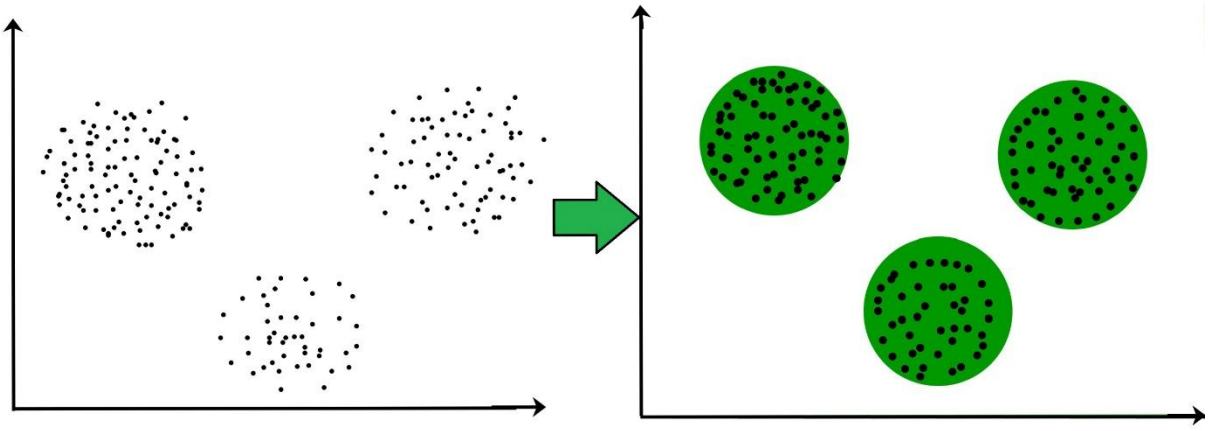
Kümeleme (Clustering)

Verilerin yakınlık, uzaklık, benzerlik gibi ölçütlere göre analiz edilerek sınıflara ayrılmasına kümeleme denir. Kümeleme, denetimsiz öğrenmenin bir yöntemidir ve birçok alanda kullanılan istatistiksel veri analizi için yaygın bir tekniktir. Denetimsiz öğrenme, veri kümesi ile çıktıların olmadığı bir öğrenme metodudur. Veri kümesindeki verileri yorumlayarak ortak noktaları bulmak ve bunları kümeleştirme işlemi yapılarak anlamlı bir veri elde edebilmektir. Sistem, öğreten olmadan öğrenmeye çalışır. Ham verileri organize verilere dönüştüren bir makine öğrenimi türüdür.

Denetimsiz öğrenmenin metodu olduğu için etiketlenmemiş verileri gruplandırır.

Kümeleme, heterojen bir veri kümesinden homojen veri noktaları grupları oluşturmayı amaçlar. Benzerliği Öklid mesafesi, Kosinüs benzerliği, Manhattan mesafesi vb. gibi bir metriğe göre değerlendirir ve ardından en yüksek benzerlik puanına sahip noktaları birlikte gruplandırır.



Yakınlığa göre 3 gruba ayrıldı. Burada kaç grup olacağı ve en uygun küme sayısını algoritmanın kendisi belirler.

Kümelemede sadece veriler vardır onlar hakkında bilgi verilmez. Bu verilerden sonuçlar çıkarılmaya çalışılır. Veriler arasındaki ilişkilere göre kümelenir.

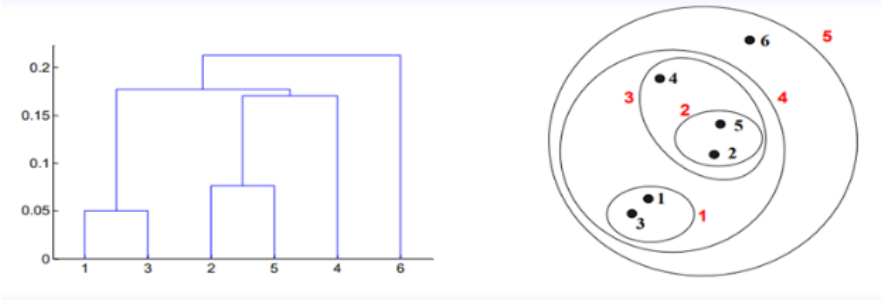
Kümeleme Algoritması Türleri

- Centroid Tabanlı Kümeleme
- Yoğunluğa Dayalı Kümeleme
- Bağlantı Tabanlı Kümeleme (Hiyerarşik kümeleme)
- Dağıtım Tabanlı Kümeleme

Burada Hiyerarşik Kümelemeyi inceleyelim.

Hiyerarşik Kümeleme

Hiyerarşik kümeleme algoritmaları 2 kategoriye ayrılır: yukarıdan aşağıya veya aşağıdan yukarıya. Aşağıdan yukarıya algoritmalar, her veri noktasını başlangıçta tek bir küme olarak ele alır ve ardından tüm kümeler tüm veri noktalarını içeren tek bir kümede birleştirilene kadar küme çiftlerini art arda birleştirir (veya toplar). Bu nedenle aşağıdan yukarıya hiyerarşik kümeleme, hiyerarşik kümelemeli kümeleme (hierarchical agglomerative clustering) veya HAC olarak adlandırılır. Bu küme hiyerarşisi bir ağaç (veya dendrogram) olarak temsil edilir. Ağacın kökü, tüm örnekleri toplayan benzersiz kümedir, yapraklar yalnızca bir örnek içeren kümelerdir. Algoritma adımlarına geçmeden önce bir örnek için aşağıdaki grafiğe bakın



Hiyerarşik kümeleme, küme sayısını belirlememizi gerektirmez ve hatta bir ağaç oluşturduğumuz için hangi küme sayısının en iyi görüneceğini seçebiliriz. Ek olarak, algoritma mesafe ölçüsü seçimine duyarlı değildir; hepsi eşit derecede iyi çalışma eğilimindeyken, diğer kümeleme algoritmalarında uzaklık ölçüsü seçimi kritiktir. Hiyerarşik kümeleme yöntemlerinin özellikle iyi bir kullanım durumu, temeldeki verilerin hiyerarşik bir yapıya sahip olması ve hiyerarşiyi kurtarmak istemenizdir; diğer kümeleme algoritmaları bunu yapamaz. Hiyerarşik kümelemenin bu avantajları, K-Ortalamalarının ve GMM'nin doğrusal karmaşıklığından farklı olarak, O (n^3) zaman karmaşıklığına sahip olduğundan, daha düşük verimlilik pahasına gelir. Hiyerarşik kümeleme, veri noktalarını kümelemek için denetimsiz bir öğrenme yöntemidir. Algoritma, veriler arasındaki farklılıkları ölçerek kümeler oluşturur. Denetimsiz öğrenme, bir modelin eğitilmesi gerekmediği ve bir "hedef" değişkene ihtiyacımız olmadığı anlamına gelir. Bu yöntem, tek tek veri noktaları arasındaki ilişkiyi görselleştirmek ve yorumlamak için herhangi bir veri üzerinde kullanılabilir.

Kullanım Alanları:

- Tıp' ta benzer semptomları olan hastalara ait görüntüler üzerinde inceleme yapılırken,
- Suça eğilimli yerlerinin belirlenmesinde,

- Şüpheli durum tespitinde, normal akışın dışında kalan yani normal grupların, dışında kalan durumların tespitinde,
- Deprem için tehlikeli bölgelerin belirlenmesinde de kullanılır

En Sık Kullanılan Algoritmaları:

1. *K-means Clustering*
2. *Hierarchical Clustering*
3. *DBSCAN (Density-Based Spatial Clustering of Applications with Noise)*
4. *Gaussian Mixture Models (GMM)*
5. *Agglomerative Clustering*
6. *Spectral Clustering*
7. *Mean Shift Clustering*
8. *Affinity Propagation*
9. *OPTICS (Ordering Points To Identify the Clustering Structure)*
10. *Birch (Balanced Iterative Reducing and Clustering using Hierarchies)*