****

**YILDIZ TEKNİK ÜNİVERSİTESİ**

**KİMYA-METALÜRJİ FAKÜLTESİ**

**MATEMATİK MÜHENDİSLİĞİ BÖLÜMÜ**

**VERİ MADENCİLİĞİNE GİRİŞ**

**ÖDEV RAPORU**

Öğretim Görevlisi: Dr.Öğr.Üyesi Nilgün GÜLER BAYAZIT

İlker Efil 22052610

Abdulkadir Sarıhan 19052095

Berkay Altınay 20052025

Grup Adı: GR\_PersonalityAnalysis\_01

|  |  |
| --- | --- |
| **İÇİNDEKİLER** **Sayfa** | |
| 1. GİRİŞ ...................................................................................................................................................................... | 3 |
| 2. VERİ SETİ | 4 |
| 2.1 Önitelikler | 4 |
| 2.2 Veriye Ön Bakış | 5 |
| 3. VERİ ÖN İŞLEME VE ÖZELLİK ÇIKARIMI  ................................................................................................................................ | 7 |
| 3.1 Değişkenlerin İncelenmesi | 8 |
| 3.2 Encoding İşlemi Ve Aykırı Değerler | 11 |
| 4. MODELLERİN KURULMASI | 13 |
| 4.1 Kmeans Modelinin Kurulması  .................................................................................................................................... | 13 |
| 4.2 Modelin Yorumlanması | 14 |
| 4.3 Kmedoids Modelinin Kurulması | 17 |
| 5. SONUÇ | 19 |
| KAYNAKÇA | 20 |

**1. GİRİŞ**

Müşteri Kişilik Analizi, bir şirketin ideal müşterilerinin ayrıntılı bir analizidir. Bir işletmenin müşterilerini daha iyi anlamasına yardımcı olur ve farklı müşteri türlerinin özel ihtiyaçlarına, davranışlarına ve endişelerine göre ürünleri değiştirmelerini kolaylaştırır.

Müşteri kişilik analizi, bir işletmenin ürününü farklı müşteri segmentlerinden hedef müşterilerine göre değiştirmesine yardımcı olur. Örneğin, bir şirket, yeni bir ürünü şirketin veri tabanındaki her müşteriye pazarlamak için para harcamak yerine, hangi müşteri segmentinin ürünü satın alma olasılığının yüksek olduğunu analiz edebilir ve ardından ürünü yalnızca o belirli segmentte pazarlayabilir.

**2. VERİ SETİ**

Bu projede kullanılacak veri seti veri bilimi platformu Kaggle’dan alınmış Customer Personality Analysis adlı veri setidir.

Seçilen veri seti, bir şirketin satış yaptığı müşterileri ile ilgili etiketlerden oluşmaktadır ve 2240 veri noktası içerir. Bu etiketler dört kategoriden oluşmaktadır. İlk kategori; müşterilerin eğitim durumu, kimlik numaraları, medeni durumları vb. etiketlerden oluşmaktadır. İkinci kategori, müşterilerin belli ürünlere son iki yılda ne kadar para harcadığının bilgisini içerir. Üçüncü kategori, müşterilerin kampanyalı kaç ürün aldığının ve ürünü kaçıncı kampanyada aldığının bilgisini içerir. Son kategori ise satın alma işleminin hangi platform aracılığı ile gerçekleştiğinin bilgisini içerir. Veri toplamda 2040 veri noktası için 29 sütun içermektedir. Müşteri segmentlerini özetlemek için kümeleme yapılması gerekir.

**2.1 Öznitelikler**

**Müşteriler**

* ID: Müşterinin benzersiz tanımlayıcısı
* Year\_Birth: Müşterinin doğum yılı
* Education: Müşterinin eğitim düzeyi
* Marital\_Status: Müşterinin medeni durumu
* Income: Müşterinin yıllık hane geliri
* Kidhome: 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 şikâyette bulunduysa 1, aksi durumda 0

**Ürünler**

* MntWines: Son 2 yılda şaraba harcanan miktar
* MntFruits: Son 2 yılda meyvelere 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 miktar

**Kampanyalar**

* NumDealsPurchases: İndirimli yapılan satın alma sayısı
* AcceptedCmp1: Müşteri 1. kampanyadaki teklifi kabul ederse 1, aksi takdirde 0
* AcceptedCmp2: Müşteri 2. kampanyadaki teklifi kabul ederse 1, aksi halde 0
* AcceptedCmp3: Müşteri 3. kampanyadaki teklifi kabul ederse 1, aksi takdirde 0
* AcceptedCmp4: Müşteri 4. kampanyadaki teklifi kabul ederse 1, aksi takdirde 0
* AcceptedCmp5: Müşteri 5. kampanyadaki teklifi kabul ederse 1, aksi takdirde 0
* Response: Müşteri son kampanyadaki teklifi kabul ettiyse 1, aksi halde 0

**Yer**

* NumWebPurchases: Şirketin web sitesi aracılığıyla yapılan satın alma sayısı
* NumCatalogPurchases: Katalog kullanılarak yapılan satın almaların sayısı
* NumStorePurchases: Doğrudan mağazalarda yapılan satın almaların sayısı
* NumWebVisitsMonth: Geçtiğimiz ay şirketin web sitesine yapılan ziyaretlerin sayısı

**2.2 Veriye Ön Bakış**

Değişkenlerin hangi türde olduğuna bakıldı.

**metin, ekran görüntüsü, çizgi, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu**

**metin, menü, ekran görüntüsü, doküman, belge içeren bir resim

Açıklama otomatik olarak oluşturuldu**

Hangi değişkenin ne kadar eşsiz değeri olduğuna bakıldı.

**metin, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu**

Veride eksik değer olup olmadığını kontrol edilmişir.

**metin, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu**

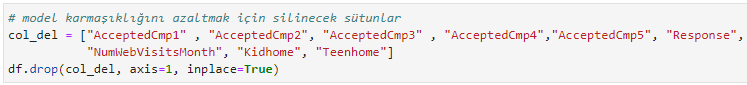
**3. VERİ ÖN İŞLEME VE ÖZELLİK ÇIKARIMI**

**metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu**Eksik veriler silindi, müşterilerin kayıt yılı zaman cinsinden yeniden tanımlandı. En sonda ise kayıt yılının maksimum değeri bulundu, bu değer müşterinin yaşını bulmak için kullanıldı.

metin, ekran görüntüsü, yazı tipi, doküman, belge içeren bir resim

Açıklama otomatik olarak oluşturulduKarmaşıklığı azaltmak için örneğin ayrı ayrı tutulan “Kidhome” ve “Teenhome” değerleri “Children” etiketinde birleştirildi. Bu şekilde birlikte bir anlam ifade eden değişkenler aynı etiket altında toplandı. Bu sayede yorumlamaya daha açık değerler elde edildi.

Yeni etkiketler ürettiğmiz için genel model karmaşıklığını azaltmak için silinecek sütunlar ise şunlardır:

metin, yazı tipi, ekran görüntüsü, çizgi içeren bir resim

Açıklama otomatik olarak oluşturulduİşe yaramayacak sütunlar da silindi.

**3.1 Değişkenlerin İncelenmesi**

Veri setindeki kategorik ve nümerik değerleri ayrı ayrı incelemek için aşağıdaki yazdığımız kod parçasıyla ayrıldı.

**metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu**

metin, ekran görüntüsü, yazı tipi, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldu

Kategorik değişkenlerin dağılımları;

metin, ekran görüntüsü, dikdörtgen, ekran, görüntüleme içeren bir resim

Açıklama otomatik olarak oluşturuldumetin, yazı tipi, ekran görüntüsü, beyaz içeren bir resim

Açıklama otomatik olarak oluşturuldumetin, ekran görüntüsü, dikdörtgen, ekran, görüntüleme içeren bir resim

Açıklama otomatik olarak oluşturuldumetin, yazı tipi, ekran görüntüsü, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldumetin, ekran görüntüsü, diyagram, çizgi içeren bir resim

Açıklama otomatik olarak oluşturulduNümerik değişkenlerin dağılımı;

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldumetin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldumetin, ekran görüntüsü, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, diyagram içeren bir resim

Açıklama otomatik olarak oluşturuldudiyagram, ekran görüntüsü, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin, ekran görüntüsü, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturulduekran görüntüsü, metin, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, diyagram içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin, ekran görüntüsü, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldumetin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldumetin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldumetin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldudiyagram, ekran görüntüsü, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldumetin, ekran görüntüsü, diyagram, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Açıklama otomatik olarak oluşturuldudiyagram, metin, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldumetin, ekran görüntüsü, diyagram, dikdörtgen içeren bir resim

Açıklama otomatik olarak oluşturuldu

**3.2 Encoding İşlemi Ve Aykırı Değerler**

Kategorik olarak ayırdığımız değişkenler label encoder ile 0 ve 1 olarak dönüştürüldü.

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldumetin, yazı tipi, çizgi, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu

Boxplot ile aykırı değer gözlemi yapıldı.

metin, ekran görüntüsü, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

Income değişkeninde aykırı değerler görülmektedir.

Aykırı değerleri interquartile istatistik yöntemi ile baskılandı.

metin, ekran görüntüsü, yazı tipi, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin, ekran görüntüsü, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

Yukarıdaki görselden anlaşılacağı üzere “Income” etiketinin maksimum değeri değişmiştir.

Bu bağlamda veri setimiz son halini aldı. Bundan sonraki aşamada planlanan modellerin kurulum aşamasına geçildi.

**4. MODELLERİN KURULMASI**

Veri setimizin özelliklerine baktığımızda kümeleme (clustering) yapılması gerekmektedir. Bunun için de belli başlı kümeleme algoritmaları mevcuttur. Kmeans en sık kullanılan kümeleme algoritmalarındandır. Uygulanması kolaydır. Büyük ölçekli verileri hızlı ve etkin şekilde kümeleyebilir. Bir diğer kullancağımız kümeleme algoritmamız da Kmedoids’dir.

**4.1 Kmeans Modelinin Kurulması**

**metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu**

Kmeans’ in modelinin 3 kümeye ayrılması uygun görüldü. Kümeleme dağılımı grafikteki gibidir. 1. küme için 782, 2. küme için 738 ve 3. küme için 692 adet gözlem bulunmaktadır.

**metin, ekran görüntüsü, diyagram, dikdörtgen içeren bir resim

Açıklama otomatik olarak oluşturuldu**

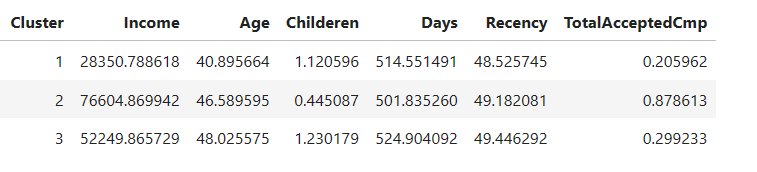
**4.2 Modelin Yorumlanması**

Kümelerin özniteliklerinin ortalamalarına bakıldı.

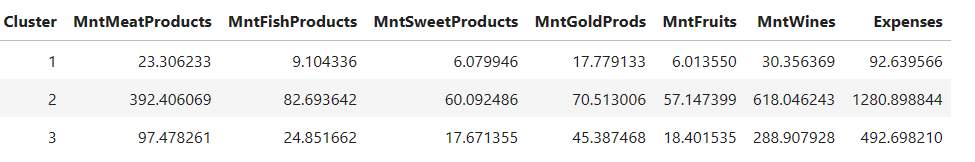
**metin, yazı tipi, çizgi, beyaz içeren bir resim

Açıklama otomatik olarak oluşturuldu**

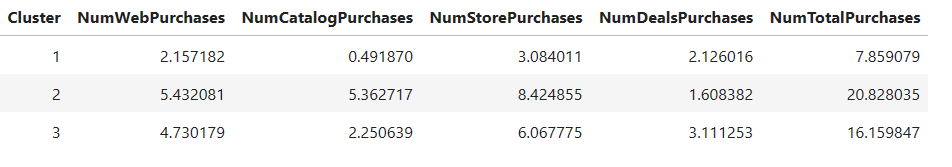
Kümelerin gelir, yaş, çocuk sayısı, müşterilik süreleri, son alışveriş zamanı ve kaçıncı indirimi kabul ettikleri değerlerinin ortalamaları aşağıdaki grafiktedir.

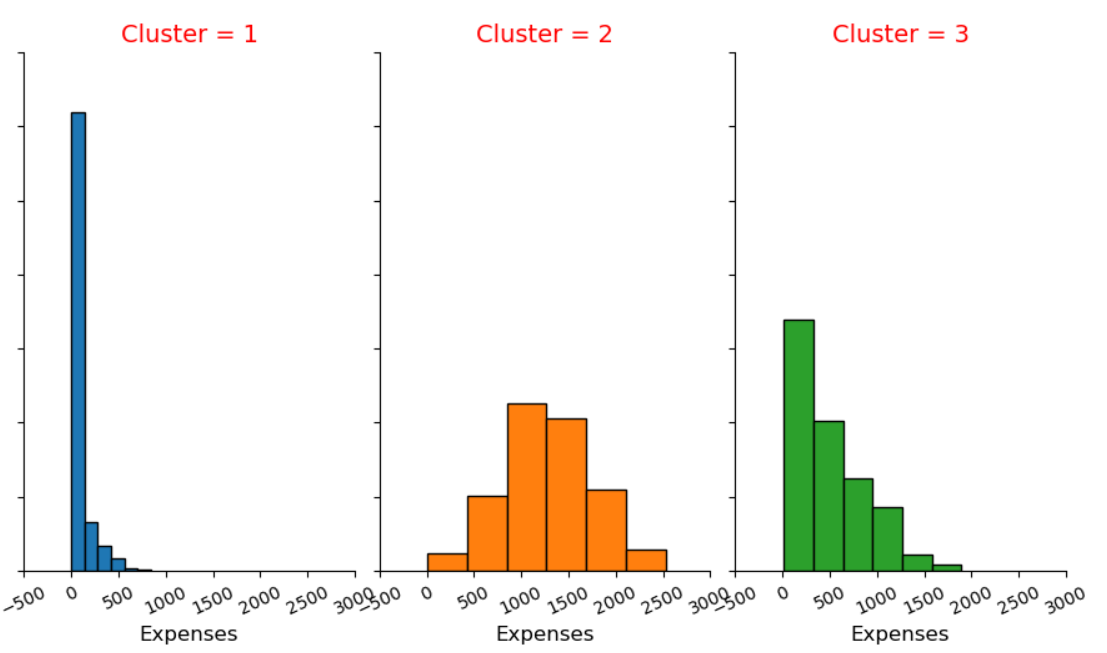


Kümelerin hangi ürüne ne kadar harcama yaptıklarının ve toplam harcamalarının ortalamaları aşağıdaki grafiktedir.



Kümelerin alışverişlerini yapma araçlarının, indirimli alışveriş sayısı değerlerinin ve bu değerlerin toplamının ortalamaları aşağıdaki grafiktedir.





Kümeler harcamalar bakımından yorumlandığında 2. kümenin en aktif olduğu, 3. kümenin orta aktif olduğu ve 1. kümenin az aktif olduğu görülmektedir.

Kümeler birbirlerine hesaplanan ortalama değerlere ve dağılımlarına göre kıyaslandığında aşağıdaki karakeristik özellikler ortaya çıkmaktadır:

* Genel olarak 2. kümenin çocuk sayısı diğer kümelere kıyasla daha azdır.
* Kümelerin müşterilik sürelerine ve son alışverişlerinden kaç gün geçtiğine bakıldığında aralarında önemli bir fark görülmemektedir.
* Eğitim açısından bakıldığında, 1. kümenin daha eğitimsiz olduğu ve diğer iki kümenin birbirine yakın olduğu görülmektedir.
* 1. küme daha genç bir yaş ortalamasına sahipken, 2. kümede yaşlı müşterilerin daha yoğunlukta olduğu görülmektedir. 3. kümenin müşterileri diğer iki kümeye göre daha yoğun bir orta yaş nüfusuna sahiptir.

metin, ekran görüntüsü, yazı tipi, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldu

* Kümelerin her bir ürüne harcamalarının tüm harcamalarına oranı hesaplandığında, 1. kümenin ortalama gelirlerinin düşük olmasına rağmen altına harcama yapma oranının diğer kümelerin altın harcama oranlarından fazla olduğu görülmektedir.Yine aynı grafiğe bakıldığında, her 3 küme için en çok harcama yapılan ürünün şarap ve et ürünleri olduğu ve 3. kümenin şarap harcaması oranının diğer kümelerden daha fazla olduğu görülmektedir.

metin, yazı tipi, çizgi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

* Alışveriş yapma araçlarının kullanılma sıklıklarının oranları hesaplandığında 1. ve 3. kümelerin indirimli alışveriş yapma oranları 2. kümeninkinden bir hayli yüksektir ve 2. küme neredeyse hiç indirimli alışveriş yapmamaktadır.
* 2. kümenin katalogdan alışveriş yapma oranı 1. kümeninkinden oldukça yüksek ve 3. kümeninkinin ise neredeyse 2 katıdır.
* İnternet ve mağazaların kullanım oranları her 3 küme için hemen hemen aynıdır.

**4.3 Kmedoids Modelinin Kurulması**

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin, ekran görüntüsü, dikdörtgen, diyagram içeren bir resim

Açıklama otomatik olarak oluşturulduKmedoids’in modelinin de 3 kümeye ayrılması uygun görüldü, method olarak ‘pam’ seçildi. Kümeleme dağılımı grafikteki gibidir. 1. küme için 734, 2. küme için 771 ve 3. küme için 707 adet gözlem bulunmaktadır.

diyagram, metin, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, çizgi içeren bir resim

Açıklama otomatik olarak oluşturulduKümeler harcamalar bakımından yorumlandığında kmedoids modelinde 1. kümenin en aktif olduğu, 3. kümenin orta aktif olduğu ve 2. kümenin az aktif olduğu görülmektedir.

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturulduBu bağlamda her iki modelin değişkenlerinin en aktif küme bazında ortalama farklarına bakıldığında aralarında önemli bir fark bullunmamaktadır.

Bu durum, diğer iki küme için de gözlenmiştir. Dolayısıyla her iki modelin de büyük oranda aynı şekilde sonuç verdiği görülmüştür.

Sonuç olarak, kmeans ve kmedoids modellerinin çıktılarına baktığımızda, kmedoids için kmeans’de yapılan çıkarımlar geçerlidir.

**5. SONUÇ**

Kullanılan algoritmalar karşılaştırıldığında, küme sayısının her iki algoritma için aynı seçilmesinden dolayı bu algoritmaların veriyi çok büyük oranda aynı şekilde kümelediği görülmüşür. Dolayısıyla böyle bir veri seti ve problem için her ikisinin de uygun olduğu sonucuna varılmıştır.

Kümeler incelendiğinde, hangi kümenin ne özellikleri taşıdığı anlaşılmamaktadır. Dolayısıyla görselleştirme, ortalama hesapları, oran hesapları veriden bilgi çıkarımı için büyük önem taşımıştır.

Kümeler incelendiğinde, tahmin edileceği üzere kümelerin gelir düzeyi arttıkça yaptıkları harcamaların miktarları da artmıştır. Aynı şekilde harcama seviyesi arttıkça yaş ortalamasının da arttığı görümüştür, fakat orta yaştaki grubun harcama düzeyi en yaşlı gruptan daha fazladır.

Önceden tahmin edilemeyecek bir çıkarım olarak ise, az aktif grupların harcamalarının büyük bir kısmını altın harcamalarının oluşturması olduğu gözlemlenmiştir. Benzer şekilde her 3 küme için şarap ve et ürünlerinin en çok harcama yapılan ürünler olduğu görülmektedir.

Daha uzun süredir müşteri olanların daha fazla harcama yapmış olmaları beklentisine karşın her 3 küme için müşterilik süreleri büyük oranda benzerlik göstermektedir.

**KAYNAKÇA**

https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis/data

https://pandas.pydata.org/

<https://scikit-learn-extra.readthedocs.io/en/stable/generated/sklearn_extra.cluster.KMedoids.html>

https://seaborn.pydata.org/

https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

<https://medium.com/@ali.soleymani.co/beyond-scikit-learn-is-it-time-to-retire-k-means-and-use-this-method-instead-b8eb9ca9079a>

https://matplotlib.org/