevrimiçi perakende şirketinin belirli bir dönem boyunca gerçekleştirdiği işlemleri içermektedir. Bu veri seti, müşteri davranışlarını, satın alma alışkanlıklarını ve ürün performansını analiz etmek için kapsamlı bilgiler sunar. Veri seti, fatura numarası (InvoiceNo), ürün kodu (StockCode), ürün açıklaması (Description), satın alınan ürün miktarı (Quantity), fatura tarihi ve saati (InvoiceDate), ürün birim fiyatı (UnitPrice), müşteri kimlik numarası (CustomerID) ve müşterinin ikamet ettiği ülke (Country) gibi çeşitli sütunları içermektedir.

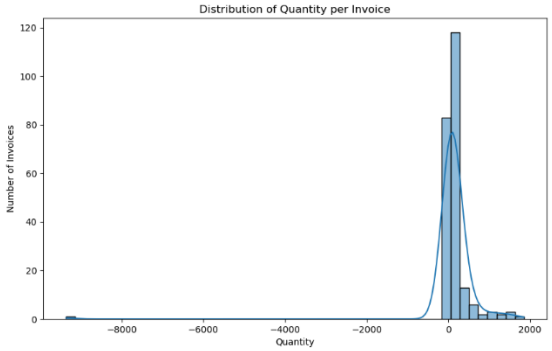


Şekil-1.a: *Ülke bazında toplam harcama grafiği*

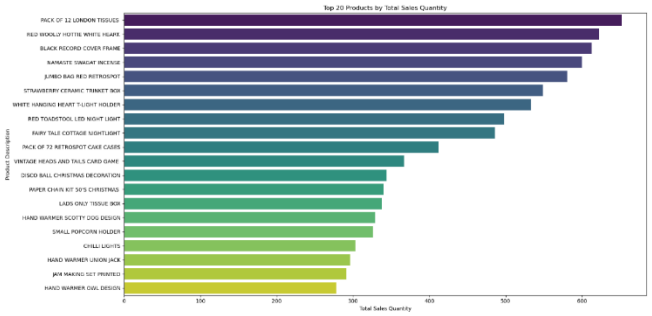


Şekil-1.b: *Müşteri bazında toplam harcama grafiği*

Bu veri seti, müşteri segmentasyonu, satış analizleri, stok yönetimi ve pazarlama stratejileri geliştirmek için kullanılabilir. Özellikle, müşteri davranışlarını ve alışveriş alışkanlıklarını anlamak için değerli bilgiler sunar. Müşterileri alışveriş alışkanlıklarına göre segmentlere ayırmak, kişiselleştirilmiş pazarlama stratejileri geliştirmek için önemlidir. Veri seti, müşteri segmentasyonu için gerekli olan tüm bilgileri sağlar. Ayrıca, satış trendlerini ve ürün performansını analiz etmek, en çok ve en az satan ürünleri belirlemek için de kullanılabilir. Hangi ürünlerin ne kadar sıklıkla ve hangi miktarlarda satıldığını analiz etmek, etkin stok yönetimi için kritik öneme sahiptir. Müşteri alışkanlıklarına ve satın alma eğilimlerine göre hedeflenmiş pazarlama kampanyaları oluşturmak için de bu veri seti büyük bir değer taşımaktadır. Şekil-1 ve Şekil-2'deki müşterilerin harcama davranışlarına ilişkin çeşitli grafikler sunulmaktadır.



Şekil-2.a: *Fatura bazında sipariş sayıları grafiği*



Şekil-2.b: *En iyi 20 müşteriye ait toplam satış adedi grafiği*

Veri setindeki örnek bir kayıt, bir müşterinin 1 Aralık 2010 tarihinde 6 adet "WHITE HANGING HEART T-LIGHT HOLDER" ürünü satın aldığını ve bu işlemi İngiltere'den gerçekleştirdiğini göstermektedir. Bu tür bilgiler, müşteri davranışlarını ve satış trendlerini anlamak için kullanışlıdır.

Sonuç olarak, Online Retail veri seti, perakende sektöründeki müşteri davranışlarını ve satış performansını analiz etmek için geniş kapsamlı bilgiler sunar. Veri setinin doğru analiz edilmesi, işletmelere stratejik kararlar alma ve

rekabet avantajı sağlama konusunda yardımcı olabilir. Bu veri seti, çeşitli veri bilimi ve makine öğrenimi teknikleriyle işlenerek, derinlemesine içgörüler elde edilmesine olanak tanır.

3. Önerilen Model

Veri önişleme adımları, veri analizi ve modelleme sürecinde kritik bir rol oynar ve verinin kalitesini, doğruluğunu ve modelin performansını doğrudan etkiler. Bu çalışmada, Online Retail veri seti üzerinde uygulanan önişleme adımları detaylandırılacaktır. İlk olarak, 'TotalSpending' değişkeni oluşturulmuştur. Bu değişken, her bir ürün için 'TotalPrice' ve 'Quantity' sütunlarının çarpımı ile hesaplanmıştır ve müşterilerin toplam harcamalarını temsil etmektedir.

Sonraki adımda, veriler müşteri düzeyinde özetlenmiştir. Bunun için 'CustomerID' sütununa göre gruplama yapılmış ve her bir müşteri için çeşitli özellikler hesaplanmıştır. 'TotalSpending' sütununun toplamı alınarak müşterilerin toplam harcamaları hesaplanmıştır. 'InvoiceNo' sütununda benzersiz fatura sayısı hesaplanarak her bir müşterinin alışveriş sıklığı ('Frequency') belirlenmiştir. 'TotalPrice' sütununun ortalaması alınarak her bir müşterinin ortalama sepet büyüklüğü ('AvgBasketSize') hesaplanmıştır. Son olarak, 'StockCode' sütununda benzersiz ürün sayısı hesaplanarak her bir müşterinin satın aldığı ürün çeşitliliği ('ProductVariety') belirlenmiştir.

Veri önişleme adımlarının son aşamasında, negatif ve sıfır değerler içeren kayıtlar veri setinden çıkarılmıştır. Bu adım, veri kalitesini artırmak ve analizlerin doğruluğunu sağlamak amacıyla gerçekleştirilmiştir. Sonuç olarak, bu önişleme adımları ile temizlenmiş ve özetlenmiş bir veri seti elde edilmiştir. Bu veri seti, müşteri davranışlarının analizi ve segmentasyonu gibi ileri analizler için uygun hale getirilmiştir. Bu önişleme süreci, veri analizi ve modelleme çalışmalarında kritik bir adımdır ve elde edilen sonuçların güvenilirliğini artırmak için titizlikle uygulanmalıdır.

Bu bölümde, veri kümesi üzerinde bulanık mantık tabanlı kümeleme yöntemi uygulanarak yapılan işlemler ve kullanılan kurallar detaylandırılmaktadır. İlk olarak, veri seti ölçeklendirilmiş ve 'TotalSpending', 'Frequency', 'AvgBasketSize' ve 'ProductVariety' özellikleri kullanılarak normalize edilmiştir. Normalize edilen veriler, bulanık kümeleme sistemine giriş olarak tanımlanmıştır.

Bulanık mantık tabanlı kümeleme modeli, her bir giriş değişkeni (TotalSpending, Frequency, AvgBasketSize ve ProductVariety) için üçlü üyelik fonksiyonları (low, medium, high) tanımlayarak, her bir değişkenin çeşitli üyelik seviyelerine sahip olmasını sağlar. Çıktı değişkeni olarak tanımlanan 'Cluster' değişkeni de benzer şekilde üçlü üyelik fonksiyonlarına sahiptir ve her küme için düşük, orta ve yüksek seviyelerde üyelik fonksiyonları belirlenmiştir.

Kurallar sistemi, giriş değişkenlerinin çeşitli kombinasyonlarına dayanarak veri noktalarının hangi kümeye ait olduğunu belirler. Bu çalışmada, toplamda yedi kural tanımlanmıştır:

1. Kural 1: 'TotalSpending', 'Frequency', 'AvgBasketSize' ve 'ProductVariety' düşük olduğunda ('low'), veri noktası düşük kümeye ('low') atanır.

2. Kural 2: 'TotalSpending', 'Frequency', 'AvgBasketSize' ve 'ProductVariety' orta seviyede olduğunda ('medium'), veri noktası orta kümeye ('medium') atanır.

3. Kural 3: 'TotalSpending', 'Frequency', 'AvgBasketSize' ve 'ProductVariety' yüksek olduğunda ('high'), veri noktası yüksek kümeye ('high') atanır.

4. Kural 4: 'TotalSpending' düşük ('low') ve 'Frequency' yüksek ('high') olduğunda, veri noktası orta kümeye ('medium') atanır.

5. Kural 5: 'TotalSpending' yüksek ('high') ve 'Frequency' düşük ('low') olduğunda, veri noktası orta kümeye ('medium') atanır.

6. Kural 6: 'AvgBasketSize' yüksek ('high') ve 'ProductVariety' düşük ('low') olduğunda, veri noktası orta kümeye ('medium') atanır.

7. Kural 7: 'AvgBasketSize' düşük ('low') ve 'ProductVariety' yüksek ('high') olduğunda, veri noktası orta kümeye ('medium') atanır.

Kontrol sistemi bu kuralları kullanarak oluşturulmuş ve her veri noktası için bu kurallar uygulanarak küme tahminleri yapılmıştır. Veriler ölçeklendirilmiş haliyle kontrol sistemine girilmiş ve her veri noktası için bulanık çıkarım yapılarak hangi kümeye ait olduğu belirlenmiştir. Kümeleme sonuçları, tahmin edilen kümelerle birlikte veri setine eklenmiştir.

Kümeleme performansını değerlendirmek için Silhouette skoru, Calinski-Harabasz skoru ve Davies-Bouldin skoru olmak üzere üç farklı metrik kullanılmıştır. Silhouette skoru, veri noktalarının kendi kümeleri ile diğer kümeler arasındaki benzerlik farkını ölçerken, Calinski-Harabasz skoru kümelerin yoğunluğunu ve ayrılığını değerlendirir. Davies-Bouldin skoru ise kümeler arası benzerliği ölçer. Bu metrikler, farklı küme sayıları için hesaplanarak en iyi performansı gösteren model belirlenmiştir. Bu süreç bulanık mantık tabanlı kümeleme modelinin oluşturulması, uygulanması ve performans değerlendirmesini kapsamaktadır. Farklı küme sayıları için en iyi performansı gösteren modeller belirlenmiş ve bu modellerin sonuçları detaylandırılmıştır. Bu analiz, müşteri segmentasyonu ve diğer veri analizi uygulamaları için değerli içgörüler sunmaktadır.

K-Means kümeleme algoritması, verileri belirli sayıda kümeye ayıran ve her bir küme için bir merkez belirleyen bir kümeleme yöntemidir. Bu algoritma, veri noktalarını en yakın küme merkezine atayarak çalışır ve bu işlem iteratif olarak küme merkezleri stabilize olana kadar tekrarlanır. K-Means algoritması, genellikle büyük veri kümelerinde kümeleme yapmak için kullanılır ve veri analizinde yaygın bir yöntemdir.

Bu çalışmada, K-Means kümeleme algoritması, Online Retail veri seti üzerinde uygulanmıştır. İlk adımda, veri seti ölçeklendirilmiş ve 'TotalSpending', 'Frequency', 'AvgBasketSize' ve 'ProductVariety' özellikleri kullanılarak normalize edilmiştir. Normalize edilen veriler, K-Means algoritmasına giriş olarak kullanılmıştır. Belirli bir küme sayısı için K-Means algoritması uygulanmış ve veri noktaları en yakın küme merkezine atanarak kümelere ayrılmıştır. Her bir veri noktası, ilgili küme merkezine atanarak veri setine 'Cluster' sütunu eklenmiştir.

K-Means algoritmasının farklı küme sayıları için performansını optimize etmek amacıyla belirli bir küme aralığında çeşitli denemeler yapılmıştır. Her küme sayısı için algoritma çalıştırılmış ve yukarıda belirtilen metrikler hesaplanmıştır. En iyi sonuçlar, en yüksek Silhouette skoru ile belirlenmiştir. Bu süreçte, her küme sayısı için en iyi performansı gösteren modeller kaydedilmiş ve bu modellerin sonuçları detaylandırılmıştır. Bu analiz, müşteri segmentasyonu ve diğer veri analizi uygulamaları için değerli içgörüler sunmaktadır.

K-Means kümeleme algoritması, veri setine uygulanarak performans değerlendirmesi ve optimizasyonu gerçekleştirilmiştir. Bu yöntem, büyük veri kümelerinde hızlı ve etkili kümeleme yapabilme kapasitesi nedeniyle veri analizi uygulamalarında sıklıkla tercih edilmektedir. Farklı küme sayıları için en iyi performansı gösteren modeller belirlenmiş ve bu modellerin sonuçları detaylandırılmıştır. Bu süreç, müşteri davranışlarını anlamak ve stratejik kararlar almak için önemli içgörüler sunmaktadır.

Kümeleme Kıyaslama Metrikleri

Kümeleme algoritmalarının performansını değerlendirmek için çeşitli metrikler kullanılır. Bu metrikler, kümelerin ne kadar iyi ayrıldığını, kümeler içindeki veri noktalarının ne kadar tutarlı olduğunu ve kümeler arası benzerliği değerlendirir. En yaygın kullanılan kıyaslama metrikleri arasında Silhouette skoru, Calinski-Harabasz skoru ve Davies-Bouldin skoru yer almaktadır. Bu bölümde, bu üç metriğin hesaplama yöntemleri ve kullanım alanları detaylandırılacaktır.

4.1. Silhouette Skoru

Silhouette skoru, bir veri noktasının kendi kümesi ile diğer kümeler arasındaki benzerlik farkını ölçer. Her bir veri noktası için hesaplanan Silhouette değeri, tüm veri noktaları için ortalamalar alınarak genel Silhouette skoru elde edilir. Eşitlik 1'de Silhouette skoru hesaplama yöntemi verilmiştir.

$$S(i) = \frac{b(i) - a(i)}{\max\big(a(i), b(i)\big)} \qquad (1)$$

Burada $S(i)$ veri noktası i için Silhouette skorunu, $a(i)$ veri noktası i 'nin kendi kümesi içindeki ortalama mesafesini, $b(i)$ veri noktası i'nin en yakın komşu küme içindeki ortalama mesafesini ifade etmektedir.

Silhouette skoru, -1 ile 1 arasında bir değer alır. Değer 1'e yaklaştıkça, veri noktası iyi bir şekilde kümelenmiştir. 0'a yakın değerler, veri noktasının kümeler arasında olduğunu ve kötü bir şekilde kümelendiğini gösterir. Negatif değerler ise veri noktasının yanlış kümelendiğini gösterir.

4.2. Calinski-Harabasz Skoru

Calinski-Harabasz skoru, kümeler arasındaki ayrımı ve kümeler içindeki yoğunluğu ölçen bir metriktir. Yüksek değerler, iyi bir kümeleme yapısını gösterir. Eşitlik 2'de Calinski-Harabasz skoru hesaplama yöntemi verilmiştir.

$$CH = \frac{\sum_{k=1}^{K} n_k(\mu_k - \mu)^2}{\sum_{k=1}^{K} \sum_{i \in C_k}(x_i - \mu_k)^2} \times \frac{N-K}{K-1} \qquad (2)$$

Burada, CH Calinski-Harabasz skorunu, N Toplam veri noktası sayısını, K Küme sayısını, n_k k. kümedeki veri noktası sayısını, μ_k k. kümenin merkezini, μ tüm veri noktalarının merkezini, x_i i. veri noktasını, C_k k. kümeyi ifade etmektedir.

Calinski-Harabasz skoru, kümeler arasındaki uzaklıkların kümeler içindeki uzaklıklara oranını ölçer. Yüksek bir CH skoru, iyi ayrılmış ve yoğun kümeleri gösterir.

4.3. Davies-Bouldin Skoru

Davies-Bouldin skoru, her bir kümenin diğer kümelerle olan benzerliğini ölçer. Daha düşük değerler, daha iyi bir kümeleme yapısını gösterir. Eşitlik 3'te Davies-Bouldin skoru hesaplama yöntemi verilmiştir.

$$DB = \frac{1}{K}\sum_{i=1}^{K} \max_{j \neq i} \left(\frac{d_i + d_j}{d_{ij}}\right) \qquad (3)$$

Burada DB Davies-Bouldin skorunu, K küme sayısını, d_i i. kümenin ortalama dağılımını (kümeler içindeki ortalama mesafe), d_{ij} i ve j kümeleri arasındaki mesafeyi ifade etmektedir.
Davies-Bouldin skoru, her bir kümenin diğer kümelerle olan benzerliğini ölçer. Düşük bir DB skoru, kümelerin birbirinden iyi ayrıldığını ve kümeler içindeki benzerliğin yüksek olduğunu gösterir.

Silhouette skoru, kümelerin ne kadar iyi ayrıldığını ve kendi içinde ne kadar tutarlı olduğunu gösterir. Yüksek Silhouette skorları, iyi bir kümeleme yapısının göstergesidir. Calinski-Harabasz skoru, kümelerin yoğunluğunu ve ayrılığını ölçerek, yüksek değerlerin iyi bir kümeleme yapısını gösterdiği bir metriktir. Davies-Bouldin skoru ise kümeler arası benzerliği değerlendirir ve düşük değerler, iyi bir kümeleme yapısının göstergesidir. Bu metrikler, farklı açılardan kümeleme kalitesini değerlendirir ve her biri farklı kümeleme özelliklerine odaklanır. Bu nedenle, en iyi kümeleme yapısını belirlerken bu metriklerin birlikte değerlendirilmesi önemlidir. Uygulamanın gereksinimlerine ve veri setinin yapısına bağlı olarak, bu metriklerin ağırlıkları ve önem dereceleri değişebilir. Dolayısıyla, hangi metriklerin daha önemli olduğuna karar vererek en uygun kümeleme yöntemini seçmek gereklidir.

5. Bulgular ve Tartışma

Çizelge-1'de, FCM kümeleme algoritması için çeşitli metriklerle değerlendirilen farklı küme sayılarının performans skorları sunulmaktadır. Silhouette skoru, Calinski-Harabasz skoru ve Davies-Bouldin skoru olmak üzere üç ana performans metriği incelenmiştir.

Silhouette skoru, bir kümenin ne kadar sıkı bir şekilde kümelendiğini ve diğer kümelerden ne kadar iyi ayrıldığını ölçer. Silhouette skorunun yüksek olması, daha iyi bir kümeleme kalitesini gösterir. Çizelgeye bakıldığında, 8 küme

sayısının en yüksek Silhouette skoruna (0.465) sahip olduğu görülmektedir. Bu durum, 8 kümenin veri noktalarını en iyi şekilde ayrıştırdığını ve kümeler içinde tutarlı olduğunu göstermektedir.

Çizelge 1. *FCM algoritması kümeleme sonuçları*

	Küme Sayısı						
	3	**4**	**5**	**6**	**7**	**8**	**9**
Silhouette	0,411	0,437	0,404	0,445	0,422	0,465	0,419
Calinski-Harabasz	124,332	122,948	95,778	117,892	128,450	120,165	228,348
Davies-Bouldin	0,894	0,869	0,877	1,059	0,665	1,067	0,569

Calinski-Harabasz skoru, kümelerin yoğunluğunu ve ayrılığını değerlendiren bir başka metriktir. Skorun yüksek olması, daha iyi bir kümeleme kalitesine işaret eder. Çizelge-1 incelendiğinde, 9 küme sayısının en yüksek Calinski-Harabasz skoruna (228,348) sahip olduğu görülmektedir. Bu durum, 9 küme sayısının kümeler arası ayrılığı en iyi sağladığını ve kümeler içindeki yoğunluğun yüksek olduğunu göstermektedir.

Davies-Bouldin skoru ise kümelerin ne kadar iyi ayrıldığını ve kümeler arası benzerliği değerlendirir. Daha düşük Davies-Bouldin skorları, daha iyi kümeleme kalitesine işaret eder. Çizelgeye göre, en düşük Davies-Bouldin skoru (0,569) 9 küme sayısında elde edilmiştir. Bu durum, 9 küme sayısının kümeler arası ayrılığı en iyi sağladığını göstermektedir.

Sonuç olarak, Silhouette skoru en yüksek olan 8 küme sayısı, veri noktalarını iyi bir şekilde kümelese de, Calinski-Harabasz ve Davies-Bouldin skorlarına bakıldığında 9 küme sayısı daha iyi performans göstermektedir. Bu bulgular, FCM kümeleme algoritması ile yapılan analizde, en uygun küme sayısının 9 olduğunu göstermektedir. Ancak, uygulamaya ve veri yapısına bağlı olarak farklı metriklerin ağırlıkları ve önem dereceleri değişebilir. Dolayısıyla, nihai karar verirken bu metriklerin birlikte değerlendirilmesi önemlidir.

Çizelge 2. *Bulanık Mantık Tabanlı Kümeleme algoritması kümeleme sonuçları*

	Küme Sayısı						
	3	**4**	**5**	**6**	**7**	**8**	**9**
Silhouette	0,48	0,48	0,39	0,53	0,53	0,53	0,12
Calinski-Harabasz	32,74	32,74	27,20	34,88	34,88	34,88	17,80
Davies-Bouldin	1,88	1,88	2,00	1,86	1,86	1,86	2,57

Çizelge-2'de bulanık mantık tabanlı kümeleme algoritması kümeleme sonuçları görülmektedir. Alternatif fuzzy modelinin performans değerlendirmesi, üç ana metrik olan Silhouette skoru, Calinski-Harabasz skoru ve Davies-Bouldin skoru ile yapılmıştır. Silhouette skoru, kümelerin ne kadar iyi ayrıldığını ve kendi içinde ne kadar tutarlı olduğunu ölçer. Alternatif fuzzy modelde, 3 ve 4 küme sayısında 0,48 ile yüksek bir değer elde edilmiştir. En yüksek değerler ise 6, 7 ve 8 küme sayılarında 0,53 olarak kaydedilmiştir. 5 küme sayısında 0,39 ile bir düşüş gözlenirken, 9 küme sayısında 0,12 ile en düşük değer elde edilmiştir. Calinski-Harabasz skoru, kümelerin yoğunluğunu ve ayrılığını değerlendirir. 3 ve 4 küme sayısında 32,74 ile benzer değerler elde edilmiştir. En yüksek skorlar ise 6, 7 ve 8 küme sayılarında 34,88 olarak gözlenmiştir. 5 küme sayısında 27,20 ile bir düşüş, 9 küme sayısında ise 17,80 ile en düşük değer elde edilmiştir. Davies-Bouldin skoru ise kümeler arası ayrımı ve kümeler içi benzerliği değerlendirir. 6, 7 ve 8 küme sayılarında 1,86 ile en düşük değerler elde edilmiştir, bu da kümeler arası ayrımın en iyi olduğu durumu ifade eder. 3 ve 4 küme sayısında 1,88 ile benzer değerler görülmüş, 5 küme sayısında 2,00 ile bir artış, 9 küme sayısında ise 2,57 ile en yüksek değer elde edilmiştir, bu da kümeler arası ayrımın en kötü olduğu durumu belirtir.

FCM modeli ile karşılaştırıldığında, bazı metrikler açısından farklılıklar gözlenmektedir. FCM modelinde en yüksek Silhouette skoru 8 kümede 0,465 iken, bulanık mantık tabanlı kümeleme modelinde en yüksek skor 6, 7 ve 8 kümelerde 0,53 olarak elde edilmiştir. Bu durum, alternatif fuzzy modelin genel olarak daha iyi bir kümeleme kalitesi sağladığını göstermektedir. Calinski-Harabasz skoru açısından, Fuzzy C-Means modeli çok daha yüksek değerler göstermektedir; en yüksek skor 9 kümede 228,348 iken, alternatif modelde en yüksek skor 6, 7 ve 8 kümelerde 34,88 olarak gözlenmiştir. Bu, FCM modelinin kümeler arası ayrım ve yoğunluk açısından daha üstün olduğunu belirtir. Davies-Bouldin skorunda ise FCM modeli daha düşük değerler elde etmiştir; en düşük skor 9 kümede 0,569 iken, alternatif modelde en düşük skor 6, 7 ve 8 kümelerde 1,86 olarak elde edilmiştir. Bu da FCM modelinin kümeler arası ayrımda daha başarılı olduğunu göstermektedir.

Silhouette skoru açısından bulanık mantık tabanlı kümeleme modelinde daha yüksek performans sergilerken, Calinski-Harabasz ve Davies-Bouldin skorlarında FCM modeli üstünlük göstermektedir. Bu nedenle, hangi modelin kullanılacağına karar verirken, uygulamanın gereksinimlerine ve veri setinin yapısına göre her iki modelin de avantajları ve dezavantajları dikkate alınmalıdır.

Çizelge 3. *K-Means algoritması kümeleme sonuçları*

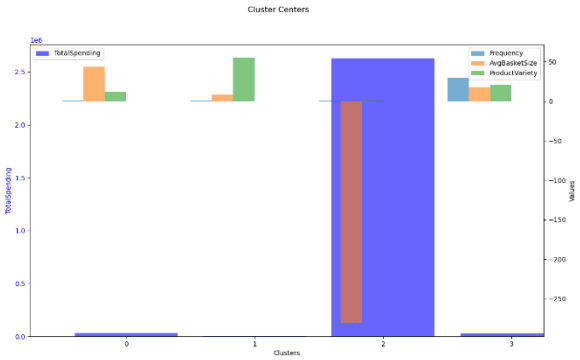
	Küme Sayısı						
	3	**4**	**5**	**6**	**7**	**8**	**9**
Silhouette	0,68	0,69	0,44	0,46	0,45	0,45	0,47
Calinski-Harabasz	124,93	119,92	148,20	192,96	206,87	234,19	304,52
Davies-Bouldin	0,33	0,27	0,51	0,55	0,56	0,54	0,54

Çizelge-3, K-Means kümeleme algoritmasının çeşitli küme sayılarındaki performansını üç ana metrikle değerlendirmektedir: Silhouette skoru, Calinski-Harabasz skoru ve Davies-Bouldin skoru. Silhouette skoru, kümelerin ne kadar iyi ayrıldığını ve kendi içinde ne kadar tutarlı olduğunu ölçer. Çizelgeye göre, 4 küme sayısında en yüksek Silhouette skoru 0,69 elde edilmiştir, bunu 3 küme sayısında 0,68 takip etmektedir. Bu durum, 3 ve 4 küme sayısının veri noktalarını en iyi şekilde ayrıştırdığını ve kümeler içinde tutarlı olduğunu göstermektedir. Ancak, 5 ve 9 küme sayıları arasında Silhouette skoru 0,44 ile 0,47 arasında değişiklik göstermekte, bu da bu küme sayılarında kümeleme kalitesinin biraz daha düşük olduğunu işaret etmektedir.

Calinski-Harabasz skoru, kümelerin yoğunluğunu ve ayrılığını değerlendirir. En yüksek değer 9 küme sayısında 304,52 olarak elde edilmiştir. Bu durum, 9 küme sayısının kümeler arası ayrımı en iyi sağladığını ve kümeler içindeki yoğunluğun yüksek olduğunu göstermektedir. Genel olarak, küme sayısı arttıkça Calinski-Harabasz skorunun da arttığı görülmekte, bu da daha fazla küme ile daha iyi bir ayrım ve yoğunluk sağlandığını göstermektedir.

Davies-Bouldin skoru ise, kümeler arası ayrımı ve kümeler içi benzerliği değerlendirir. Daha düşük Davies-Bouldin skorları, daha iyi bir kümeleme kalitesine işaret eder. Çizelgeye göre, en düşük Davies-Bouldin skoru 4 küme sayısında 0,27 olarak elde edilmiştir, bu da 4 küme sayısının kümeler arası ayrımı en iyi sağladığını göstermektedir. 3 küme sayısı da düşük bir Davies-Bouldin skoru (0,33) göstermekte, bu da iyi bir küme ayrımına işaret etmektedir. Küme sayısı arttıkça Davies-Bouldin skoru genellikle artmakta, bu da kümeler arası benzerliğin arttığını ve ayrımın zorlaştığını göstermektedir.

Genel olarak değerlendirildiğinde, K-Means algoritması için 4 küme sayısı, en yüksek Silhouette skoru (0,69) ve en düşük Davies-Bouldin skoru (0,27) ile en iyi performansı göstermektedir. Calinski-Harabasz skoru açısından 9 küme sayısı en yüksek değeri sunmakta olup, kümeler arası ayrımın ve yoğunluğun en iyi sağlandığını göstermektedir. Ancak, genel kümeleme kalitesi için 4 küme sayısının optimal olduğunu söyleyebiliriz. Bu bulgular, veri setinin yapısına ve uygulama gereksinimlerine göre model seçimini yönlendirebilir.



Şekil-3: K-Means Kümeleme Sonuçları: Küme Merkezlerinin Özelliklere Göre Dağılımı

Şekil-3'te, dört farklı kümenin (Cluster 0, Cluster 1, Cluster 2 ve Cluster 3) merkezlerinin dört temel özellik açısından (Toplam Harcama, Frekans, Ortalama Sepet Büyüklüğü ve Ürün Çeşitliliği) analizi sunulmaktadır. Grafikte iki farklı eksen kullanılarak görselleştirilmiştir: birincil eksen (sol), Toplam Harcama değerlerini gösterirken, ikincil eksen (sağ), diğer üç özelliğin değerlerini göstermektedir.

Cluster 0, düşük harcama yapan, nadiren alışveriş yapan ve sepetinde çok az çeşitlilik olan müşterileri temsil etmektedir. Bu müşteriler, toplam harcama ve alışveriş frekansı açısından en düşük değerleri göstermektedir. Cluster 1 ise, nadiren alışveriş yapan ancak sepetinde nispeten ortalama büyüklükte ve çeşitlilikte ürün bulunduran müşterileri tanımlamaktadır. Bu küme, toplam harcama açısından düşük, ancak ortalama sepet büyüklüğü ve ürün çeşitliliği açısından daha dengeli bir profil sergilemektedir.

Cluster 2, yüksek harcama yapan, ortalama sıklıkta alışveriş yapan ve sepetinde ortalama büyüklükte ve çeşitlilikte ürün bulunduran müşterileri temsil etmektedir. Bu küme, toplam harcama açısından en yüksek değerlere sahip olup, şirket için kritik öneme sahip bir müşteri segmentini ifade etmektedir. Cluster 3 ise, nadiren alışveriş yapan ancak sepetinde nispeten ortalama büyüklükte ve yüksek çeşitlilikte ürün bulunduran müşterileri temsil etmektedir. Bu küme, düşük frekans ve toplam harcama, ancak yüksek ürün çeşitliliği ile dikkat çekmektedir.

Bu kümeleme analizi, müşteri segmentlerini belirlemekte oldukça etkilidir ve farklı alışveriş davranışlarına sahip müşteri gruplarını ortaya koymaktadır. Özellikle Cluster 2'nin yüksek harcama yapan müşterileri temsil etmesi, bu grubun şirket için kritik bir öneme sahip olduğunu göstermektedir. Diğer kümeler, daha düşük harcama ve frekansa sahip müşteri segmentlerini tanımlayarak, farklı pazarlama stratejileri geliştirmek için temel oluşturabilir. Örneğin, Cluster 0 ve Cluster 1, alışveriş sıklığını ve toplam harcamayı artırmak için özel teklifler ve kampanyalarla hedeflenebilir. Bu analiz, müşteri segmentasyonu ve pazarlama stratejilerinin optimizasyonu için değerli bilgiler sunmaktadır. Farklı kümeler arasındaki belirgin farklılıklar, müşterilerin alışveriş davranışlarını anlamak ve bu davranışlara uygun stratejiler geliştirmek için kullanılabilir.

K-Means, FCM ve alternatif fuzzy modellerinin performans kıyaslamasında, her modelin kendine özgü avantaj ve dezavantajları bulunmaktadır. K-Means algoritması, 4 küme sayısı için en yüksek Silhouette skoru (0,69) ve en düşük Davies-Bouldin skoru (0,27) ile en iyi performansı göstermektedir. Calinski-Harabasz skoru açısından ise 9 küme sayısı en yüksek değeri sunmaktadır. Fuzzy C-Means modeli, genel olarak yüksek Calinski-Harabasz ve düşük Davies-Bouldin skorları ile kümeler arası ayrımda üstünlük göstermektedir. Alternatif fuzzy model ise Silhouette skorunda daha yüksek performans sergilemekte olup, belirli küme sayılarında daha iyi bir kümeleme kalitesi sunmaktadır. Bu bulgular, hangi modelin kullanılacağına karar verirken, uygulamanın gereksinimlerine ve veri setinin yapısına göre her üç modelin de avantajlarının ve dezavantajlarının dikkate alınması gerektiğini göstermektedir.

Sonuç

Bu çalışmada, müşteri segmentasyonu için K-Means, FCM ve Bulanık Mantık tabanlı kümeleme algoritmalarının performansları karşılaştırılmıştır. K-Means algoritması, yüksek Silhouette skorları ve Calinski-Harabasz indeks değerleri ile iyi tanımlanmış kümeler sunarken, FCM algoritması, veri noktalarının birden fazla kümeye ait olabilmesini sağlayarak esneklik göstermiştir. Bulanık Mantık tabanlı kümeleme, özellikler arasındaki ilişkileri

yakalama ve örtüşen kümeleri yönetme konusunda daha sofistike bir yöntem sunmuştur. Her üç algoritmanın da kendi avantajları ve dezavantajları bulunmaktadır ve seçim, uygulamanın gereksinimlerine ve veri setinin yapısına bağlı olarak yapılmalıdır. Online Retail veri seti üzerinde yapılan analizler, müşteri segmentasyonunun pazarlama stratejileri ve müşteri ilişkileri yönetimi açısından önemini vurgulamaktadır. Gelecekteki çalışmalar, farklı veri setleri ve daha ileri kümeleme teknikleri kullanarak bu bulguları genişletebilir ve müşteri segmentasyonu için daha etkili stratejiler geliştirebilir.

Kaynakça

[1] Anitha, P. and Patil, M. M. RFM model for customer purchase behavior using K-Means algorithm, Journal of King Saud University - Computer and Information Sciences, 2022, 34, pp. 1785–1792.

[2] Hu, Y. H. and Yeh, T. W. Discovering valuable frequent patterns based on RFM analysis without customer identification information, Knowledge-Based Systems, 2014, 61, pp. 76–88.

[3] Khajvand, M. , Zolfaghar, K. , Ashoori, S. , and Alizadeh, S. Estimating customer lifetime value based on RFM analysis of customer purchase behavior: Case study, Procedia Computer Science, 2011, 3, pp. 57–63.

[4] Khalili-Damghani, K. , Abdi, F. , and Abolmakarem, S. 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, 2018, 73, pp. 816–828.

[5] Griva, A. , Bardaki, C. , Pramatari, K. , and Papakiriakopoulos, D. Retail business analytics: Customer visit segmentation using market basket data, Expert Systems with Applications, 2018, 100, pp. 1–16.

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

[7] Arunachalam, D. and Kumar, N. Benefit-based consumer segmentation and performance evaluation of clustering approaches: An evidence of data-driven decision-making, Expert Systems with Applications, 2018, 111, pp. 11–34.

[8] Yoo, B. and Jang, M. A bibliographic survey of business models, service relationships, and technology in electronic commerce, Electronic Commerce Research and Applications, 2019, 33, pp. 100818.

[9] Zerbino, P. , Aloini, D. , Dulmin, R. , and Mininno, V. Big Data-enabled Customer Relationship Management: A holistic approach, Information Processing & Management, 2018, 54, pp. 818–846.

[10] Han, J. , Kamber, M. and Pei, J. Data mining: concepts and techniques, Elsevier, 2011.

[11] Brito, P. Q. , Soares, C. , Almeida, S. , Monte, A. , and Byvoet, M. Customer segmentation in a large database of an online customized fashion business, Robotics and Computer-Integrated Manufacturing, 2015, 36, pp. 93–100.

[12] Femina, B. T. and Sudheep, E. M. An Efficient CRM-Data Mining Framework for the Prediction of Customer Behaviour, Procedia Computer Science, 2015, 46, pp. 725–731.

[13] Sheng, J. , Amankwah-Amoah, J. , and Wang, X. A multidisciplinary perspective of big data in management research, International Journal of Production Economics, 2017, 191, pp. 97–112.

[14] Nguyen, D. H. , Leeuw, S. de , and Dullaert, W. E. H. Consumer Behaviour and Order Fulfilment in Online Retailing: A Systematic Review, International Journal of Management Reviews, 2018, 20, pp. 255–276.

[15] Patak, M. , Lostakova, H. , Curdova, M. , and Vlckova, V. The E-Pharmacy Customer Segmentation Based on the Perceived Importance of the Retention Support Tools, Procedia - Social and Behavioral Sciences, 2014, 150, pp. 552–562.

[16] Carnein, M. and Trautmann, H. Customer Segmentation Based on Transactional Data Using Stream Clustering, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2019, 11439 LNAI, pp. 280–292.

[17] Kaur, A. , Pal, S. K. , and Singh, A. P. Hybridization of Chaos and Flower Pollination Algorithm over K-Means for data clustering, Applied Soft Computing, 2020, 97, pp. 105523.

[18] Fu, X. , Chen, X. , Shi, Y. T. , Bose, I. , and Cai, S. User segmentation for retention management in online social games, Decision Support Systems, 2017, 101, pp. 51–68.

[19] Holý, V. , Sokol, O. , and Černý, M. Clustering retail products based on customer behaviour, Applied Soft Computing, 2017, 60, pp. 752–762.