

# Customer Personality Segmentation with Different Clustering Approaches and Profiling Analysis

Oğuz DENİZ

TOBB ETU

Department of Artificial Intelligence Engineering

Ankara, Turkey

o.deniz@etu.edu.tr

**Abstract**—Müşteri kişilik segmentasyonu, modern pazarlama stratejilerinde kritik bir rol oynamaktadır. Müşterilerin kişilik özelliklerinin karakteristiklerini analiz edebilen işletmeler ürünleri, hizmetleri ve pazarlama kampanyalarını bireysel tercihlere daha iyi uyacak şekilde özelleştirebilirler. Bu makalede, istatistiksel analiz ve makine öğrenmesi tekniklerinin birleşimi kullanılarak müşteri kişilik segmentasyonuna analitik bir sınıflandırma yaklaşımı getirilmeye çalışıldı. Yaklaşım, müşteri verilerinden tanımlayıcı-belirleyici özellikleri çıkarmayı ve farklı kişilik segmentlerini belirlemek için çeşitli kümeleme algoritmalarının uygulamalarını içermektedir.

**Index Terms**—Customer segmentation, personality traits, machine learning, marketing, personalized campaigns

## I. GİRİŞ

Günümüzün son derece rekabetçi iş dünyasında, müşterileri daha derinlemesine anlamak, etkili pazarlama stratejilerini kurgulamada ve müşteri memnuniyetini artırmada kilit bir faktör haline gelmiştir. Bireysel tabanlı yaklaşımlar, müşteri tercihlerini hedeflediği ve ürün önerilerini, hizmet planlamasını kişiye özel hale getirmeyi amaçladığı için işletmeler tarafından tercih edilmektedir. Bu kişiselleştirmenin temelini, müşterilerin kişilik özelliklerine ve önemli verilerine dayalı olarak sınıflandırılması oluşturmaktadır.

Organizasyonlar, kişilik segmentasyonu ile benzer tanımlayıcı özelliklere sahip belirli müşteri sınıflarını belirleyerek daha hedefli ve spesifik pazarlama kampanyaları geliştirebilirler; bu da müşteri etkileşimini, sadakati ve işletmenin başarısını geliri artırır.

Bu makalede, müşteri kişilik segmentasyonuna yönelik kişisel verilere dayalı istatistiksel analizleri, makine öğrenmesi teknikleriyle birleştiren yeni bir yaklaşım geliştirildi. Müşteri verilerinden kişilikle ilişkilendirilmiş özelliklerin kapsamlı bir şekilde çıkarılmasını ve k-means tabanlı kümeleme algoritmasının kullanılmasını içeren bir yöntem sunuldu. Algoritmanın başarımına dayalı çeşitli profillemeye teknikleriyle analizin, bütünsel bir çerçevede sunulması hedeflendi. Başarım metrikleri olarak silhouette score, davies bouldin score ve calinski harabasz score kullanıldı.

Makine öğrenmesi modeli geliştirilirken aşağıdaki aşamalarla ilerlendi:

- 1-Verilerin İncelenmesi ve Hazırlanması
- 2-PCA ve Standardizasyon
- 3-Elbow Method Uygulaması ve 10 Farklı Kümeleme Algoritmasının Denenmesi

- 4-Değerlendirme
- 5-Hiperparametre Ayarı
- 6-Denetimli Öğrenme Algoritmasının Sonuçlara Uygulanması
- 7-Segmentasyon ve Profillemeye Analizleri

## II. VERİ SETİ, VERİ ÖZELLİKLERİ, ÖZNİTELİKLER

Kümeleme modelinin kullanımından önce verisetinin ve özniteliklerin anlamları incelendi.

Veri Kaynağı: <https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis>

### A. Verilerin İncelenmesi ve Hazırlanması

Veri Seti Bilgisi:

- ID: Müşterinin benzersiz tanımlayıcısı
- Year Birth: Müşterinin doğum yılı
- Education: Müşterinin eğitim düzeyi
- Marital Status: Müşterinin medeni durumu
- Income: Müşterinin yıllık hanehalkı geliri
- Kidhome: Müşterinin evindeki çocuk sayısı
- Teenhome: Müşterinin evindeki ergen sayısı
- Dt Customer: Müşterinin şirkete kaydolma tarihi
- Recency: Müşterinin son satın alma işleminden bu yana geçen gün sayısı
- Complain: Müşterinin son 2 yılda şikayette bulunup bulunmadığı
- MntWines: Son 2 yılda şaraba harcanan tutar
- MntFruits: Son 2 yılda meyveye harcanan tutar
- MntMeatProducts: Son 2 yılda ete harcanan tutar
- MntFishProducts: Son 2 yılda balığa harcanan tutar
- MntSweetProducts: Son 2 yılda tatlılara harcanan tutar
- MntGoldProds: Son 2 yılda altına harcanan tutar
- NumDealsPurchases: İndirimli olarak yapılan satın alma işlemlerinin sayısı
- AcceptedCmp1: Müşterinin 1. kampanyada teklifi kabul ettiyse 1, etmediyse 0
- AcceptedCmp2: Müşterinin 2. kampanyada teklifi kabul ettiyse 1, etmediyse 0
- AcceptedCmp3: Müşterinin 3. kampanyada teklifi kabul ettiyse 1, etmediyse 0
- AcceptedCmp4: Müşterinin 4. kampanyada teklifi kabul ettiyse 1, etmediyse 0

- AcceptedCmp5: Müşterinin 5. kampanyada teklifi kabul ettiyse 1, etmediyse 0
- Response: Müşterinin son kampanyada teklifi kabul ettiyse 1, etmediyse 0
- NumWebPurchases: Şirketin web sitesi üzerinden yapılan satın alma işlemlerinin sayısı
- NumCatalogPurchases: Katalog kullanılarak yapılan satın alma işlemlerinin sayısı
- NumStorePurchases: Mağazalardan doğrudan yapılan satın alma işlemlerinin sayısı
- NumWebVisitsMonth: Son ayda şirketin web sitesine yapılan ziyaret sayısı

Bu aşamada veri setine genel bir inceleme yapıldı. Missing valuelar tespit edildi ve drop edildi. Dt Customer verisi formatlandı, Enrollment Time olarak müşterinin şirkete kaç yıldır üye olduğu şeklinde değiştirildi. Kategorik öznitelikler tespit edildi. Education özniteliği OrdinalEncoding stratejisi ile sıralı olarak encode edildi. Marital status özniteliğinin değerleri kişinin bekar veya birlikte olma durumuna göre Has partner özniteliğine binary olarak çevrildi. Year Birth özniteliğinden Age özniteliğine çevrim yapıldı. Diğer özniteliklerin dağılımlarının görselleştirilmesi gerçekleştirildi. Şikayette bulunmuş müşteriler arasında ortak noktalar arandı. Genelinin orta gelir düzeyli, 70li yaşlarda, üniversite mezunu olduğu tespit edildi. Korelasyon matrisi oluşturuldu.

KidHome ve TeenHome özniteliklerinin değerleri toplanarak toplam çocuk sayısını belirten Total Children özniteliği oluşturuldu. MntWines, MntFruits, MntMeatProducts, MntFishProducts, MntSweetProducts, MntGoldProds özniteliklerinin değerleri toplanarak toplam harcanan miktarı belirten Total Spent özniteliği oluşturuldu. NumWebPurchases, NumCatalogPurchases, NumStorePurchases, NumDealsPurchases özniteliklerinin değerleri toplanarak toplam satın alma sayısını temsil eden Num Total Purchases özniteliği oluşturuldu. Kişinin Has Partner durumuyla Total Children özniteliğinin değerleri toplanarak toplam aile nüfusunu belirten Family Size özniteliği oluşturuldu. Artık kullanımına gerek kalmıyan YearBirth, ID, Enrollment Start, Z CostContact, Z Revenue öznitelikleri veri setinden çıkarıldı.

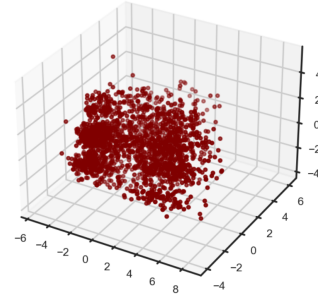
Ardından kümeleme algoritmasının performansını artırması adına Income ve Age özniteliklerinde outlier tespitleri yapıldı ve eliminasyonları gerçekleştirildi.

### B. PCA ve Standardizasyon

Kümeleme algoritmasının performansını artırmak ve outliera yakın verilerin etkisini düşürebilmek için StandardScaler yardımıyla tüm veri seti standardize edildi.

Ardından kümeleme algoritmalarının sonuçlarının görselleştirilebilmesi, birbirine yakın verilerin ayırıcı özelliklerinin daha iyi temsil edilebilmesi ve algoritmaların daha kısa zamanda yakınsayabilmeleri için PCA (Principal Component Analysis) uygulanarak veri seti 3 boyutlu uzaya indirildi.

A 3D Projection Of Data In The Reduced Dimension

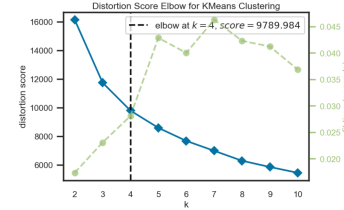


PCA sonucu veri setinin 3 boyutlu uzayda görünümü

## III. KULLANILAN MODELLER

### A. Elbow Method Uygulaması ve 10 Farklı Kümeleme Algoritmasının Denenmesi

Kümeleme algoritmasının en önemli parametresi olan küme sayısının belirlenmesi için Elbow Method yaklaşımı kullanıldı. Optimum küme sayısı 4 olarak tespit edildi.



Elbow Method İllüstrasyonu

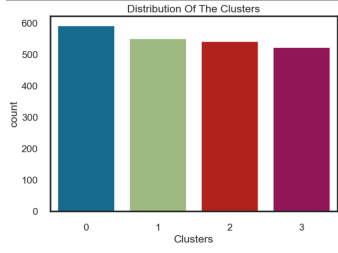
Ardından sırasıyla 10 adet kümeleme algoritması problem üzerinde denendi ve sonuçlarının dağılımı (hem nümerik hem 3 boyutlu uzayda) çıkarıldı.

Problemin doğası gereği hiyerarşik kümeleme algoritmalarının da partitional kümeleme algoritmalarının da iyi çalışabileceği göz önüne alınarak sırasıyla aşağıdaki kümeleme algoritmaları uygulandı:

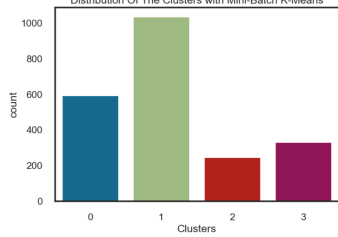
- 1-Klasik K-Means Algoritması
- 2-Agglomerative Clustering Algoritması
- 3-Affinity Propagation Algoritması
- 4-BIRCH Algoritması
- 5-DBSCAN Algoritması
- 6-Mini-Batch K-Means Algoritması
- 7-Mean Shift algoritması
- 8-OPTICS Algoritması
- 9-GaussianMixture Algoritması
- 10-K-Means++ Algoritması

### B. Değerlendirme

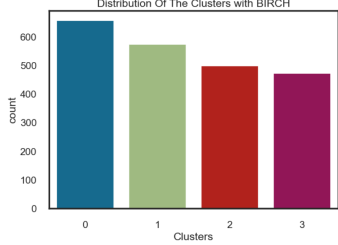
Uygulamalar sonucundaki dağılımlara bakıldığında en dengeli (balanced) dağılımın Klasik K-Means, K-Means++, AC, BIRCH, Mini-Batch K-Means, Gaussian Mixture algoritmalarında gerçekleştiği tespit edildi.



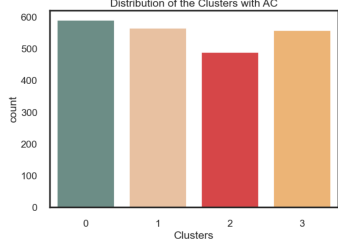
KMeans++ algoritmasının sonuçlarının dağılımı



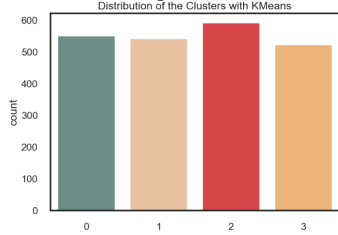
Mini Batch KMeans algoritmasının sonuçlarının dağılımı



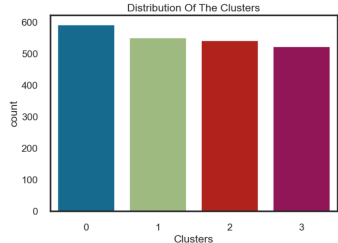
BIRCH algoritmasının sonuçlarının dağılımı



Agglomerative Clustering algoritmasının sonuçlarının dağılımı

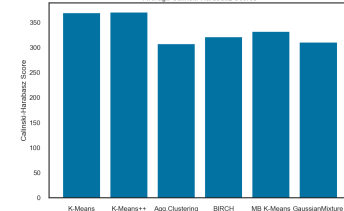


Klasik KMeans algoritmasının sonuçlarının dağılımı

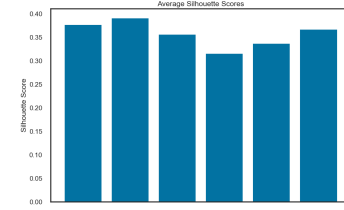


KMeans++ algoritmasının sonuçlarının dağılımı  
Ardından bu algoritmaların her biri için 5 katlı Çapraz

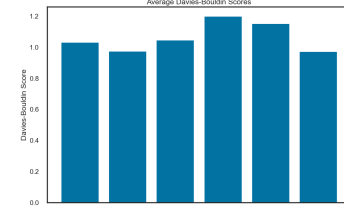
Doğrulama tarzı bir yaklaşım ile veri seti eşit parçalara bölünerek 5 farklı run gerçekleştirildi. Her run için silhouette score, davies bouldin score ve calinski harabasz score değerleri ayrı ayrı tutuldu. Sonrasında karşılaştırma için ortalama değerleri kullanıldı ve görselleştirilmeleri yapıldı.



Calinski-Harabasz Metriği özelinde algoritmaların performansı



Silhouette Score Metriği özelinde algoritmaların performansı



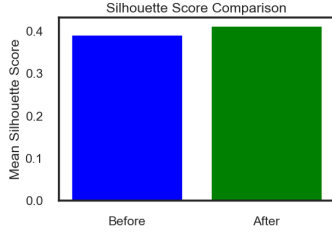
Davies-Bouldin Metriği özelinde algoritmaların performansı

### C. Hiperparametre Ayarı

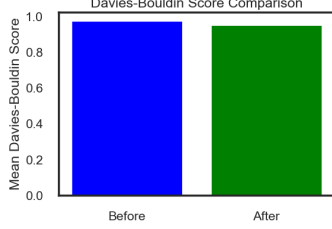
Bu aşamada ilgili 3 metrik özelinde en iyi performans gösteren algoritma seçildi (Silhouette Score metriğinde en yüksek, Davies-Bouldin metriğinde en düşük, Calinski-Harabasz metriğinde en yüksek) ve hiperparametre ayarı yapıldı.

Hiperparametre ayarı için GridSearch, Genetic Algorithm ve Random Search araçları kullanıldı. İlgili üç araç için de optimum küme sayısı 3 olarak sonuçlandı.

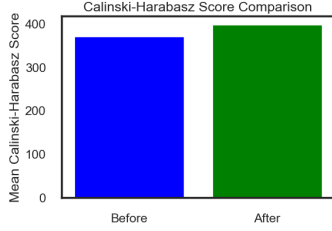
Bu aşamada Elbow Method ile elde edilen 4 küme değeri için, başarımları metrikleri açısından 3 kümeye göre daha dezavantajlı olduğu, buna karşın 3 kümeye karşı daha fazla bilgi içerebildiği gözlemlendi. Hiperparametre ayarından elde edilen 3 küme değeri için ise 4 kümeye kıyasla daha az bilgi içerebileceği ancak içerdiği bilgilerin daha genelleyici olabileceği ihtimali görüldü. Bu tradeofflar gözlemlenerek olabildiğince derin bir analiz yapabilmek adına hem 3 küme, hem 4 küme için segmentasyon uygulandı.



Hiperparametre ayarından sonraki Silhouette Score değışimi



Hiperparametre ayarından sonraki Davies-Bouldin Score değışimi



Hiperparametre ayarından sonraki Calinski-Harabasz Score değışimi

#### D. Denetimli Öğrenme Algoritmasının Sonuçlara Uygulanması

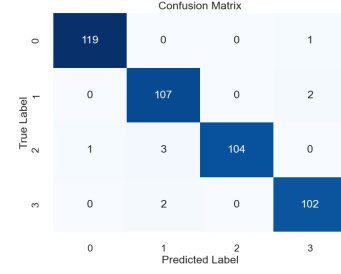
Bu aşamada halihazırda kümeleme işlemi tamamlanmış veriye bir Denetimli Öğrenme algoritması olan Multi Layer Perceptron algoritmasının uygulanmasının, ayırımın rahat yapılabilirliğinin bir ölçümü olarak kullanılması yer alır.

Hiperparametre ayarımı bölümünde belirtildiği gibi MLP algoritması, hem 3 sınıflı hem de 4 sınıflı kümelenebilir uygulandı. Sonuçların karmaşıklık matrisleri çıkarıldı ve Accuracy değerleri hesaplandı.

Model	Accuracy (%)
3 sınıf için MLP	0.9886621236801147
4 sınıf için MLP	0.9795918464660645



3 Sınıflı MLP için Karmaşıklık Matrisi



4 Sınıflı MLP için Karmaşıklık Matrisi

#### E. Segmentasyon ve Profillem Analizleri

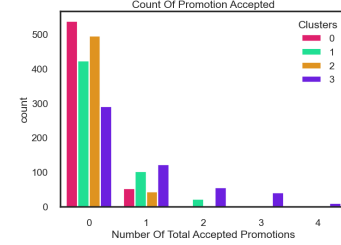
Segmentasyon ve profillem analizleri, Hiperparametre Ayar bölümünde bahsedildiği üzere, hem bilgi kaybetmemek adına 4 sınıflı duruma, hem de daha genel bir durumu görebilmek adına 3 sınıflı duruma uygulandı.

Analizler sonucu elde edilen bilgiler sırasıyla şu şekildedir:

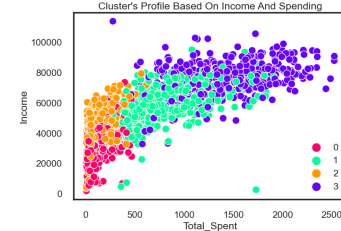
1) 4 Sınıflı Kümeleme için Profillem: 4 Sınıflı kümeleme sonuçları analiz edilirken önce sınıfların dağılımları incelendi. Herhangi bir imbalance sorunu tespit edilmedi. Sınıfların Toplam Harcama seviyelerine karşılık Gelir seviyeleri incelendi, Campaign kabul etme durumları görselleştirildi ve Toplam harcama aralıkları ile birlikte sonuçlar gruplara göre sınıflandırıldı.



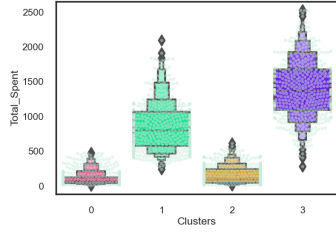
4 sınıflı kümeleme için sınıfların dağılımı



Campaign kabul etme durumları

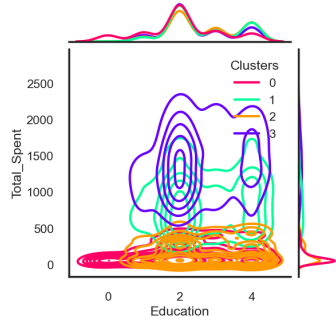


Toplam Harcama seviyesine karşılık Gelir Seviyeleri

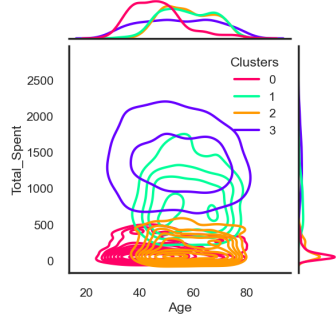


Toplam Harcama Aralıkları

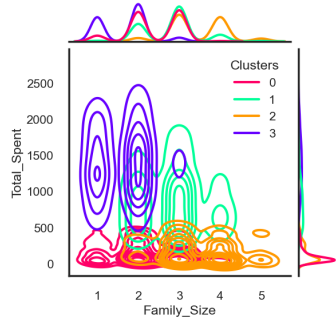
Ardından müşterinin karakterini ve segmentini tanımlayabilecek belirleyici öznitelikler Kidhome, Teenhome, EnrollmentTime, Age, TotalChildren, FamilySize, Education, HasPartner olarak belirlendi ve hepsine karşılık Toplam Harcama miktarları incelendi. Her birinin ayrı ayrı grafikleri oluşturularak olabildiğince detaylı tanımlamalar yakalanmaya çalışıldı. Bunlardan bazıları aşağıda gösterilmiştir.



Eğitim Seviyesine bağlı dağılım



Yaşa bağlı dağılım



Aile Nüfusuna göre dağılım

4 sınıflı kümelemenin sonuçları son olarak şu şekilde oluştu:

#### Grup 0:

- 1- Ya hiç çocuğu yok ya da 1 tane var, aile nüfusu genellikle 2-3 kişi.
- 2- Evdeki teen sayısı en fazla 1, genelde 0.
- 3- Şirkete ortalama 10+ yıldır üyeler.
- 4- Genel yaş aralıkları 25-50.
- 5- Eğitim seviyeleri yaygın olarak her kesimden
- 6- Partner bilgisi yaygın olarak tek tip değil, single da together da mevcut.
- 7- Gelir seviyeleri 5000-40000 arası, harcama aralıkları 0-500.
- 8- Campaign acceptance rate'i çok düşük

#### Grup 1:

- 1- Geneli çocuk sahibi değil. aile nüfusu genellikle 2-3 kişi.
- 2- Evdeki teen sayısı genelde 1.
- 3- Şirkete ortalama 10 yıldır üyeler.
- 4- Genel yaş aralıkları 40-80.
- 5- Eğitim seviyeleri yaygın Graduation ve Phd, özetle geneli üniversite mezunu.
- 6- Partner bilgisi yaygın olarak tek tip değil, single da together da mevcut.
- 7- Gelir seviyeleri 42000-78000 arası, harcama aralıkları 500-1800.
- 8- Campaign acceptance rate'i orta-düşük.

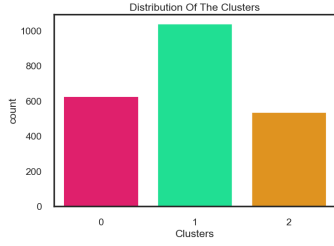
#### Grup 2:

- 1- Geneli 1 çocuk sahibi, aile nüfusu genellikle 3-4 kişi.
- 2- Evdeki teen sayısı en fazla 2, genelde 1.
- 3- Şirkete ortalama 9+ yıldır üyeler.
- 4- Genel yaş aralıkları 40-80.
- 5- Eğitim seviyeleri yaygın olarak üniversite mezunu.
- 6- Partner bilgisi single olanlar olsa da geneli together.
- 7- Gelir seviyeleri 40000-75000 arası, harcama aralıkları 0-500.
- 8- Campaign acceptance rate'i çok düşük.

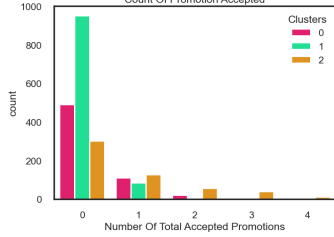
#### Grup 3:

- 1- Çok büyük çoğunluğunun çocuğu yok, aile nüfusu genellikle 1-2 kişi.
- 2- Evdeki teen sayısı genelde 0.
- 3- Şirkete ortalama 10 yıldır üyeler.
- 4- Genel yaş aralıkları 25-80.
- 5- Eğitim seviyeleri yaygın olarak Graduation, üniversite mezunu.
- 6- Partner bilgisi single olanlar da together olanlar da var.
- 7- Gelir seviyeleri 65000-90000 arası, harcama aralıkları 550-2500.
- 8- Campaign acceptance rate'i yüksek

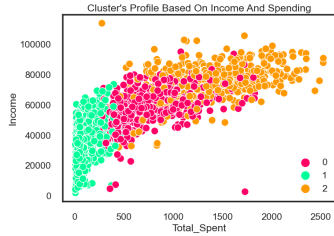
2) 3 Sınıflı Kümeleme için Profilleme: 3 Sınıflı kümeleme sonuçları analiz edilirken yine sınıfların dağılımları incelendi. 2.sınıfın baskınlığıyla imbalance tespit edildi. Aynı şekilde sınıfların Toplam Harcama seviyelerine karşılık Gelir seviyeleri incelendi, Campaign kabul etme durumları görselleştirildi ve Toplam harcama aralıkları ile birlikte sonuçlar gruplara göre sınıflandırıldı.



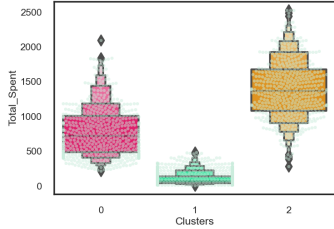
3 sınıflı kümeleme için sınıfların dağılımı



Campaign kabul etme durumları

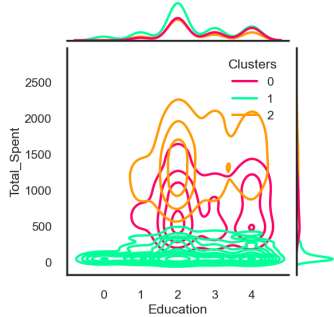


Toplam Harcama seviyesine karşılık Gelir Seviyeleri

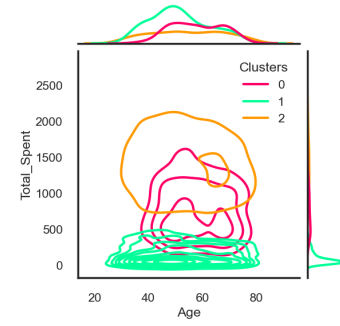


Toplam Harcama

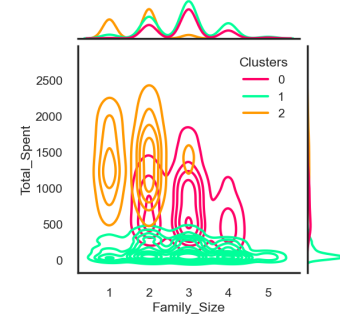
Tekrardan müşterinin karakterini ve segmentini tanımlayabilecek belirleyici öznitelikler olan Kidhome, Teenhome, EnrollmentTime, Age, TotalChildren, FamilySize, Education, HasPartner değerleri alındı ve hepsine karşılık Toplam Harcama miktarları incelendi. Bunlardan bazıları aşağıda gösterilmiştir.



Eğitim Seviyesine bağlı dağılım



Yaşa bağlı dağılım



Aile Nüfusuna göre dağılım

3 sınıflı kümelemenin sonuçları ise son olarak şu şekilde oluştu:

#### Grup 0:

- 1- Ya hiç çocukları yok ya da en fazla 1 tane var, aile nüfusu genellikle 3 kişi.
- 2- Evdeki teen sayısı en fazla 2, genelde 1.
- 3- Şirkete ortalama 10+ yıldır üyeler.
- 4- Genel yaş aralıkları 38-80.
- 5- Eğitim seviyeleri yaygın olarak üniversite mezunu.
- 6- Partner bilgisi single olanlar olsa da geneli together.
- 7- Gelir seviyeleri 35000-80000 arası, harcama aralıkları 400-1700.
- 8- Campaign acceptance rate'i düşük.

#### Grup 1:

- 1- Çocuk sayısı en az 0 en fazla 2, geneli 1 çocuk sahibi, aile nüfusu genellikle 2-5 kişi.
- 2- Evdeki teen sayısı genelde 0-1.
- 3- Şirkete ortalama 10 yıldır üyeler.
- 4- Genel yaş aralıkları 25-80.
- 5- Eğitim seviyeleri yaygın olarak her seviyeden.
- 6- Partner bilgisi single-together eşit dağılmış.
- 7- Gelir seviyeleri 0-65000 arası, harcama aralıkları 0-500.
- 8- Campaign acceptance rate'i düşük.

#### Grup 2:

- 1- Çok büyük bir kısmı çocuk sahibi değil, aile nüfusu genellikle 1-2 kişi.
- 2- Evdeki teen sayısı en fazla 1, genelde 0.
- 3- Şirkete ortalama 10 yıldır üyeler.
- 4- Genel yaş aralıkları 30-80.
- 5- Eğitim seviyeleri yaygın olarak üniversite mezunu.

- 6- Partner bilgisi single-together eşit dağılmış.
- 7- Gelir seviyeleri 60000-90000 arası, harcama aralıkları 500-2500.
- 8- Campaign acceptance rate'ı yüksek.

#### IV. SONUÇLAR

Bu çalışma, müşteri kişilik segmentasyonunun modern pazarlama stratejilerindeki rolünü ve etkisini incelemeyi amaçlamıştır. Elde edilen sonuçlar ve yapılan analizler, işletmelerin kişilik özelliklerine dayalı müşteri segmentasyonunun müşteri katılımını artırma ve kişiselleştirilmiş hizmet sunma konularında önemli bir araç olduğunu vurgulamaktadır.

##### A. Yapılanlar ve Kazançlar

Kişilik özelliklerine dayalı segmentasyon, müşterilerin tercihlerini daha iyi anlamak ve kişiselleştirilmiş pazarlama stratejileri geliştirmek için kullanılabilir. Elde edilen sonuçlar, bu yaklaşımın müşteri etkileşimini artırabileceğini ve müşteri memnuniyetini yükseltebileceğini yönünde kullanışlı olabileceğini göstermektedir.

##### B. Eksikler ve Gelecek Çalışmalar

Bu çalışmada ele alınmayan bazı noktalar da bulunmaktadır. Örneğin, müşteri segmentlerinin uzun vadeli davranışlarını izlemek ve ölçmek için daha fazla çalışma gerekmektedir. Ayrıca, farklı endüstrilere veya sektörlere özgü uygulamaların incelendiği daha geniş kapsamlı araştırmalar da yapılabilir. Ek olarak elde edilen sonuçlar, psikometrik bir yaklaşımla birleştirilerek ve nöropazarlama alanındaki teorilerden yararlanılarak segmentasyon performansı artırılabilir.

##### C. Gelecek Çalışmalar

Bu çalışmanın sonuçları, gelecekteki araştırmalar için çeşitli fırsatlar sunmaktadır. Özellikle, daha fazla veri toplama, farklı algoritmaların karşılaştırılması ve farklı kişilik özelliklerinin etkisinin daha derinlemesine analizi gibi konuları ele alan çalışmalar yapılabilir.

Sonuç olarak, bu çalışma, müşteri kişilik segmentasyonunun pazarlama stratejilerini zenginleştirebileceğini ve ayrıntılı analizler sunarak müşteri katılımını artırabileceğini göstermiştir. Gelecekteki çalışmalar, bu yaklaşımın daha da geliştirilmesi ve özelleştirilmiş pazarlama çözümlerinin tasarlanması yönünde ilerleyebilir.

#### V. REFERENCES

##### REFERENCES

- [1] Customer Data for Personality Segmentation, Kaggle Dataset, Available: <https://www.kaggle.com/dataset-link>, 2023.
- [2] S. P. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.
- [3] M. Abadi *et al.*, "TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems," arXiv preprint arXiv:1603.04467, 2015.
- [4] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [5] R. C. Blattberg and J. Deighton, "Database Marketing: Analyzing and Managing Customers," *Springer Science & Business Media*, 2001.
- [6] I. T. Jolliffe, "Principal Component Analysis," *Springer*, 2002.
- [7] Scikit-learn Contributors, *GridSearchCV: Exhaustive search over specified parameter values for an estimator*, Scikit-learn Documentation, Available: [https://scikit-learn.org/stable/modules/generated/sklearn.model\\_selection.GridSearchCV.html](https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html)
- [8] P. J. Rousseeuw, "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.
- [9] D. L. Davies and D. W. Bouldin, "A Cluster Separation Measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224–227, 1979.
- [10] T. Calinski and J. Harabasz, "A Dendrite Method for Cluster Analysis," *Communications in Statistics*, vol. 3, no. 1, pp. 1–27, 1974.