GENETİK ALGORİTMA VE K-ORTALAMALAR ALGORİTMASININ TAVSİYE SİSTEMLERİ İÇİN UYGULANMASI

APPLICATION OF GENETIC ALGORITHM AND K-MEANS ALGORITHM FOR RECOMMENDER SYSTEMS

MERVE POSLU

PROF. DR. SEVİL BACANLI

Tez Danışmanı

PROF. DR. GÜVENÇ ARSLAN

Tez Eş Danışmanı

Hacettepe Üniversitesi

Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliğinin

İstatistik Anabilim Dalı için Öngördüğü

YÜKSEK LİSANS TEZİ olarak hazırlanmıştır.

2023

ÖZET

GENETİK ALGORİTMA VE K-ORTALAMALAR ALGORİTMASININ TAVSİYE SİSTEMLERİ İÇİN UYGULANMASI

Merve POSLU

Yüksek Lisans, İstatistik Bölümü

Tez Danışmanı: Prof. Dr. Sevil BACANLI

Tez Eş Danışmanı: Prof. Dr. Güvenç Arslan

Eylül 2023, 76 sayfa

Müşteri memnuniyetine verilen önemin her geçen gün artması ile birlikte işletmeler, müşterilerin ilgi alanlarını ve tercihlerini keşfetmeyi hedefler. Dolayısıyla uygun ürün veya hizmetleri sağlamaları bulunduğumuz rekabet ortamında oldukça kritiktir. Günümüzde de dijitalleştirilmiş hizmetlere önemli yatırımlar yapılmaktadır. Örneğin e-ticaret alanı, COVID-19 pandemisini avantaja çevirerek önemli gelişmelerin olduğu bir alan oluşturmuştur. E-ticaretteki bu gelişmelerden biri de tavsiye sistemlerinden en iyi şekilde faydalanılmasıdır.

Tavsiye sistemleri, kullanıcı ile öğeler arasındaki etkileşimi izleyerek çalışır. Bu etkileşimlerin ortaya konmasında ise kümeleme yöntemleri kullanılarak, benzer özelliklere sahip kullanıcıları veya öğeleri gruplandırabilir. Ayrıca, kümeleme yöntemleri kullanarak tavsiye sistemlerindeki bazı dezavantajlar giderilebilir. Bu dezavantajların giderilmesi noktasında, optimize edilmiş kümeleme

süreçleri önemli bir rol oynamaktadır. Optimize edilmiş kümeleme süreçleri, yeni öğeler ve yeni kullanıcılar için daha iyi başlangıç önerileri sağlayabilir. Aynı zamanda da veri seyrekliği vb. sorunları ele alabilir.

Genetik algoritmaların optimizasyon alanında kullanımı da oldukça popülerdir. Temel olarak doğal seleksiyon ve genetik prensiplere dayanan etkili optimizasyon algoritmaları olarak, çeşitli alan ve uygulamalarda başarıyla kullanılan sezgisel tekniklerden biridir. Genetik algoritmada, çözüm dizilerinden oluşan bir başlangıç popülasyonu, çaprazlama ve mutasyon gibi genetik operatörler kullanılmaktadır. Genetik algoritmanın kümeleme süreçlerinin optimizasyonunda kullanılması, tavsiye sistemlerine olumlu etkiler sağlayabilmektedir. Bu şekilde tavsiye sistemleri için daha iyi kümeleme sonuçları, veri seyrekliğinin azaltılması, kişiselleştirilmiş öneriler, soğuk başlangıç probleminin azaltılması ve daha iyi çeşitlilik sağlama gibi olumlu etkiler sağlanabilmektedir. Bu yaklaşım, tavsiye sistemlerinin performansını artırarak kullanıcı memnuniyetini ve öneri sonuçlarının etkinliğini arttırmayı hedeflemektedir.

Tez çalışmasında, öncelikle geleneksel tavsiye sistemleri ve dezavantajlı durumları incelenmiştir. Ardından kümeleme yöntemleri ve tavsiye sistemlerinde kullanımı detaylandırılarak sonrasında genetik algoritma tanıtılmıştır. Genetik algoritmalar, pazarlama veri kümesinde K-ortalamalar kümeleme işlemlerini optimize etmek için uygulanmış olup performansı K-ortalamalar kümelemesi ile karşılaştırılmıştır. Bu çalışma, genetik algoritma kümeleme tekniğini kullanarak (GA-KOK) kullanıcıları daha iyi profillemek ve tavsiye çıktılarını iyileştirerek tavsiyelerin oluşturulmasını sağlamaktır.

Anahtar kelimeler: Tavsiye sistemleri, Kümeleme, Genetik algoritma

ABSTRACT

APPLICATION OF GENETIC ALGORITHM AND K-MEANS ALGORITHM FOR RECOMMENDER SYSTEMS

Merve POSLU

Master of Science, Department of Statistics

Supervisor: Prof. Dr. Sevil BACANLI

Co- Supervisor: Prof. Dr. Güvenç ARSLAN

September 2023, 76 pages

As there is an increasing emphasis on customer satisfaction, businesses are aiming to discover their customers' interests and preferences. Therefore, providing suitable products or services is crucial in today's competitive environment. Significant investments are being made in digitized services today. For example, the e-commerce sector has seen significant developments, taking advantage of the COVID-19 pandemic. One of these developments in e-commerce is the effective utilization of recommendation systems.

Recommendation systems work by tracking interactions between users and items. Cluster analysis methods are used as a preprocessing step to reveal these interactions, grouping users or items with

similar characteristics. Additionally, using clustering methods can help address some of the disadvantages in recommendation systems. In addressing these disadvantages, optimized clustering processes play a significant role. Optimized clustering processes can provide better initial recommendations for new users and items while addressing the data sparsity issue.

The use of genetic algorithms in optimization is also quite popular. Based on the principles of natural selection and genetics, genetic algorithms are effective optimization algorithms used successfully in various fields and applications. In genetic algorithms, a starting population consisting of solution sequences is used, along with genetic operators like crossover and mutation. Using genetic algorithms to optimize clustering processes can have a positive impact on recommendation systems. This approach can lead to better clustering results in recommendation systems, reduced data sparsity, personalized recommendations, mitigating the cold start problem, and enhancing diversity. This approach aims to improve the performance of recommendation systems, ultimately enhancing user satisfaction and the effectiveness of recommendation results.

In this thesis, we first examined traditional recommendation systems and their drawbacks. Then, clustering methods and their use in recommendation systems were detailed, followed by the introduction of genetic algorithms. Genetic algorithms were applied to optimize K-means clustering processes in the marketing dataset, and their performance was compared to K-means clustering. The aim of this study is to use the genetic algorithm clustering technique (GA-KOK) to better profile users and improve recommendation outputs, thereby enhancing the process of recommendation generation.

Keywords: Recommendation systems, Clustering, Genetic algorithm

TEŞEKKÜR

Yüksek lisans dönemi ve tez süresi boyunca ilgileri ve tecrübelerini benimle paylaşarak çok kıymetli yönlendirmeleri ile bana destek olan Sayın Prof. Dr. SEVİL BACANLI ve Prof. Dr. GÜVENÇ ARSLAN'a,

Tezin değerlendirilmesi aşamalarında yönlendirici yorumları ile öneriler sunan Sayın jüri üyelerim Prof. Dr. Sevgi YURT ÖNCEL ve Doç. Dr. Duygu İÇEN'e

teşekkürlerimi sunarım.

Tezimin hazırlanması sürecinde bana destek olan annem ve motivasyon sağlayıp umut veren sevdiklerime teşekkür ederim.

Merve Poslu

Eylül 2023, Ankara.

İÇİNDEKİLER

ŞEKİLLER DİZİNİ

ÇİZELGELER LİSTESİ

SİMGELER VE KISALTMALAR LİSTESİ

Simgeler

A, B	Kosinüs benzerliğinde temsili noktalar
u	Kulanıcı
i	Öğe

Kısaltmalar

GA	Genetik Algoritma
GA-KOK	Genetik Algoritma K-ortalamalar Kümeleme
İF	İşbirlikçi Filtreleme
SOM	Self Organizing Maps
LBA	Location Based Advertising
CACF-GA	Context-Aware Collaborative Filtering using Genetic Algorithm
KNN	K-nearest neighbors
PM	Prefer Matrix
SVD++	Singular Value Decomposition
CF	Collaborative Filter

1.GİRİŞ

Günümüzde büyük verinin hızlı artışı sebebiyle hizmetlerin dijitalleşmesi pazarlama, ticaret ve bilişim gibi birçok alanda önemli hale gelmektedir. Özellikle COVID-19 pandemisi sürecinde ve sonrasında kişilerin, hizmete online erişme olanakları önem kazanmış olup uzaktan satın alma alışkanlıkları birçok gelişim alanı açığa çıkarmıştır [1]. E-ticaret alanında farklı üretici ve satıcıları aynı platformda buluşturan "Marketplace" kavramı da öne çıkarak ticari akışa önemli ölçüde katkıda bulunmuş ve çeşitlilik kazandırmıştır. Dolayısıyla bu durum rekabeti de beraberinde getirmektedir. Benzer bir hizmet veya ürün için ne kadar çok sağlayıcı olursa, rekabet edilen alanda da performansın yüksek olması kritiktir.

Ticaret alanında birçok hizmet ve satış kanalının dijitalleşme anlamında önemli kaynaklar ayırarak kendilerini rekabetçi koşullara adapte edip performanslarını ileriye taşımayı hedefledikleri görülmektedir. Bunun yanı sıra sosyal uygulamalarda da daha çok kullanıcı kazanımı için birçok yöntem kullanırlar. Bu yöntemlerden tavsiye sistemleri yaygın kullanılmaktadır. Tavsiye sistemleri, kullanıcıların geçmiş tercihlerini, davranışlarını ve diğer özelliklerini analiz ederek onların potansiyel veya gelecekteki tercihlerini tahmin eden yöntemlerdir. Tavsiye sistemleri kullanıcıların birçok alanda kullanılan öğelere ya da hizmetlere ait kapsamlı bilgiler arasında gezinmesine yardımcı olabilecek güçlü bir araçtır. Elektronik işlemler sayesinde internet üzerinden alınan hizmetler, çevrimiçi ticaret/alışveriş, çevrimiçi kütüphane ve öğrenme gibi birçok uygulamada tavsiye sistemleri kullanılmaktadır. Bunların yanı sıra oldukça popüler olan Netflix, Amazon, Youtube, Instagram gibi sosyal ağlarda da doğru öneri ile kullanıcıya en iyi deneyimi sağlayarak ekran sürelerinin her geçen gün daha da uzatılması hedeflenmektedir. Bu hedefte ise tavsiye sistemleri kilit rol üstlenmektedir.

Farklı platformlarda her geçen gün kullanıcı ve öğe artışı kompleks problemleri de ortaya çıkarmaktadır. Bu problemler karşısında çok fazla parametre ile karşı karşıya kalınan günümüz şartlarında, güncel ve pratik çözüm metotları üretmek önemli hale gelmektedir. Evrimsel algoritmalar (evolutionary algorithm), kolayca çözülemeyen problemlere iyileştirilmiş ve yaklaşık

çözümler sağlamak için kullanılmaktadır. Bu sebeple tavsiye sistemleri için kullanılan yöntemlerde evrimsel algoritmaların kullanımı da giderek artmaktadır. Evrimsel algoritma türlerinden genetik algoritmalar (GA), büyük ve karmaşık uzaylarda arama yapabilen stokastik arama teknikleridir. Çözüm üretirken biyolojik evrimin işleyişini rol model alarak doğal seçilim ve genetik kurallara dayandırılmış sürekli iyileşen çözümler üretir. Çözümler için bir uygunluk (fitness) fonksiyonu ve yeni çözümlerin oluşturulması için yeniden kopyalama (recombination), değiştirme (mutation) gibi operatörleri kullanır [2]. GA yaygın optimizasyon yöntemleriyle karşılaştırıldığında parametre kümesi yerine bir çözümü temsil etme yöntemi olarak bir bit, sayı, karakter dizisi gibi parametre takımının kodlanmış haliyle çözüme gidilir. GA çeşitli gerçek dünya uygulamalarında kullanım alanı bulmaktadır. Sinir ağlarında genetik optimizasyon olarak kullanımının yanı sıra görüntü segmentasyonu gibi konularda karmaşık optimizasyon problemlerini çözmede ve görüntü analizinin farklı alanlarında GA'dan faydalanılır [3]. Başka bir örnek olarak kablosuz sensör ağlarında, GA uygunluk fonksiyonu yardımıyla tüm operasyonel aşamalar optimize edilmekte ve hatta özelleştirmek için kullanılmaktadır [4]. Tıp bilimi başta olmak üzere finansal piyasalar, üretim sistemleri, ekonomi ve ticaret gibi sektörlerde de kullanım örnekleri yaygındır.

Satın alma veya hizmet talebi öncesinde tavsiye arayışında bulunma eğilimi, son yıllarda yaygınlaşmaktadır. Bu eğilim, büyük miktardaki bilgi, ürün ve hizmet gelişimi ile birlikte artar ve gerçek zamanlı olarak performans gösterebilen güçlü ve ölçeklenebilir tavsiye sistemleri oluşturmayı daha zorlu hale getirir. Tavsiye sistemlerinin ölçeklenebilirliğini artırmak ve zaman karmaşıklığını azaltmak için yaygın bir yaklaşım, kullanıcıların profillerine ve benzerliklerine dayalı olarak kümeleme yapmaktır. Kümelemede doğru küme sayısının seçimi, aykırı değerler ve başlangıç noktaların rastgele seçilmesi gibi dezavantajlar da bulunmaktadır. Bu dezavantajları gidermek ya da var olan çözüm performansını iyileştirmek için GA'nın kullanıldığı çalışmalar yaygınlaşmıştır. Zaman içinde farklılaşan problemlere ve girdilere karşı uygun olan yöntemler ile önemli çıktılar elde edilerek literatürde çeşitli yaklaşımlarla çözümler sunulmuştur [5].

Kim ve Ahn [6], çalışmasında GA' yı tavsiye sisteminde K-ortalama kümelemesinde optimal veya alt-optimal başlangıç elemanlarını seçmek için kullanır. Bu çalışmada, çevrimiçi alışveriş pazarını

etkin bir şekilde gruplandırma yapmak için genetik algoritmalara dayalı yeni bir kümeleme algoritması (GA K-ortalamalar) önerilmiştir. GA ile kümeleme tekniğinin, ilgili kümeleri daha etkin bir şekilde oluşturduğu ortaya konmuştur. Uygulama kısmında ise başlangıç elemanları GA tarafından optimize edilen K-ortalama kümelemesi gerçek bir e-ticaret gruplandırma çalışmasına uygulanmıştır. Sonrasında ise GA K-ortalamalar sonuçları, basit bir K-ortalamalar algoritması ve kendi kendini organize eden haritalar (Self Organizing Map-SOM) sonuçlarıyla karşılaştırılmıştır. Sonuçlar, GA K-ortalamalar kümelemesinin, diğer tipik kümeleme algoritmalarına kıyasla gruplandırma performansını iyileştirebileceğini göstermiştir. Ek olarak, çalışmada önerilen modelin öneri sistemleri için bir ön işleme aracı olarak kullanışlılığı doğrulanmıştır.

Bobadilla ve diğerleri [7], tavsiye sistemlerinde yürütülen işbirlikçi filtreleme süreçlerinde geçerli olan, kullanıcılar arasındaki benzerliği ölçmek için bir ölçü sunmuşlardır. Önerilen metrik, değerlerin ve ağırlıkların basit bir doğrusal kombinasyonu ile formüle edilmiştir. Değerler, aralarında benzerliğin elde edildiği her bir kullanıcı çifti için hesaplanırken, ağırlıklar yalnızca bir kez hesaplanır ve bir genetik algoritmanın, her bir tavsiye sisteminden gelen verilerin spesifik doğasına bağlı olarak tavsiye sisteminden ağırlıkları çıkardığı önceki bir aşamadan faydalanır. Elde edilen sonuçlar, tahmin kalitesi, öneri kalitesi ve performansında önemli gelişmeler sunmaktadır.

Dao ve diğerleri [8], hem kullanıcının tercihlerine hem de etkileşimin bağlamına dayalı olarak konum tabanlı reklamcılık (Location based Advertising -LBA) için genetik algoritma (Context-Aware Collaborative Filtering using Genetic Algorithm-CACF-GA) kullanarak Bağlama Duyarlı İşbirliğine Dayalı Filtreleme adını verdikleri yeni bir öneri modeli önermişlerdir. Sistemlerinde, bağlam benzerlik değerleri kümesini optimize etmek için GA uygulanmıştır.

Salehi ve diğerleri [9], çalışmasında e-öğrenme sistemindeki tavsiyenin kalitesini iyileştirmek için nitelik tabanlı filtreleme ve genetik tabanlı karma bir tavsiye sistemi önermektedir. Öğrenen taraf için materyallerin örtük veya gizli niteliklerinin ağırlıklarını optimize etmek adına GA kullanmışlardır. Sunulan sistemin iki ana modülü vardır: açık öznitelik tabanlı önerici ve örtük

öznitelik tabanlı önerici. Birinci modülde, materyallerin öğrenen için örtük veya gizli özniteliklerinin ağırlıkları GA'da kromozom olarak kabul edilmekte ve bu algoritma ağırlıkları tarihsel derecelendirmeye göre optimize etmektedir. Ardından, öğrencilerin görüşlerini temsil eden optimize edilmiş ağırlık vektörleri örtük öznitelikleri kullanılarak En Yakın Komşu Algoritması (KNN) tarafından öneri oluşturulur. İkincisinde, çok boyutlu bir bilgi modelinde öğrenme materyallerinin açık niteliklerine dayalı olarak öğrenenlerin ilgilerini modelleyebilen tercih matrisi (Prefer Matrix-PM) tanıtılır. Ardından, PM'ler arasında yeni bir benzerlik ölçüsü tanıtılır ve KNN tarafından öneriler oluşturulur. Deneysel sonuçlar, önerilen yöntemin doğruluk ölçümlerinde mevcut algoritmalardan daha iyi performans gösterdiğini ve soğuk başlatma ve seyreklik gibi bazı sorunları hafifletebileceğini göstermektedir.

Maghsoudi ve diğerleri [10], çalışmasında ilk olarak K-ortalamalar yöntemi ile temel bir veri kümeleme yöntemi anlatılmış ardından GA ile K-ortalamalar yöntemini geliştirmek için "GA-Clustering" olarak adlandırdıkları öneri modelini tanıtmışlardır. Son olarak "GA-Clustering" bilinen bazı veri setleri üzerinde incelenmiştir. Sonuçlar GA kümelemesinin K-ortalamalar kümeleme algoritmasından önemli ölçüde daha iyi kümelediğini göstermiştir.

Lv ve diğerleri [11], ilişkisel özelliklerin öneri sürecine değerli bilgiler katabileceğini düşünerek, bu nedenle işlenebilir bir öneri sistemi çerçevesi geliştirmişlerdir. Bu çerçevede, alan ontolojisindeki ilişkisel verileri entegre etmişler ve önerileri oluşturmak için GA kullanılmıştır. Deneysel sonuçlar, seyreklik ve soğuk başlangıç sorunlarıyla başa çıkma yöntemlerinde ve önerilerin doğruluğunda belirgin iyileştirmeler olduğunu göstermektedir.

Alhijawi [12], yapmış olduğu "The Use of the Genetic Algorithms in the Recommender Systems" adlı tez çalışmasında öğeye dayalı semantik benzerlik, n-kriterler ve çoklu filtreleme kriterleri fikirlerini genetik tabanlı öneri sistemi ile birleştiren yeni bir teknik sunmaktadır. Aktif kullanıcıya en iyi öğe listesini tahmin etmek için GA kullanılır. Sonuç olarak, popülasyondaki her birey bir aday öneri listesini temsil eder. Her liste, kalitesini ölçmek için üç teste tabi tutulur. Önerilen

sistem, seyreklik ve soğuk başlangıç problemlerinin etkisini hafifletir ve tavsiye sistemini, bir benzerlik ölçüsü kullanmaya gerek kalmadan veya hibrit sistem tarafından sağlanan herhangi bir ek bilgiye ihtiyaç duymadan tavsiye üretebilir hale getirir.

Seyrek [13], çalışmasında tavsiye sistemlerine genel bir bakış sunmaktadır. Ayrıca uygulama kısmında işbirlikçi ve içerik bazlı filtreleme kullanarak hibrit bir film öneri sistemi oluşturulmaktadır. Açık kaynaklı veriseti olan MovieLens üzerinde, içerik tabanlı filtreleme, işbirlikçi filtreleme ve birleşiminden oluşan hibrit filtreleme algortimaları ile tahminler yapılmıştır. Kullanılan birleştirme algoritmasındaki hata oranını minimize etmek için GA'dan faydalanılmıştır. Çalışma sonucunda, önerilen hibrit film tavsiye sisteminin daha düşük hata yüzdesiyle sonuçlar verdiği bulunmuştur.

Anwar ve Uma [14], İşbirlikçi Filtreleme (CF) ve Tekil Değer Ayrıştırma (Singular Value Decomposition - SVD++) yöntemini kullanarak bir öneri sistemi geliştirmek için yeni bir yaklaşım sunmaktadır. Bu çalışma, kullanıcı ve ürün arasındaki benzerliği derecelendirme matrislerinden kosinüs benzerliği kullanarak bulma, eksik derecelendirmeleri matris ayrıştırma yöntemi kullanarak tahmin etme ve kullanıcı tercihine göre en iyi N öğe önerme gibi mevcut öneri sistemlerini genişletme amacı taşımaktadır. Önerilen yaklaşım, MovieLens veri kümesi ile test edilmiş olup önerilen yöntemlerin performans analizi hata metrikleri ile gerçekleştirilmiştir. Sonuçlar, önerilen yaklaşımın çapraz doğrulama yapıldığında daha düşük bir hata oranı verdiğini göstermektedir.

Sultan ve diğerleri [15], öğretmenler için özellikle kısa bir süre içinde test kağıtları oluşturmanın zorluğunu problem olarak ele almış ve GA ile K-ortalamalar algoritmasını birleştiren bir yöntem önermişlerdir. Bu yöntem , Bloom Taksonomisi'ne uygun bir test kağıdı oluşturmak için kullanılır ve K-ortalamalar yöntemi, aynı özelliklere sahip soruları bir araya getirerek soruları altı farklı gruba böler. Bu, soruların aynı özelliklere sahip tekrarlarını önleme sorununu ele alır. Önceki adımlara dayalı olarak, sınav kağıdında GA'nın Bloom'un altı seviyesine göre en iyi sınavı

seçmesini kolaylaştırır. Bu yaklaşım, birinci aşamada soru bankasını oluşturur, ikinci aşamada soruları benzer özelliklere sahip altı gruba böler ve üçüncü aşamada soruların başlangıç nüfusu oluşturulurken tekrarlanmamasını sağlamak için rastgele bir sıralama işlevi kullanır. Çalışmada tüm soru türlerini içeren 800 soruluk bir soru bankası kullanılmıştır. Bu nedenle önerilen yöntem, sınav kağıtları oluşturma sürecini otomatize etme ve soruların tekrarlanmasını engelleme konularında önemli bir katkı sağlamaktadır.

Qomariyah ve diğerleri [16], ikili tercih GA'ya dayalı bir gezi öneri sistemi yaklaşımını gezi planlayıcı öneri sistemi üzerinde uygulayarak açıklamaktadırlar. Bu çalışma, internet kullanıcılarının büyük artışının işletmeler için hizmetlerini tanıtmak için büyük bir fırsat olabileceği bir dönemde, özellikle e-turizm gibi diğer e-hizmetler için kişiselleştirilmiş bir öneri sistemi yaklaşımını önermektedir. E-turizm alanında çiftli tercih toplama yöntemiyle kişiselleştirilmiş bir öneri sistemi geliştirmek amacıyla GA ile çiftli kullanıcı tercihi toplama yaklaşımının bir kombinasyonunu kullanılmıştır. Çiftli tercih toplama yönteminin, puanlama yöntemine göre avantajları, birçok çalışmada gösterilmiştir ve bunlar arasında bir derecelendirme numarasının tutarsızlığını ve karmaşıklığını azaltma bulunmaktadır. Ayrıca, önerilen sistemi incelemek üzere 24 katılımcı üzerinden kullanıcı değerlendirme çalışması gerçekleştirilmiş ve bu çalışmada kullanılan 201 cazibe merkezini içeren POI veri setini yayınlanmıştır. Bu çalışmanın temel katkısı, e-turizm alanında kişiselleştirilmiş öneri sistemleri için GA ile çiftli tercih toplama yöntemini kullanarak etkili bir yaklaşım sunmasıdır.

Tez çalışmasında, tavsiye sistemleri için literatürde yer alan filtreme ve kümeleme yaklaşımları incelenmiştir. Yaygın kullanılan filtreleme yöntemlerinin bazı dezavantajlarına karşın gözetimsiz öğrenme olan kümeleme yönteminde K-ortalamalar algoritması, GA ile geliştirilerek genetik algoritma K-ortalamalar kümelemesi (GA-KOK) çalışılmış ve basit K-ortalamalar kümeleme ile performansı karşılaştırılmıştır. Bu geliştirme sonucunda, GA-KOK ilk uygun çözümde durmaksızın en uygun çözüm için optimize kümeler elde ederek kullanıcıların daha iyi profillenmesi ve bu sayede kümeler için tavsiye çıktılarının iyileştirilmesini amaçlamaktadır.

2. TAVSİYE SİSTEMLERİ

Tavsiye sistemleri, birçok farklı kaynaktan gelen birbiriyle ilişkili ya da ilişkisi ortaya çıkarılacak verileri analiz sürecine dâhil ederek kullanıcıların gideceği noktayı önceden tahmin edip bu bilgilerden hareketle kullanıcının ilgileneceği öğeleri öngörerek çeşitli yaklaşımlar üretir. Bu yaklaşımlarda, bir kullanıcının tercihlerini belirlemek için girdi olarak örneğin kullanıcı tarafından satın alınan ürünler, dinlenen müzikler veya izlenen filmler gibi bilgileri kullanır. Bu bilgilere dayanarak, kullanıcının henüz karşılaşmadığı ürünler, yeni filmler veya daha önce duymadığı müzikler gibi öğeler hakkında tahminler yapar. Kullanıcılara öneriler sunmanın yanı sıra bir dizi öğeyi tek tek veya grup halinde sıralamak için tahminlerde bulunurlar.

2.1. Öneri Algoritması

Bir tavsiye sisteminin geliri artırmasının birçok yolu vardır. En basit yol çapraz satıştır; kullanıcılara ek öğeler önererek, kullanıcının başlangıçta amaçladığından daha fazlasını satın alma olasılığını artırırız. Çapraz satış için kümeleme yöntemi, öncelikle kullanıcıların veya ürünlerin benzer özelliklere sahip olanlarını gruplara ayırmak amacıyla kullanılır. Bu gruplandırma, benzer demografik özelliklere sahip kullanıcıları veya benzer özelliklere sahip ürünleri aynı kümeye yerleştirir. Ayrıca sistemde henüz izi olmayan yeni kullanıcılar için, benzer profillere sahip kullanıcılarla kümelenmesi sağlanarak aynı kümeye ait diğer kullanıcıların ilgilendiği ürünleri keşfetmesi sağlanabilir. Bu sayede hem var olan kullanıcılara hem de yeni kullanıcılara ilgilerine uygun çapraz satış önerileri sunulabilir. Kullanıcının hem platformda geçirdiği vakit artacaktır hem de ilgili bir öğeyi kaçırmamaya çalışarak ayrıca ilgilenmesi olası ürünlerle beraber daha çeşitli öneri listeleri sunulabilir.

Genel bir öneri algoritması, Çizelge 2.1'de gösterildiği üzere formüle edilebilir. Çizelge 2.1'de öneri problemi, bir dizi derecelendirilmemiş öğe (i)'nin kullanıcı (u) tarafından ilgilenme olasılığı en yüksek olan öğelerin ($i*$) bulunmasından oluşur [12].

Çizelge 2.1. Genel Tavsiye Algoritması

Başla

1. u kullanıcısının bütün derecelendirilmemiş i öğeleri için, bir algoritma kullanarak tahmin oranı hesaplanır.

2. $i*$ öğeleri en yüksek tahmin edilen oranlarla önerilir.

Bitir

2.2. Tavsiye Sistemlerinde Kullanılan Yöntemler

Tavsiye sistemlerinde ilgili öneriyi yapmak için veri havuzlarından kayda değer bilgiler çıkarılması noktasında yaklaşımlar da farklılaşmaktadır. Bu önerilerin şekillenme süreci ise platforma göre değişiklikler göstermektedir.

Kullanıcıya tavsiye edilmek üzere bir öneriler çıktı listesi oluşturmak için sırasıyla bilgi çıkarımı (retrieval), filtreleme (filtering), skorlama (scoring) ve sıralama (ordering) olarak gerçekleştirilen dört ana görev vardır. Görevler ve örnek kullanım alanları Çizelge 2.2'de verilmiştir [17].

Çizelge 2.2. Tavsiye Sistemlerinde Bilgi Çıkarımı Aşamaları

Örnek Kullanım	Bilgi Çıkarımı	Filtreleme	Skorlama	Sıralama
Online Kanallar	Satın alınan, görüntülenen öğeler bulma	Stokta olmayan öğelerin çıkarılması	Bir kullanıcının bir ürünü satın alma olasılığının tahminlenmesi	Öğelerin fiyat noktalarına göre yeniden sıralanması
Canlı Servisler	Farklı satırlara/konulara dayalı öğeleri bulma	Kullanıcının ülkesi için mevcut olmayan öğelerin çıkarılması	Öğe başına kullanıcının akış süresinin tahminlenmesi	Tür dağılımlarına uyacak şekilde öneriler düzenlenmesi
Sosyal Medya	Kullanıcının ağında yeni gönderiler bulma	Kullanıcın engellediği ya da sessize aldığı gönderilerin çıkarılması	Kullanıcının etkileşime geçebileceği öğelerin tahminlenmesi	Farklı kaynaklardan gönderilerin sıralanması
Müzik Keşfi	En yakın komşu aramasına dayalı benzer şarkılar bulma	Kullanıcıların daha önce dinlediği parçaların çıkarılması	Bir kullanıcının bir şarkıyı dinleme olasılığının tahminlenmesi	Puan, benzerlik, BPM vb. arasında değiş tokuş yapılması

- **Bilgi çıkarımı (retrieval)**, giriş isteğinde yer alan bir bilgi alt kümesini alır ve bunu, tüm seçenekler kümesini, kullanıcı tarafından seçilme şansı olan çok daha küçük bir seçenekler kümesine hızla daraltmak için kullanır. Çevrimiçi hizmetin veri tabanında belki de milyonlarca öğe olabilir ve tavsiye sisteminin her kullanıcı için hepsini puanlaması mümkün olmayacaktır. Bu hızlı ön seçim, hız gereksinimleri tarafından motive edilir. Kullanıcı için sorunsuz bir hizmet önemli olacağından tavsiye sistemi, çıktı önerilerini, talebi aldıktan sonra birkaç milisaniye içinde sunmalıdır. Ortaya çıkan önceden seçilmiş öğelerin listesi yüzlerce, belki de on binlerce öğe içerebilir bu, tavsiye sisteminin önceden seçilmiş tüm öğeleri zaman içinde ayrı ayrı puanlamasına izin veren bir sayıdır.
- **Filtreleme (filtering)**, bilgi çıkarımı aşamasında oluşturulan önceden seçilmiş öğelerin listesine bir dizi kural uygular, böylece başka bir amaca hizmet etmeyen öğeler çıkarılır (Örneğin, kullanıcı tarafından zaten satın alınan ürünleri listeden çıkarmak).

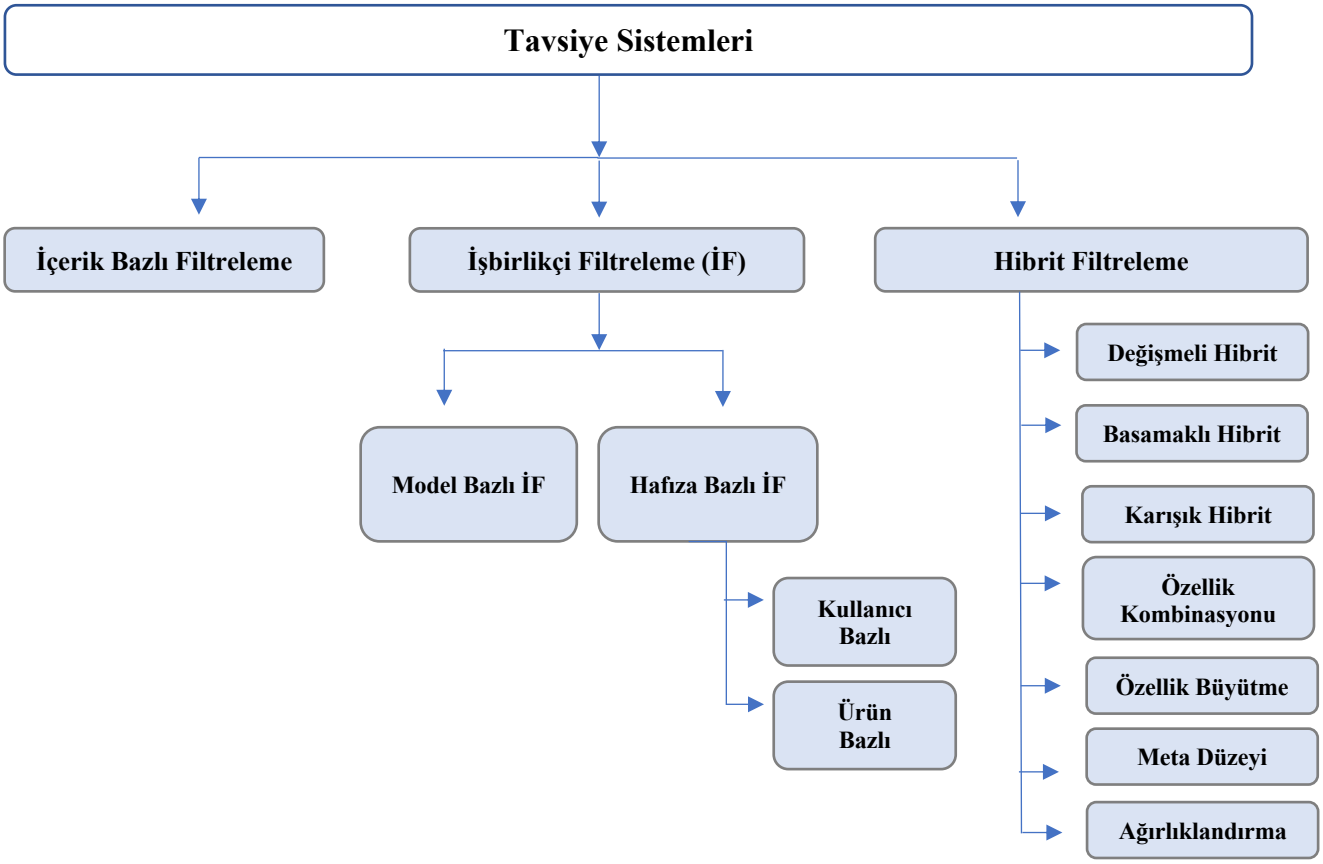
- **Skorlama (scoring)**, filtrelenmiş öğe listesindeki her bir öğeye bir puan (kullanıcıların tavsiye sistemiyle geçmişteki etkileşimi sırasında toplanan) atamak için geçmiş verileri kullanarak eğitilmiş bir makine öğrenimi modeli kullanır. Bu puan, kullanıcının her bir öğeye olan ilgisini ölçmeli, başka bir deyişle, kullanıcının öğeler arasından seçim yaparken kişisel tercihini yansıtmalıdır. Bu aşamada, puanın oluştuğu zamanı dikkate alarak puanın doğruluğunu tercih ederiz, bu nedenle giriş sorgusundan ilgili tüm bilgiler ve kullanıcı hakkında mevcut tüm bilgiler dikkate alınır. Puanlamayı gerçekleştiren makine öğrenim modelinin bir örneği olarak girdi isteğinden gelen öğe verilerini sunduktan sonra, öğe için tıklanma olasılığıyla birlikte tıklama tahmini üreten derin bir sinir ağı olabilir. Tıklama olasılığı daha sonra öğenin çıktı puanı olur.
- **Sıralama (ordering)**, öğenin puanını dikkate alır ve bunu, öğe listesi tavsiye sistemi çıktısına teslim edilmeden önce öğeleri sıralamak için kullanır. Ayrıca bu aşamada, hizmetin iş mantığı gereksinimlerine göre öğelerin çıktı listesi istenen sayıda öğeye kısaltılabilir.

Tavsiye sistemlerinin temel unsurları, bilgi çıkarımı aşamasının merkezindeki makine öğrenimi modelleri ve puanlama aşamasında sıralamayı gerçekleştiren modellerdir. Kümeleme yöntemi ise bu süreçte bilgi çıkarımı aşamasında önemli bir rol oynar. Tavsiye sistemi için kümeleme yönteminde, benzer ilgi alanlarına veya özelliklere sahip kullanıcılar veya ürünler gruplandırılır. Bu gruplar, daha sonra makine öğrenimi modellerinin eğitimi ve önerilerin oluşturulmasında kullanılır. Özellikle bilgi çıkarımı ve puanlama aşamalarında kümeleme sonuçları, kullanıcıların ilgi alanlarına daha hassas bir şekilde odaklanılmasını ve daha kişiselleştirilmiş önerilerin sunulmasını sağlar. Dolayısıyla, tavsiye sistemi başarısını ve kullanıcı memnuniyetini büyük ölçüde kümeleme kalitesi ve verimliliği belirler.

Tavsiye sistemlerinin orijinal sınıflandırması, filtreleme algoritmasına dayanmaktadır. Bu sınıflandırma, tavsiye sistemlerini üç kategoriye ayırır. Mevcut kaynaklar, kullanıcı verileri (demografik özellikler), öğe verileri (anahtar kelimeler, ürün kategorileri) ve kullanıcı-ürün

değerlendirmeleri olabilmektedir. Modern tavsiye sistemlerinde doğrusal modeller ve derin sinir ağları gibi gelişmiş yöntemleri kullanılır.

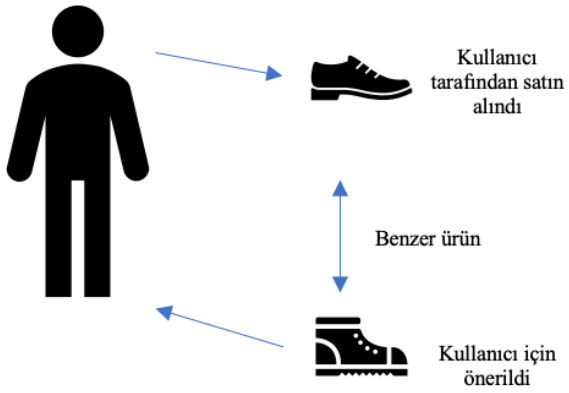
Tavsiye sistemlerinde kullanılan yöntemler Şekil 2.1'de verilmiştir [18].



Şekil 2.1. Tavsiye Sistemleri Yöntemleri

2.2.1. İçerik Bazlı Filtreleme

İçerik Bazlı (Content-Based) filtrelemedeki temel hedef, ürünleri belirli anahtar kelimeler veya kategorilerle gruplandırarak kullanıcının tercihlerini anlamak ve benzer öğeleri önermek için bu terimleri veritabanında araştırmaktır. Şekil 2.2'de de verildiği gibi kullanıcının geçmiş davranışlarından ve öğelere ait içerik bilgisinden yararlanarak önceki öğeye benzer bir öğe önerilmektedir.



Şekil 2.2. İçerik Bazlı Filtreleme

Filtreleme için ilk adımda bir özellik matrisi oluşturulur. Özellik matrisi, kategori, uzunluk, içerik vb. gibi öğelerle ilgili özellikleri içermelidir. Sistem, kullanıcının tercihlerine, geçmişine ve sistemle geçmişteki etkileşimlerine dayalı olarak bir kullanıcı profili oluşturur. Kullanıcı, aynı özellik alanında temsil edilmelidir. Kullanıcıyla ilgili özellikler hem açık hem de örtük olabilir. Açık bir özellik, kullanıcının profilinde belirli bir kategoriyi tercih ettiğini belirtmesi olabilirken, örtük bir özellik, kullanıcının geçmişte yalnızca belirli bir renkteki ürünleri satın alması olabilir. Özellik matrisi hazır olduğunda bir benzerlik metriği seçilmelidir. Benzerlik metriği, örneğin nokta çarpımının (dot product) kullanılması ile benzerlik puanının nasıl ölçülmesi gerektiğini basitçe açıklamaktadır. Sistem, özellik vektörlerini karşılaştırarak her bir öğe ile kullanıcı profilinde ilgi alanlarındaki benzerliği hesaplar. Öğeleri benzerlik puanlarına göre kullanıcıya önerir ve burada

en yüksek puana sahip öğeler en iyi öneriler olarak kabul edilir. İçerik bazlı filtreleme diğer kullanıcıları dikkate almaz, sadece diğer öğeleri dikkate alır. İçerik bazlı filtreleme ile kullanıcılar tarafından bir değerlendirme olmadan da yeni ürünler önerilebilir. Veri tabanında kullanıcının açıkça tercihlerinin belirtilmemesi önerilerin doğruluğu etkilemez. Kullanıcılar kendi profillerini paylaşmadan da öneriler alabilirler.

Bir kullanıcı profili oluşturmak için sistem çoğunlukla iki tür bilgiye odaklanmaktadır [19].

1. Kullanıcının tercih ettiği bir model

2. Kullanıcının tavsiye sistemiyle etkileşiminin geçmişi

Genel olarak, içerik bazlı filtrelemenin matematiksel çalışma prensibi, öğeleri ve kullanıcıları özellik vektörleri olarak temsil etmeyi, her bir özelliğe önemine göre ağırlık atamayı, bir benzerlik işlevi kullanarak öğeler ve kullanıcı ilgi alanları arasındaki benzerliği hesaplamayı ve en yüksek değere dayalı öneriler üretmeyi içermektedir.

2.2.2. İşbirlikçi Filtreleme

İşbirlikçi (Collaborative) filtreleme, tavsiye sistemleri tarafından kullanıcılara geçmiş davranışlarına ve diğer benzer kullanıcıların davranışlarına dayalı olarak kişiselleştirilmiş öneriler sağlamak için kullanılan bir tekniktir ve literatürde Breese ve arkadaşlarının çalışmasında kullanılmıştır [19]. Tavsiye sistemleri arasında popülerdir ve yaygın kullanılan bir yaklaşımdır [20]. Bu filtreleme yönteminde kullanıcı ve ürün benzerliklerini ölçmede ve değerlendirmede farklı algoritmalar kullanılmaktadır. Benzer kullanıcıların benzer öğeleri tercih etmesi gerçeğine dayanmaktadır [21].

İlk adım, kullanıcılar ve tercihleri hakkında veri toplamaktır. Bu veriler, kullanıcı tercihlerini anlamak için kullanılabilecek derecelendirmeleri, incelemeleri veya diğer ilgili bilgileri içerebilir. Toplanan veriler daha sonra, her satırın bir kullanıcıyı temsil ettiği ve her sütunun tavsiye edilebilecek bir öğeyi temsil ettiği bir kullanıcı-öğe matrisi oluşturmak için kullanılır. Matristeki

girdiler, kullanıcılar tarafından öğeler için verilen derecelendirmeleri temsil eder. Benzerlik ölçümleri, benzer tercihlere sahip kullanıcıları belirlemek için kullanılır. Bu ölçümler, Kosinüs Benzerliği'ni, Pearson Korelasyon Katsayısı'nı veya herhangi bir başka uygun benzerlik ölçüsünü içerebilir. Benzer kullanıcılar belirlendikten sonra sistem, benzer kullanıcıların geçmişte beğendiği öğelere dayalı öneriler üretebilir. Sistem, önerileri daha da hassaslaştırmak için ürün bazlı filtreleme veya içerik bazlı filtreleme gibi ek teknikler de kullanabilir. İşbirlikçi filtrelemedeki son adım, önerileri zaman içinde iyileştirmek için kullanıcı geri bildirimlerini kullanmaktır. Bu, kullanıcılardan önerilen öğeler hakkında geri bildirim isteyerek ve bu geri bildirimi önerileri iyileştirmek için kullanarak yapılabilir. İşbirlikçi filtreleme model bazlı ve hafıza bazlı yöntemler olarak Şekil 2.1'de gösterildiği gibi ikiye ayrılmaktadır.

1) Model Bazlı İşbirlikçi Filtreleme

Model bazlı İF yöntemler, kullanıcıların geçmiş alışveriş ya da gezinme izlerini kullanarak ürünlere olan eğilimlerini veya beğeni skorlarını tahmin etmeye çalışan bir model oluşturur ve kullanıcı-ürün etkileşimlerindeki örüntüleri yakalamayı hedefler. Bu model, kullanıcının tercihlerini anlamak ve önerilerde bulunmak için kullanılır.

Model oluşturma süreci, bayes ağı, kümeleme yöntemi ve kural tabanlı yaklaşımlar gibi farklı makine öğrenme algoritmaları ile gerçekleştirilir.

- Bayes ağ modeli, işbirlikçi filtreleme problemi için olasılıksal bir model formüle eder.
- Kurala dayalı yaklaşım, birlikte satın alınan öğeler arasındaki ilişkiyi bulmak için birliktelik kuralı keşif algoritmaları uygular ve ardından öğeler arasındaki ilişkinin gücüne dayalı olarak öğe önerisi üretir [22].
- Kümeleme yöntemi, benzerlik ölçümleri kullanılarak, kullanıcılar veya ürünler benzer gruplara ayrılmaktadır. Bu gruplar, daha sonra model bazlı işbirlikçi filtreleme algoritmaları için temel oluşturur. [23].

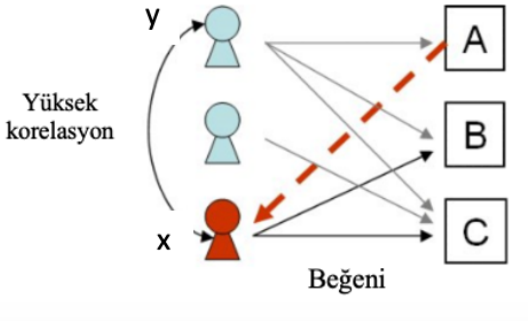
2) Hafıza Bazlı İşbirlikçi Filtreleme

Hafıza bazlı İF'de, ürün ve kullanıcı bazlı komşuluk değerlendirmeleri sonucunda tavsiyeler üretilir. Bu filtrelemede, tavsiyeyi oluşturmak için kullanılan benzerlik değerlerini hesaplamak için geçmiş kullanıcı derecelendirme verilerini temel alırken; model tabanlı yöntemde ise tavsiyeyi oluşturmak için bayesci sınıflandırma, bulanık algoritma, kümeleme yöntemi ve genetik algoritma gibi bir model oluşturmak için geçmiş kullanıcı derecelendirme verilerini kullanır [24]. Genetik algoritmayı kullanan işbirliğine dayalı tavsiye sistemleri temel olarak üç yönde modellemede kullanılmıştır: kümeleme [25-27], hibrit kullanıcı modelleri [28-30] ve genetik algoritma tabanlı modellemeler [31,32]. Hafıza bazlı İF, kullanıcı bazlı ve ürün bazlı olmak üzere ikiye ayrılmaktadır.

Kullanıcı bazlı işbirlikçi filtreleme: Tercihlerin belirli bir kullanıcıyla önemli ölçüde ilişkili olduğu tavsiye sistemidir. Kullanıcıların benzer tercihlerinden yola çıkar. Veri tabanındaki benzer tercihleri yapmış diğer kullanıcıları bulup satın alma ya da beğenme potansiyeli olan diğer öğeleri kullanıcıya öneren sistemdir.

Kullanıcı bazlı İF, Şekil 2.3'te de gösterildiği gibi kullanıcılar arasındaki istatistiksel benzerlik kavramlarına dayalı öneriler sunar [33]. Benzer müşteriler yüksek korelasyona göre ilişkilendirilmeye çalışılır ve müşterilerin seçtiği ürünlere dayanarak öneri ürünler sunulmaktadır.

Örneğin, benzer alışveriş davranışı gösteren iki farklı kullanıcı x ve y olarak tanımlansın. A,B ve C ise farklı elbise çeşitleri varsayıldığında, x kişisi yeni bir elbise bakıyorken geçmiş elbise seçimleri veya beğenileri doğrultusunda benzer profildeki y kişisinin seçimleri ile benzerliğini ölçen sistem, y kişisinin alıp x kişisinin henüz almadığı elbiseyi (A) önerecektir. Bu öneri gerçekleşirken x kullanıcısının y kullanıcısına benzer skorlarına göre, pearson korelasyonu veya kosinüs benzerliği yöntemleri kullanılarak tahminleme yapılmaktadır.



Şekil 2.3. Kullanıcı Bazlı İşbirlikçi Filtreleme

Pearson korelasyon katsayısı (r) ile x ve y kullanıcısı arasındaki benzerlik hesaplaması Eşitlik (2.1)'de gösterilmiştir. Pearson katsayısı değerleri -1 ile 1 arasındadır. Değerin 1'e yaklaşması x ve y kullanıcıları arasındaki benzerliğin arttığını gösterir [34].

$$r(x,y) = \frac{\sum_{i\in I}(R_{x,i}-\bar{R}_x).(R_{y,i}-\bar{R}_y)}{\sqrt{\sum_{i\in I}(R_{x,i}-\bar{R}_x)^2}.\sqrt{\sum_{i\in I}(R_{y,i}-\bar{R}_y)^2}} \quad (2.1)$$

I : Aktif kullanıcı ile diğer kullanıcı arasında ortak olarak değerlendirilen öğelerin kümesi

Rx,i: Aktif kullanıcının i öğesine ait değerlendirmesi

Ry,i: Diğer kullanıcının i öğesine ait değerlendirmesi

$\overline{R_y}$: y kullanıcısının öğelere verdiği puanların ortalaması

$\overline{R_x}$: x kullanıcısının öğelere verdiği puanların ortalaması

Diğer adım ise aktif kullanıcı ile örüntü yakalanan başka kullanıcılar için komşuluk seçimidir. K-en yakın komşu algoritması (KNN) bu adımda sık kullanılmaktadır. Algoritma sonucunda aktif kullanıcı ile en yüksek benzerliği gösteren k tane komşu seçimi yapılır.

Son adım olarak da aktif kullanıcıya daha önce değinmediği öğeler için, k tane komşu üzerinden davranış skorları hesaplanır. Bu tahminde k tane komşu kullanıcının ilgili ürünle olan ilişkileri sonucundaki skorların ağırlıklı ortalaması denklem Eşitlik (2.2)'deki gibi alınır.

$$R_{x,i} = \overline{R_x} + \frac{\sum_{y \in U}(R_{y,i} - y) \,.\, r(x,y)}{\sum_{y \in U} r(x,y)} \tag{2.2}$$

x: Aktif kullanıcı

y: Diğer kullanıcılar

U: Aktif kullanıcının komşularının bulunduğu küme

$r(x,y)$*:* x kullanıcısı ile y kullanıcısı arasındaki pearson katsayısı

Rx,i: Aktif kullanıcının i öğesine ait değerlendirmesi

Ry,i: Diğer kullanıcının i öğesine ait değerlendirmesi

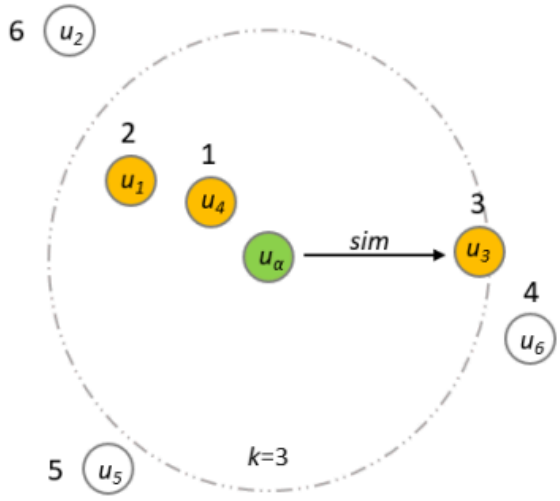
$\overline{R_y}$: y kullanıcısının öğelere verdiği puanların ortalaması

$\overline{R_x}$: x kullanıcısının öğelere verdiği puanların ortalaması

Tahmin hesaplamasında ise derecelendirme matrisi (rating matrix) kullanılır. Şekil 2.4'te gösterildiği üzere satırlar kullanıcıları, sütunlar ise öğeleri temsil eder. Matris girişleri (a), kullanıcıların öğelere sağladığı "Bilinen" derecelendirmeleri içerir. "Bilinmeyen" derecelendirmeler soru işaretleriyle temsil edilir. Ayrıca Şekil 2.4'te komşuluk seçimi de sembolize edilmiştir (b). u_α aktif kullanıcıyı ve r_α aktif kullanıcının henüz keşfetmediği öğeler için beğeni skoru tahminini göstermektedir [35].

	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8
u_1	?	4.0	4.0	2.0	1.0	2.0	?	?
u_2	3.0	?	?	?	5.0	1.0	?	?
u_3	3.0	?	?	3.0	2.0	2.0	?	3.0
u_4	4.0	?	?	2.0	1.0	1.0	2.0	4.0
u_5	1.0	1.0	?	?	?	?	?	1.0
u_6	?	1.0	?	?	1.0	1.0	?	1.0
u_α	?	?	4.0	3.0	?	1.0	?	5.0
r_α	3.5	4.0			1.3		2.0	

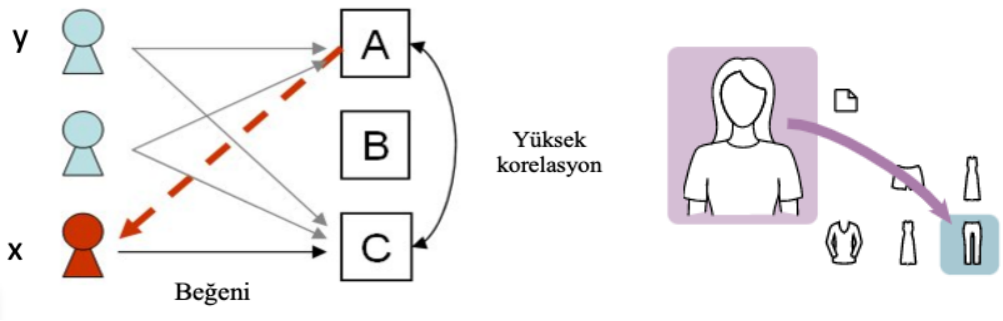
(a)



(b)

Şekil 2.4. Kullanıcı – Ürün Skor Matris ve Tahmin Hesaplaması (a), Komşuluk Seçimi (b)

Ürün bazlı işbirlikçi filtreleme: Kişilerin satın aldığı, görüntülediği, puanladığı ürünlerin benzerliklerini ve diğer kullanıcıların tercihlerini dikkate alan bir filtreleme yapar. Kişinin daha önceden baktığı ya da beğendiği ürünlere göre skorlama yaparak buna benzer ürünleri kişiye sunar. İçerik bazlı filtreleme tekniği ile ürün bazlı İF tekniği arasındaki fark; içerik bazlı filtreleme tekniği öğelerin meta verilerinden yararlanarak öğeler önerirken, ürün tabanlı İF tekniğinde ilgili kullanıcının satın aldığı ürünler, incelediği içerikler ile birlikte diğer kullanıcıların tercihlerini dikkate alarak öneriler sunulmaktadır.



Şekil 2.5. Ürün Bazlı İşbirlikçi Filtreleme

Şekil 2.5'te de gösterildiği gibi ürün bazlı İF'de, ürün benzerliğine odaklı olarak algoritma çalışır [33]. Müşteriler arasında benzerlik bulmak yerine ürünlerin özelliklerine göre benzer özellikteki öğeler gösterilir. Her öğe için bir profil oluşturulur (A,B ve C) ve kullanıcıların (x,y) tercih ettiği ürünlerin profili, kullanıcıların tercihlerini temsil eder. Ürünler arasındaki benzerlikler hesaplanarak kullanıcı profilindeki ürünlerin benzerleri belirlenir ve en benzer ürünler öneri olarak sunulur.

Örneğin bir kullanıcı pantolon satın almak istiyorsa aranılan özelliklerdeki benzer pantolonlar tavsiye edilecektir. Bu pantolonun özelliklerine dayanarak tavsiyeler sunması, kullanıcı için kişiselleştirilmiş bir deneyim sağlamaktadır. Ürün bazlı işbirlikçi filtrelemede kullanıcılar arasındaki farklar dikkate alınmaz bu da kullanıcıların benzersiz tercihlerinin göz ardı edilmesine sebep olur.

Aktif kullanıcının değerlendirdiği öğeleri, kullanıcı – ürün matrisinden alarak bir ürün benzerlikleri modeli ortaya koyar. Bunun için ilk adımda aktif kullanıcının öğelere verdiği puanlar veya yorumlar gibi veriler toplanır. Sonrasında öğelerin benzerlikleri hesaplanır. İki öğenin özellikleri (fiyat, marka, tasarımı, renk, beden vb.) benzerlik skoru hesaplamak için kullanılabilir. Benzerlik skoru, Pearson Korelasyonu Eşitlik (2.1), Kosinüs Benzerliği Eşitlik (2.3) veya Öklit Uzaklığı Eşitlik (2.4) gibi istatistiksel yöntemlerle hesaplanabilir.

Eşitlik (2.3)'te Kosinüs benzerliği, iki öğeyi ve derecelendirmelerini vektörler (a,b) olarak tanımlar. $a_i \; ve \; b_i$ her bir i'nci değişken değerini temsil etmektedir ve aralarındaki benzerlik bu vektörler arasındaki açı olarak Eşitlik (2.3)'teki gibi tanımlanmaktadır.

$$kosinüs(a,b) = \frac{a.b}{||a||||b||} = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2}\sqrt{\sum_{i=1}^{n} b_i^2}} \tag{2.3}$$

Eşitlik (2.4)'te tanımlandığı gibi öklid uzaklığı ise iki nokta arasındaki doğrusal uzaklığı temel alır. Bu iki nokta arasındaki mesafenin iki boyutlu uzayda X ve Y koordinatları vardır. Örneğin bir M noktasının koordinatları (X_M, Y_M) ve bir L noktasının koordinatları (X_L, Y_L) olarak temsil edildiğinde M ile L noktası arasındaki öklid mesafesi Eşitlik (2.4)'teki şekli ile ifade edilmektedir.

$$\text{d(M,L)} = \sqrt{(X_L - X_M)^2 + (Y_L - Y_M)^2} \tag{2.4}$$

2.2.3. Hibrit Filtreleme

Hibrit filtreleme, önerilerin doğruluğunu artırmak veya geleneksel tavsiye sistemlerinin sınırlamalarıyla (örneğin, yeni ürünler veya yeni kullanıcılar gibi sorunlar) kısmen başa çıkabilmek için bilgi temelli yöntemlerin birleştirildiği bir tekniktir. Herhangi bir bireysel sistemin daha az dezavantajıyla daha iyi performans elde etmek için iki veya daha fazla öneri tekniği birleştirilmektedir. Örneğin, içerik tabanlı filtreleme ve işbirliği tabanlı filtreleme yöntemleri hibrit bir yaklaşımla birleştirilebilir. İçerik tabanlı filtreleme, kullanıcının geçmiş tercihlerine veya profiline dayalı olarak benzer öğeleri önerirken, işbirliği tabanlı filtreleme, kullanıcıların benzer tercihlere sahip diğer kullanıcıların önerilerini dikkate alır. Hibrit filtreleme, bu iki yaklaşımı birleştirerek hem içerik benzerliğini hem de kullanıcı davranışlarını göz önünde bulundurarak daha hassas öneriler sağlamayı hedefler. Çizelge 2.3'te hibrit filtrelemede kullanılmış olan bazı kombinasyon teknikleri ve kısaca tanımları verilmiştir [36].

Çizelge 2.3. Hibrit Teknikler

Hibrit Metot	**Tanım**
Değişmeli	Sistem, mevcut duruma bağlı olarak öneri teknikleri arasında geçiş yaptığı metottur.
Basamaklı	Bu yöntemde, adayların kaba bir sıralamasını oluşturmak için bir öneri tekniği kullanılırken, ikinci teknik yalnızca ek düzeltme gerektiren maddelere odaklanır.
Karışık	Birkaç tavsiye ediciden gelen önerilerin aynı anda sunulduğu yöntemdir.
Özellik Kombinasyonu	Farklı veri kaynaklarından gelen öneri özelliklerinin birlikte tek bir öneri algoritmasına atılması tekniğidir.
Özellik Büyütme	Bir öğenin derecelendirmesini veya sınıflandırmasını oluşturmak için bir teknik kullanılır ve bu bilgi daha sonra bir sonraki öneri tekniğinin işlenmesine dahil edilmesidir.
Meta Düzeyi	Bir tavsiyeci tarafından öğrenilen model, diğerine girdi olarak kullanılır.
Ağırlıklandırma	Birkaç öneri tekniğinin skorlarının ya da oylamalarının tek bir öneri oluşturmak için bir araya getirilmesidir.

Farklı çalışmalar, hibrit tekniklerin geleneksel tekniklere kıyasla daha doğru tavsiyeler sağladıklarını göstermek amacıyla işbirlikçi ve içerik bazlı filtreleme ile karşılaştırmıştır. Yapılan bu karşılaştırmalar hibrit filtrelemenin daha etkili sonuçlar verdiğini göstermektedir [36-37].

2.3. Tavsiye Sistemlerinde Karşılaşılan Sorunlar

Tavsiye sistemleri kullanılarak doğru ve iyi kalitede tavsiye oluşturmak için bazı zorluklarla karşılaşılabilmektedir. Model oluşturma sürecinde genellikle seyreklik (sparsity), ölçeklenebilirlik (scalability), soğuk başlatma (cold start), kararsız kullanıcılar (gray sheep), benzerlik (synonymy), yanlış yönlendirme (shilling attacks), gecikme (latency), popülerlik yanlılığı (popularity bias) gibi problemler ile karşılaşılmaktadır. Temel sorunlar aşağıda açıklanmıştır [38].

- Seyreklik sorunu: Öneri kalitesini etkileyen ve tahmin yapmak için gereken hesaplama süresini artıran, tavsiye sisteminin karşılaştığı önemli bir sorundur. Bu sorun, işbirlikçi filtrelemede görülür, çünkü çoğu kullanıcı öğelerin çoğunu derecelendirmez ve bu derecelendirmeler seyrektir. Başka bir deyişle, kullanıcı-öğe derecelendirme matrisi birkaç sıfır olmayan giriş içerir.
- Soğuk Başlangıç sorunu: Yeni bir kullanıcıya tavsiyede bulunmak, yeni ürünler tavsiye etmek veya yeni bir tavsiye sistemi tarafından öneri oluşturmak zordur. Öğe veya kullanıcı hakkında bilgi eksikliği bu soruna yol açmaktadır. İçerik bazlı filtreleme kullanılan tavsiye sistemi, soğuk başlatma problemini önler çünkü geçmiş derecelendirme bilgilerine bağlı değildir. Oysa işbirlikçi filtreleme kullanılan tavsiye sistemi bu durumda kaliteli bir öneri üretemez.
- Ölçeklenebilirlik: Bilgi miktarı muazzam bir şekilde artmaya devam etmektedir. Örneğin, çok fazla kullanıcının olduğu ve her gün daha da artmaya devam eden platformlarda öğe çeşitliliği de artacak ve tercihler daha da karmaşık hale gelecektir. Bu karmaşıklık beraberinde işlem yoğunluğuna sebep olarak ölçeklenebilirlik sorunu ortaya çıkmaktadır. Ölçeklenebilirlik sorununun çözümünde genellikle boyut azaltma ve kümeleme yöntemleri uygulanmaktadır.
- Aşırı Uzmanlaşma sorunu: Tavsiyenin kullanıcı profiline göre üretildiği durumlarda "Aşırı Uzmanlaşma" sorunu ortaya çıkmaktadır. Bu, kullanıcının yeni öğeler keşfetmesini engeller.
- Kararsız kullanıcılar: İşbirlikçi filtrelemede yaşanan "Gray Sheep" sorunu, ilgili kullanıcıya benzer hiçbir kullanıcının bulunmaması/eşlenmemesi sonucunda ortaya çıkmaktadır. Sorunun çözümünde genellikle, kullanıcı profili ve öğe özelliklerinin kullanımı ile içerik bazlı filtreleme uygulanmaktadır.
- Benzerlik: Benzer öğelerin sistem üzerinde farklı tanımlanmaları sonucunda ortaya çıkan benzerlik problemi, tavsiye sistemi uygulamalarının karar verme sürecinde doğruluğunu etkileyebilir.

- Yanlış Yönlendirme: Bu sorun, sahte profil hesaplarına sahip kişilerin tavsiye sistemi uygulamalarına yaptığı saldırıların sonucunda ortaya çıkmaktadır. Burada, tavsiye sistemlerinde yapılan saldırılarda bazı ürünlerin popülaritesi artabilir veya azalabilir. Bu durum sonucunda, kişisel ürün önermede sistemlerin güvenirliliği azalmaktadır.
- Gecikme: İşbirlikçi filtreleme ile geliştirilen tavsiye sistemleri uygulamalarında, veri tabanına yeni ürünlerin sık sık eklenmesi sonucunda gecikme sorunu ortaya çıkmaktadır. Burada eklenen yeni öğelerin derecelendirme bilgisine sahip olmaması sebebiyle ilgili ürünler öneri listelerinde yer alamamaktadır. Yaşanan sorunu çözmek ve uygulamaların performansını artırmak için kümeleme, kullanıcıların veya öğelerin önceden hesaplanmış benzerlik değerleri, hızlı bir öneri veri tabanı oluşturarak gecikmeyi azaltabilir.
- Popüler yanlılığı: İşbirlikçi filtrelemede yaşanan problemlerden biridir, popüler öğeler popüler olmayan öğelere göre daha çok öne çıkarılır ve ilgili popüler öğeler aşırı tavsiye edilir. Burada ilgili kullanıcıların ilgilenebilecekleri ancak daha az popüler olması nedeniyle daha az tavsiye edilen öğeler doğru öneri listelerin oluşturulmasında dezavantaj oluşturabilir. Tavsiye uygulamalarında önerilen öğelerin görünürlüğünü artırmak amacıyla sahte hesaplar tarafından, belirli öğelerin derecelendirmeleri artırabilir, sahte incelemeler eklenebilir. Bu durum tavsiye sistemi uygulamalarının doğruluğunu olumsuz etkileyecektir.

3. KÜMELEME YÖNTEMLERİ VE TAVSİYE SİSTEMİNDE KULLANIMI

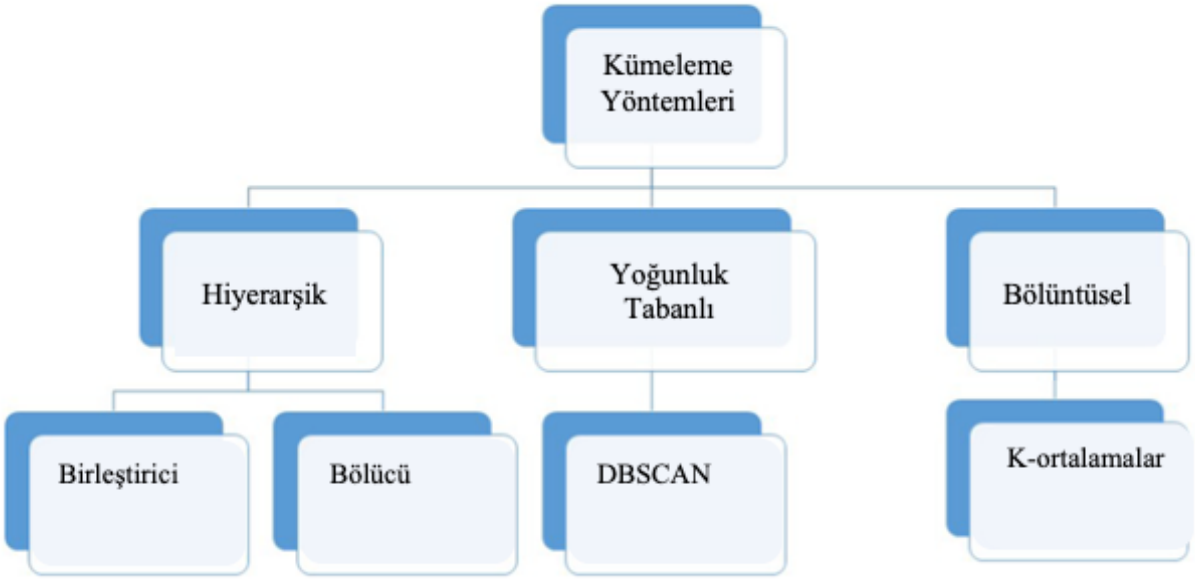
Veri madenciliği ve istatistiksel analiz gibi alanlarda kullanılan kümeleme yöntemleri benzer niteliklere sahip veri noktalarını gruplandırmak için kullanılmaktadır. Kümeleme, etiket veya önceden alınmış bilgi içermeyen verileri gruplama yeteneğine sahip bir dizi teknik sunan bir yöntemdir. Bir veri kümesi, bir dizi nesneden oluşmaktadır. Her bir nesne, bir dizi özellikten oluşur. Kümeleme yöntemleri, her bir grubun özellikleri diğer grup elemanlarına göre daha fazla benzerlik gösteren bir şekilde nesnelerin belirli bir sayıda kümesini bulmayı amaçlayan algoritmalardan oluşur.

Kümeleme yöntemleri çeşitli alanlardaki problemlere basit ve hızlı çözümler sunarak maliyet ve zaman anlamında önemli tasarruflar sağlayabilmektedir. Veri analizinin birçok yönünde faydalıdır ve uygulama alanları oldukça geniştir. Sağlık sektöründe, hastaları belirli semptomlara veya teşhislere göre kümelere ayırarak tedavi stratejilerini optimize etmek mümkündür. Sosyal ağ analizinde, kullanıcıları ilgi alanlarına veya ilişkilere göre gruplandırmak, sosyal ağların yapısını anlamak ve kullanıcı davranışlarını tahmin etmek için önemlidir. Veri madenciliği ve makine öğreniminde, veri noktalarını benzer özelliklere sahip gruplara ayırarak daha kesin sınıflandırmalar veya tahminler elde edilebilir. Ayrıca, kümeleme yöntemleri, tavsiye sistemlerinin bilgi çıkarım aşamasında ve tavsiye sistemlerindeki dezavantajlı özel durumları çözmede kullanılabilen bir yaklaşımdır. Bu sayede, kullanıcıların tercihlerini analiz ederek benzer içerikleri gruplandırabilir ve daha özelleştirilmiş tavsiyeler sunabilir.

3.1. Kümeleme Yöntemleri

Kümeleme yöntemlerinde, nesneleri içsel benzerlikleri doğrultusunda gruplandırmak ve aynı zamanda birbirleriyle farklılık gösteren gruplar şeklinde düzenlemek için uzaklık matrisleri kullanılmaktadır. Bu yöntemler için pek çok farklı algoritma geliştirilmiştir. Kümeleme yöntemleri bölüntüsel, hiyerarşik ve yoğunluk tabanlı kümeleme şeklinde temel kategoriler olup, her bir kategori altında farklı teknikler bulunmaktadır. Bu tekniklerdeki ortak hedef, küme içinde yüksek

benzerlik ve kümeler arasında ise düşük benzerlik olarak nesnelerin gruplandırılmasıdır. Kümeleme yöntemlerinin temel yapısı Şekil 3.1'de sunulmuştur [39].



Şekil 3.1. Kümeleme Yöntemleri

3.1.1. Hiyerarşik Kümeleme

Hiyerarşik kümeleme, veri noktalarını başlangıçta tek bir küme olarak ele alır ve ardından benzer veri noktalarını birleştirerek hiyerarşik bir ağaç yapısı oluşturur. Bu ağaç, veri noktalarının nasıl gruplandığını ve alt seviyedeki kümelemelerin nasıl birleştirildiğini görsel olarak temsil eder. Aynı zamanda, bölücü ve birleştirici algoritmalar, bir başlangıçta tüm veri noktalarının bir küme içinde olduğu hiyerarşik bir yapının yukarıdan aşağıya veya aşağıdan yukarıya doğru bölünerek veya birleştirilerek oluşturulduğu özel hiyerarşik kümeleme yaklaşımlarını ifade eder. Bu tür algoritmaların, verinin karmaşıklığını anlamada, alt kümelerin yapısını analiz etmede ve grupların ilişkilerini görselleştirmede kullanımı yaygındır [40].

3.1.2. Yoğunluk Tabanlı Kümeleme

Yoğunluk tabanlı kümeleme, veri noktalarını yoğunluk ve komşuluk ilişkileri temelinde gruplara bölen bir kümeleme yöntemidir. Bu yaklaşım, bir veri noktasının etrafındaki yoğunluğa dayalı komşularını ve minimum komşu sayısını kullanarak veriyi kümelere ayırır. Bu sayede farklı şekillerde ve boyutlarda kümeleri tanımlayabilir, aykırı değerlere direnç gösterebilir ve daha karmaşık veri yapılarını işleyebilir. DBSCAN, her veri noktasının etrafında bölgeler tanımlayarak ve veri uzayındaki yoğun bölgeleri belirleyerek çalışır. Bu algoritma, önceden küme sayısını belirlemeyi gerektirmez ve özellikle verinin temel yapısının bilinmediği senaryolarda kullanışlıdır. Yoğunluk tabanlı kümeleme, heterojen ve gürültülü veri kümelerinin belirlenmesi, küme sayısı hakkında önceden bilgiye sahip olunmadığı durumlar ve veri gruplarının yapısını anlama gibi birçok uygulama alanında etkili bir araç olarak kullanılmaktadır [41].

3.1.3. Bölüntüsel Kümeleme

Bölüntüsel kümeleme yaklaşımında küme sayısı genellikle önceden tanımlanmıştır. Ayrıca, bir küme içerisindeki nesnelerin kümedeki uzaklığını minimize ederek küme oluşturma ve optimize etme amacı taşıyan bir amaç fonksiyonu bulunur. Bu nedenle, küme sayısı önce kullanıcı tarafından sezgisel bir şekilde belirlenir ve değerlendirilir ardından optimizasyon için yinelemeli bir prosedür uygulanır. Bölüntüsel kümelemede yaygın kullanılan algoritma ise K-ortalamalar algoritmasıdır.

- **K-ortalamalar Algoritması**

K-ortalamalar algoritmasının temel mantığı, veri kümesinin içerdiği n veri öğesini, giriş olarak belirtilen K sayısına sahip kümeler halinde bölünmesi yaklaşımına dayanmaktadır. K-ortalamalar algoritması bir veri kümesi üzerinde uygulandığında ise, araştırmacı tarafından belirlenmiş olan 'K' sayısına ($1 < K < n$) kadar kümeler oluşturulur. Gerçekleştirilen bölüntü işleminin temel hedefi, oluşturulan küme içindeki öğelerin birbirine en yakın ve benzer olduğu, ancak farklı kümeler arasındaki öğelerin mümkün olduğunca farklı ve benzersiz olduğu bir yapı oluşturmaktır.

K-ortalamalar algoritmasının uygun başlangıç merkez noktalarını seçmek için herhangi bir mekanizması yoktur. Bununla birlikte, farklı başlangıç merkez noktalarının seçilmesi, özellikle örneklem birçok aykırı değer içerdiğinde, kümeleme sonuçlarında büyük farklılıklar oluşturabilir. K-ortalamalar algoritması, başlangıçta rastgele seçilen merkez noktalarıyla başlar ve bu merkez noktalarını kullanarak verileri kümelere ayırır. Rastgele seçilen başlangıç merkez noktaları, algoritmanın farklı yerel optimumlara doğru ilerlemesine neden olabilir. Yani, farklı başlangıç noktaları kullanıldığında, algoritma farklı kümeleme sonuçlarına ulaşabilir. Bu nedenle, geleneksel K-ortalamalar algoritmasında uygun başlangıç merkez noktalarının seçilmesi çok önemlidir. K-ortalamalar kümeleme süreci aşağıdaki gibidir:

Adım 1) Seçilen küme sayısına (K) sahip ilk merkez noktalar rastgele seçilir ve bir başlangıç kümelenme oluşur. $\{C_1,\ C_1, C_2, C_3, \dots, C_K\}$ şeklinde K kümelere ayrılmıştır. Burada X_{K} değeri, C_K kümesine ait i'nci örneği ve n ise veri kümesinin içerdiği veri öğesi sayısını sembolize eder. C_K kümesinin vektör ortalaması (M_K), Eşitlik (3.1)'deki şekilde ifade edilir [42].

$$M_K = \frac{1}{n_K} \sum_{i=1}^{n_K} X_{iK} \tag{3.1}$$

Adım 2) Küme içi değişimler ise "Karesel Hata Formülü" kullanılarak hesaplanmaktadır. C_K için karesel hata formülü Eşitlik (3.2)'de verilmiştir. Tüm kümeler için karesel hata, Eşitlik (3.3)'de gösterildiği gibi küme içindeki değişimlerin toplamıdır. Karesel hata yönteminin amacı, verilen K değerine bağlı olarak E_K^2 değerini minimize eden K tane kümeyi bulmaktır.

$$e_i^2 = \sum_{i=1}^{n_K} (X_{iK} - M_K)^2 \tag{3.2}$$

$$E_K^2 = \sum_{i=1}^{K} {e_i}^2 \tag{3.3}$$

Adım 3) Her kayıt en yakın merkeze atanarak küme oluşturulur.

Adım 4) Küme sayısı aynı tutularak her kümenin yeni merkezi hesaplanır.

Adım 5) Kümeler değişmeyi bırakana veya durma koşulları sağlanana kadar Adım 2 ve Adım 3 yinelenir.

K-ortalamalar algoritması, uygulama kolaylığı ve basitliği nedeniyle popüler olmuştur. Bununla birlikte, bazı eksiklikleri de vardır [42]. İlk olarak, örtüşen kümelerde iyi sonuç vermeyebilir ve kümeler aykırı değerler tarafından merkezden çekilebilir. Kümeleme sonucu ilk merkez noktalara bağlı olabilir, ancak ilk merkez noktaları optimize edecek bir mekanizma yoktur. GA, arama işlemini tek bir başlangıç noktası yerine bir nokta kümesi üzerinden gerçekleştirir ve K-ortalamalar algoritması üzerinde uygulanarak optimum veya alt-optimal başlangıç merkez noktalarını seçmek için kullanılabilir. GA, bu optimizasyon problemini çözmek için evrimsel prensipleri kullanır. İlk olarak, başlangıçta rastgele oluşturulan bir dizi küme merkezi noktasıyla başlar. Her bir nokta, bir kümeyi temsil eder. Ardından, GA bu merkezi noktaların konumlarını iteratif olarak geliştirir ve en iyi çözümü bulmak için evrimsel süreçler uygular.

3.2. Tavsiye Sistemlerinde Kümeleme Yöntemi

Tavsiye sistemlerinde kümeleme yöntemi, kullanıcıları veya öğeleri benzer özelliklere sahip gruplara ayırmak amacıyla kullanılmaktadır. Benzer özelliklere sahip kullanıcılar veya öğeler aynı grupta toplanarak daha etkili tavsiyelerin oluşturulması hedeflenir. Kümeleme yöntemleri, büyük veri kümeleri ve karmaşık ilişkilerle karşılaşılan durumlarda tavsiye sistemlerinin performansını artırabilir ve verimliliğini sağlayabilir. Dolayısıyla, kullanıcıların veya öğelerin benzer davranışlarını ve tercihlerini daha iyi anlamak ve daha iyi öneriler sunmak mümkün olabilir.

Aynı zamanda tavsiye sistemlerinde kümeleme yönteminin kullanımıyla belirli dezavantajlar aşılabilmektedir. Özellikle büyük ve karmaşık veri setleri içerisinde benzer davranış ve tercihlere sahip kullanıcı veya öğeleri gruplandırmak, bu veri setlerindeki düzensizliği ve karmaşıklığı ele

almayı sağlar. Kümeleme yöntemiyle, veri setindeki kalabalıklığı ve düzensizliği azaltarak daha anlamlı gruplar oluşturabilir, bu sayede tavsiyeler daha tutarlı ve etkili hale gelebilir. Özellikle yeni kullanıcılar veya yeni öğeler için ortaya çıkan soğuk başlangıç sorununu aşmada yardımcı olabilir. Yeni gelen verileri mevcut gruplara ekleyerek veya benzer davranışlara sahip olanlarla ilişkilendirerek daha iyi tavsiyeler sunulabilir. Bu şekilde kümeleme yöntemi, tavsiye sistemlerinin performansını artırarak kullanıcı deneyimini geliştirme potansiyeline sahiptir. Tavsiye sistemlerinde kümeleme yöntemi genellikle kullanıcı tabanlı ve içerik tabanlı olmak üzere iki teknik altında sınıflandırılabilir.

Kullanıcı tabanlı kümeleme tekniğinde, benzer tercihlere sahip kullanıcılar gruplandırılır. Bir kullanıcının davranışları ve tercihleri, diğer benzer kullanıcılarınkilerle karşılaştırılır. Bu sayede, bir kullanıcının tercihleri temel alınarak benzer kullanıcıların tavsiyeleri verilebilir. Kullanıcı Tabanlı Kümeleme tekniğinin işleyişi şu adımları içermektedir:

Adım 1) Veri Toplama: Bu adımda kullanıcıların demografik bilgileri, önceki tercihleri veya davranışlarına dair veriler toplanır. Bu veriler, tavsiye sisteminin kullanıcılar arasındaki benzerlikleri anlamasına yardımcı olur.

Adım 2) Benzerlik Hesaplaması: Kullanıcı verileri arasındaki benzerliği hesaplamak için çeşitli yaklaşımlar kullanılır. Örneğin, kosinüs benzerliği veya Pearson korelasyonu gibi metrikler kullanılabilir. Bu hesaplamalar, kullanıcıların benzerliklerini sayısal bir değerle ifade ederek benzerlik derecelerini belirlemeye yardımcı olur.

Adım 3) Kümeleme: Benzerlik matrisleri veya uzaklık metrikleri kullanılarak kullanıcılar gruplara ayrılır. Bu gruplar, benzer tercih ve özelliklere sahip kullanıcıları içerir. Kullanıcılar arasındaki benzerlik derecesine göre gruplar oluşturulur.

Adım 4) Tavsiye Oluşturma: Kullanıcının tercih ettiği bir öğe veya içeriğe benzer tercihleri olan diğer kullanıcıların tercih ettiği öğeler analiz edilir. Benzer kullanıcıların tercih ettiği içerikler, ilgili kullanıcıya tavsiye olarak sunulabilir.

İçerik tabanlı kümeleme tekniği ise benzer özelliklere sahip ürünlerin veya içeriklerin gruplandırılmasıdır. Örneğin, bir e-ticaret tavsiye sisteminde, benzer türdeki ürünler aynı kümede toplanabilir. Kullanıcının daha önce tercih ettiği içeriklere dayalı olarak yine benzer olan fakat henüz satın almadığı ürünler önerilebilir. İçerik tabanlı kümeleme algoritmasının işleyişindeki adımlar ise aşağıda verildiği gibidir:

Adım 1) İçerik Özelliklerinin Belirlenmesi: Tavsiye edilecek öğelerin içerik özellikleri belirlenir. Bu özellikler, öğelerin niteliklerini temsil eden ve analiz edilen verilerdir. Örneğin, bir e-ticaret tavsiye sistemi için ürün kategorisi, markası, ürün bedeni gibi özellikler belirlenebilir.

Adım 2) Özellik Vektörleri Oluşturma: Her bir öğe için belirlenen içerik özellikleri, bir vektör olarak temsil edilir. Bu vektörler, öğelerin özelliklerini sayısal değerlerle ifade eder. Özellikler arasındaki benzerlikler bu sayısal vektörler üzerinden hesaplanır.

Adım 3) Benzerlik Hesaplaması: Öğeler arasındaki benzerliği hesaplamak için çeşitli yaklaşımlar kullanılır. Örneğin, kosinüs benzerliği veya Pearson korelasyonu gibi metrikler kullanılabilir. Bu hesaplamalar, öğelerin içerik özelliklerine dayalı olarak benzerlik derecelerini belirlemeye yardımcı olur.

Adım 4) Kümelenme: Benzerlik matrisleri veya uzaklık metrikleri kullanılarak öğeler gruplara ayrılır. Bu gruplar, benzer içerik özelliklerine sahip öğeleri içerir. Öğeler arasındaki benzerlik derecesine göre gruplar oluşturulur.

Adım 5) Tavsiye Oluşturma: Kullanıcının tercih ettiği bir öğe veya içeriğe benzer içerik özelliklerine sahip diğer öğeler analiz edilir. Benzer içerik özelliklerine sahip öğeler, ilgili kullanıcıya tavsiye olarak sunulabilir.

4.GENETİK ALGORİTMA

Genetik Algoritma (GA), ilk olarak John Holland tarafından 1975 yılında evrimsel süreçlere dayanan güçlü ve etkili araştırma algoritmaları olarak ortaya konulmuş ve evrim yasaları içinde en iyileme problemleri için kullanılmıştır. Holland, çalışmasında genetik popülasyon modelleri geliştirmiştir. Ayrıca Holland çalışmalarını doğadaki canlıların hayatta kalma prensibinden yaralanarak makine öğrenimi üzerine yoğunlaştırmıştır. Lawrence Davis [43], ilk defa GA'nın kısıtlı optimizasyon problemlerinde etkili olarak kullanıldığı çalışması ile literatürde dikkat çekmiştir. Ayrıca yayımlanan bir başka çalışmasında gerçek yaşam problemlerini GA'ya uygulayarak çözümlemiştir.

4.1. Genetik Algoritma Terminolojisi

GA, tıp alanındaki genetik bilimini temel aldığından birbirine yakın terminolojileri bulunmaktadır. Kromozom, gen, alel (allele), konum (locus), fenotip (phenotype) ve genotip (genotype) terimleri aynı zamanda genetik algoritmadaki sayı dizilerini ve ifade etmek için de kullanılan terimlerdir.

Gen: Bir gen, her bir bireyin problemdeki çözüm adayının belirli bir özelliğini veya parametresini temsil eder. Genetik algoritma içindeki bir kromozomun parçasıdır.

Alel: Genlerin taşıdıkları değerdir, her alelin bilgisinin tutulduğu kısım konum olarak adlandırılmaktadır.

Kromozomlar: Popülasyondaki her bireyin ismi kromozomdur. Bir çözüm adayının tamamını temsil eden genlerin bir dizisidir.

Popülasyon: Verilen problemin tüm olası (kodlanmış) çözümlerinin bir alt kümesidir. Popülasyon, başlangıçta rastgele oluşturulan çözüm adaylarını içerir ve genetik operatörlerle (çaprazlama, mutasyon) birlikte evrimleşir.

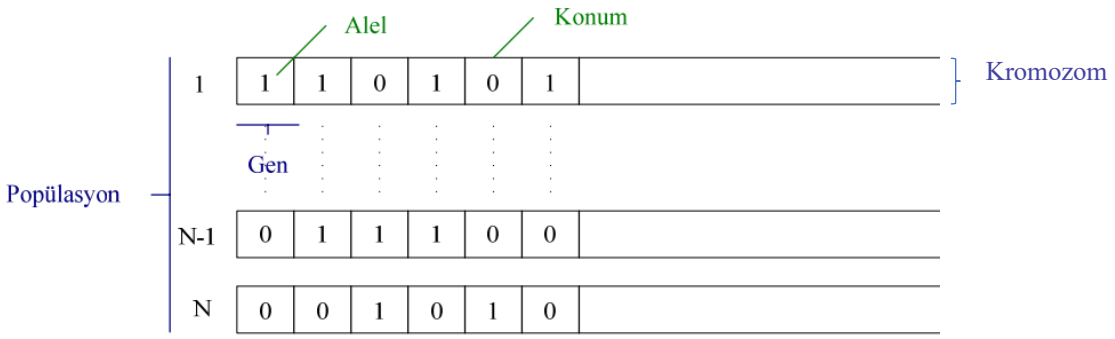
Fenotip: Fenotip, bir organizmanın gözlemlenebilir fiziksel özelliklerini ifade ederken, genetik algoritmalarda ise bir çözüm adayının özelliklerini temsil eder.

Genotip: Bir bireyin genetik materyalini temsil eder. Genetik algoritmalarda, bir bireyin genotipi, olası bir çözümü ifade eden bir dizi özellik veya parametre kümesidir. Bu özellikler, genellikle bir bit dizisi veya gerçel sayılar gibi belirli bir formatta ifade edilir.

Fitness Fonksiyonu: Fitness fonksiyonu, bir çözüm adayının ne kadar iyi veya kötü olduğunu değerlendiren bir ölçüdür. Fitness fonksiyonu, genellikle optimize edilmek istenen hedef fonksiyona veya performans kriterine dayanır. Genetik algoritma, fitness fonksiyonunu kullanarak çözüm adaylarının kalitesini değerlendirir ve gelecek nesillere aktarır.

İterasyon : Matematiksel olarak fonksiyon yineleme işlemidir. İterasyonlar, genetik algoritmanın evrim sürecini simgeler.

Genler birleşerek kromozomları meydana getirmektedir. Gen, yapay sistemlerde bilgiyi içeren karakter ya da özellikler olarak sembolize edilmektedir. Genlerin içerdikleri bilgiye alel denir ve her alelin bilgisinin saklandığı yere konum adı verilir. Yapay sistemlerde alel, bir özelliğin değerini temsil ederken konum ise bu özelliğin dizi üzerindeki konumunu belirtir. Popülasyon yapısıyla ilgili bir örnek Şekil 4.1'de gösterilmiştir [44].



Şekil 4.1. Genetik Algoritmada Popülasyon Yapısı

4.2. Genetik Algoritma Parametreleri

GA parametreleri, algoritma performansı ve sonuçları üzerinde büyük bir etkiye sahip olan ayarlanabilir değerlerdir. Bu parametreler, problem özelliklerine ve optimizasyon hedeflerine bağlı olarak ayarlanır. Genellikle deneme-yanılma veya optimizasyon yöntemleri kullanılarak en uygun parametre değerleri belirlenir. Ayrıca, problem özelliklerini dikkate almak ve özel durumlara uyacak şekilde parametreleri ayarlamak önemlidir.

İlk olarak gerekli parametrelerin belirlenmesi ya da kodlanması aşamasıdır. Bilginin yani kromozomların kodlanmasında genetik modelin çalışma performansı için doğru yapılması önemlidir. İkili, permütasyon, değer ve ağaç kodlama çeşitleri bulunmaktadır:

- İkili kodlamada, her kromozom ikili diziye sahiptir { 0, 1 }. Bu sıradaki her bit, çözümün belirli bir özelliğini temsil eder veya tüm dizi bir sayıyı temsil eder. Bu yapı, kolay ve hızlı işlemler için uygundur.

- Permütasyon kodlama, sıralama problemlerinde kullanılan bir yaklaşımdır. Burada her kromozom, sayıları belirli bir sıra ile temsil eder. Permütasyon kodlama, gezgin satıcı ve çizelgeleme problemleri gibi durumlarda kullanışlıdır.

- Değer kodlama, yaklaşımında ise değerler gerçek sayılar, karakterler veya nesneler olabilir. Karmaşık verilerin kullanıldığı problemlerde, ikili kodlamanın zor olduğu durumlarda doğrudan değer kodlaması kullanılabilir. Ancak, bu tür bir kodlamada, probleme özgü yeni çaprazlama ve mutasyon yöntemleri geliştirmek gerekmektedir. Bu şekilde, gerçek sayılar gibi karmaşık verilerin etkili bir şekilde temsil edilebilmesi sağlanır.

- Ağaç kodlama çeşidi de değişken programlar veya ifadeler için kullanılan bir yaklaşımdır. Örneğin, genetik algoritmalarda ağaç kodlama yöntemi kullanılarak her kromozom, belirli nesnelerin (fonksiyonlar veya programlama dilindeki komutlar gibi) bir ağacını temsil eder.

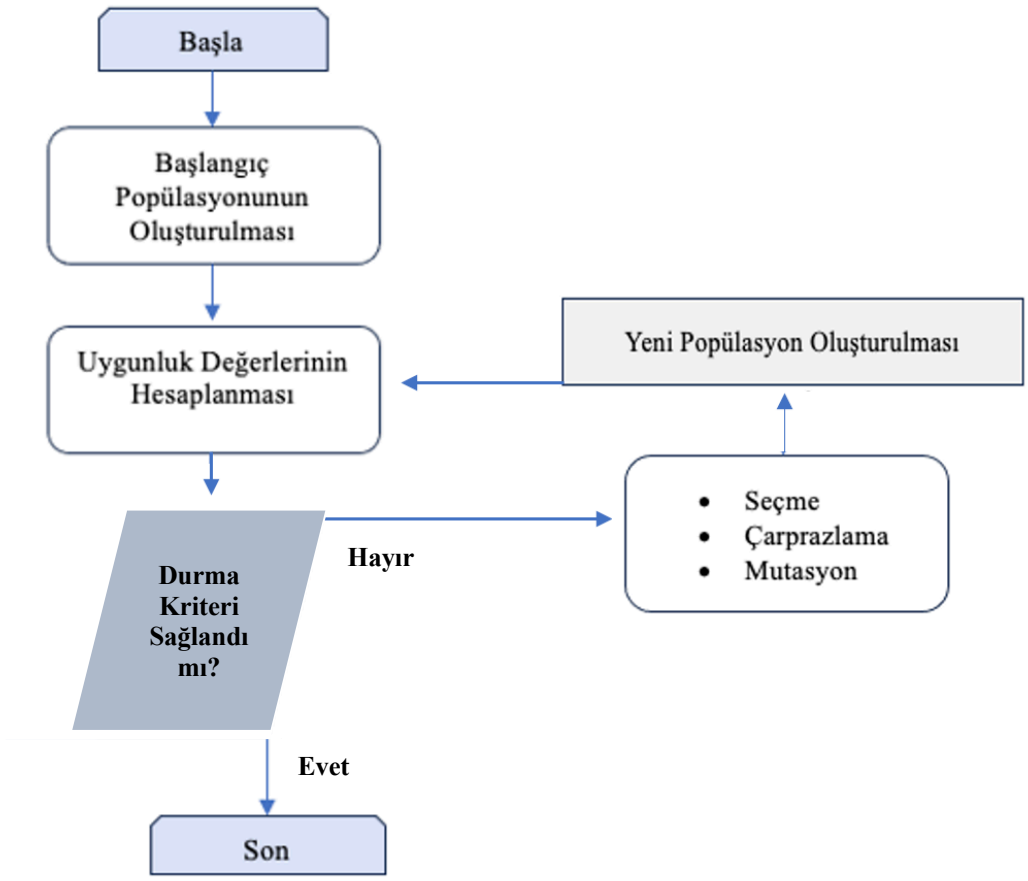
Modellemede kullanılan veri seti için genlerin gerçek sayı değerlerini temsil ettiği şekilde bırakılmıştır. Bu temsilde, her bir gen bir parametrenin değerini ifade eder ve genler gerçek sayı değerleriyle temsil edilir. Gerçek sayılarla kodlama, sürekli değerli problemlerde yaygın olarak kullanılır.

Popülasyon boyutu, her bir iterasyonda işlem gören çözüm adaylarının sayısını belirler. Daha büyük bir popülasyon genellikle daha fazla genetik çeşitlilik sağlar, ancak hesaplama süresini de arttırmaktadır [45]. Bu hesaplama süresi de göz önüne alınarak, uygulama kısmında her bir nesildeki popülasyonun boyutu 20 olarak girilmiştir. Her bir nesildeki kromozomların sayısı bu parametre ile belirlenmektedir.

Çaprazlama oranı, seçilen kromozomlar arasında genetik materyalin alışverişinin gerçekleşme olasılığını belirler. Yüksek çaprazlama oranı, genetik çeşitliliği artırırken, düşük çaprazlama oranı daha çok yeni çözüm yerine daha fazla mevcut çözümün kullanılmasına neden olur. Mutasyon oranı, her bir genin rastgele olarak mutasyona uğrama olasılığını belirler. Mutasyon, genetik çeşitliliği koruma ve yeni çözümler keşfetme potansiyelini artırır. Düşük mutasyon oranı, genetik materyalde daha az değişiklik yapılmasını sağlar. Çaprazlama ve mutasyon, genetik algoritmanın genetik çeşitlilik sağlamasına yarayan iki temel operatördür. Çaprazlama oranını seçerken mutasyon oranıyla dengelenmesine önem verilmiş olup 0.2 olarak mutasyon oranı; 0.7 olarak da çaprazlama oranı denemeler sonucunda bulunmuş ve performans çıktılarına bakılarak daha iyi sonuçlar elde edilmiştir.

4.3. Genetik Algoritma Akışı

Parametre belirleme süreci sonrasında GA genellikle aşağıdaki adımları izleyen bir döngü içinde çalışır, bu döngü Şekil 4.2.'deki akış diyagramında gösterilmiştir [46].



Şekil 4.2. Genetik Algoritma Akış Diyagramı

İlk adım olarak parametreler belirlendikten sonra başlangıç popülasyonunun oluşturulması gelmektedir. Bu çalışmada GA, optimal başlangıç noktaların birleşimini bulmayı amaçlar.

GA'da popülasyon yapısı oluşturulurken kromozom kodlama yapılır. Optimal başlangıç noktalarını aramak için her bir kromozom için değer kodlaması kullanılmıştır.

Bir sonraki adımda başlangıç popülasyonundaki bireylerin her birinin uygunluk değerleri hesaplanması gerekmektedir. GA'da kullanılan uygunluk fonksiyonu, optimizasyon problemlerindeki amaç fonksiyonu ile aynı işlevi görmektedir. Popülasyon kümesindeki değerlere yönelik uygunluk (Amaç Fonksiyonu) değerleri hesaplanmasını ifade eder. Uygunluk fonksiyonu bir kromozomun istenen şartlara ne kadar uygun olduğunun ölçülmesini sağlamaktadır.

GA'da seçme, çaprazlama ve mutasyon olmak üzere üç adet evrim operatörü mevcuttur [47]. Seçilim işleminde, seçim yöntemleri ile bireyler içerisinde yeni nesle aktarılacaklar seçilecektir. Popülasyon içindeki bireylerin uygunluk fonksiyonuna dayalı değerlerine göre, optimal çözüme daha yakın olan bireylerin seçilme olasılığı artar. Seçilen bireyler arasında evrimsel işlemler gerçekleştirilir. GA, kümeleme için iki veya daha fazla çözümün (küme merkezi pozisyonları) birleştirilmesiyle yeni çözümler üretmek için çaprazlama operatörünü kullanır. K-ortalamalar için bu çaprazlama işleminde iki farklı çözüm seçilerek belirli bir kesme noktası seçilir ve bu noktadan önceki küme merkezleri birleştirilir. Örneğin, çaprazlama noktası bir boyutun sonunda seçilirse, birinci çözümün ilk n-1 boyutu ile ikinci çözümün son boyutu birleştirilir. K-ortalamalar için mutasyon işleminde ise rasgele küme merkezi seçilir, belirli bir aralık içinde rastgele bir sayı ekleyerek veya çıkararak mevcut çözüm bir miktar değiştirilir. Yeni oluşan neslin uygunluk fonksiyonları tekrar hesaplanarak bir sonraki nesle aktarılacak olan bireyler, uygunluk değerlerine göre yeniden seçim sürecine dahil olurlar [47, 48]. Düşük uygunluk değerine sahip olan bireyler, seçilme ihtimalinin az olmasından dolayı belli adımdan sonra tamamen nüfusun dışında kalırlar. Yeni popülasyon uygunluk değeri daha yüksek olan bireylerden oluşmaktadır.

Belirli bir durma kriteri sağlanana kadar (örneğin, belirli bir zaman ya da iterasyon sayısı geçene kadar) döngü tekrar eder. Bu şekilde her aşamada popülasyon içi yüksek uygunluğa sahip olan

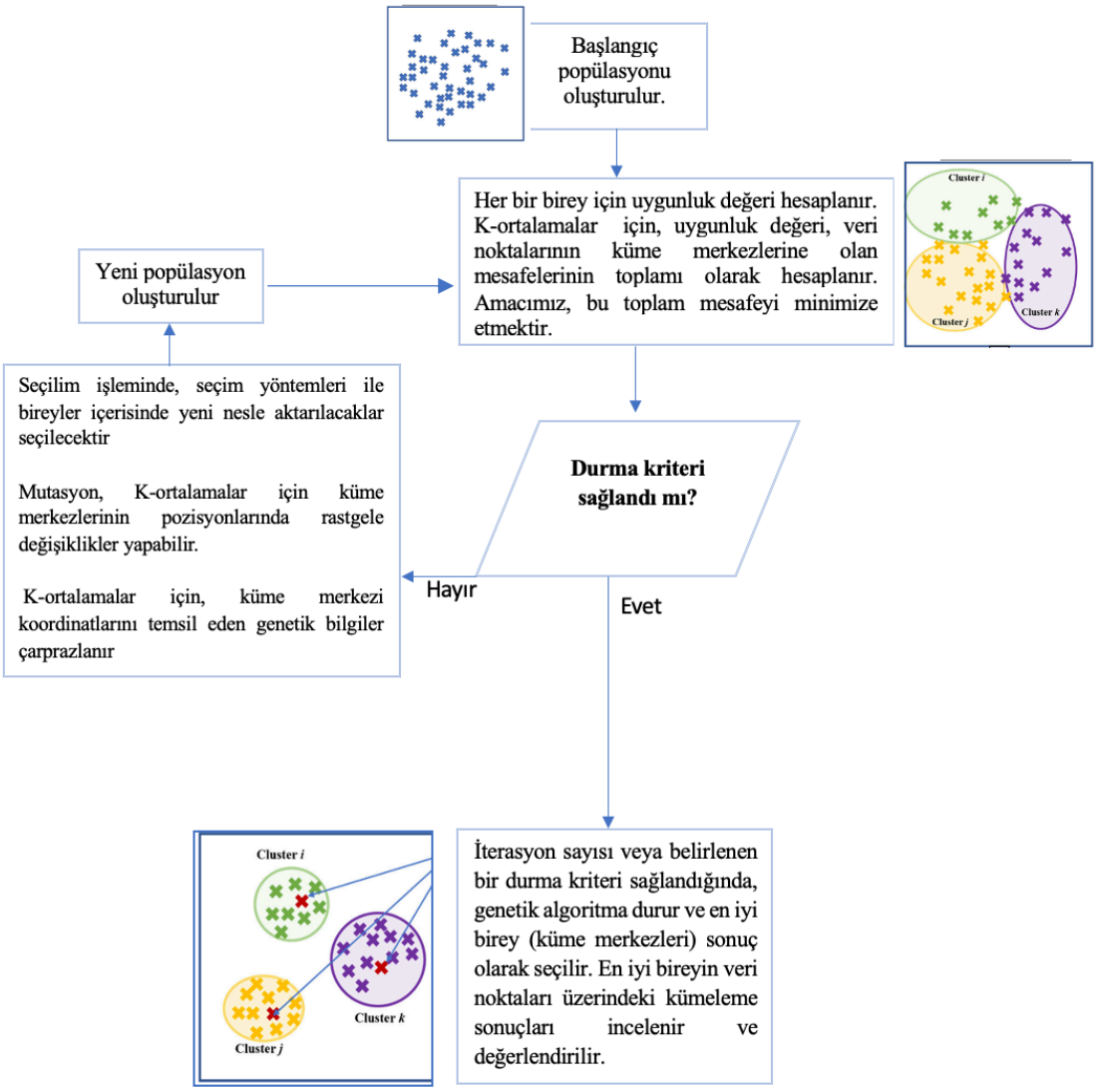
bireyler korunarak arama uzayında optimal bir çözüm bulunur. En iyi performans gösteren çözüm adayı, son popülasyondan seçilir ve en uygun çözümün bulunması ile işlem sonlanmaktadır.

4.4. Genetik K-ortalamalar Algoritması

GA'nın en iyileme problemlerindeki başarısı ve K-ortalamalar algoritmasının kümeleme problemlerinde etkili olması araştırmacılar adına bu iki yöntemi birleştirmek için teşvik edici olmuştur. Genetik K-ortalamalar algoritması, geleneksel K-ortalamalar algoritmasını evrimsel prensiplerle birleştirerek yeni bir yaklaşım sunar [49]. Bu yöntem, K-ortalamalar algoritmasının kümelerin merkezlerini optimize etmek için GA prensiplerini kullanmasını sağlar. Aynı zamanda küme merkezlerinin en iyi konumunu bulmak için genetik operatörlerin (seçilim, çaprazlama, mutasyon) ve uygunluk fonksiyonunun birleşimini kullanır.

GA, K-ortalamalar kümeleme algoritmasıyla birlikte birçok farklı problemin çözümünde kullanılabilir. Örneğin, müşteri gruplandırması, pazar analizi, sosyal ağ analizi gibi veri madenciliği uygulamalarında kullanılabilir. GA ve K-ortalamalar, örüntü tanıma problemlerinde de bir araya getirilebilir. Ayrıca, optimal yol problemlerinde de örneğin şirketin belirli noktalar arasında en kısa ve en ekonomik rota için çözüm aradığı bir koşulda GA, genetik materyali temsil eden rotaları optimize ederken, K-ortalamalar algoritması belirli rotaları gruplandırarak daha iyi sonuçlar elde edilebilir.

Kümeleme için genetik algoritmanın temel adımları, birey temsili ve popülasyon oluşturma, uygunluk hesaplama, seçim, çaprazlama ve mutasyon içerir. Her birey, bir özellik alt uzayını temsil eder. Uygunluk, bireyin temsil ettiği özellik uzayına göre kümeleme sonucunu temsil eder. Uygunluk ne kadar büyükse, bu tür özellik alt uzayında veri o kadar yoğun olur ve kümeleme sonuçları o kadar iyi olabilmektedir. GA K-ortalamalar kümeleme adımları Şekil 4.3'te verilmiştir [50].

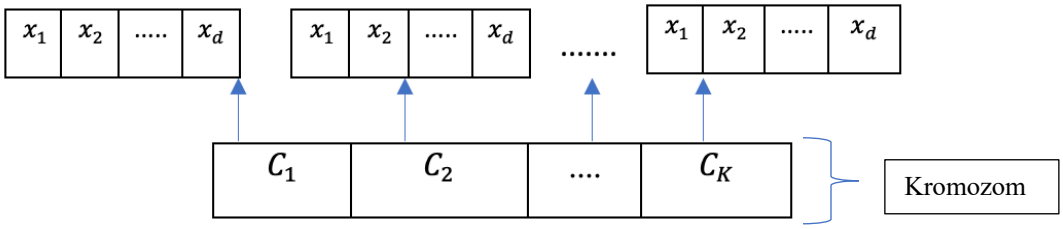


Şekil 4.3. GA K-ortalamalar Kümeleme Adımları

1. Başlangıç Populasyonu Oluşturma: İlk adımda genetik algoritmanın çalışması için başlangıç populasyonu oluşturulur. Her bir birey, bir kümelenme çözümünü temsil eder. Bu bireyler rastgele seçilen kromozomlara (örneklem noktalarının hangi kümelere ait olduğunu gösteren vektörler) denk gelebilir. Başlangıç popülasyonu, çözüm uzayından rastgele seçilen kromozomlardan oluşur. Bu nedenle, başlangıç popülasyonunda küme sayısının (K), popülasyon büyüklüğü (n) ile çarpımı

kadar kromozom seçilir. Her bir kromozom küme saysısı (K) ile veri kümesindeki açıklayıcı değişkenlerin boyutunun (d) çarpımı uzunluğunda bir vektördür. GA, bir dizi nesil boyunca K-ortalamalar algoritmasının ürettiği kümelerin kalitesini arttıran doğru küme merkezlerini seçer.

GA kümeleme için kromozom temsili Şekil 4.4'te gösterilmiştir [51].



Şekil 4.4. GA Kümeleme için Kromozom Temsili

2. Uygunluk Fonksiyonu: Her bir bireyin ne kadar iyi bir çözüm olduğunu ölçen bir uygunluk fonksiyonu belirlenir. Bu fonksiyon, K-ortalamalar algoritmasının özelliğine uygun bir şekilde, her bir veri noktasının merkeze olan uzaklıklarını temel alabilir. Uygunluk değeri, her bir veri noktasının merkezlere olan uzaklıklarının optimize edilen kümeleme çözümleri içindeki dağılımını değerlendirerek, veri noktalarının etkili bir şekilde ayrışmasını teşvik eden bir kriter olarak işlev görür. Durma kriteri sağlanmadığında genetik operatörlerle işleme devam edilir, sağlandığında ise akış sonlanır.

Bu çalışma için uygunluk fonksiyonu tanımlarken, tüm veri noktaları ile tüm küme merkezleri arasındaki öklid mesafesi temel alınır. Uygunluk değeri, kromozomun ne kadar iyi olduğunu ölçer. Bunun için bir "for" döngüsünde her veri en az öklid mesafesine göre kümeye atanmaktadır. Her kümedeki tüm mesafelerin toplamının hesaplanması için başka bir döngü kümelerden geçerek uzakları toplar ve toplamın tersi uygunluk değerini vermektedir. Toplamın tersi, uygunluk fonksiyonunu bir maksimizasyon fonksiyonu yapmak için hesaplanır. Eşitlik (4.1)'de formül yer almaktadır [52].

$$\text{Uygunluk} = \frac{1}{\sum_{k=1}^{N_C} \sum_{j=1}^{N_k} \sqrt{\sum_{i=1}^{F} (C_{ki} - P_{ji})^2}} \quad (4.1)$$

N_C: Küme sayısıdır.

N_k: Küme içindeki örnek sayısıdır.

F: örnekleri temsil eden özelliklerin sayısıdır.

C_k: Küme merkezidir.

P_j: Örnek veri noktasıdır.

3. Seçilim: Seçim operatörü, uygunluk değerine dayalı olarak kromozomların seçilmesini sağlar. Genellikle yüksek uygunluk değerine sahip kromozomlar daha fazla seçilme şansına sahiptir. Temel bir seçim yöntemi, rulet tekerleği seçimidir. Bu yöntemde, her kromozomun seçilme olasılığı, toplam uygunluk değerine göre belirlenir. Daha iyi uygunluğa sahip bireyler, gelecek nesillere daha fazla aktarılır.

4. Çaprazlama ve Mutasyon: Genetik algoritmalarda çaprazlama, farklı kromozomların özelliklerini birleştirerek yeni kromozomları oluşturur. Bu işlem, genetik materyalin farklı kombinasyonlarıyla daha iyi çözümler elde edilmesine yardımcı olur. Örneğin, iki kromozom arasında rastgele kesme noktaları belirlenerek parçalar değiştirilir. Mutasyon ise kromozomlardaki özelliklerin rastgele değiştirilmesini sağlar. Bu, popülasyondaki çeşitliliği artırarak yeni ve potansiyel olarak daha iyi çözümler keşfetmeyi amaçlar. Her iki operatör de genetik algoritmanın nesiller boyunca çözüm alanını keşfetmesine yardımcı olur, farklı özelliklerin birleştirilmesi ve rastgele değişikliklerle daha iyi sonuçlar elde edilmesini sağlar.

5. Yeni Popülasyon Oluşturma: Çaprazlama ve mutasyon sonucunda elde edilen yeni bireyler, mevcut popülasyondan daha iyi bir sonuç elde etmek için yeni popülasyonu oluşturur.

6. Durma Koşulu: Belirli bir iterasyon sayısına veya uygunluk değerlerinin istenen seviyeye ulaştığı bir noktaya kadar süreç tekrar edilir.

7. Sonuç Analizi: Son iterasyonda elde edilen en iyi bireyin kromozomu kullanılarak, her bir veri noktasının hangi küme için seçildiği belirlenir. Bu, verilerin optimal kümelenmiş halini elde etmenizi sağlar.

Bu akış, GA'nın K-ortalamalar algoritmasını destekleyerek daha iyi bir çözüme ulaşmasını sağlar. GA, çeşitli kombinasyonlar ve mutasyonlar yoluyla farklı kümeleme çözümlerini keşfedebilir ve uygunluk fonksiyonuna göre en iyi çözümü yakalayabilir. Bunun yanı sıra K-ortalamalar algoritması ile GA kümelemesi farklı yaklaşımlara dayalıdır ve veri kümeleme problemlerini farklı açıdan ele almaktadır. Her iki yöntemin beraber kullanımı ise problemin özelliklerine ve amaçlara bağlıdır. Problemin boyutu, veri tipleri, istenen sonuçlar ve hesaplama kaynakları, yöntemin kullanımını etkileyen faktörler arasında yer almaktadır. Çizelge 4.1'de ise K-ortalamalar kümeleme ve GA kümeleme yaklaşımlarının kıyaslaması yer almaktadır [53].

Çizelge 4.1. K-ortalamalar Kümeleme ve GA Kümeleme Kıyası

K-ortalamalar Kümeleme	**GA Kümeleme**
Bölümleme Tabanlı Yöntem	Evrim Tabanlı Yöntem
Girdi: k, veri kümesi, rastgele seçilmiş k merkez noktası	Girdi: k, veri kümesi, popülasyon sayısı, rastgele seçilmiş kromozom, iterasyon sayısı
Amaç: Toplam karesel uzaklığı en aza indirme	Amaç: Her veri noktasının küme merkezine olan uzaklıklarının toplamını en aza indirme
Sonlandırma koşulu: Yeni küme merkezlerinde değişiklik yok ise	Sonlandırma koşulu: Maksimum iterasyon sayısına ulaşıldı ise
Son kümeleme yerel optimuma yakınsayabilir	GA, global arama yaklaşımlarına dayalı olarak ve örtük paralellikle çalışır
Zaman karmaşıklığı: (n*k*d*i) n = veri sayısı k = küme sayısı d = veri boyutu i= iterasyon sayısı	Zaman karmaşıklığı: (tmax*p*n*k*d) n = veri sayısı k = küme sayısı d = veri boyutu tmax = maksimum iterasyon sayısı p = popülasyon sayısı

4.5. Genetik Algoritma Kümelemede K-En Yakın Komşu Kavramı

K-En Yakın Komşu (KNN), denetimli öğrenme tekniklerinden biridir ve sınıflandırma problemlerini çözmek için kullanılan bir yöntemdir. Bu algoritma, yeni bir veri noktasını sınıflandırmak için yakınındaki eğitim örneklerine dayanır. Veri noktasının sınıflandırılması için uzaklık metriği kullanılarak en yakın "k" eğitim örneği belirlenir. Yeni veri öğesi, eğitim kümesindeki diğer veri noktalarıyla karşılaştırılarak aralarındaki hesaplanan uzaklık değerine göre, en uygun sınıf etiketi atanır. KNN, daha fazla veriye veya öğrenme parametresine ihtiyaç duymadan, veri setindeki doğal yapıları anlamak ve tahmin yapmak için kullanılır [54].

GA kümelemede, en yakın komşu kavramı genellikle iki şekilde kullanılabilir. İlk olarak veri noktaları, genetik algoritmanın popülasyonunda bireyler olarak temsil edilir. Bu bireyler, genetik operatörler (çaprazlama, mutasyon) ile evrim geçirirken, en yakın komşuluk ilişkilerini kullanarak yeni bireyler üretebilir. Yani, yeni bireyler oluşturulurken, en yakın komşu bireylerden alınan özellikler veya genetik materyal kullanılabilir.

Diğer bir yöntemde ise uygunluk değeri hesaplamasında kullanılabilmektedir. GA sürecinde her bireyin bir uygunluk değeri hesaplanır. Bu uygunluk değeri, bireyin ne kadar iyi olduğunu veya hedef fonksiyonu ne kadar iyi optimize ettiğini gösterir. Örneğin, bireylerin uygunluk değerleri, en yakın komşu veri noktalarına olan uzaklıkların toplamı gibi bir metrik kullanılarak hesaplanabilir.

Kim ve Ahn [6], ortaya koydukları GA temelli kümeleme algoritmasını online kullanıcı üzerinde müşterilerin özelliklerini baz alarak gruplandırmada kullanarak, diğer geleneksel kümeleme algoritmalarına daha iyi sonuçlar elde etmişlerdir. Ayrıca tavsiye modelleri için önişleme aracı olarak kullanışlı olabileceğini ortaya koymuşlardır. Çalışmada kullanılan araştırma verilerinde hem ikili kromozom kodlaması hem de değer kodlaması yapılmış olup bütün değişkenler kullanılmıştır.

Tez çalışmasında, GA-KOK yöntemi Kim ve Ahn'ın önerdiği modeli temel alarak çalışılmış ve açık kaynaklı bir veri kümesi üzerinde iki boyutlu uzayda küme merkezi noktaları üzerinde genetik işlemler uygulanarak kümeleme süreçleri değerlendirilmiştir. Bu çalışmada farklı olarak, müşteri değişkenlerine iki boyut üzerinde değer kodlaması yapılarak popülasyon oluşturulmuştur. Ayrıca açık veri seti kullanımı dolayısıyla tavsiye sistemi oluşturmak yerine test müşterileri için kümelenmiş olan eğitim müşterilerindeki komşuluklarına bakılarak tavsiye üretimi yapılmış ve incelenmiştir. Kümeleme süreçlerindeki merkez noktaların evrimsel işleyişinde ise yeni bir kütüphane olan PyGAD kütüphanesi kullanılmıştır [55]. PyGAD kütüphanesi paralel işleme avantajıyla bulut tabanlı çözümlerde hızlı çalışmayı sağlar. Bunun yanı sıra, kapsamlı bir dökümantasyonu ve sürekli geliştirilen versiyonlarla birçok probleme uygulanabilir. Parametre seçimleri farklı problem türleri için uyarlanabilir. Kullanıcılar, problem gereksinimlerine göre farklı seçilim, çaprazlama ve mutasyon yöntemleri seçebilirler.

Bu bağlamda, GA-KOK yönteminde bireylerin uygunluk değerlerini test ederken "en yakın komşu" kavramı kümelemeye dahil olmayan müşterilerin komşuluklarına göre atanması olası olan küme merkezlerine uzaklıkları bakımından metrik olarak kullanılabilir. Bu uygunluk değeri, bir bireyin müşteri bölümlendirme görevini ne derece başarılı bir şekilde gerçekleştirdiğini yansıtarak, müşterilerin aynı grupta yer alan diğer müşterilere kıyasla ne kadar benzer olduğunu ölçme amacına hizmet eder [50]. Dolayısıyla, her bir bireyin uygunluk değerini hesaplamak için, söz konusu bireyin temsil ettiği müşteri grubunun, bu grubun diğer müşteri üyelerine olan benzerliğini nicel olarak ifade eden bir ölçüm olarak kullanılabilir. Bu sayede, her bir bireyin kümelenme performansı, daha geniş bir perspektiften değerlendirilebilir ve elde edilen uygunluk değerleri, GA'nın optimize etmeye çalıştığı hedef fonksiyonunun bir yansıması olarak işlev görebilir.

5. MODELLEME ÇALIŞMASI

5.1. Veri Seti

Açık kaynak olarak Kaggle'dan alınan "Marketing_Campaign" veri kümesi üzerinde önceki bölümlerde değinildiği şekilde K-ortalamalar ve GA kullanılarak GA-KOK yöntemi uygulanmıştır [55]. K-ortalama algoritması ile oluşturulan küme kalitesi, rasgele olarak seçilen ilk başlangıç merkez noktalara bağlıdır. Farklı başlangıç noktaların seçimi, kümeleme sonuçlarında büyük farklılıklar oluşturabilir. GA kümeleme yaklaşımı ile başlangıç noktalarını seçerek kümeleme optimizasyonu sağlamak ve bu sayede tavsiye sistemleri için daha iyi kümeleme sonuçları ortaya koymak modellemenin ana motivasyonunu oluşturmaktadır. Bu sayede kümelerin kalitesi ve önerilerin uygunluğunun iyileştirilmesi hedeflenmektedir. Veri seti seçiminde, açık kaynaktan erişilebilme avantajının yanı sıra GA'nın kümelemede uygulanabilmesi için yeterli değişken içermesi ve müşteri özelliklerinin gerçek e-ticaret verilerine benzerliği etkin olmuştur.

"Marketing_Campaign" veri kümesi 2240 tekil kullanıcı ve 3'ü kategorik; 26'sı sayısal olan 29 değişkenden oluşmaktadır. Bu değişkenler müşteri bilgileri, demografik ve sosyo-ekonomik bilgileri içermektedir. Müşterilerin satın alma davranışlarını ise gerçekleştirdiği satın alma işlemleri, son satın alma işleminden bu yana geçen gün sayısı, promosyon detayları ve satın alma kanalı bazında değişkenler oluşturmaktadır. Çizelge 5.1'de detayları yer almaktadır [56].

Çizelge 5.1. Kullanıcı, Ürün, Promosyon ve Kanal Değişkenleri

Değişkenler	
Musteri Bilgileri	ID: Müşterinin tekil tanımlayıcısı Year_Birth: Müşterinin doğum yılı Education: Müşterinin eğitim düzeyi Marial_Status: Müşterinin medeni durumu Income: Müşterinin yıllık hane geliri Children: Müşterinin evindeki çocuk sayısı Teenhome: Müşterinin evindeki gençlerin sayısı Dt_Customer: Müşterinin şirkete kayıt tarihi Recency: Müşterinin son satın alma işleminden bu yana geçen gün sayısı Complain: Müşteri son 2 yılda şikayette bulunduysa 1, aksi halde 0
Ürün Detayları	MntWines: Son 2 yılda şaraba harcanan miktar MntFruits: Son 2 yılda meyvelere harcanan miktar MntMeatProducts: Son 2 yılda ete harcanan miktar MntFishProducts: Son 2 yılda balığa harcanan miktar MntSweetProducts: Son 2 yılda tatlılara harcanan miktar MntGoldProds: Son 2 yılda altına harcanan miktar
Promosyon Detayları	NumDealsPurchases: İndirimli satın alma sayısı AcceptedCmp1: Müşteri 1. kampanyada teklifi kabul ederse 1, aksi takdirde 0 AcceptedCmp2: Müşteri 2. kampanyada teklifi kabul ederse 1, aksi halde 0

	AcceptedCmp3: Müşteri 3. kampanyada teklifi kabul ederse 1, aksi halde 0 AcceptedCmp4: Müşteri 4. kampanyada teklifi kabul ederse 1, aksi halde 0 AcceptedCmp5: Müşteri 5. kampanyada teklifi kabul ederse 1, aksi halde 0 Response: Müşteri son kampanyada teklifi kabul ettiyse 1, aksi takdirde 0
Satın Alma Kanalı	NumWebPurchases: Şirketin web sitesi aracılığıyla yapılan satın alma sayısı NumCatalogPurchases: Katalog kullanılarak yapılan satın alma sayısı NumStorePurchases: Doğrudan mağazalarda yapılan satın alma sayısı NumWebVisitsMonth: Geçen ay şirketin web sitesine yapılan ziyaretlerin sayısı

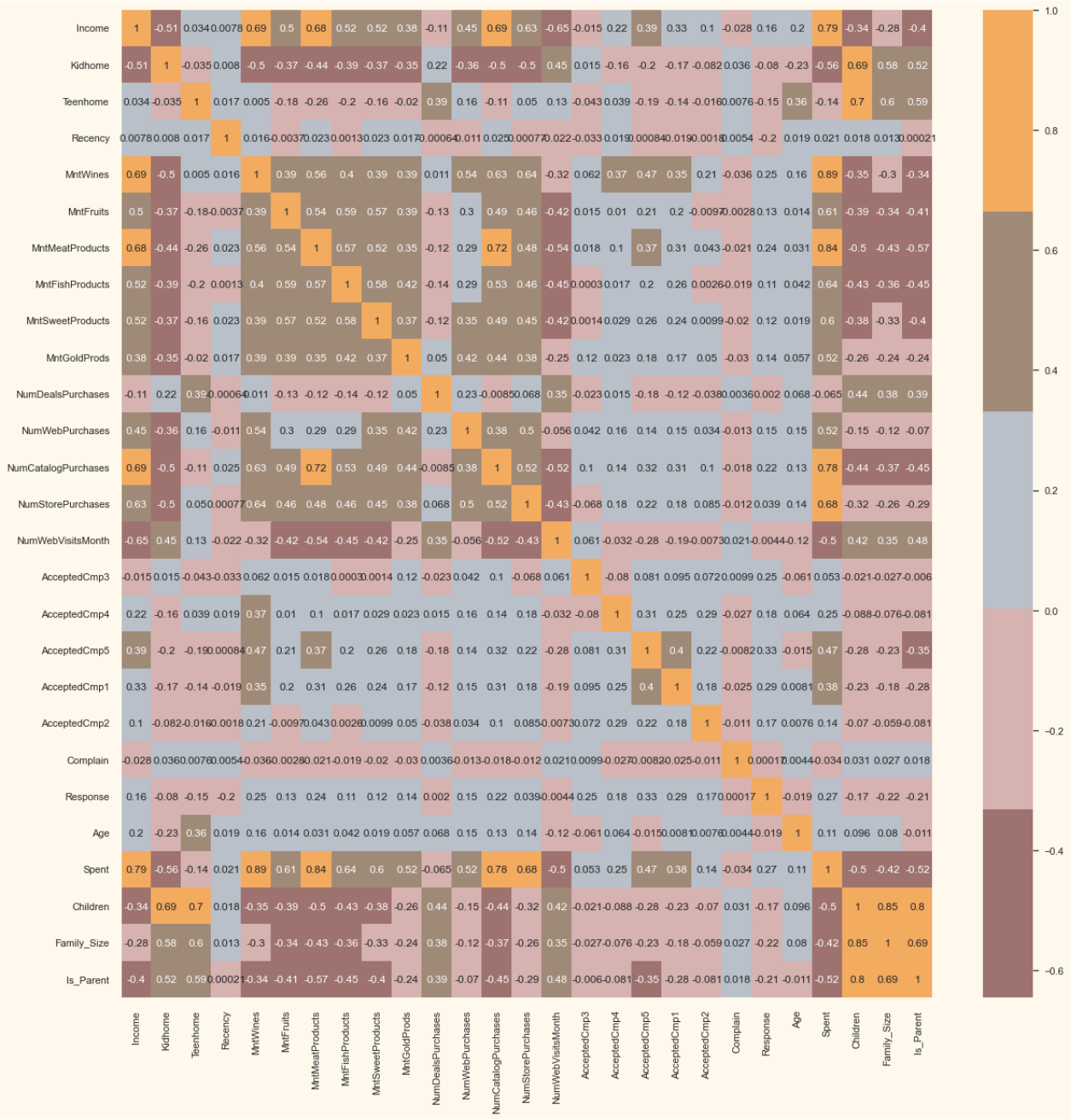
Veri kümesinde ilişkili değişkenler arasında yüksek bir korelasyon varsa, bunlardan sadece birini kullanmak veya birleştirerek yeni bir değişken oluşturmak performansı etkileyen faktörlerdendir. Bu sebeple veri setinde de modelde kullanılmak üzere bazı değişkenler ön işleme sokularak özet değişkenler olarak yeniden oluşturulmuştur. "Kidhome", "Teenhome" değişkenleri "Children" değişkeninde birleştirilmiştir. Müşterilerin doğum tarihleri veri kümesinin yüklenme yılı olan 2014 baz alınarak o yıldaki yaşları dikkate alınmış ve "Age" değişkenine dönüştürülmüştür. Müşterilerin ürün detayları ise 6 ürün kategorisine harcanan miktarın toplamı olarak 'Spent' değişkeni oluşturulmuştur. Müşterilerin medeni durumu "Marital_Status" ise 8 farklı etiketten oluşurken sadeleştirilerek "Alone" veya "Partner" olarak 2 kapsamlı kategoride gruplandırılmıştır. Medeni durum çocuk sayısı ile birleştirilerek "Family Size" değişkeni elde edilmiştir. Bu işlemler sonrasında modelde kullanılan değişkenler ise yaş, gelir, harcama, ailedeki kişi sayısı, müşterinin son satın alma işlemi üzerinden geçen süre, web sitesi aracılığıyla yapılan satın alma sayısı, geçen ay şirketin web sitesine yapılan ziyaretlerin sayısı ve indirimli satın alma işlemleridir.

5.2. Yöntem

Modelde kullanılmak istenen veri seti CSV dosyası olarak bulunmaktadır. Jupyter Notebook ortamında Python 3.8.3 sürümü ile veri yükleme, veri ön işleme, normalleştirme-standartlaştırma, boyut azaltma gibi işlemler sonrasında kümeleme ve optimizasyon işlemleri yapılmıştır.

Bu çalışmada GA için kullanılmak üzere PyGAD (Python Genetic Algorithm) kütüphanesi tercih edilmiştir. PyGAD, genetik algoritma oluşturmak ve makine öğrenimi algoritmalarını optimize etmek için açık kaynaklı bir Python kitaplığıdır. PyGAD, diğer Python kütüphaneleriyle kolayca entegre edilebilir. Özellikle NumPy, Pandas veya Matplotlib gibi veri analizi ve görselleştirme kütüphaneleriyle uyumludur. GA kullanarak optimizyon problemlerini çözmek için farklı çaprazlama, mutasyon ve ebeveyn seçimi gibi parametre özelleştirmelerini de desteklemektedir [55]. Dolayısıyla sağladığı avantajlardan dolayı bu çalışmada GA için Pygad kütüphanesi tercih edilmiştir. Ayrıca GA'da küme başlangıç nokta seçimini kolaylaştıran dolayısıyla da kümeleme sürecini geliştiren kodların tanımlanabilme olanağı çalışmalarda avantaj sağlamaktadır. PyGAD kütüphanesinin ilk sürümü 2020 yılında yayınlanmıştır. Günümüzde de GA kullanarak optimizasyon problemlerini çözmek için Python geliştiricilerine hizmet vermektedir.

Veri kümesinin hazırlanması aşamasının ilk adımında, farklı yapılardaki veriler düzenlenmiş, aykırı veriler tespit edilmiş, normalleştirme ve standartlaştırma yapılmış, boş gelen değerler ise medyan ile değiştirilmiştir. Aykırı verilerin az olması nedeniyle veri kümesinden çıkarılıp, 2236 tekil kullanıcı üzerinden modelleme çalışması yapılmıştır. Veri setinde kümeleme yapmak için bir sonraki adımda ise değişken seçimi yapılmıştır. İyi bir değişken seçimi, algoritmanın performansını artırabilir ve daha anlamlı kümeleme sonuçları elde edilmesini sağlayabilir. Değişkenler arasındaki ilişkiyi incelemek için korelasyon matrisi kullanılması, K-ortalamalar kümeleme algoritması için değişken seçimi ve analizi açısından da önemlidir. Şekil 5.1.'de korelasyon matrisi, ısı haritası grafiğinde değişkenler arasındaki korelasyonlar görsel olarak gösterilmiştir.



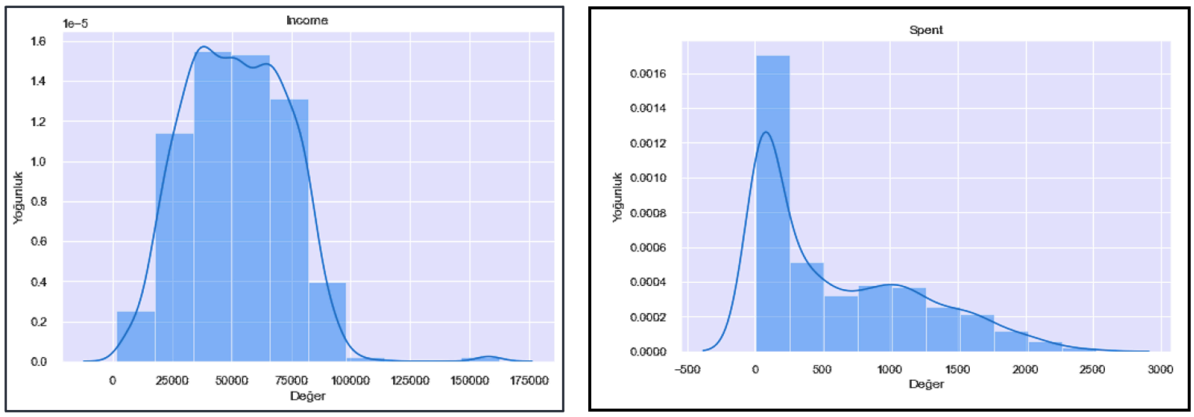
Şekil 5.1. Korelasyon için Isı Haritası

Veri setindeki değişkenlerin ilişkilerine bakıldığında gelir ile harcama değişkeni arasında pozitif güçlü bir ilişki olduğu görülmektedir. Ayrıca her kategori için harcama miktarının "Spent" değişkeni ile ilişkisi de yüksektir bu nedenle modelde yeniden oluşturduğumuz özet bilgi olan

"Spent" değişkeni yer almıştır. "Kidhome" ve "Teenhome" ile oluşturduğumuz çocuk sayısı değişkeni arasında da güçlü bir ilişki tespit edilmiştir. Ayrıca müşterilerin medeni durumu ve çocuk sayısı birleştirilerek elde edilen ailedeki kişi sayısı "Family Size" değişkeni arasında da güçlü pozitif ilişki bulunmaktadır. Bu sebeple de modelde "Family Size" yer alarak evdeki kişi sayısını özetleyen bir değişken olarak ele alınmıştır. Seçilen özellikler ve özet istatistikleri Çizelge 5.2'de ve gelir-harcama dağılımı ise Şekil 5.2'de yer almaktadır.

Çizelge 5.2. Seçilen Özellikler ve Özet İstatistikleri

	Income	Recency	NumDeals Purchases	NumWeb Purchases	NumWebVisits Month	Age	Spent	Family_Size
count	2236.0	2236.0	2236.0	2236.0	2236.0	2236.0	2236.0	2236.0
mean	51952.6	49.1	2.3	4.1	5.3	45.1	606.0	2.6
std	21411.5	29.0	1.9	2.8	2.4	11.7	601.9	0.9
min	1730.0	0.0	0.0	0.0	0.0	18.0	5.0	1.0
25%	35502.5	24.0	1.0	2.0	3.0	37.0	69.0	2.0
50%	51381.5	49.0	2.0	4.0	6.0	44.0	396.5	3.0
75%	68275.8	74.0	3.0	6.0	7.0	55.0	1045.5	3.0
max	162397.0	99.0	15.0	27.0	20.0	74.0	2525.0	5.0



Şekil 5.2. Gelir ve Harcama Dağılımı

Seçilen özelliklerin istatistiklerine göre 2236 müşterinin yaş ortalaması 45.1'dir. En küçük müşteri 18 yaşında iken; yaşı en yüksek müşteri ise 74 yaşındadır. Aile büyüklüğünün ortalaması 2.6'dır. Veri setinde aileler ortalama 3 kişiliktir. Gelir grafiğinde ise yoğunluk 50.000 civarındadır. Harcama grafiğine bakıldığında müşterilerin son 2 yılda ortalama harcaması 600 civarındadır.

5.3. Uygulama

Uygulama, GA-KOK yöntemi ile küme merkez noktalarının optimize edilmesi ve bu sayede müşterilerin optimal kümelenmesini sağlamayı amaçlar. Bu amaçla veri setinde daha sonra test müşterisi olarak kullanılma amacıyla, veri seti eğitim (%70) ve test kümesi (%30) şeklinde ayrılmıştır. Böylelikle 671 test ve 1565 eğitim verisi olarak bölünmüştür. Eğitim veri seti ile yöntem kısmında açıklanan ön işlemler sonrasında K-ortalamalar algoritması ve Genetik K-ortalamalar algoritması ayrı ayrı kümelemede uygulanmıştır. Performans karşılaştırması bakımından karesel hata toplam değeri, test müşterilerin kümelenmiş olan eğitim müşterilerdeki komşuluklarına göre küme merkezlerine uzaklıklarının toplamı ve küme içi homojenlik; kümeler arası heterojenlik gibi kriterler değerlendirilmiştir. Değerlendirme sonucunda GA-KOK yöntemin en iyi küme merkez çözümleri ve düşük karesel hata toplamı sonuçlarıyla, tavsiye üretimi GA-KOK üzerinden gruplandırılan müşteriler ile yapılmıştır.

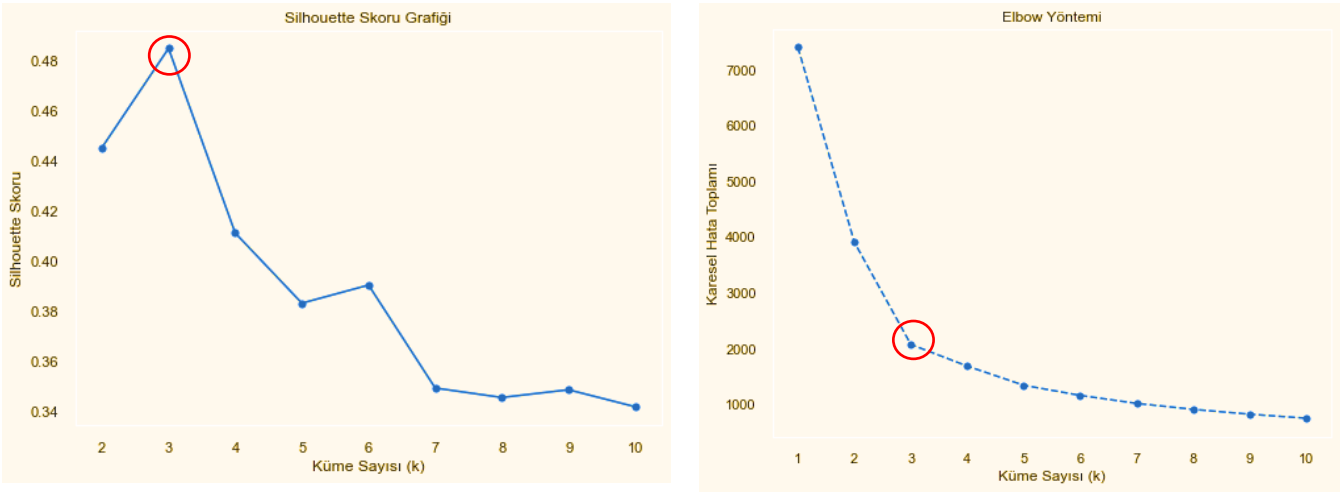
Kümelemede yer almayan test müşterilerden üç veri noktası rasgele seçilerek örnek olarak temsil edilmiş ve her iki kümeleme yönteminde de iki boyutlu uzayda kümelenmeleri görselleştirilmiştir. Daha sonrasında GA-KOK yönteminde, eğitim veri setinde kümelere atanmış komşuluklarına bakarak ilgili kümeye atanması olası olan test müşterisine önerilerde bulunmak için örneklendirilmiştir.

K-ortalamalar algoritmasında rasgelelik içeren bileşenlerine bağlı olarak, her çalıştığında farklı kümeleme sonuçları elde edilebilir. GA-KOK yöntemi ile daha iyi başlangıç noktaları ile her çalıştırmada genetik operatörler sayesinde daha iyi performans sağlayarak kümeleme kalitesini arttırmıştır.

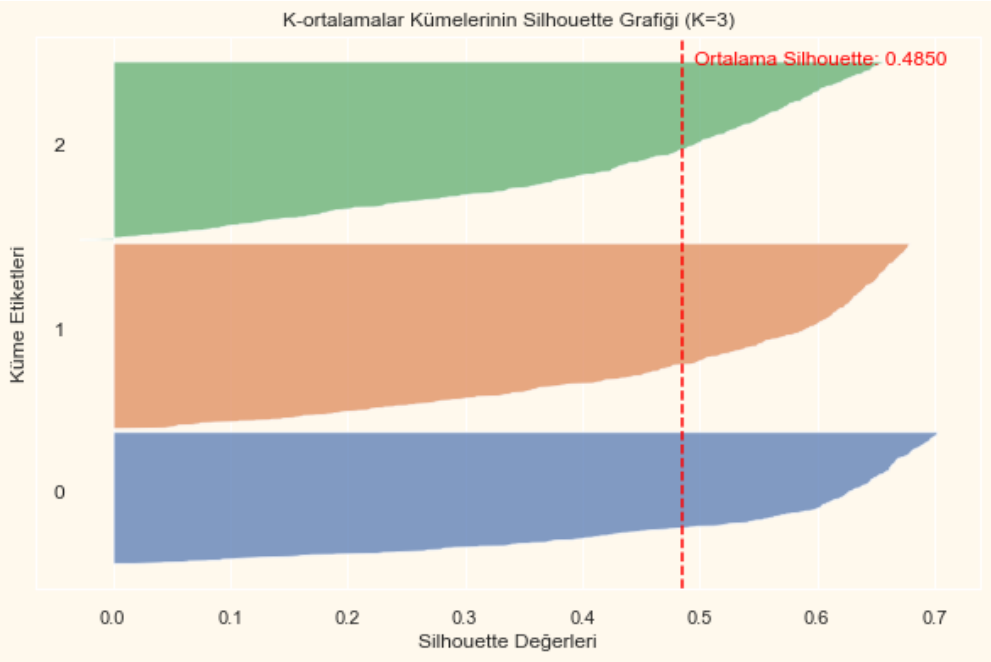
5.3.1. K-ortalamalar Algoritması ile Kümeleme

Silhouette grafiği ve karesel hata toplamı, bir kümeleme algoritmasının performansını değerlendirirken kullanılan önemli araçlardır. Daha yüksek Silhouette değeri homojen ve doğru bir kümeleme sonucunu göstermektedir. Karesel hata toplamı ise kümeleme algoritmasının hedefi olan kümelerin içindeki veri noktalarının birbirlerine ne kadar yakın olduğunu gösterir.

Kümeleme için belirlenen küme sayısı “Silhouette Score” ve “Elbow” metodu ile Şekil 5.3.’te gösterildiği gibi en iyi küme sayısı K =3 olarak bulunmuştur.

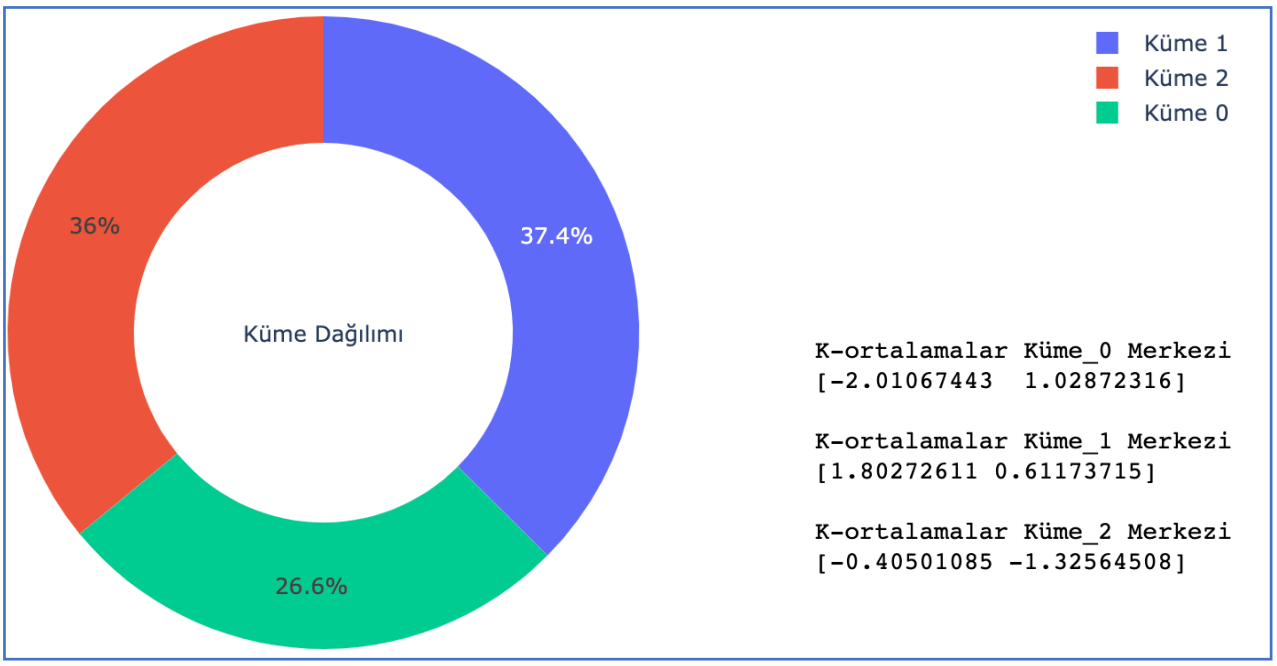


Şekil 5.3. Silhouette Skor ve Elbow Medotu ile Küme Sayısı



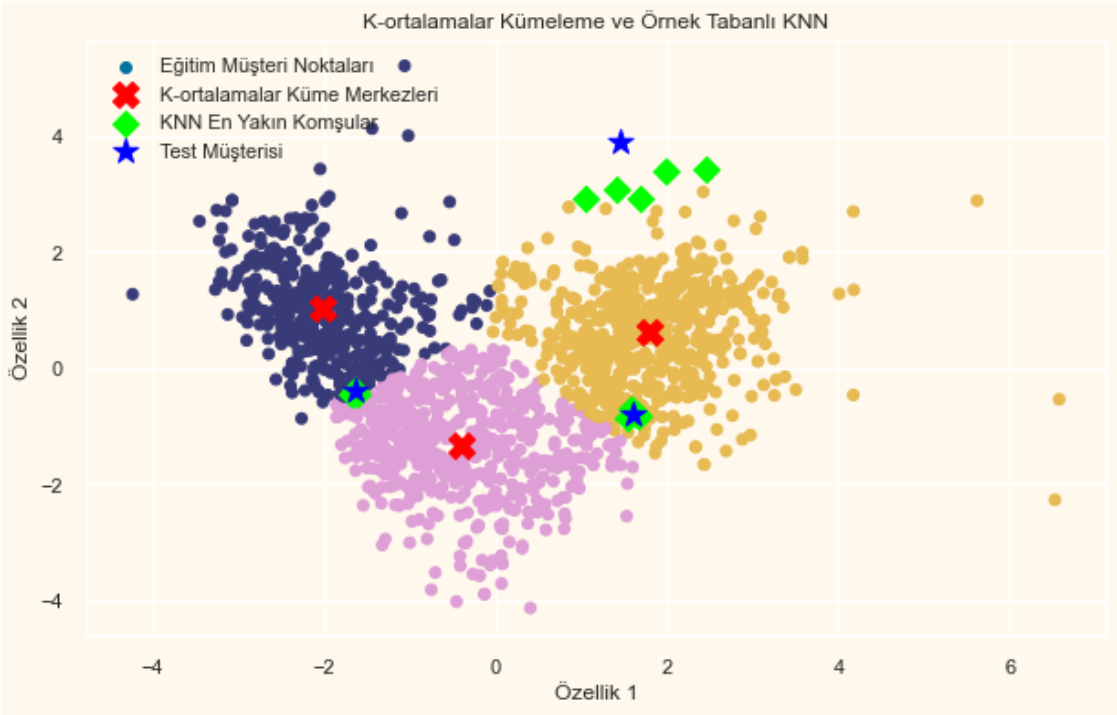
Şekil 5.4. Silhouette Değerleri ve Ortalama Silhouette Skoru

K=3 küme sayısı üzerinden yapılan kümeleme sonucunda, ortalama Silhouette değeri 0.48 iken kümeleme algoritmasının performansını gösteren küme içi karesel hata toplam değeri 2069.43 olarak bulunmuştur. Şekil 5.5.'te ise küme dağılımları ve küme merkezleri yer almaktadır.



Şekil 5.5. K-ortalamalar Küme Dağılımı ve Küme Merkezleri

Kümeleme görselleştirmesinde örneklendirmek adına, test veri setinden alınan rastgele üç test müşteri yeni müşteri gibi varsayılarak en yakın komşuluk algoritması ile K-ortalamalar kümelemesinde 5 yakın komşusuna göre kümeye ataması yapılmıştır. Test müşterilerin özelliklerine göre 3 küme merkezine öklid uzaklığı hesaplandığında ilk test müşteri 1.4 ile en yakın uzaklık olan küme 1'e, ikinci test müşteri 1.45 ile küme 0'a ve üçüncü test müşteri 3.3 ile küme 1'e atanması Şekil 5.6.'da gösterilmiştir.



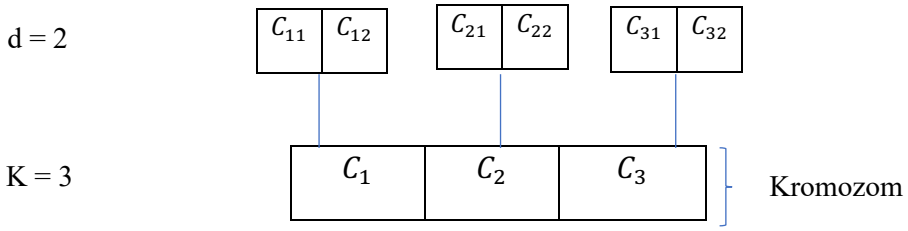
Şekil 5.6. K-ortalamalar Kümelemesi ve Test Müşteri Ataması

5.3.2. Genetik Algoritma K-ortalamalar ile Kümeleme

GA-KOK yönteminde GA kullanma hedefi, her bir veri noktasının kendi küme merkezine olan uzaklıklarının karelerinin toplamını minimize etmek ve aynı zamanda ilk küme merkezlerine daha az duyarlı olmakla birlikte, en uygun çözüme ulaşma yeteneğinden yararlanmaktır. Bu hedefte GA evrimsel süreci sayesinde en iyi başlangıç noktaları yerel optimuma düşmeden bulabilecek ve her iterasyonda en uygun küme merkezlerini arayarak süreci geliştirecektir. Kümelemede başlangıç nokta seçimi küme merkezlerini ve dolayısıyla kümelenme kalitesini etkiler. Her iterasyonda daha iyi sonuç arayan GA kümeleme, rasgele seçilen başlangıç noktalarıyla çalışan K-ortalamalar kümelemeye göre daha performanslı çalışmaktadır. Bu şekilde, daha homojen ve birbirine benzer veri noktalarından oluşan kümeler elde edilebilir.

Ele alınan kümeleme probleminde, problem çözümü kümelerin merkez koordinatlarıdır. GA kümelemede ise küme sayısı K-ortalamalar algoritmasında K=3 olarak verilmiştir. Açıklayıcı değişkenler standartlaştırılıp iki boyutlu uzayda gösterimi açısından boyut indirgenmiştir. Tanım olarak, $(1 < C_i < n)$ küme merkezini 2 boyutlu özellik uzayında temsil eden bir kromozomu ifade edersek kodlama biçimi Şekil 5.7.'deki gibi ifade edilebilir [54].

- Bir kromozom K*d uzunluğunda bir vektördür. Burada K küme sayısı ve d ise boyut sayısıdır**.**



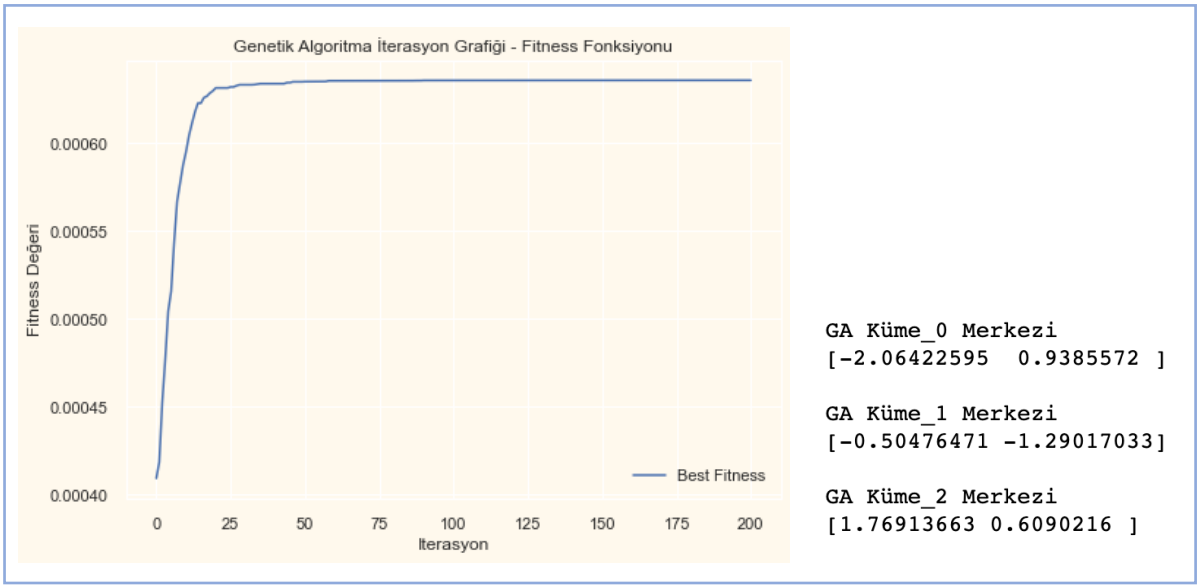
Şekil 5.7. GA Kümeleme için Kromozom Kodlaması

GA'da kullanılan evrim operatörleri ve parametreler Bölüm 4'te açıklanmıştır. GA'da düşük mutasyon oranı, mevcut popülasyonun genetik çeşitliliğini korurken, yüksek çaprazlama oranı yeni çözümlerin hızla üretilmesini sağlar. Bu denge, algoritmanın hızlı yakınsamasına ve farklı çözüm alanlarını keşfetmesine yardımcı olur. Çaprazlama ve mutasyon oranları için mutasyon oranını düşük; çarprazlamayı ise yüksek tutulması dikkate alınarak seçilmiştir. Mutasyon oranı 0.2; çaprazlama oranı 0.7 olarak verilmiştir. Hiperparametreleri deneme-yanılma yolu ile bulunmuş olup başlangıç değerlerinden başlayarak, algoritmanın performansını etkileyen hiperparametreleri değiştirip sonuçları izleyerek en iyi kombinasyon bu oranlarla belirlenmiştir. Uygulama kısmında iterasyon sayısı seçilebilir bir değer olarak ara yüze tanımlanmıştır. Farklı iterasyon sayılarının genellikle kural bulma sürecinde etkili olduğu gözlenmiştir. Uygulamada kullanılan genetik operatörler ve parametre değerleri Şekil 5.8.'de gösterilmiştir.

Parametre	Değer
Küme sayısı	3
Her nesildeki popülasyon sayısı	20
İterasyon sayısı	200
Mutasyon oranı	0.2
Çarprazlama oranı	0.7

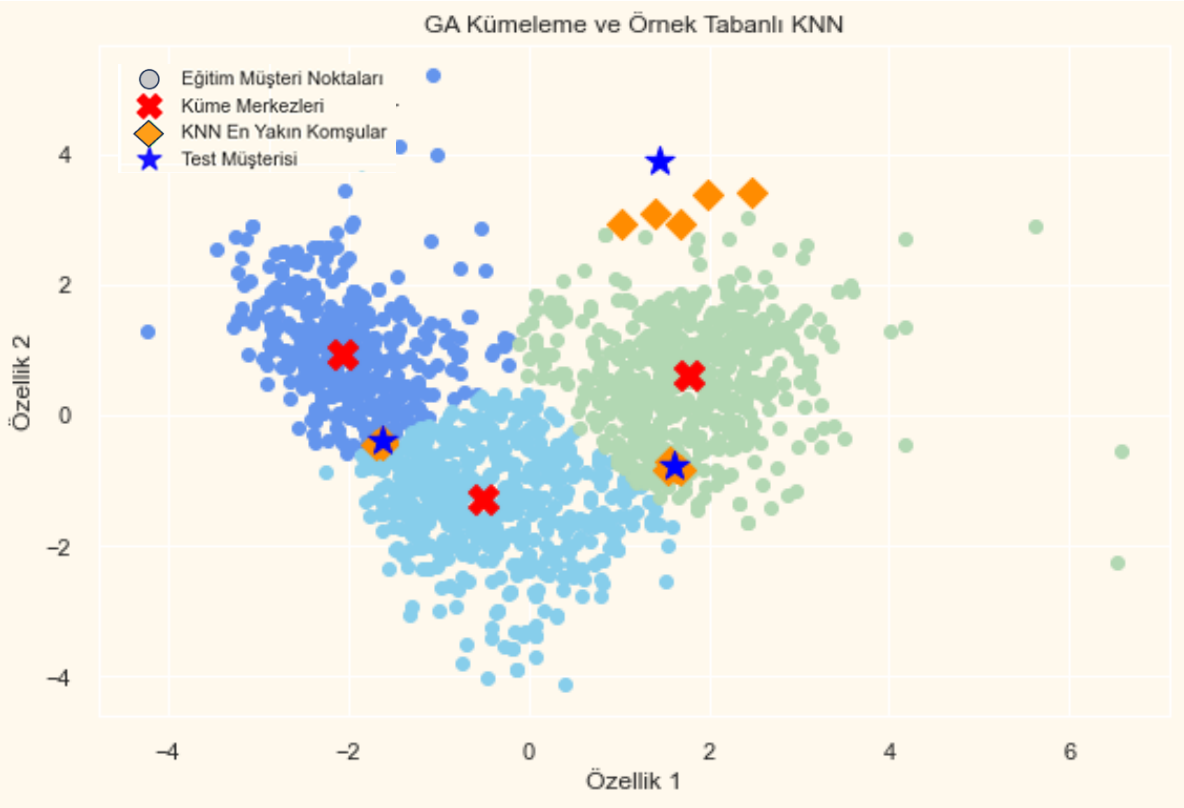
Şekil 5.8. GA Parametre Değerleri

Bölüm 4.4'te seçilen parametreler ve uygunluk fonksiyonu tanımlanmıştır. Bu parametreler ile GA kümelemede çalıştırıldığında, bu fonksiyonu kullanarak bireyler arasında rekabeti sağlar ve daha iyi sonuçlar üreten bireyleri seçme, çaprazlama ve mutasyon işlemlerine tabi tutar. Şekil 5.9.'da, popülasyonun iterasyon sayısı boyunca daha iyi uyum sağlayan bireylerle geliştiği ve en iyi çözüme yaklaştığı gösterilmiştir. En iyi çözüm 174 jenerasyon sonunda bulunmuş olup uygunluk değeri 0.000635 olarak sonuçlanmıştır.



Şekil 5.9. GA Uygunluk- İterasyon Grafiği ve Küme Merkezleri

GA küme merkezleri ile veri noktaları arasındaki mesafelerin toplamı sonucunda karesel hata toplam değeri 1585.35 olarak bulunmuştur. GA'da kümelenmeler ve test müşterilerin en yakın komşuluğa bakarak 3 küme merkezine öklid uzaklığı hesaplandığında ilk test müşteri 1.4 ile en yakın uzaklık olan küme 2'ye, ikinci test müşteri 1.3 ile küme 0'a ve üçüncü test müşteri 3.3 ile küme 2'ye atanması Şekil 5.10.'da gösterilmiştir.



Şekil 5.10. GA Kümelemesi ve Test Müşteri Ataması

5.3.3. Değerlendirme

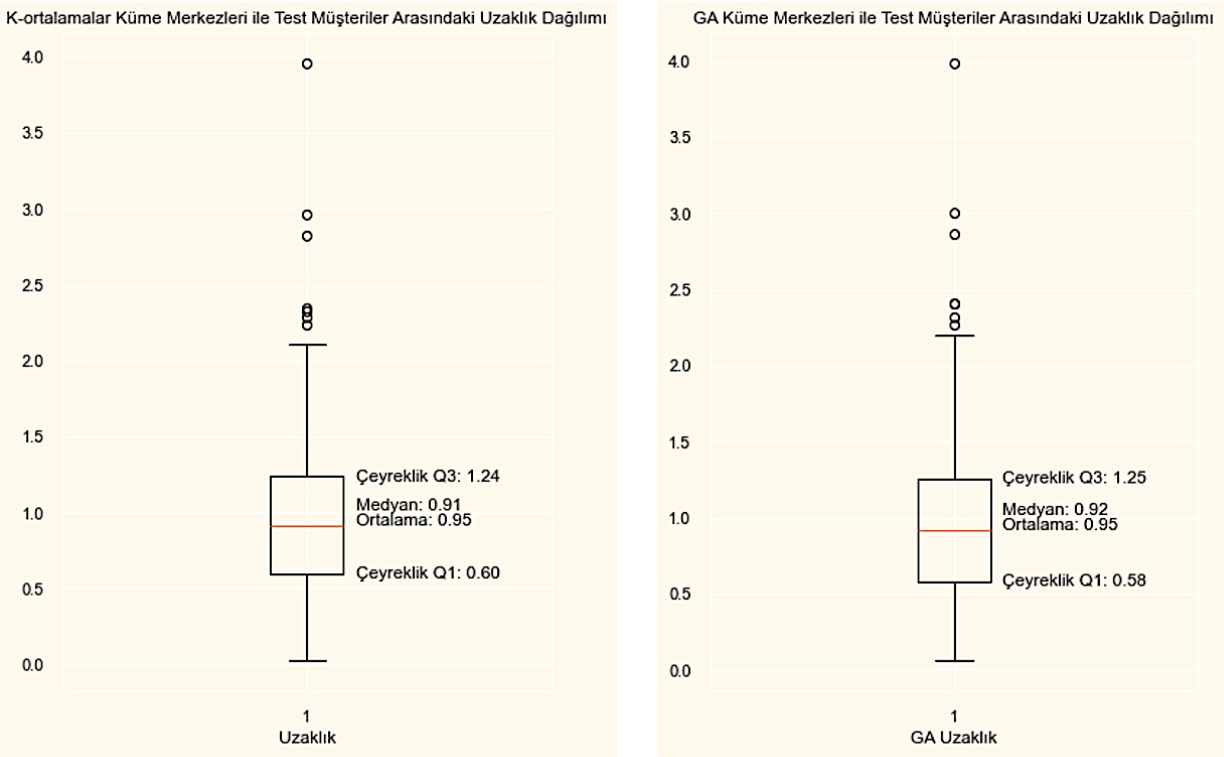
Kümeleme performans ölçüsü olarak karesel hata toplam değerlerine Çizelge 5.3'te bakıldığında, GA-KOK ile kümeleme yapıldığında küme içi karesel hata toplam değeri 1585.35 değeri ile geleneksel K-ortalamalar kümelemeden daha düşük çıkmıştır. GA-KOK yönteminin geleneksel K-Ortalama kümeleme yöntemine göre daha düşük bir küme içi karesel hata toplamına sahip olduğunu göstermektedir. Her bir veri noktasının kendi kümesinin merkezine daha yakın olduğu ve böylece kümeleme işleminin basit K-ortalamalar kümelemesine göre daha daha iyi gruplandırıldığı çıkarsanmaktadır.

Ayrıca test müşterisi için, 5 en yakın komşusuna göre atandığı küme merkezleri her iki algoritmada karşılaştığında da GA K-ortalamalarda merkeze çok daha yakın atandığı görülmektedir. 671 test müşterisinin küme merkezlerine uzaklık toplamı karşılaştırıldağında ise K-ortalamalar kümelemede 636.76 iken GA kümelemede ise 634.08 bulunmuştur.

Tüm test müşterilerinin küme merkezlerine olan toplam uzaklığı karşılaştırıldığında, GA-KOK yöntemi daha düşük bir toplam uzaklık değerine sahiptir. Bu, GA'nın küme merkezlerini daha iyi bir şekilde yerleştirdiği ve veri noktalarını daha iyi bir şekilde kümelediği anlamına gelmektedir. Şekil 5.11'te kutu grafikleri ile dağılımı gösterilmiştir.

Çizelge 5.3. K-ortalamalar ve GA-KOK Karşılaştırması

Kümeleme Algoritması	Küme içi karesel hata toplamı	Test müşterilerin komşuluğa göre atanan küme merkezler i arasındaki mesafeler toplamı
K-ortalamalar	2069.43	636
GA-KOK	1585.35	634

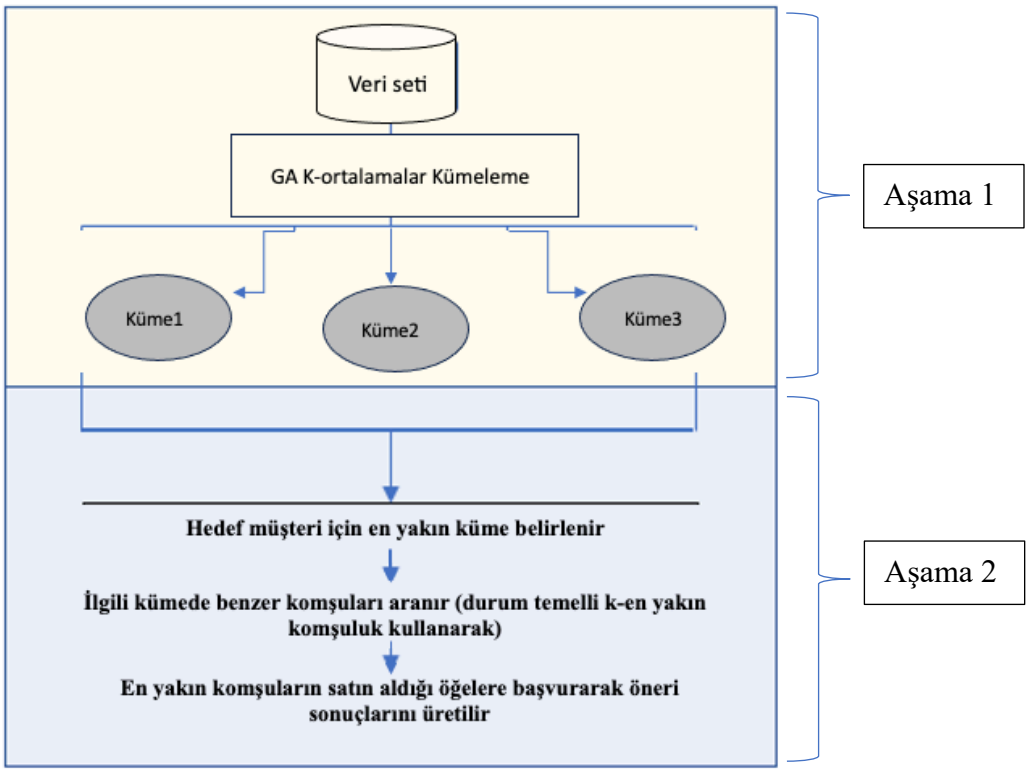


Şekil 5.11. K-ortalamalar Kümeleme ve GA-KOK Test Müşteri Kutu Grafiği

İkinci bölümde değinilen, tavsiye sistemlerinin karşılaştığı bazı problemlerde kümeleme ile belli stratejiler oluşturulması önemli hale gelmektedir. Örneğin yeni müşteri durumunda satın alma alışkanlıklarının ya da öğe değerlendirmelerinin olmaması problem yaratmaktadır. Bu sorunu çözmek için K-ortalama kümeleme yöntemi kullanılarak kullanıcılar benzer demografik gruplara ayrılır. Her bir grup, en benzer demografik yapıya sahip kullanıcılardan oluşur. Bu şekilde, hedef kullanıcıya ürün önerileri yapılırken önce hangi gruba ait olduğu belirlenir ve ardından aynı gruptaki kullanıcıların en çok satın aldığı ürünler önerilir. Şekil 5.12'de açıklandığı gibi, bu süreç iki aşamadan oluşur:

Aşama 1, daha kaliteli gruplar oluşturmayı içerir. Bu aşama, GA kullanılarak K-ortalama kümeleme algoritmasının başlangıç noktalarının seçilmesiyle gerçekleştirilir.

Aşama 2, hedef müşterilerin, aynı gruptaki kullanıcıların en çok satın aldığı ürünleri önerir.



Şekil 5.12. GA Kümeleme ile Tavsiye Üretim Akışı

Tavsiye sistemlerinde, yeni müşterilere tavsiyelerde bulunurken, sadece belirli bir müşteri kümesi içindeki verileri kullanmak, zaman açısından daha verimli bir yaklaşım sunabilir. Bu yaklaşım, daha az veri ile çalışmanın avantajlarına işaret ederken, özellikle modeldeki müşteri girdilerine odaklanarak bu kümelenmenin nasıl gerçekleştirildiği açıklanabilir. Kümeleme işlemi, müşterilerin satın alma geçmişlerine veya derecelendirme verilerine dayanabilir. Ancak, bu tür verilere sahip olmayan müşterilerle başa çıkmak bazı zorluklar doğurabilir. Bu noktada, kümeleme stratejisi devreye girer ve bu müşterilere özgü tavsiyeler üretmek için kullanılabilir. GA gibi teknikler, bu süreçte önemli bir rol oynayabilir.

Profil çıkarırken seçilen değişkenlerin önemi büyüktür. Hangi özelliklerin müşteri profilini belirlemede etkili olduğu, tavsiye sistemlerinin başarısını büyük ölçüde etkiler. Bu nedenle, doğru değişkenlerin seçimi, verimli ve etkili bir tavsiye üretimi için kritik bir adımdır.

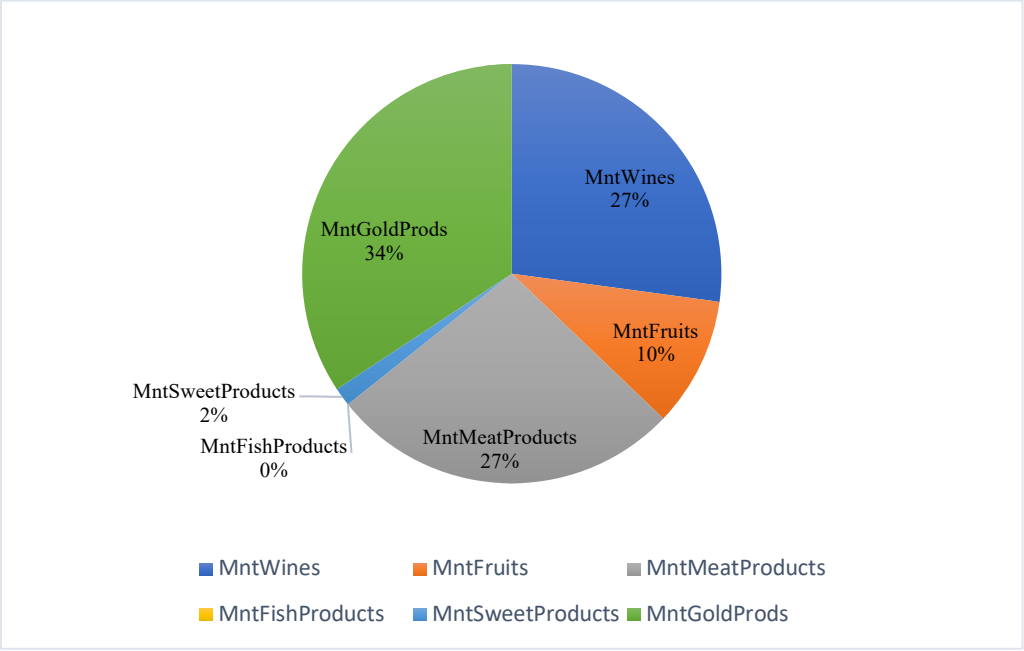
Hedef test müşterisinin 5 en yakın komşusuna göre atandığı kümeler ayrı ayrı incelenmiştir. Test müşterisinin demografik bilgileri bakımından, GA küme 1 için en yakın komşuluklarına bakarak çok benzer müşteriler olduğu ortaya çıkmaktadır. Bu kümenin demografik yapısına Şekil 5.13'te bakıldığında yaş ortalaması 40, eğitim durumlarının lisans üstü ve online kanal ziyaretleri diğer kanallara göre daha yoğunlukludur.

Çizelge 5.4'te eğitim kümesinde bulunan ve test müşterilerine en yakın komşuların satın alma davranışlarının grafiği yer almaktadır. Bu kümedeki müşterilerin sıklıkla altın, şarap ve et ürünleri aldığı fakat balık ürünlerinin henüz tercih edilmediği ve tatlı ürünlerinin de az alındığı çıkarımı yapılmaktadır. Bu çıkarımla birlikte bu kümedeki müşterilere balık ve tatlı ürünleri müşteri benzerliğine dayanarak önerilebilir. Henüz satın alma davranışı göstermeyen fakat demografik olarak bu kümeye dahil olma potansiyelindeki yeni müşterilere ise şarap, et ürünleri önerilebilir.

Bu doğrultuda ise GA-KOK ile yapılan kümelemedeki komşulara göre tavsiyelerde bulunulması daha etkili pazarlama stratejileri için tercih sebebidir.

Çizelge 5.4. Test Müşterisi GA-KOK

Index	Income	Recency	NumDeals Purchases	NumWeb Purchases	NumWebVIsits Month	Age	Spent	Family Size	Education Postgraduate
1424	43269	61	1	1	8	42	19	3	1
477	38961	60	1	2	7	41	70	3	1
1423	34596	48	1	1	8	40	23	3	1
965	34596	48	1	1	8	40	23	3	1
1076	41014	65	1	1	7	36	20	3	1
950	42767	53	1	3	8	43	131	3	1



Şekil 5.14. GA Küme_1 Ürün Grafiği

6. SONUÇ VE TARTIŞMA

Kümeleme yöntemleri, tavsiye sistemlerinin temelini oluşturarak benzer özelliklere sahip kullanıcıları veya öğeleri gruplandırma konusunda önemli bir rol oynar. Bu gruplandırma sayesinde tavsiye sistemleri, daha doğru ve etkili öneriler sunabilir. Özellikle büyük veri kümeleri ve karmaşık ilişkilerin olduğu durumlarda, kümeleme yöntemleri tavsiye sistemlerinin performansını artırarak verimlilik sağlar. Bununla birlikte GA, kümeleme sürecini optimize etmek ve tavsiye sistemleri için benzer kullanıcılara daha etkili tavsiyeler sunmak için kullanılabilmektedir. Bu şekilde, kullanıcıların ihtiyaçlarına daha iyi cevap veren ve kişiselleştirilmiş öneriler sunan bir tavsiye sistemi oluşturulabilir.

Tez çalışmasında, ilk olarak geleneksel tavsiye sistemleri ve karşılaşılan problemler ele alınıp, kümeleme yöntemleri ve genetik algoritma tanıtılmıştır. Ardından K-ortalamalar algoritmasıyla kümeleme yapılmış olup optimize etmek için GA kullanarak ayrıca bir kümeleme yöntemi de (GA-KOK) oluşturulmuştur. K-ortalamalarda olduğu gibi GA ile kümeleme de aynı aşamaları takip etmektedir. Fakat GA kümeleme, başlangıçta bir kümeleme çözümü oluşturur ve ardından bu çözümleri genetik operatörler olan çaprazlama, mutasyon ve seçilim işlemleriyle geliştirir. Bu, farklı kümeleme çözümleri yaratılmasına ve aralarından en iyi olanının seçilmesine olanak tanır. GA kümeleme, belirli bir hedef işlevini optimize etmeyi amaçlar. Bu hedef işlevi, kümeleme sonuçlarının kalitesini değerlendiren bir ölçüttür, küme içi ve kümeler arası uzaklıkların dengelenmesini hedefler.

Modelleme çalışmasında, iki kümelemedeki performans karşılaştırıldığında, GA-KOK yönteminin karesel hata toplam değeri %27 daha düşük çıkmıştır. Aynı zamanda GA-KOK, K-ortalamalar yöntemiyle yapılan kümelemeye kıyasla, benzer demografik yapıya ve alışveriş davranışlarına sahip müşterilere daha uygun atamalar yapma eğilimindedir. Komşu müşterilerin satın aldığı ürünlere bakarak, test müşterisi için ürün tercihi tahminlemesi, indirimli satın alma olasılığı ya da

satın alma yaparken hangi kanala yatkın olduğu gibi çıkarımlar yapılabilmektedir. Bu yöntemde, GA-KOK ile komşulara dayalı tavsiyelerde bulunmak, daha etkili müşteri profillemesi ve beraberinde pazar payı rekabet avantajı kazandırabilir. Bu nedenle, GA-KOK, karmaşık problemlerde ve özellikle büyük veri setlerinde kullanılarak gelecek çalışmalar için veri analitiği ve tavsiye sistemlerinde tercih edilen bir yaklaşım olabilir. Bunun yanında GA-KOK, tavsiye sistemlerinde bir bileşen olarak kullanılabilir ancak tek başına yeterli olmayabilir. Daha geniş ve farklı veri kümesi boyutları üzerinde test edilmesi ve sonuçların genelleştirilebilirliğinin değerlendirilmesi gerekebilir. Ayrıca, farklı hiperparametre ayarlarının etkisi daha detaylı bir şekilde incelenebilir ve en iyi performansı elde etmek için bu ayarların optimize edilmesi sağlanabilir.

Bununla birlikte, bu sonuçların istatistiksel olarak anlamlı olup olmadığını değerlendirmek için daha fazla hipotez testi ve çapraz doğrulama çalışmaları gerekebilir. Son olarak, GA-KOK yönteminin gerçek dünya uygulamalarında nasıl performans gösterdiğini anlamak için saha çalışmalarına odaklanmak da önemlidir. Kişiselleştirme, hesaplama gücü, açıklanabilirlik ve dinamik değişkenler gibi faktörleri ele almak için diğer öneri algoritma teknikleri ve veri madenciliği yaklaşımları ile birleştirilmesi daha verimli olacaktır.

K-ortalamalar algoritması yerine GA-KOK yönteminin avantajı, hata, gürültü veya anormallikler gibi çeşitli veri sorunlarına karşı dirençli olma yeteneği ile istikrarlı sonuçlar üretebilir ve performansını koruyabilir. Geleneksel yöntemlere göre daha iyi sonuçlar verdiği göz önüne alındığında, bu yaklaşımın gelecekteki kümeleme ve veri analizi çalışmalarında önemli bir potansiyele sahip olduğu söylenebilir. Aynı zamanda tavsiye sistemleri için, belirli bir uygulama veya kullanıcı grubunun ihtiyaçlarına uygun çözümler üretmek istendiğinde çok değerlidir. GA'nın genetik operatörleri, çaprazlama ve mutasyon, farklı veri yapıları ve hedeflere uyacak şekilde özelleştirilebilir. Bu nedenle, GA, her türlü veriye ve probleme özgü bir şekilde ayarlanabilir olması ve tavsiye sistemlerinde hem veri kalitesi sorunlarına karşı daha dayanıklı hem de özelleştirme yeteneği sayesinde güçlü bir destekleyici olarak gelecek çalışmalar için dikkat çekmektedir.

7. KAYNAKLAR

[1] Y., Çiçek, ve H., Muzaffaer, The Impact of Covid-19 Pandemic Crisis on Online Shopping, AYBU Business Journal, 1 **(2021)** 116-25.

[2] D., Goldberg, Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wesley, **1989.**

[3] O., Üstün, Genetik Algoritma Kullanılarak İleri Beslemeli Bir Sinir Ağında Etkinlik Fonksiyonlarının Belirlenmesi, Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi, 15 (**2009**) 395-403.

[4] K.S., Yıldırım, T.E., Kalaycı, A., Uğur, Optimizing Coverage in a K-Covered and Connected Sensor Network Using Genetic Algorithms, 9th WSEAS International Conference on Evolutıonary Computıng (EC'08), Bulgaria, May 2-4, 2008, Sofia, Bulgaria, **2008,** p. 42.

[5] A., Lambora, K. Gupta and K. Chopra, Genetic Algorithm - A Literature Review, 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), Faridabad, India, **2019**, p. 380-384.

[6] K., Kim and H., Ahn, A Recommender System using GA K-Means Clustering in An Online Shopping Market, Expert Systems with Applications, 34 **(2008)** 1200-1209.

[7] J., Bobadilla, F., Ortega, C., Hernando and A., Gutierrez, Recommender Systems Survey, Journal Of Knowledge-Based Systems, 46 (**2013**), 109-132.

[8] T., Dao, S., Jeong and H., Ahn, A Novel Recommendation Model of Location-Based Advertising: Context-Aware Collaborative Filtering Using GA Approach, In Expert Systems with Applications 39 (**2012)** 3731–3739.

[9] M., Salehi, M., Pourzaferani and S., Razavi, Hybrid Attribute-based Recommender System for Learning Material using Genetic Algorithm and a Multidimensional Information Model, Egyptian Informatics Journal, 14 (**2014)** 67–78.

[10] Maghsoudi vd., Representing the New Model for Improving K-Means Clustering Algorithm based on Genetic Algorithm, The Journal of Mathematics and Computer Science, 2 **(2011)** 329-336.

[11] G., Lv, C., Hu and S., Chen, Research on Recommender System Based on Ontology and Genetic Algorithm, Neurocomputing, 187 (**2016)** 97.

[12] B., Alhijawi, The Use of the Genetic Algorithms in the Recommender Systems, Master's Thesis, Faculty of Graduate Studies at the Hashemite University, Jordan, **2017.**

[13] H., Seyrek, Genetik Algoritma ile Ağırlıklandırılmış Hibrit Bir Film Öneri Sistemi, Yüksek Lisans Tezi, Harran Üniversitesi Fen Bilimleri Enstitüsü, Şanlıurfa, **2020.**

[14] T., Anwar, G., Srivastava and V., Uma, Implementing Recommendation System Using Collaborative Filtering and Singular Value Decomposition (SVD)++, International Journal of Information Technology & Decision Making (IJITDM), 20 **(2021)** 1075-1093.

[15] S., Sultan, A., Manal, A., Nashat, A Framework for Automatic Exam Generation based on k-means and Genetic Algorithm, International Journal of Computer Applications, 183 **(2021)** 18-23.

[16] N., Qomariyah, D., Kazakov, A Genetic-Based Pairwise trip Planner Recommender System, Journal of Big Data, 8 (**2021)** 1-23.

[17] Veronika Geltner, Recommendation Systems fort the Second time: You Will Learn How Their Architecture Is Built, https://blog.seznam.cz/en/2022/05/recommendation-systems-for-the-second-time-you-will-learn-how-their-architecture-is-built/ (Erişim tarihi: **13 Temmuz 2022**).

[18] Ş., Güler, Öneri Sistemleri ve E-ticarette Öneri Sistemlerinin Kullanımı, Yüksek Lisans Tezi, Sakarya Üniversitesi, Sakarya, **2019.**

[19] J., Breese, D., Heckerman and C., Kadie, Empirical Analysis of Predictive Algorithms for Collaborative Filtering, Technical Report MSR-TR-98-12, Microsoft Research, Redmond, WA. **1998.**

[20] F., Ricci, L., Rokach and B., Shapira, Recommender Systems Handbook, Springer, p. 1–34, **2015.**

[21] A., Ekman, Designing and Implementing A Recommender System for an E-Learning Platform, Master's thesis, Department of Computer Science at the Lund University , **2022**.

[22] B. M. Sarwar, G. Karypis, J. A. Konstan and J. Riedl, Analysis of Recommendation Algorithms for E-Commerce, In Proceedings of the ACM EC'00 Conference, Minneapolis, **2000**, p. 158-167.

[23] L., Ungar and D.P., Foster, Clustering Methods for Collaborative Filtering, In Workshop on Recommender Systems at the 15th National Conference on Artificial Intelligence, **1998**, p. 74.

[24] L., Baltrunas, T., Makcinskas and F., Ricci, Group Recommendation with Rank Aggregation and Collaborative Filtering, Proceedings of the 2010 ACM Conference on Recommender Systems, Belgium, October 30-November 1 2014, Leuven, Belgium, **2014**, p. 119-126.

[25] K., Kim, and H., Ahn, Using a Clustering Genetic Algorithm to Support Customer Segmentation for Personalized Recommender Systems, 13th International Conference on AI, Simulation, and Planning in High Autonomy Systems, Korea, October 2004, Jeju Island, Korea, **2004**, p. 4-7.

[26] S., Zhang, Internet of Things Services Based on Genetic K-Means Clustering Algorithm, 4th International Conference on Big Data Analytics for Cyber-Physical System in Smart City 2, Singapore, **2023**, p. 168**.**

[27] F. Zhang and H. Chang, A Collaborative Filtering Algorithm Employing Genetic Clustering to Ameliorate the Scalability Issue, Proceeding of the IEEE Int. Conf. on eBusiness Engineering (ICEBE), China, 24-26 October 2006, Shanghai, China, **2006**, p. 331-338.

[28] Al-Shamri, M. and Bharadwaj, K.: Fuzzy-Genetic Approach to Recommender Systems based on a Novel Hybrid User Model. Journal of Expert Systems with Applications, 35 (**2008)** 1386-1399.

[29] L., Gao and C. Li, Hybrid Personalizad Recommended Model Based On Genetic Algorithm, Proceeding of the 4 th Int. Conf. Wireless Communication Network Mobile Computing, IEEE**; 2008.**

[30] Ho, Y., Fong, S. and Yan, Z., A Hybrid Ga-Based Collaborative Filtering Model For Online Recommenders, Proceeding of the Int. Conf. e-Business, (**2007)** 200-203.

[31] B., Alhijawi and Y., Kilani, Using Genetic Algorithms for Measuring the Similarity Values between Users in Collaborative Filtering Recommender Systems. Proceeding of the 15th IEEE/ACIS International Conference on Computer and Information Science, Japan, 26–29 June 2016, Okayama, Japan, **2016,** p. 1–6.

[32] J., Bobadilla, F., Ortega, A., Hernando, and J. Alcala, Improving Collaborative Filtering Recommender System Results and Performance Using Genetic Algorithms, Journal of Knowledge-Based Systems, 24 (**2011)** 1310-1316.

[33] M., Kalz, H., Drachsler, J., Van Bruggen, H., Hummel, R., Koper, Wayfinding Servicesfor Open Educational Practices, International Journal of Emerging Technologies in Learning, 3 (**2008)** 24-28.

[34] H., Bulut, M., Milli, İşbirlikçi Filtreleme İçin Yeni Tahminleme Yöntemleri, Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi, 22 (**2016)** 123-128.

[35] M., Hahsler, A Framework For Developing Andtesting Recommendation Algorithms, R package version 0.2–6,Tech. Rep. 1–40, **2015.**

[36] R., Burke, Hybrid Recommender Systems: Survey and Experiments. User Model, User-Adapt. Interact, California State University, 12 (**2002)** 331-370.

[37] R., Lavetı, J., Ch, N., Supriya, N., Pal and C., Babu, A Hybrid Recommender System Using Weighted Ensemble Similarity Metrics And Digital Filters, 23rd International Conference on High Performance Computing Workshops (HiPCW) IEEE, 19-22 December, Hyderabad, **2016,** p.32-38.

[38] Y., Zhang, L., Wang, Some Challenges For Context-Aware Recommender Systems, 2010 5th International Conference on Computer Science and Education (ICCSE), **2010,** p. 362–365.

[39] MF., Kaya, M, Schoop, Analytical Comparison of Clustering Techniques for the Recognition of Communication Patterns, 31 (**2022)** 555–589.

[40] M., Demiralay, Hiyerarşik Kümeleme Metodları ile Veri Medenciliği Uygulamaları, Doktora Tezi, Marmara Universitesi, Türkiye, **2005.**

[41] S. Z., Sever, Yoğunluk Tabanlı Kümeleme Metodları Kullanılarak Paralel Veri Madenciliği Gerçekleştirilmesi, Yüksek Lisans Tezi, Maltepe Üniversitesi, Fen Bilimleri Enstitüsü, İstanbul, **2010.**

[42] G., Sariman, Veri Madenciliğinde Kümeleme Teknikleri Üzerine Bir Çalışma: K-Means ve K-Medoids Kümeleme Algoritmalarının Karşılaştırılması, Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 15 (**2011)** 192-202.

[43] L., Davis, Job Shop Scheduling with Genetic Algorithms, Journal of Software Engineering and Applications, 5 (**1985)** 136-140.

[44] Ö., Tabak, Genetik Algoritma ile Kapasiteli Servis Güzergahı Belirlenmesi ve Bir Uygulama, Yüksek Lisans Tezi, Anadolu Üniversitesi, Eskişehir, **2008.**

[45] D., Karaboğa, Yapay Zekâ Optimizasyon Algoritmaları, 7. Baskı, Nobel Yayıncılık, **2017.**

[46] T. Başkal, A., Özbek, Genetik Algoritmayla Bindirmeli Tip Kaynaklı Bağlantılarda Optimum Kaynak Kalınlığı Seçimi, Uluslararası Mühendislik Araştırma ve Geliştirme Dergisi, 2 (**2016)** 8.

[47] A., Kılıç, G., Arslan, Sıralı Küme Örneklemesi ile Kumaraswamy Dağılımı Parametrelerinin Tahmin Edilmesinde Genetik Algoritma Kullanılması, SDÜ Fen Bil Enst Der, 23 (**2019)** 367-373.

[48] C.R., Reeves, J.E., Rowe, Genetic Algorithms: Principles and Perspectives: A Guide to GA Theory, Kluwer Academic Publishers, Norwell, MA, 1 (**2002)** 63.

[49] C., Pizzuti, N., Procopio, A K-means Based Genetic Algorithm for Data Clustering. International Conference on European Transnational Education, **2016,** p. 211-222.

[50] S., Al-Janabi, A., Patel, I., AlShourbaji, Design And Evaluation of A Hybrid System for Detection and Prediction of Faults in Electrical Transformers, International Journal of Electrical Power & Energy Systems, 67 (**2015)** 324-335.

[51] Z. Dong, H. Jia, M. Liu, An Adaptive Multiobjective Genetic Algorithm with Fuzzy C-Means for Automatic Data Clustering, Mathematical Problems in Engineering, (**2018)** 1-13.

[52] A., Gad, 5 Genetic Algorithm Applications Using PyGAD, 2020. https://blog.paperspace.com/genetic-algorithm-applications-using-pygad/ (Erişim tarihi **8 Mart 2023).**

[53] R., Dash, Comparative Analysis of K-means and Genetic Algorithm based Data Clustering. International Journal of Advanced Computer and Mathematical Sciences, 3 (**2012)** 257-265.

[54] N, Suguna, K., Thanushkodi, An Improved k-Nearest Neighbor Classification Using Genetic Algorithm, International Journal of Computer Science, **2010.**

[55] PyGAD - Python Genetic Algorithm, https://pygad.readthedocs.io/en/latest/ (Erişim Tarihi **3 Ocak 2023**).

[56] Marketing Campaign Dataset, O. Parr-Rud. Business Analytics Using SAS Enterprise Guide and SAS Enterprise Miner. SAS Institute, https://www.kaggle.com/datasets/rodsaldanha/arketing-campaign (Erişim tarihi **27 Eylül 2022**).

EKLER

EK 1- Örnek Test Müşterileri Küme Ataması

Örnek 5 test için: Her test müşterisi için train veri üzerinden yapılan kümelemeye göre K-ortalamalar merkeze uzaklığı
Örnek Müşteri 1, En Yakın K-means Küme Merkezi:
Örnek Müşteri Koordinatları: [1.83339073 -0.86431448]
Uzaklık: 0.4990
K-means Küme Merkezi Koordinatları: [1.40185314 -0.61366958]
K-means Küme Etiketi: 0
Örnek Müşteri 2, En Yakın K-means Küme Merkezi:
Örnek Müşteri Koordinatları: [1.69603023 -1.55812805]
Uzaklık: 0.9892
K-means Küme Merkezi Koordinatları: [1.40185314 -0.61366958]
K-means Küme Etiketi: 0
Örnek Müşteri 3, En Yakın K-means Küme Merkezi:
Örnek Müşteri Koordinatları: [-2.00580944 -0.56771063]
Uzaklık: 0.1866
K-means Küme Merkezi Koordinatları: [-2.04995185 -0.38642985]
K-means Küme Etiketi: 1
Örnek Müşteri 4, En Yakın K-means Küme Merkezi:
Örnek Müşteri Koordinatları: [-2.89807922 0.18362493]
Uzaklık: 1.0219
K-means Küme Merkezi Koordinatları: [-2.04995185 -0.38642985]
K-means Küme Etiketi: 1
Örnek Müşteri 5, En Yakın K-means Küme Merkezi:
Örnek Müşteri Koordinatları: [-0.86419505 1.49901918]
Uzaklık: 0.8812
K-means Küme Merkezi Koordinatları: [0.01207276 1.59222221]
K-means Küme Etiketi: 2

Örnek 5 test için: Her test müşterisi için train veri üzerinden yapılan kümelemeye göre GA-KOK merkeze uzaklığı
Örnek Müşterisi 1 için En Yakın GA Küme Merkezi:
Örnek Müşteri Koordinatları: [1.83339073 -0.86431448]
Uzaklık: 0.3876
GA Küme Merkezi Koordinatları: [1.51094613 -0.64926247]
GA Küme Etiketi: 1
Örnek Müşterisi 2 için En Yakın GA Küme Merkezi:
Örnek Müşteri Koordinatları: [1.69603023 -1.55812805]
Uzaklık: 0.9275
GA Küme Merkezi Koordinatları: [1.51094613 -0.64926247]
GA Küme Etiketi: 1
Örnek Müşterisi 3 için En Yakın GA Küme Merkezi:
Örnek Müşteri Koordinatları: [-2.00580944 -0.56771063]
Uzaklık: 0.2297
GA Küme Merkezi Koordinatları: [-2.23500078 -0.55309697]
GA Küme Etiketi: 0
Örnek Müşterisi 4 için En Yakın GA Küme Merkezi:
Örnek Müşteri Koordinatları: [-2.89807922 0.18362493]
Uzaklık: 0.9912
GA Küme Merkezi Koordinatları: [-2.23500078 -0.55309697]
GA Küme Etiketi: 0
Örnek Müşterisi 5 için En Yakın GA Küme Merkezi:
Örnek Müşteri Koordinatları: [-0.86419505 1.49901918]
Uzaklık: 0.6988
GA Küme Merkezi Koordinatları: [-0.26617872 1.13756106]
GA Küme Etiketi: 2

EK 2 - Örnek Test Müşterilerin Koordinatları

K-ortalamalar: Her test müşterisi için eğitim veri üzerinden bulunan komşular		
Örnek Müşteri 1 için En Yakın Komşular:		
	col1	col2
1688	-1.878887	-0.633198
1840	-1.857879	-0.477233
1219	-1.932292	-0.587012
1634	-1.929412	-0.539653
913	-1.803340	-0.718619
Örnek Müşteri 2 için En Yakın Komşular:		
	col1	col2
1727	1.378278	0.174590
71	1.344190	0.194844
2209	1.286806	0.183221
948	1.266287	0.116029
870	1.485233	0.175481
Örnek Müşteri 3 için En Yakın Komşular:		
	col1	col2
891	-1.285641	0.621804
239	-1.315948	0.612360
1745	-1.242413	0.625902
736	-1.336477	0.653580
2234	-1.281706	0.701183
Örnek Müşteri 4 için En Yakın Komşular:		
	col1	col2
972	-2.186503	-1.192053
1077	-2.006568	-1.195914
12	-2.002811	-1.201571
1609	-2.193456	-1.141157
1215	-1.982853	-1.259612
Örnek Müşteri 5 için En Yakın Komşular:		
	col1	col2
1504	0.001635	2.784336
1513	-0.190844	2.644301
69	0.033651	2.434960
2091	-0.270359	2.681263
73	0.274722	2.470698

GA-KOK: Her test müşterisi için eğitim veri üzerinden bulunan komşular
For Test Customer1, GA Closest Neighbors:
0 1
525 -1.878887 -0.633198
576 -1.857879 -0.477233
1301 -1.932292 -0.587012
1490 -1.929412 -0.539653
111 -1.803340 -0.718619
For Test Customer2, GA Closest Neighbors:
0 1
790 1.378278 0.174590
272 1.344190 0.194844
1510 1.286806 0.183221
314 1.266287 0.116029
811 1.485233 0.175481
For Test Customer3, GA Closest Neighbors:
0 1
40 -1.285641 0.621804
634 -1.315948 0.612360
1158 -1.242413 0.625902
1109 -1.336477 0.653580
400 -1.281706 0.701183
For Test Customer4, GA Closest Neighbors:
0 1
1507 -2.186503 -1.192053
1227 -2.006568 -1.195914
841 -2.002811 -1.201571
114 -2.193456 -1.141157
1302 -1.982853 -1.259612
For Test Customer5, GA Closest Neighbors:
0 1
275 0.001635 2.784336
420 -0.190844 2.644301
1346 0.033651 2.434960
143 -0.270359 2.681263
135 0.274722 2.470698