K-Means Kümeleme Algoritması Kullanarak Alışveriş Müşteri Veri Analizi

Abdullah Jamous

Siber Güvenlik ve Kriptografi Bölümü, Yıldız Teknik Üniversitesi, Türkiye

Özet Makine öğrenimi (ML), farklı teknolojileri, öğrenme yöntemlerini ve uygulamaları kapsayan kesişen bir konudur. Bu araştırmamızda, bir alışveriş mağazasında müşteri verilerini analiz etmek için kümeleme algoritmasını kullanacağız.Verileri doğru ve doğru bir şekilde analiz edebileceğimiz en iyi yolları öğreneceğiz.Bu deney iyi sonuçlar alıyor ve bu sayede bölme işlemi yapabiliyoruz. müşterileri gruplara ayırın ve paydalarını öğrenin. onlarla paylaşılır.

I. Giriş

Makine öğrenimi, verilerden öğrenen ve görevleri çözmek için tahminler yapan bir yapay zeka alt kümesidir. Açık talimatlarla programlanmadan öğrenebilir ve üç ana algoritma türü vardır: denetimli öğrenme, denetimsiz öğrenme ve takviyeli öğrenme. Denetimli öğrenmede, etiketlenmiş veriler üzerinde bir ML algoritması eğitilir, pekiştirmeli öğrenmede, ödül ve cezaya dayalı bir algoritma ve denetimsiz öğrenmede, algoritmaya sağlanan veriler sınıflandırılmaz veya etiketlenmez.

Denetimsiz öğrenme algoritmalarına herhangi bir ipucu, öneri veya eğitim verisi sağlanmadığından, eğitim veri setindeki kalıpları ve bilgileri kendi başlarına tanımlamaları gerekir - temel olarak, derin uca atılırlar ve çözmeleri gerekir. Denetimsiz öğrenme, minimum insan müdahalesinin gerekli olduğu belirli senaryolar için idealdir ve ana kullanımlarından biri kümelemedir. Kümeleme, bir veri kümesini tespit edilen benzerliklere/kalıplara dayalı olarak gruplara (veya kümelere) bölme işlemidir.

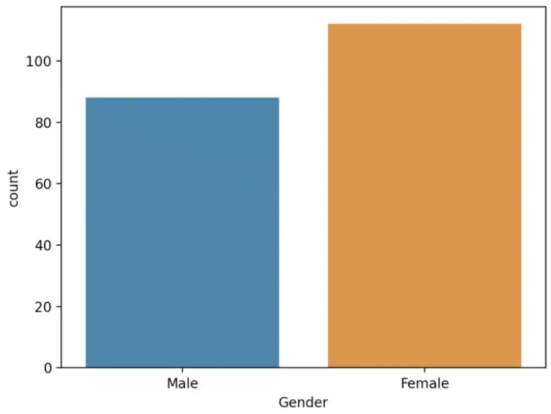
Tahmin edilecek hedef yok = denetimsiz öğrenme sorunu!

Neden Müşteri Segmentasyonu?
Müşteri segmentasyonu, pazarlama taktiklerinin optimizasyonu ve hedeflenmesinin yanı sıra, müşterinizin değerini en üst düzeye çıkarmak ve sağlanan ürünlerle deneyimlerini geliştirmek için tanımak için temel bir stratejidir. Segmentasyon, benzer özelliklere sahip müşterileri gruplandırmaktır, böylece bireysel erişim yapmak zorunda kalmadan (Google, Facebook veya Twitter gibi büyük bir şirkette neredeyse imkansız olan) iletişimlerinizi hedefleyebilir ve kişiselleştirmeyi işinize dahil edebilirsiniz. Örneğin, müşterilerinizi gelirlerine göre üç kümeye ayırırsanız, her müşteri grubuna kendileri için anlamlı olan ürünleri önerebilirsiniz. Bu genellikle çok yaygın bir kümeleme algoritması olan K-means kümeleme kullanılarak yapılır!

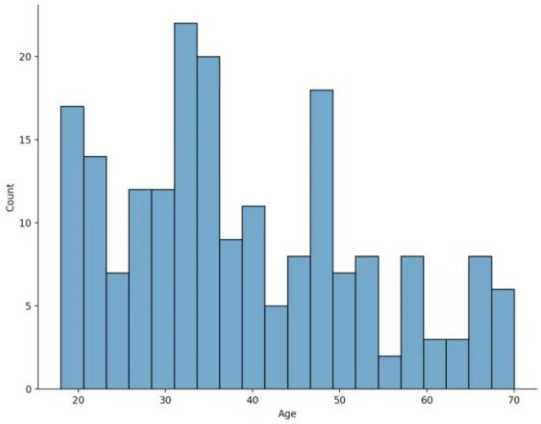
II. Yöntem

Bu projeye başlarken gerekli kütüphaneleri içe aktarabiliriz. Kütüphanelerimiz (**numpy, pandas, matplotlib, seaborn**) bu şekildedir.
Kullanılan veri seti, "alışveriş merkezi müşteri segmentasyon verileri" olarak adlandırılan Kaggle'dandı. Müşteri kimliği, yaş, yıllık gelir, harcama puanı ve cinsiyet olmak üzere 5 değişken vardır. Müşteri Kimliği, her müşterinin benzersiz tanımlayıcısı olduğundan kullanışlı değildir, bu nedenle sütun, del df[name] işlevi kullanılarak Pandas DataFrame'den silinebilir. Ek olarak veri setimizin başını yazdırabiliriz.

Toplamda 200 satır ve 4 sütun vardır. Yaş, yıllık gelir ve harcama puanının tümü sayısal veri türleridir, ancak cinsiyet kategoriktir, yani önceden işlenmesi ve cinsiyet karşılaştırmasını gösteren bir grafiğin haritasını çıkaran sayısal forma dönüştürülmesi gerekir.



Kadınların erkeklerden daha fazla olduğunu görebiliriz. Erkeklerden 20'den fazla kadın var. "Erkek" ve " Kadın " için kategorik formu kullanmayacağız, onları sayısal forma çevireceğiz, Erkek:0 ve Kadın:1.
Daha sonra, veri setimiz genelinde yaştaki varyansa bakacağız. Hedeflenen demografiyi anlamak için çok önemli olduğundan, bunu herhangi bir şirket için bilmek önemlidir. Sürekli veri değişkenlerinin dağılımını temsil eden, dağıtım grafiği olarak adlandırılan distplot() işlevini kullanabiliriz.

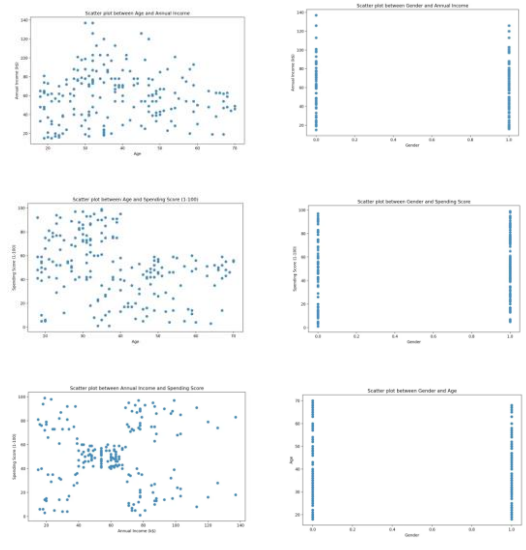


Yaş değerleri geniş bir aralığa dağılmıştır, ancak verilerimizin en iyi temsil ettiği yaş grubunun 30'ların ortası olduğunu görebiliriz.

III. Uygulama

İki Değişkenli Analiz

İki değişkenli analiz, iki değişken arasındaki ilişkinin analizini içerir. Özellikler arasındaki korelasyonu gözlemlemek için kullanılır. Bunu yapabiliriz çünkü artık her şey sayısal biçimde elimizde var. Altı dağılım grafiği yaptım. aşağıdaki gibi görüyoruz:



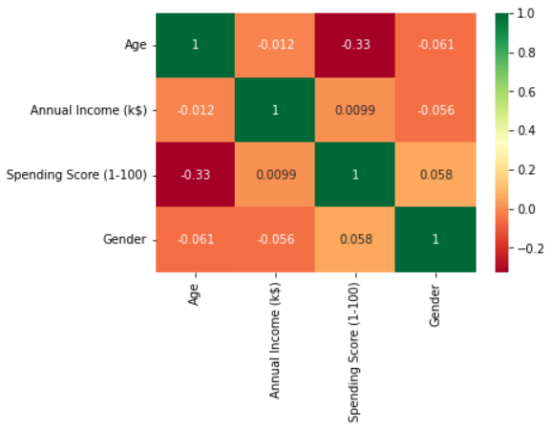
Bu diyagramlar arasında, küme algoritmasını uygulamak için güvenebileceğimiz en iyi iki özelliğin

Annual Income özelliği ve Spending Score özelliği olduğu bizim için netleşecektir. Ancak doğruluk adına HeatMap kullandım.

HeatMaping

Son yöntemimiz, birbirleriyle olan ilişkilerini anlamak için aynı anda en az üç değişken içeren verileri analiz eden çok değişkenli analizdir. Bu genellikle, bireysel değerlerin bir matris içinde yer aldığı ve renklerle görselleştirildiği verilerin grafiksel bir temsili olan bir ısı haritası ile yapılır. Dört değişkenimizi de karşılaştırabiliriz: yaş, yıllık gelir, harcama puanı ve cinsiyet.

Son satır özetlersek, df.corr() veri bağıntısı anlamına gelir ve veri çerçevesindeki sütunlar arasındaki bağıntıyı bulmaya yöneliktir, annot=True her hücrenin üzerine metin koyan veya açıklama ekleyen bir özniteliktir , ve cmap='inferno' sadece renk şemasıdır. Bu grafiği çıktı olarak oluyor:



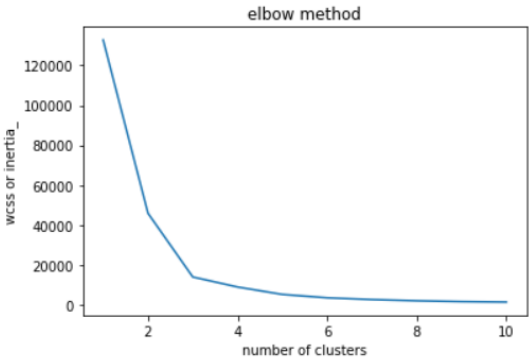
Yaşın, tüm harcama puanları, yıllık gelir ve cinsiyet ile çok negatif bir şekilde ilişkili olduğunu ve yıllık gelir ve harcama puanlarının da minimum düzeyde ilişkili olduğunu görebiliriz.

elbow

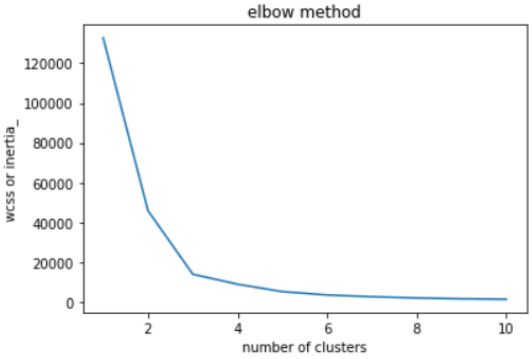
K-means kümeleme algoritmasının amacı, benzer noktaları kümelemek, dolayısıyla bir kümedeki noktaların ağırlık merkezleriyle olan mesafesini mümkün olduğunca azaltmaktır. Atalet, aynı kümeye ait iki nesne arasındaki mesafe olan küme içi mesafelerin bir ölçüsüdür ve Öklid mesafesi (iki nokta arasındaki bir çizgi parçasının uzunluğu) kullanılarak her bir veri noktası ile ağırlık merkezi arasındaki mesafe ölçülerek hesaplanır. , mesafenin karesini almak ve bu kareleri bir küme boyunca toplamak. Atalet değeri ne kadar düşük olursa, noktalar birbirine daha yakın olduğu için kümeler o kadar iyi olur.

Optimal küme sayısını bulmak için kullanabileceğimiz bir yöntem dirsek yöntemidir. Kümelerin merkezini yeniden hesaplama ve her küme için yeni merkezlere en yakın noktaları bulma adımları, atalet değeri daha fazla azaltılamayana kadar tekrarlanır. Bunu görsel olarak görmek için bir grafik veya "dirsek eğrisi" çizebiliriz, burada x ekseni küme sayısını temsil eder ve y ekseni değerlendirme ölçüsüdür.

Bazen kümelerin başlatılması uygun değilse, K-Ortalamaları kötü gruplanmış kümelerle sonuçlanabilir. Öncekimizde ilk merkezleri bulurken, randomizasyon kullanıyorduk. İlk k-merkezleri, veri noktalarından rastgele seçildi. Ancak, rasgeleleştirme çok doğru değildir çünkü adından da anlaşılacağı gibi rasgeledir. Bu nedenle, standart k-means algoritmasına geçmeden önce centroidleri başlatmak için bir prosedür belirten K-Means++ kullanıyoruz. Buradaki adımlar şunlardır: 1.) ilk küme rasgele seçilir (burada tüm merkezler yerine sadece biri seçilir), 2.) her veri noktasının zaten seçilmiş olan merkeze olan mesafesi hesaplanır, 3.) bu uzaklıkla orantılı olma olasılığı en yüksek olan yeni bir ağırlık merkezi seçilir ve 4.) kümeler seçilene kadar 2. ve 3. adımlar tekrarlanır.



Küme değeri, eylemsizlik değerinin büyük ölçüde azalmayı durdurduğu ve sabit hale geldiği yerdir. Bu grafikte Annual Income ve Spending Score özellikleri arasındaki kütle değerini gösteriyor bu da bize 5 grup oluşturabileceğimizi gösteriyor.



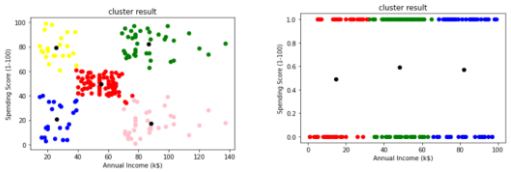
Bu grafikte Gender ve Spending Score özellikleri arasındaki kütle değerini gösteriyor bu da bize 3 grup oluşturabileceğimizi gösteriyor.

IV. Sonuçlar

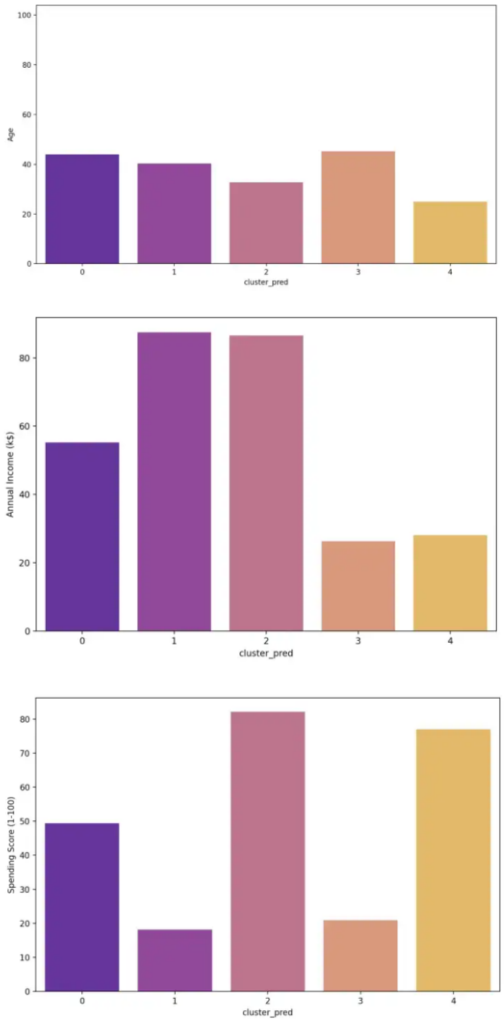
Verilerimizin 5 gruba ayrılabileceğini ve onu yorumlamaya ve fikir çizmeye başlayabileceğimizi anladık. Şu şekilde sıralayabiliriz:

- Sarı, düşük yıllık gelir ve düşük harcama puanına karşılık gelir
- Kırmızı, ortalama yıllık gelire ve ortalama harcama puanına karşılık gelir
- Yeşil, düşük yıllık gelire ve yüksek harcama puanına karşılık gelir
- Pembe, yüksek yıllık gelir ve düşük harcama puanına karşılık gelir
- Mavi, yüksek yıllık gelire ve yüksek harcama puanına karşılık gelir

Siyah nokta, kümenin merkezini gösterir.



Farklı kümelerin niteliklerini karşılaştırmak için, her kümedeki tüm değişkenlerin ortalamasını bulabiliriz. Bunu çizdiğim bazı diyagramlarla açıklayalım.



Bu, ilk kümede (küme 0) ortalama bir kişinin 25 yaş civarında olduğu ve düşük gelir ve yüksek harcama puanına sahip olduğu anlamına gelir. Yukarıdaki açıklamamıza bakarsak, bu mavi kümeye karşılık gelir!

Kümelerin Nitelikleri

Küme 0 (mor): ara yıllık gelir, orta düzey harcama puanı

- 40'ların başı
- 55 bin yıllık gelir
- Ara harcama puanı 49
- Ağırlıklı olarak kadın

Küme 1 (macenta): Yüksek yıllık gelir, düşük harcama puanı

- 30'ların sonu
- 86k yıllık gelir
- Düşük harcama puanı 17
- Cinsiyette aşağı yukarı eşit

Küme 2 (pembe): Yüksek yıllık gelir, yüksek harcama puanı

- 30'ların başı
- 85 bin yıllık gelir
- 82 ile yüksek harcama puanı
- Ağırlıklı olarak kadın

Küme 3 (turuncu): Düşük yıllık gelir, düşük harcama puanı

- 40'lı yaşların ortası
- 26k yıllık gelir
- Düşük harcama puanı 21
- Ağırlıklı olarak kadın

Küme 4 (sarı): Düşük yıllık gelir, yüksek harcama puanı

- 20'li yaşların ortası
- 26k yıllık gelir
- 78 ile yüksek harcama puanı
- Ağırlıklı olarak kadın