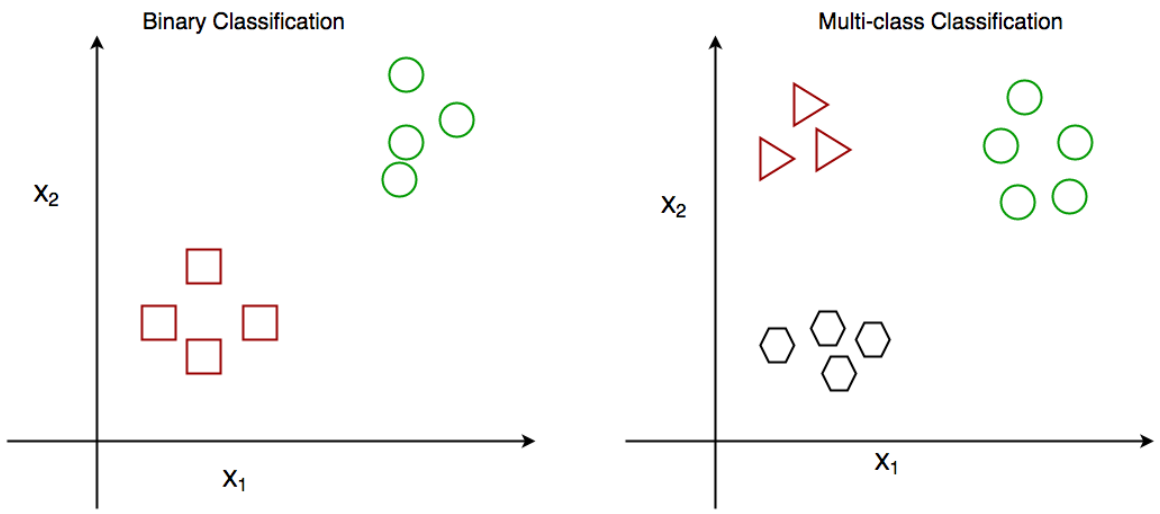
Kümeleme (clustering) algoritması nedir ?

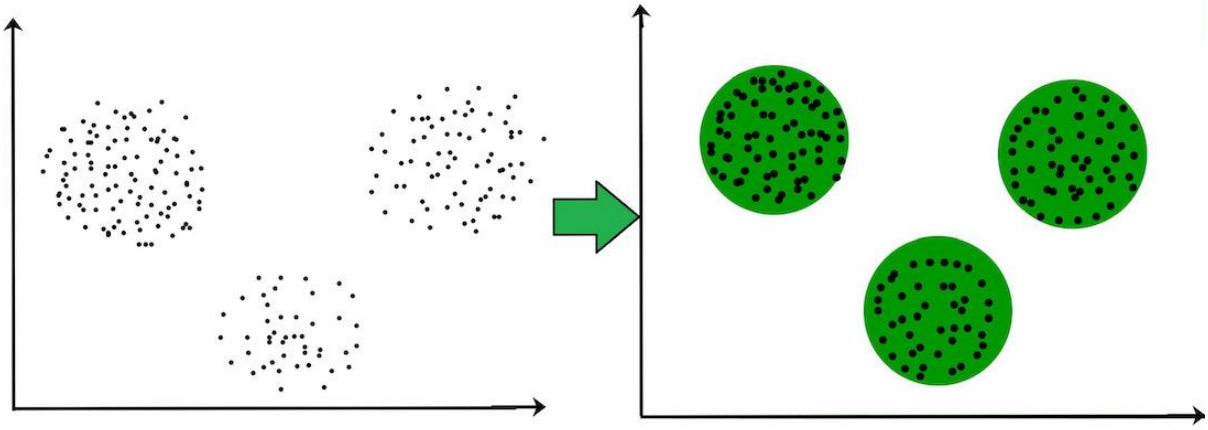
Kümeleme (clustering) algoritması, veri biliminde ve makine öğreniminde kullanılan bir tekniktir. Temel amacı, bir veri kümesini birbirine benzeyen alt gruplara (küme) ayırmaktır. Bu süreçte, her kümedeki veriler birbirine daha yakın veya benzer özelliklere sahipken, farklı kümelerdeki veriler birbirinden daha farklıdır. Kümeleme algoritmaları denetimsiz öğrenme yöntemlerinden biridir, yani bu algoritmalar, etiketlenmiş veriler olmadan çalışır.

Kümeleme Çeşitleri:
1.Hiyerarşik Kümeleme
2.Gürültülü Uygulamaların Yoğunluğa Dayalı Konumsal Kümelenmesi (DBSCAN) (DBSCAN)
3.K-means Kümeleme
4.Ağırlık Ortalama Kaydırma Kümelemesi
5.Gauss Karışım Modelleri (GMM) kullanarak Beklenti-Maksimizasyon (EM) Kümeleme



Şekil 1.1

Şekil 1.1'de görüldüğü gibi sınıflandırma yapılacak verilerin etiketi bellidir. Kare, çember ve altıgen şekilleri verilerin etiketlerini belirtmektedir. Veriler bu etiketler neticesinde sınıflandırılmıştır.



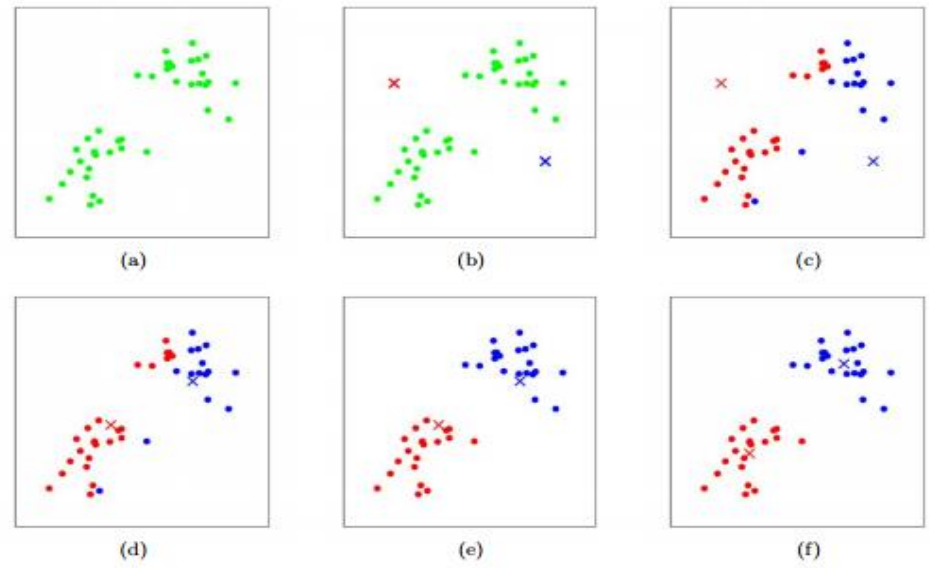
Şekil 1.2

K-Means:

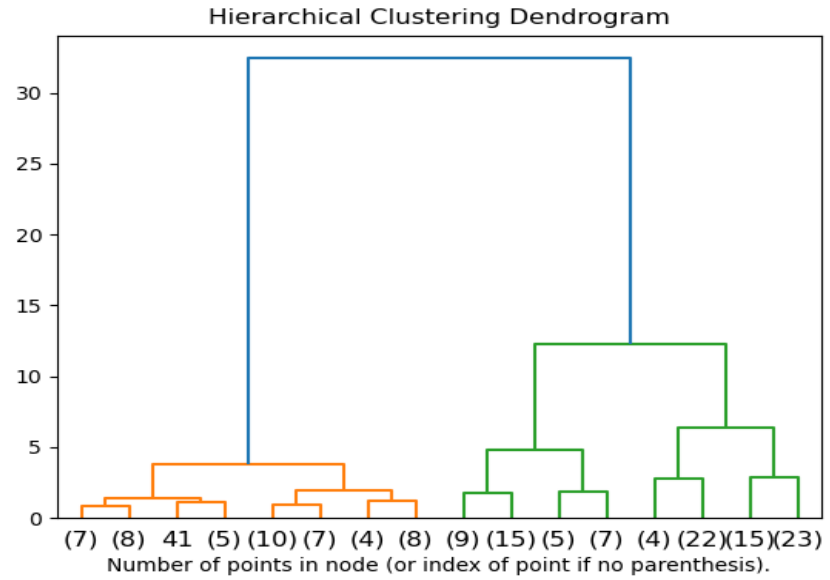
Temelinde, verileri benzerliklerine göre kümelemeyi amaçlar. Benzerlik, veriler arasındaki uzaklığa göre belirlenmektedir. Uzaklığın az olması benzerliği arttırır. Giriş parametresi olarak ‘k’ sayısı verilmelidir. Bu sayı örneklemin kaç adet kümeye ayrılacağını belirtir.
En sık kullanılan kümeleme algoritmalarındandır. Uygulanması oldukça kolaydır. Büyük ölçekli verileri hızlı ve etkin bir şekilde kümeleyebilir.
Çalışma mantığı şu şekildedir:

1. ‘k’ adet küme merkezi rastgele seçilir.
2. Merkez dışındaki veriler mesafelerine göre kümelendirilir.
3. Yapılan kümelendirmeye göre yeni küme merkezleri belirlenir.
4. Kararlı hale gelene kadar 2. ve 3. adım tekrarlanır.

‘k’ sayısının işlem başlamadan belirlenmesi duruma göre avantaj ya da dezavantaj oluşturabilir.
kümelenmiştir.



Hierarchical Clustering
Hiyerarşik kümeleme, iç içe kümeleri art arda birleştirerek veya bölerek oluşturan genel bir kümeleme algoritmaları ailesidir. Bu küme hiyerarşisi bir ağaç (veya dendrogram) olarak temsil edilir. Ağacın kökü, tüm örnekleri toplayan benzersiz kümedir, yapraklar yalnızca bir örnek içeren kümelerdir.



Şekil 1.8 Hiyerarşik Kümeleme dendogramı

Hiyerarşik kümeleme stratejileri genellikle iki türe ayrılır.

- **Aglomeratif** : Bu bir " aşağıdan yukarıya " yaklaşımdır. Her gözlem kendi kümesinde başlar ve hiyerarşide yukarı doğru çıkıldıkça küme çiftleri birleştirilir.
- **Bölücü** : Bu bir " yukarıdan aşağıya " yaklaşımdır. Tüm gözlemler tek bir kümede başlar ve hiyerarşide aşağı doğru hareket ettikçe bölmeler yinelemeli olarak gerçekleştirilir.

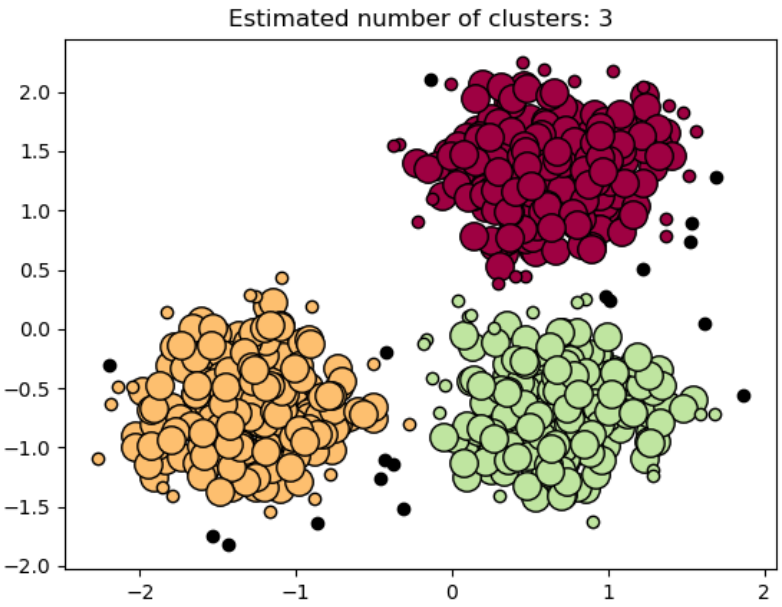
DBSCAN
Yoğunluğa dayalı bir algoritmadır. Birbirine çok yakın olan noktaları (birçok yakın komşuya sahip noktalar) birlikte gruplandırır. Düşük yoğunluklu (en yakın komşuları çok uzakta olan) bölgelerde bulunan noktaları ise tek başına bulunan aykırı noktalar olarak işaretler. DBSCAN, bilimsel literatürde en çok alıntı yapılan algoritmadır. Algoritma, kümeleri belirlerken ε (eps) ve minimum points (minPts) parametrelerini kullanır. ε (eps), noktanın komşularını arayacağı uzaklığı belirtir. Minimum points (minPts) ise bir kümenin oluşabilmesi için, noktalar arasında kaç tane komşuluk olması gerektiğini belirtir.
Algoritmanın çalışma mantığı şu şekilde özetlenebilir:

1. Her noktanın ε (eps) komşuluğundaki noktaları bulun ve minPts sayısından daha fazla komşu bulunursa, o noktayı çekirdek nokta olarak belirleyin.

2. Bağlı bileşenler arasında çekirdeği olmayan tüm noktaları göz ardı ederek, grafik üzerindeki noktalar.
3. Çekirdeği olmayan noktalar, çekirdeği olan kümelerdeki noktalardan herhangi birine ε (eps) uzaklıktaysa, o kümeye atanır.

Küme sayısının önceden belirlenmesinin gerekmemesi, aykırı değerlere karşı dayanıklı olması avantajlarıdır.
Dezavantajları ise:

- MinPts-ε kombinasyonu tüm kümeler için uygun şekilde seçilemediğinden, yoğunluklardaki büyük farklılıklarla veri kümelerini iyi bir şekilde kümeleyemez.
- Veri ve ölçek iyi anlaşılmadıysa, anlamlı bir mesafe eşiği ε seçmek zor olabilir.
- Tamamen belirleyici değildir: Birden fazla kümeden erişilebilen sınır noktaları, verilerin işlenme sırasına bağlı olarak her iki kümenin de parçası olabilir.



Şekil 1.10 DBSCAN uygulanarak kümelenmiş bir veri seti
Bu yazıda, kümeleme ile sınıflandırma arasındaki farka ve en çok kullanılan kümeleme algoritmalarına göz atmış olduk

KAYNAK: https://www.geeksforgeeks.org/machine-learning/