**Grup-2 :**

**Büşra Çalışkan-Melike Fatma Alkan- İrem Koyunlu -Esmanur Ateş-İrem Alkoyun**

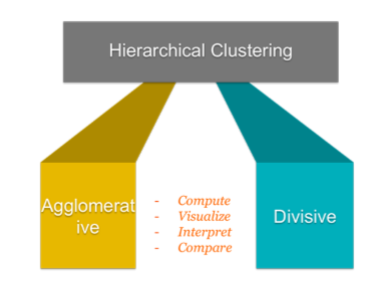
**HİYERARŞİK KÜMELEME VE DENDROGRAM**

1. **PROJENİN AMACININ TANITILMASI**

**Projemizin amacı kullandığımız verideki birbirine benzerlik gösteren gözlemleri kümeleme mantığını kullanarak dendrogramlar yardımı ile görselleştirmektir. Bu kümelemeyi yaparken en uygun k değerini kullanmayı amaçladık. İki küme arasındaki mesafeyi ölçerken kullandığımız yöntemlerden hangisinin en iyi olduğunu inceledik.**

1. **ANLATILAN KONUNUN TANITILMASI**

* **HİYERARŞİK KUMELEME ANALİZİ**

****

**Hiyerarşik kümeleme [veya hiyerarşik küme analizi (HCA)], benzerliklerine göre nesneleri gruplandırmak için kümelemeyi bölümlemeye alternatif bir yaklaşımdır. Bölümleme kümelemenin aksine, hiyerarşik kümeleme, üretilecek küme sayısını önceden belirtmek için gerekli değildir. Hiyerarşik kümeleme iki türe ayrılabilir:**

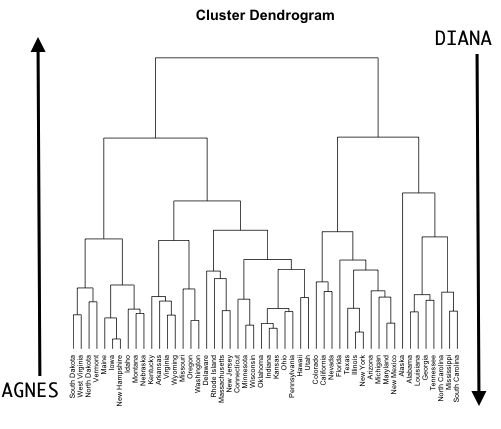
**AGGLOMERATIVE CLUSTERING;**

**• Her gözlemin başlangıçta kendi (yaprak) kümesi olarak kabul edildiği aglomeratif kümeleme. Daha sonra, en benzer kümeler, yalnızca tek bir büyük küme (kök) olana kadar art arda birleştirilir.(AGNES)**

**DIVISE CLUSTERING;**

**• DIANA (Bölünme Analizi) olarak da bilinir ve yukarıdan aşağıya doğru çalışır. Algoritma AGNES'in ters sırasıdır. Tüm nesnelerin tek bir kümeye dahil edildiği kok ile başlar. Yinelemenin her adımında, en heterojen küme ikiye ayrılır. İşlem, tüm nesneler kendi kümelerinde olana kadar yinelenir.**

**Hiyerarşik kümelemenin sonucu, dendrogram olarak da bilinen nesnelerin ağaç tabanlı bir temsilidir.**

****

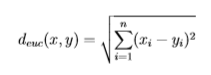
**Dendrogram, bir seviyedeki kümelerin bir sonraki seviyelerde kümeler oluşturmak için birleştirildiği çok düzeyli bir hiyerarşidir. Bu, bir veri nesnelerinin uygun gruplarını oluşturmak için ağacın kesileceği seviyeye karar vermeyi mümkün kılar.**

**Bu bölümde, nasıl görselleştirileceğini göstereceğiz. Dendrogramlar kullanarak nesneler arasındaki benzerliği hiyerarşik kümeleme algoritmalarını tanımlayarak başlıyoruz. Hiyerarşik kümelenmenin sonuçlarını hesaplamak ve görselleştirmek için r scriptleri sağlıyoruz. Sonra, dendrogramları gruplara nasıl keseceğimizi göstereceğiz. İki dendrogramın nasıl karşılaştırılacağını da göstereceğiz. Ek olarak, büyük veri kümelerinin dendrogramlarını işlemek için çözümler sunacağız.**

* **MESAFELERİ ÖLÇME YÖNTEMLERİ**

**Mesafe ölçütlerinin seçimi kümelenmede kritik bir adımdır. İki elementin (x, y) benzerliğinin nasıl hesaplandığını belirler ve kümelerin şeklini etkileyecektir. Mesafe ölçümleri için klasik yöntemler, aşağıdaki gibi tanımlanan Öklid ve Manhattan mesafeleridir:**

1. **Euclidean distance:**

****

**Matematikte Pisagor bağlantısı kullanılarak bulunan iki nokta arasındaki mesafe ölçüm birimidir. Buna göre iki boyutlu düzlemde iki nokta arasındaki mesafe basitçe iki noktanın x ve y koordinatlarının ayrı ayrı farklarının hipotenüs’üne eşittir. Euclidian uzayda iki nokta arasındaki mesafeyi en kısa yol olarak hesaplar.**

1. **Manhattan distance:**

**C:\Users\ASUS\Desktop\istyazılımfinal\2.METHOD.PNG**

**Öklid geometrisinin olağan mesafe fonksiyonunun veya metriğinin, iki nokta arasındaki mesafenin Kartezyen koordinatlarının mutlak farklılıklarının toplamı olduğu yeni bir metrik ile değiştirildiği bir geometri şeklidir.**

* **AGGLOMERATIVE HİYERARŞİK KÜMELENME ADIMLARI**

**İki gözlem kümesi arasındaki farklılığı nasıl ölçeriz? Bu soruyu cevaplamak için bir dizi farklı küme aglomerasyon yöntemi (yani bağlantı yöntemleri) geliştirilmiştir. En yaygın yöntem tipleri şunlardır:**

* **Maksimum veya tam bağlantı: İki küme arasındaki mesafe, küme 1'deki öğeler ve küme 2'deki öğeler arasındaki tüm çift yönlü mesafelerin maksimum değeri olarak tanımlanır. Daha kompakt(sıkı, yoğun) kümeler üretme eğilimindedir.**
* **Minimum veya tek bağlantı: Küme 1'deki elemanlar ile Küme 2'deki elemanlar arasındaki tüm çift farklılıkları hesaplar ve bu farklılıkların en küçüğünü bir bağlantı kriteri olarak görür. Uzun, “gevşek” Kümeler üretme eğilimindedir.**
* **Ortalama veya ortalama bağlantı: İki küme arasındaki mesafe, küme 1'deki öğeler ile küme 2'deki öğeler arasındaki ortalama mesafe olarak tanımlanır.**
* **Centroid bağlantısı: İki küme arasındaki mesafe, küme 1 için centroid (uzunluk p değişkenlerinin ortalama vektörü) ve küme 2 için centroid arasındaki mesafe olarak tanımlanır.**
* **Ward'ın minimum varyans yöntemi: Küme içindeki toplam varyansı en aza indirir. Her adımda, minimum küme mesafesi olan kümeler çifti birleştirilir.**
* **HIYERARSIK KUMELEMENE ZAMAN KULLANILMALI?**
* **Hiyerarşik kümeleme büyük verileri iyi işleyemez.**
* **Hiyerarşik kümeleme yaparken üretilecek küme sayısını önceden belirtmek gerekli değildir. Küme sayısına dendrogramlar yardımı ile karar verilebilir bu sebeple Hiyerarşik kümeleme bu konu bakımından daha avantajlıdır.**

**Hiyerarşik Kümeleme kullanım alanları;**

* **Pazarlama**
* **Genetik**
* **Tıp**
* **Veri Madenciliği**
* **Ödev benzerlikleri üzerinden kopya gruplarını bulma**
* **Boyut indirgemek amacıyla …**

1. **PROJENİN KODLAMA, VERİ TOPLAMA SÜRECİNİ ANLATAN BİR DOKÜMAN OLUŞTURULMASI**

* **KULLANILAN VERİ SETİ;**

**USArrests 1973'te 50 ABD eyaletinin her birinde saldırı, cinayet ve tecavüz için tutuklama istatistiklerini içeren veri setini kullanacağız. Aynı zamanda veri kentte yaşayan nüfusun yüzdesini de içermektedir. Kaynak:** [**https://www.kaggle.com/deepakg/usarrests**](https://www.kaggle.com/deepakg/usarrests)

**- Değişkenler:**

**Murder: Verideki Cinayet istatistiği değişkenidir.**

**Assault: Verideki Saldırı istatistiği değişkenidir.**

**UrbanPop: Verideki Kentsel popülasyon değişkenidir.**

**Rape: Verideki Tecavüz istatistiği değişkenidir.**

* **VERİ HAZIRLAMA**

**Genel olarak R'de bir küme analizi yapmak için, veriler aşağıdaki gibi hazırlanmalıdır:**

**Verilerdeki eksik değerler kaldırılmalı veya tahmin edilmelidir.**

**Değişkenleri karşılaştırılabilir hale getirmek için veriler standartlaştırılmalıdır (yani ölçeklendirilmelidir). Standardizasyonun değişkenleri ortalama 0 ve standart sapma 1 olacak şekilde dönüştürmekten ibaret olduğunu biliyoruz(Scale kodu ile yapıyoruz)**

* **Kodlama kısmında kullanacağımız paketler:**

**library(tidyverse): Veri bilimi için tasarlanmış r paketlerinden biridir.**

**library(cluster): Kümeleme algoritmaları için kullanılır.**

**library(factoextra): factoextra paketi bütün bu algoritmaları ggplot2 destegi ile elegant gorseller olarak verebilmektedir.**

**library(dendextend): İki dendrogramı karşılaştırmak için kullanılır.**

**library(ggdendro): Bu, 'ggplot2' kullanarak dendrogramlar için görselleştirme paketidir.**

**library(ade4): Öklid keşif yöntemleri çerçevesinde çevreyle ilgili ve çevresel verileri analiz etmek için veri analiz fonksiyonları içerir.**

**library(ggraph): Ggplot2'de uygulanan grafik dilbilgisi, tablo veri girişine bağlı olması nedeniyle grafik ve ağ görselleştirmeleri için yetersizdir. Diğer paketlerle destekliyoruz.**

**library(igraph): Ağ analizi için bir kütüphane. Hem deterministtik hem de skolastik grafikler oluşturmak için birçok işlev vardır.**

**library(dplyr): Tidyverse içinde bir pakettir, veriyi düzenlememizi sağlar.**

**library(RColorBrewer): RColorBrewer, güzel grafikler oluşturmak için kullanıma hazır renk paletleri içeren bir R paketidir.**

**library(leaflet): Etkileşimli Haritalar için en popüler açık kaynaklı JavaScript kütüphanelerinden biridir.**

**library(maps): Coğrafi haritalar için R paketi.**

**library(pvclust): Hiyerarşik küme analizindeki belirsizliği değerlendirmek için çok ölçekli bir bootstrap yeniden örnekleme uygulamasıdır.**

* **Kaynaklar:**

**Kullanılan Kitap:** **Kassambara, A. (2017). Practical guide to cluster analysis in R: Unsupervised machine learning (Vol. 1). STHDA. Wiley, M., & Wiley, J. F. (2019). Advanced R Statistical Programming and Data Models: Analysis, Machine Learning, and Visualization. Apress.**

**Web sitesi:**

**[1] https://uc-r.github.io/hc\_clustering**

**[2] https://www.datanovia.com/en/lessons/comparing-cluster-dendrograms-in-r/**

**[3] https://www.datanovia.com/en/lessons/examples-of-dendrograms-visualization/**

**[4] http://www.sthda.com/english/wiki/beautiful-dendrogram-visualizations-in-r-5-must-known-methods-unsupervised-machine-learning**

**[5] https://cran.r-project.org/web/packages/ggdendro/vignettes/ggdendro.html**

**[6] https://stackoverflow.com/questions/59841156/r-dendrogram-parent-child-clusters**

**[7] https://rpubs.com/gaston/dendrograms**

**[8] https://books.google.com.tr/books?id=7EpADwAAQBAJ&pg=PA13&lpg=PA13&dq=USArrests+amazing+dendrogram+with+r&source=bl&ots=FKb\_VXL-Pm&sig=ACfU3U3JUOyvbVTNx-YzNqDpbZ5j0l9bcA&hl=tr&sa=X&ved=2ahUKEwi7tfeYzZLqAhXNRBUIHeoiCNkQ6AEwEHoECAYQAQ#v=onepage&q=USArrests%20amazing%20dendrogram%20with%20r&f=false**

**[9] https://koalaverse.github.io/homlr/notebooks/21-hierarchical-clustering.nb.html**

**[10] https://bradleyboehmke.github.io/HOML/hierarchical.html**

**[11] https://rpubs.com/fernanks/HierarchicalClustering**

**[12] https://bradleyboehmke.github.io/HOML/hierarchical.html#determining-optimal-clusters**

**[13]** [**https://books.google.com.tr/books?id=7EpADwAAQBAJ&pg=PA13&lpg=PA13&dq=ggraph+USArrest&source=bl&ots=FKb-O-J7Mi&sig=ACfU3U0-191kJl4\_itj1M-\_jBXZjofu7ww&hl=tr&sa=X&ved=2ahUKEwj\_tq2Zu5zqAhWqA2MBHSa2DqIQ6AEwAHoECAcQAQ#v=onepage&q=ggraph%20USArrest&f=false**](https://books.google.com.tr/books?id=7EpADwAAQBAJ&pg=PA13&lpg=PA13&dq=ggraph+USArrest&source=bl&ots=FKb-O-J7Mi&sig=ACfU3U0-191kJl4_itj1M-_jBXZjofu7ww&hl=tr&sa=X&ved=2ahUKEwj_tq2Zu5zqAhWqA2MBHSa2DqIQ6AEwAHoECAcQAQ#v=onepage&q=ggraph%20USArrest&f=false)

**[14]**

[**https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/#:~:text=6.-,Difference%20between%20K%20Means%20and%20Hierarchical%20clustering,but%20K%20Means%20clustering%20can.&text=While%20results%20are%20reproducible%20in,2D%2C%20sphere%20in%203D**](https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/#:~:text=6.-,Difference%20between%20K%20Means%20and%20Hierarchical%20clustering,but%20K%20Means%20clustering%20can.&text=While%20results%20are%20reproducible%20in,2D%2C%20sphere%20in%203D)

**[15]**

[**https://analyticsindiamag.com/comparison-of-k-means-hierarchical-clustering-in-customer-segmentation/**](https://analyticsindiamag.com/comparison-of-k-means-hierarchical-clustering-in-customer-segmentation/)

**Projemizin Githup Linki:**

[**https://github.com/iremkoyunlu/Hieraltical-Clustering-and-K-Means/blob/master/IstatikselYazilimlarGRUP2.Rmd**](https://github.com/iremkoyunlu/Hieraltical-Clustering-and-K-Means/blob/master/IstatikselYazilimlarGRUP2.Rmd)