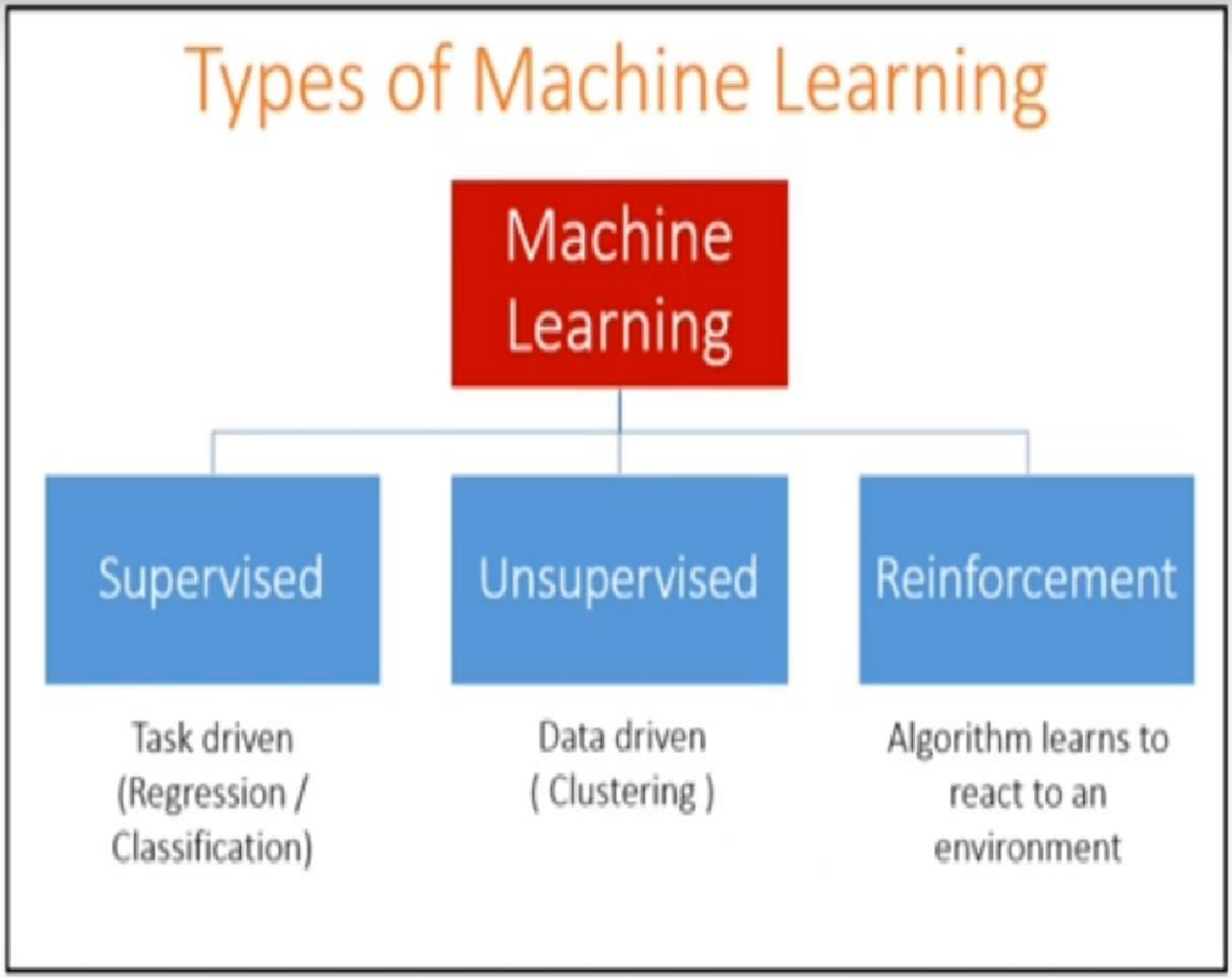
Lec. 8

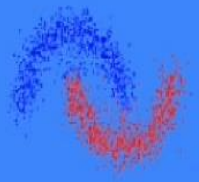
Clustering Algorithms

Assist. Prof. Dr. Saad Albawi

ما هو التعليم بدون إشراف Unsupervised ML

أقسام الـ ML :





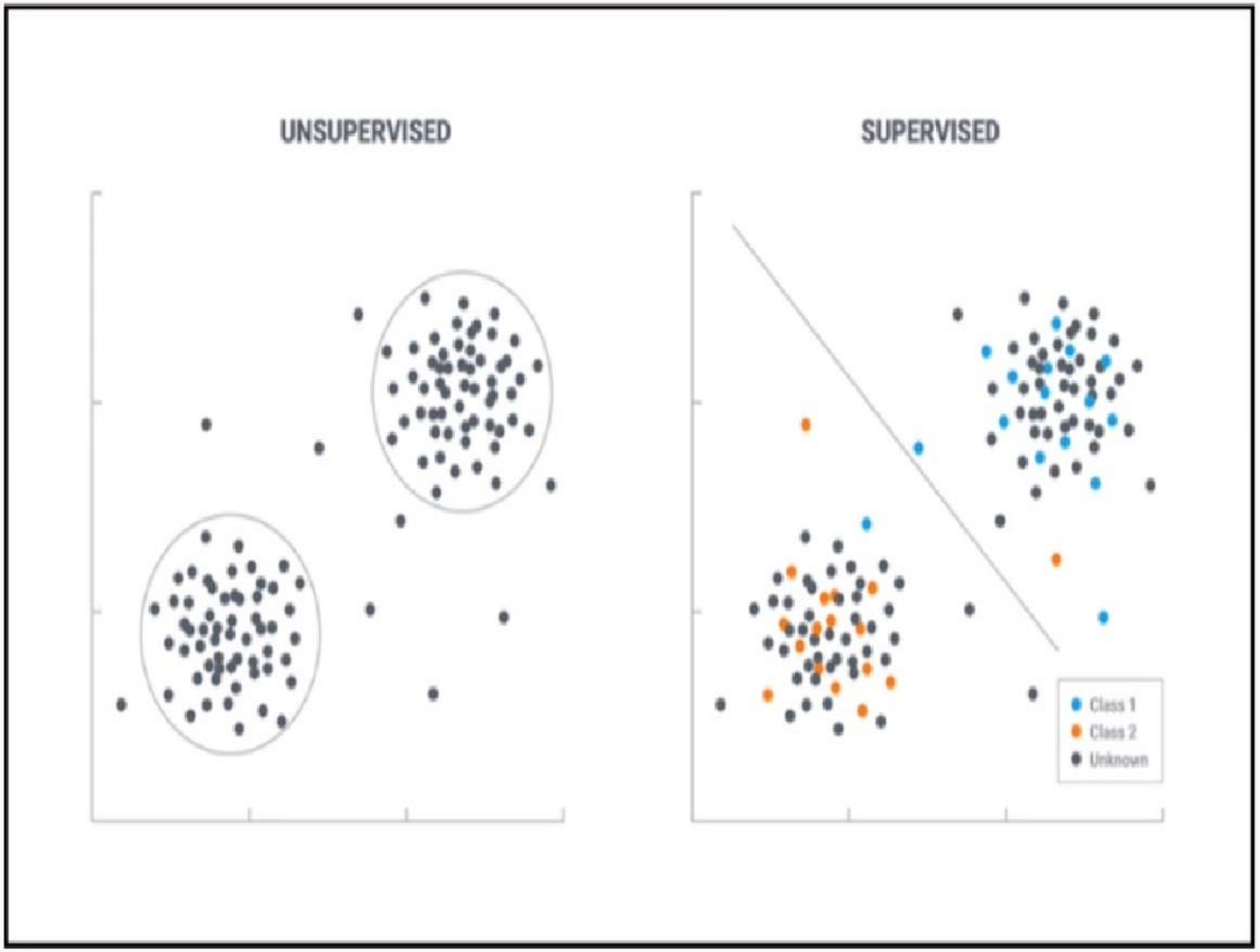
ما هو التعليم بدون إشراف Unsupervised ML

الفرق بين Supervised Vs Unsupervised :

- الفرق الحاسم بينهما أن التعليم بإشراف , هناك بيانات لها output بقيم y أما بون إشراف فهي ليس لها output وليس لها قيم y
- فإذا كان في بيانات التدريب الف طالب ,لدي معلومات عنهم (input X) , ولدي معلومة هل تم قبولهم أم لا (output y) فهذا تعليم بإشراف
- واذا كان لدي هناك الف عميل لدي شركة سامسونج , ولدينا بيانات عنهم (input X) لكن لا نعرف هل سيقومو بالشراء ام لا , ونريد تقسيمهم لمجعوعات , فهذا تعليم دون إشراف
- فالفارق الأساسي لدي بيانات التدريب , هل لدي فيها output y ام لا

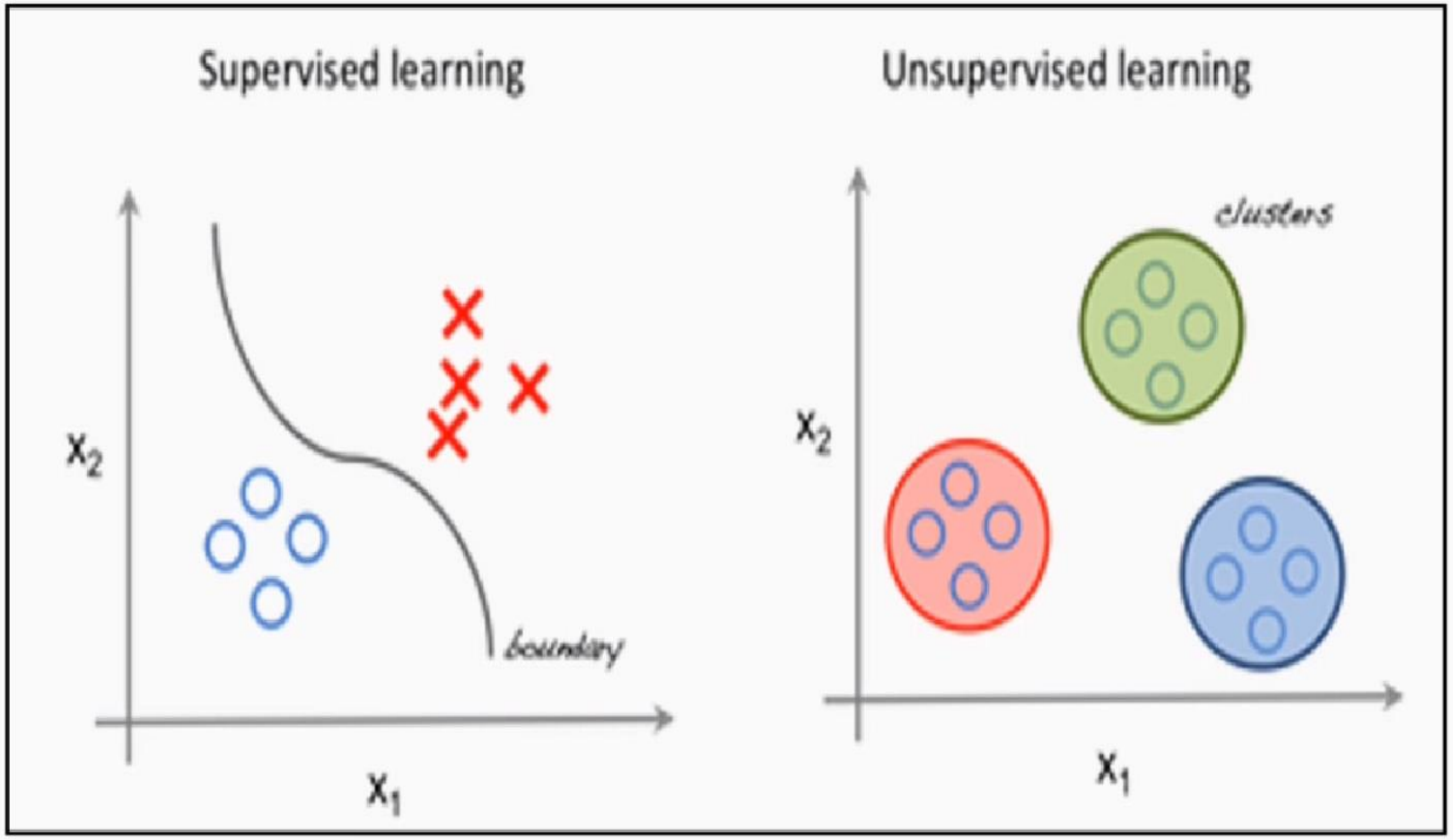
ما هو التعليم بدون إشراف Unsupervised ML

الفرق بين Supervised Vs Unsupervised :



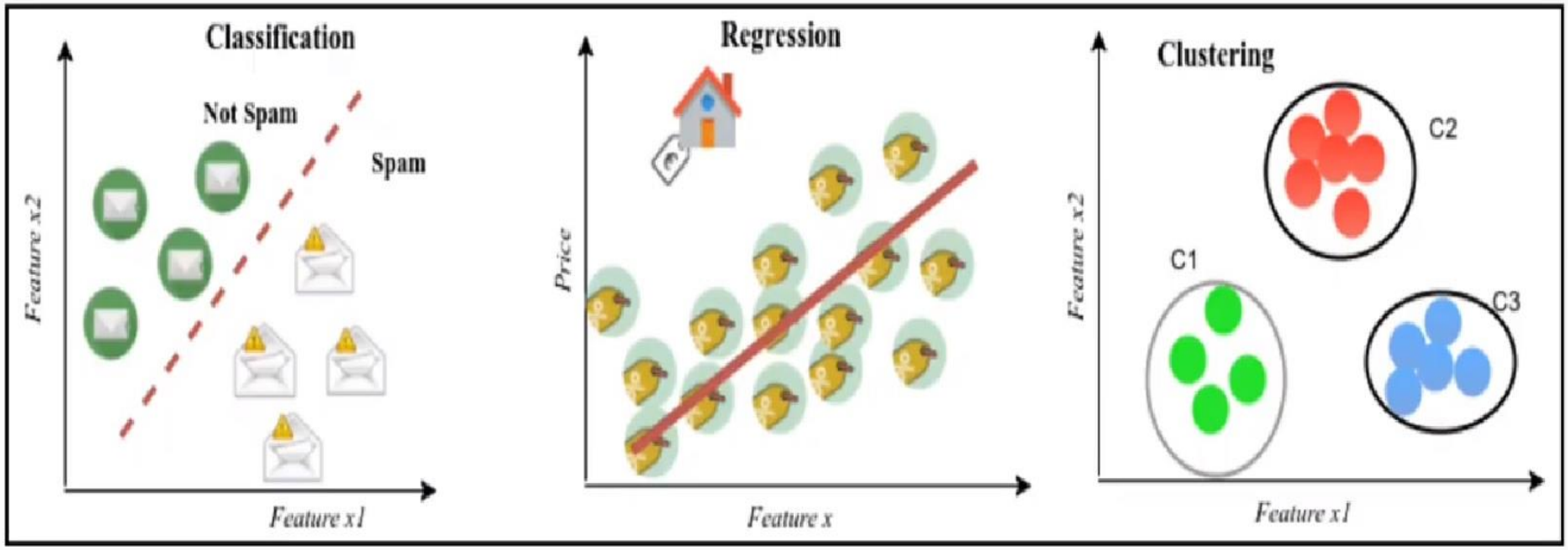
ما هو التعليم بدون إشراف Unsupervised ML

الفرق بين Supervised Vs Unsupervised :



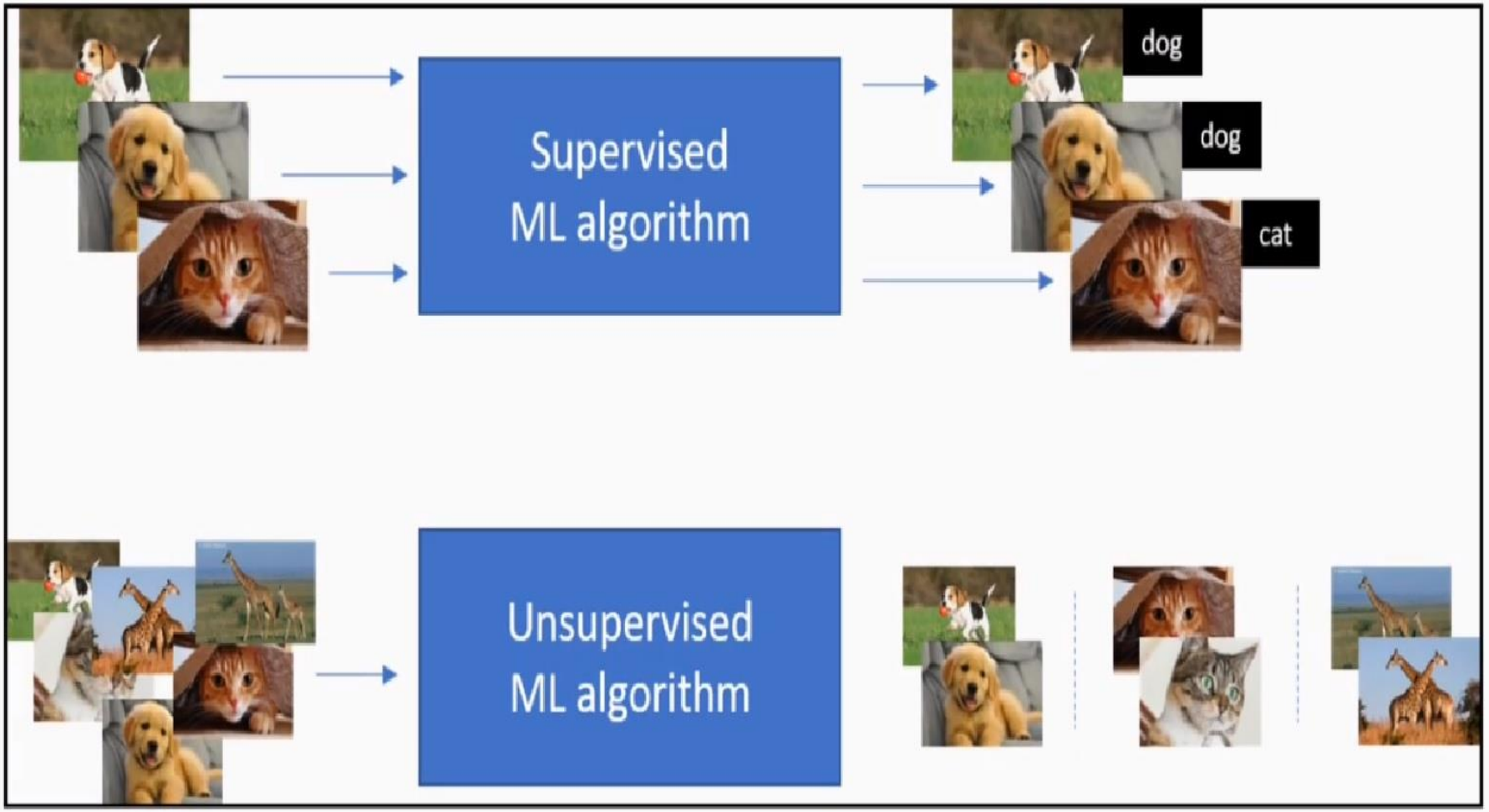
ما هو التعليم بدون إشراف Unsupervised ML

الفرق بين Supervised Vs Unsupervised :



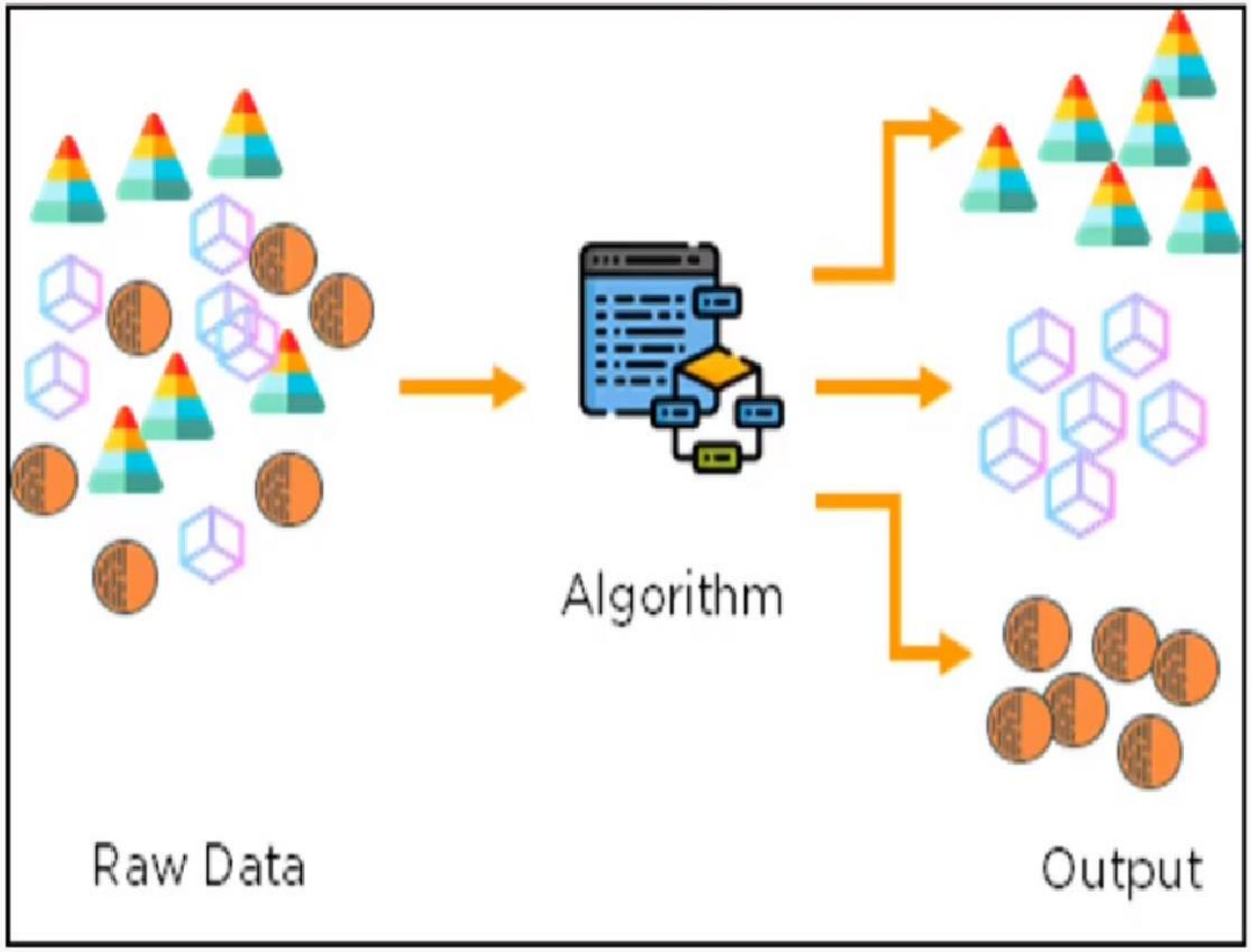
ما هو التعليم بدون إشراف Unsupervised ML

الفرق بين Supervised Vs Unsupervised :



ما هو التعليم بدون إشراف Unsupervised ML

الفرق بين Supervised Vs Unsupervised :



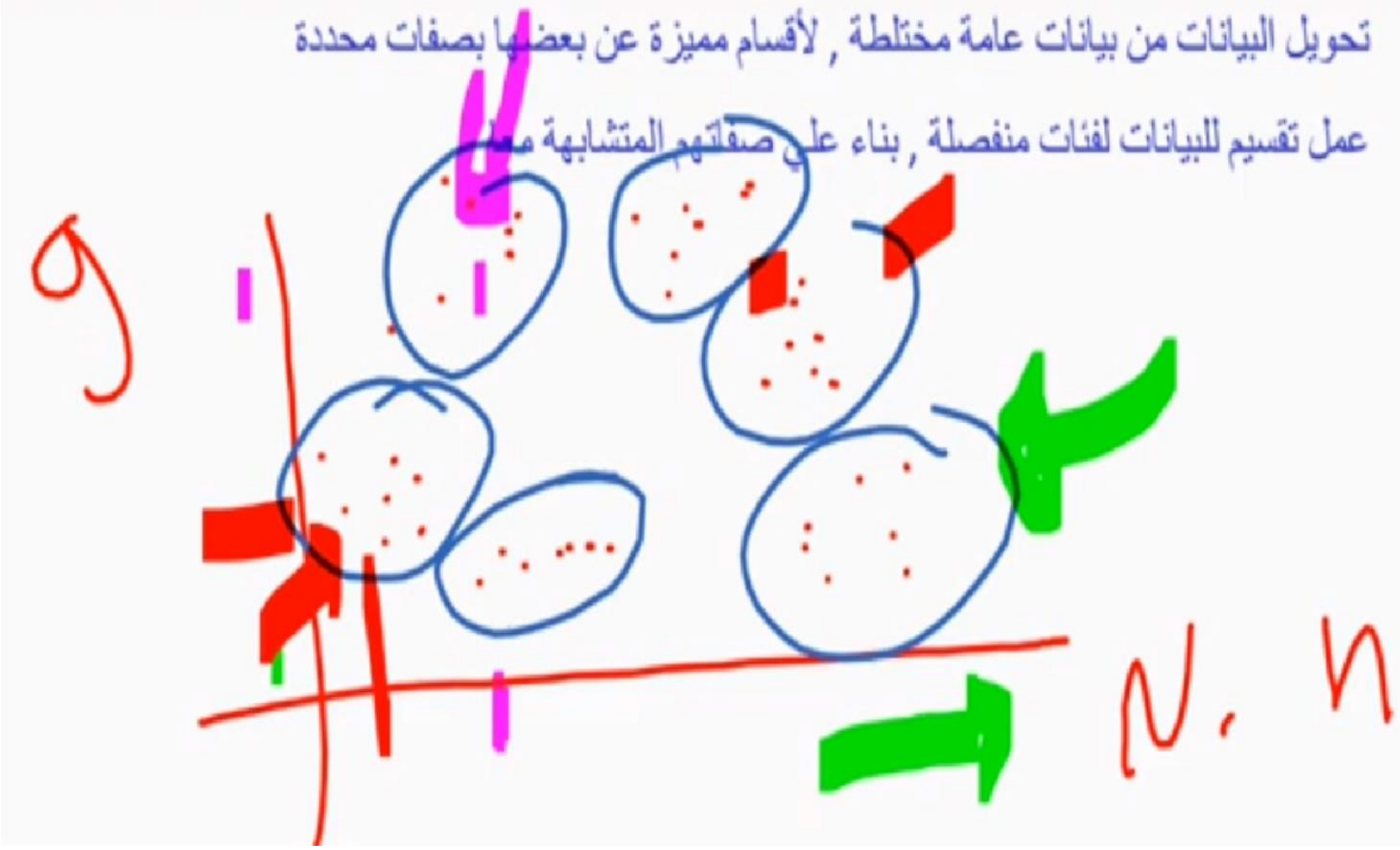
مفهوم الـ Clustering

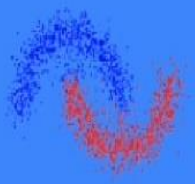
المعني :

- تحويل البيانات من بيانات عامة مختلطة , لأقسام مميزة عن بعضها بصفات محددة
- عمل تقسيم للبيانات لفئات منفصلة , بناء علي صفاتهم المتشابهة معا

المعني :

- تحويل البيانات من بيانات عامة مختلطة , لأقسام مميزة عن بعضها بصفات محددة
- عمل تقسيم للبيانات لفئات منفصلة , بناء على صفاتهم المتشابهة معا

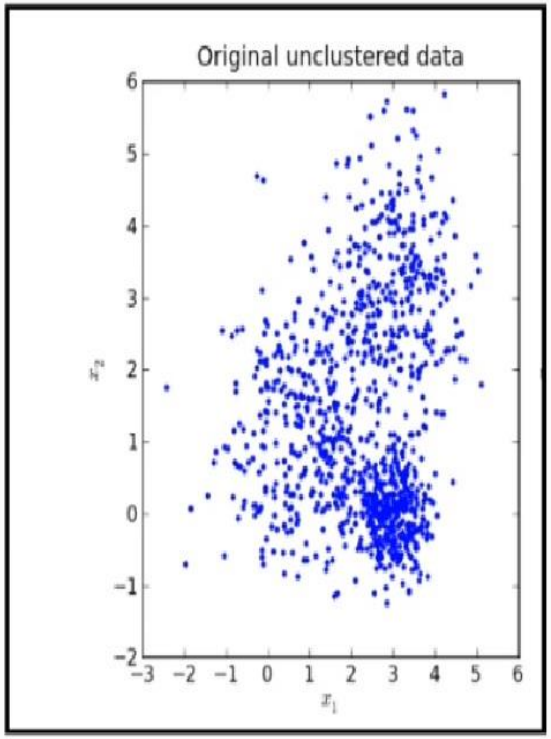


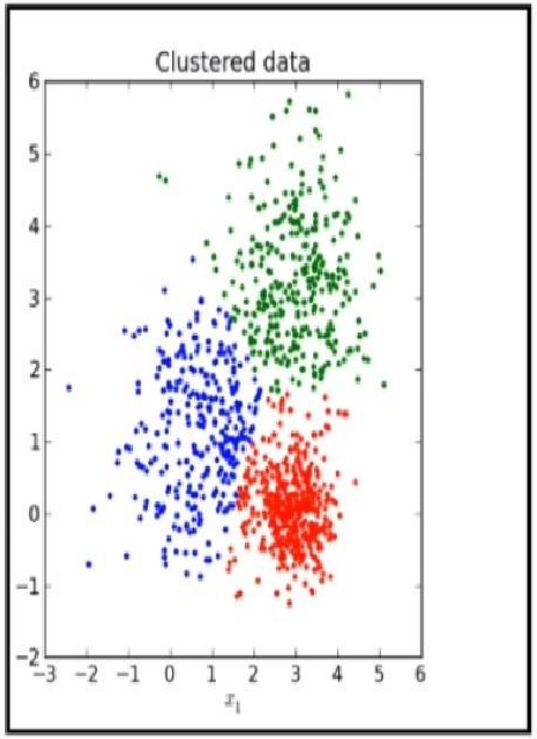


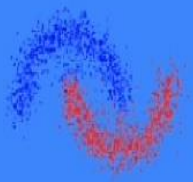
مفهوم الـ Clustering

المعني :

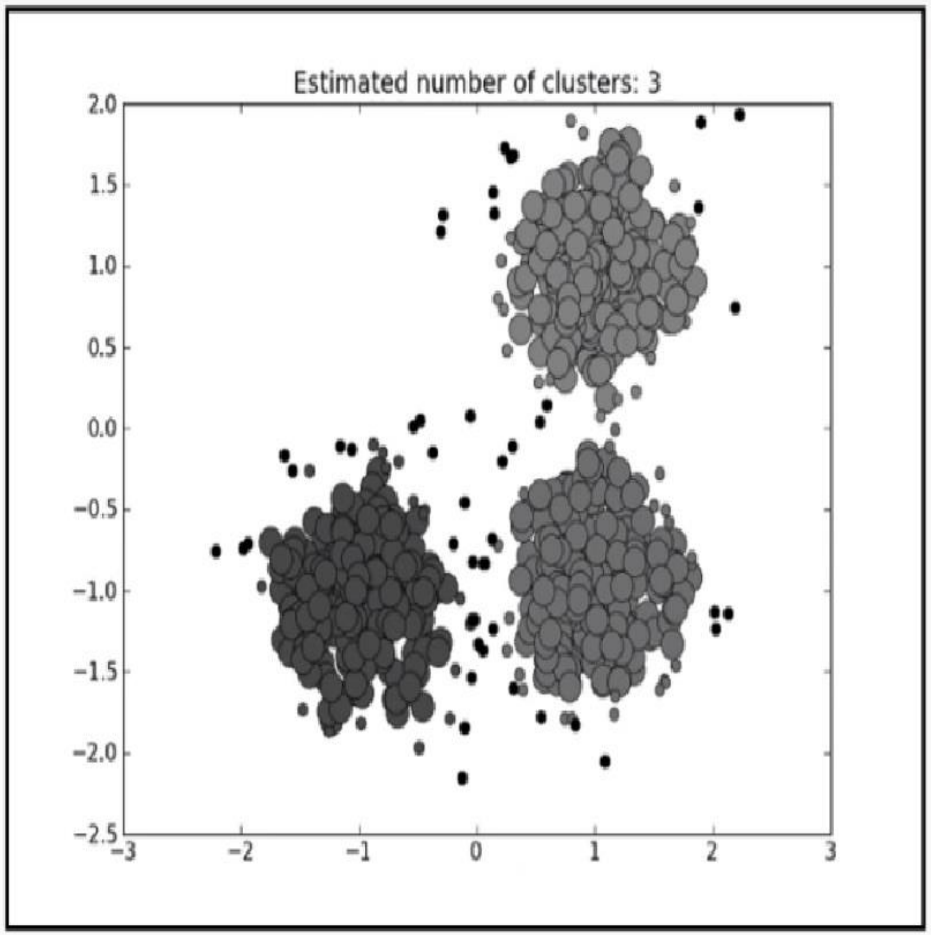
- تحويل البيانات من بيانات عامة مختلطة , لأقسام مميزة عن بعضها بصفات محددة
- عمل تقسيم للبيانات لفئات منفصلة , بناء علي صفاتهم المتشابهة معا

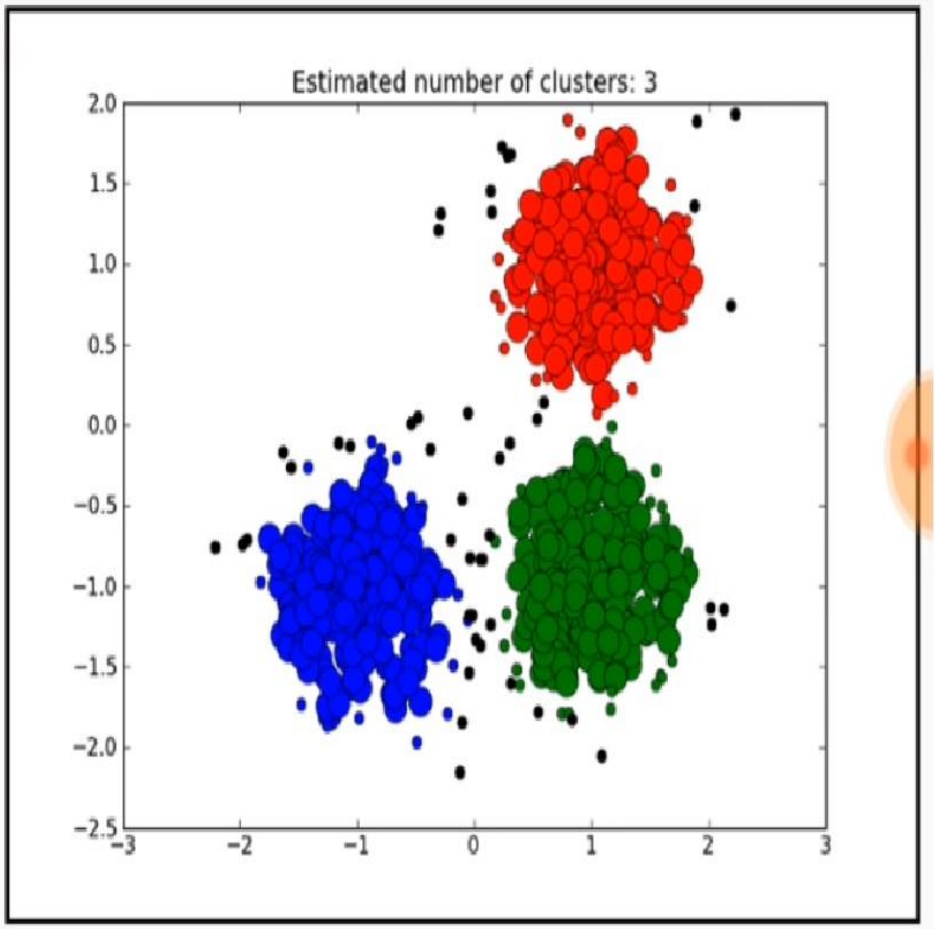




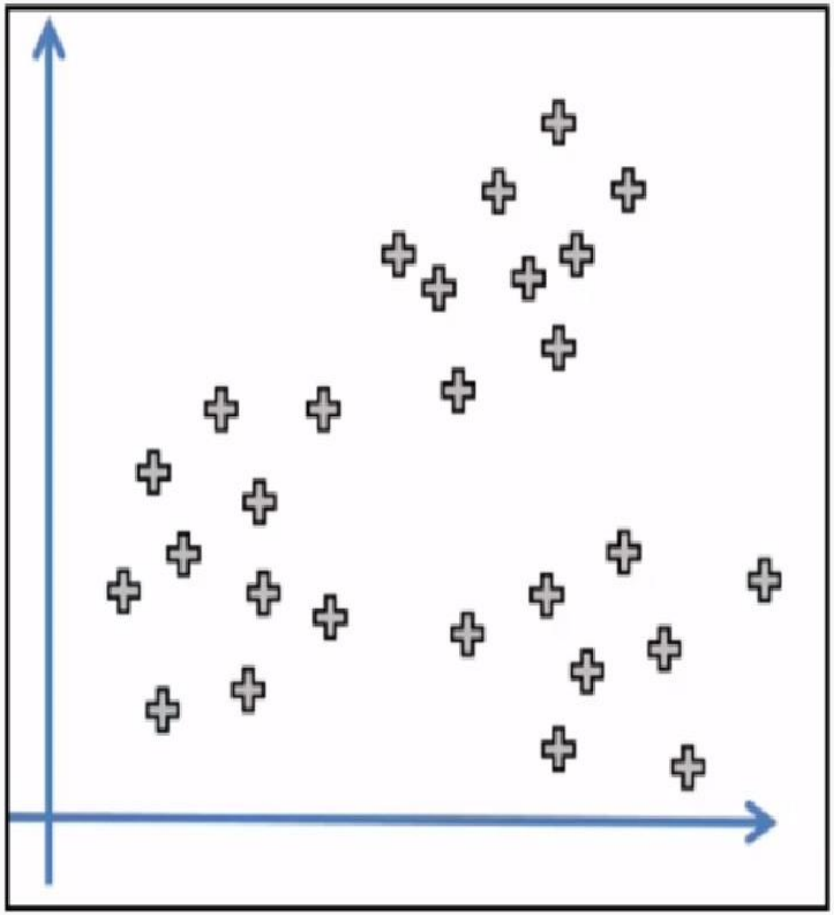


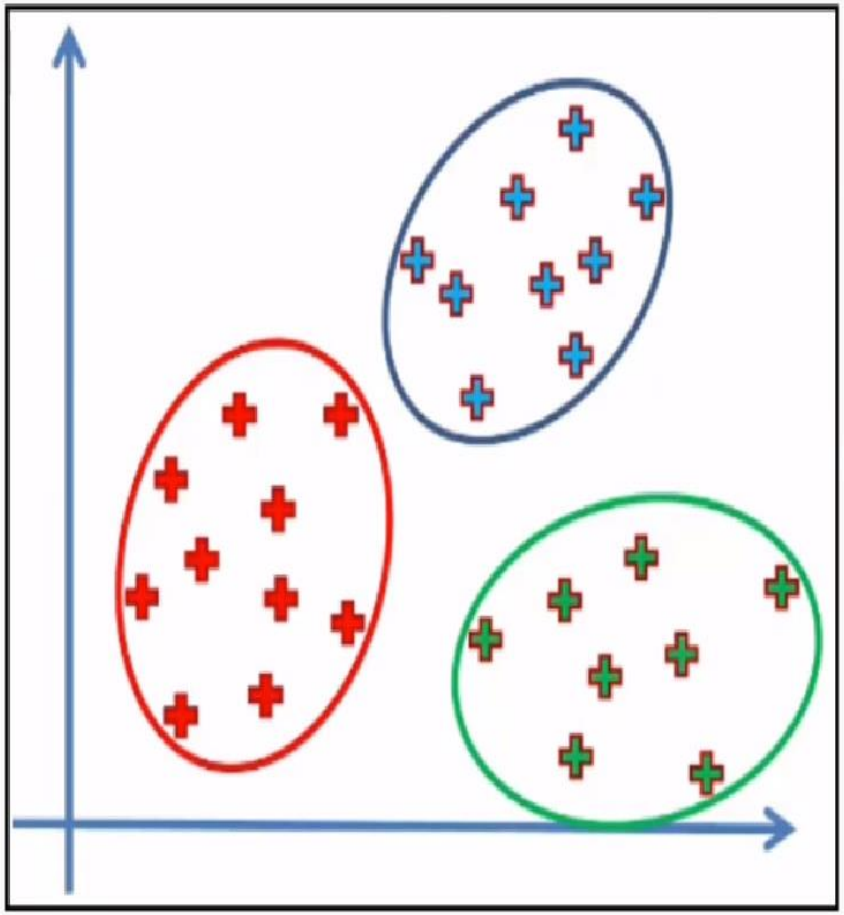
مفهوم الـ Clustering





مفهوم الـ Clustering



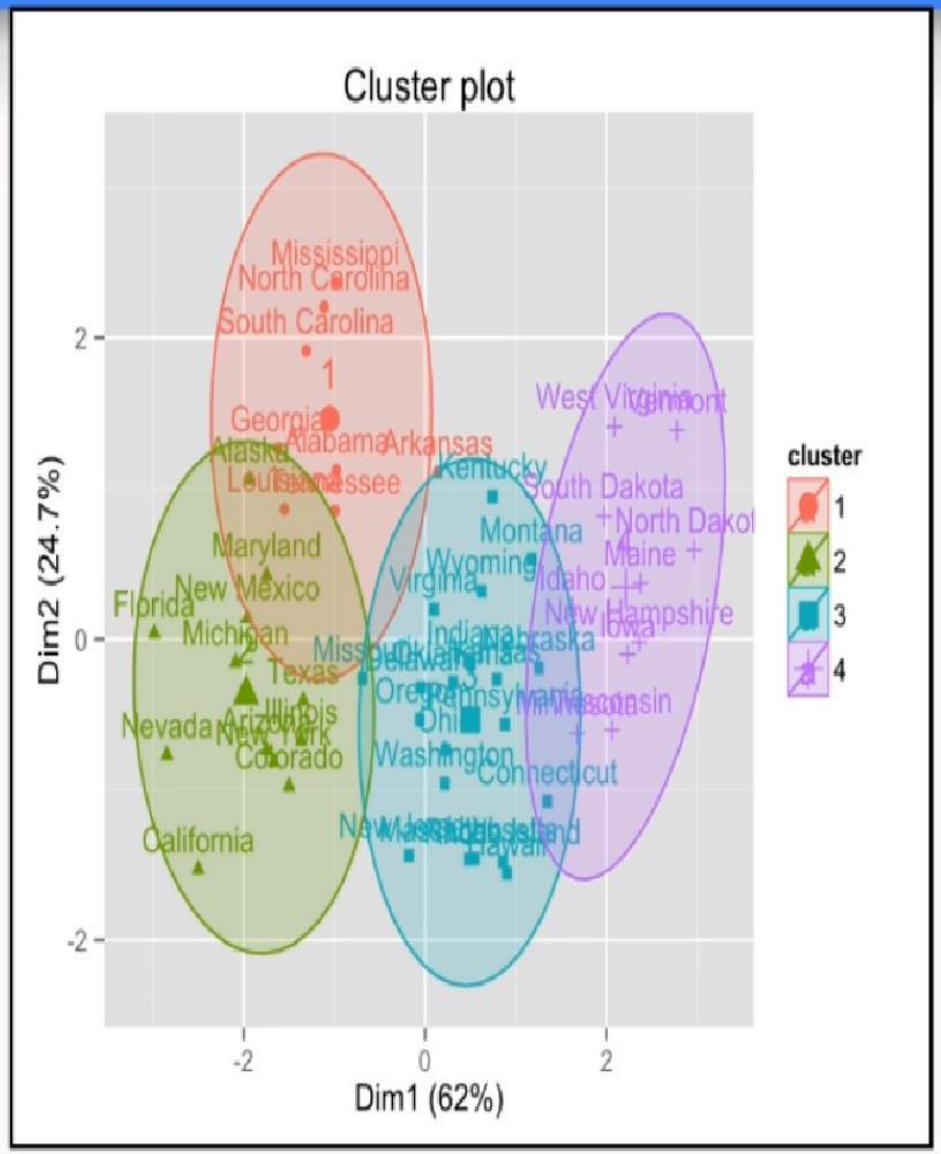


مفهوم الـ Clustering

تطبيق هام : تقسيم العملاء . .

مفهوم الـ Clustering

مبدأ التداخل



Segmentation as Clustering

- Agglomerative clustering
  - Start with each point in a separate cluster
  - At each iteration, merge two of the “closest” clusters
- Divisive clustering
  - Start with all points grouped into a single cluster
  - At each iteration, split the “largest” cluster

Segmentation as Clustering

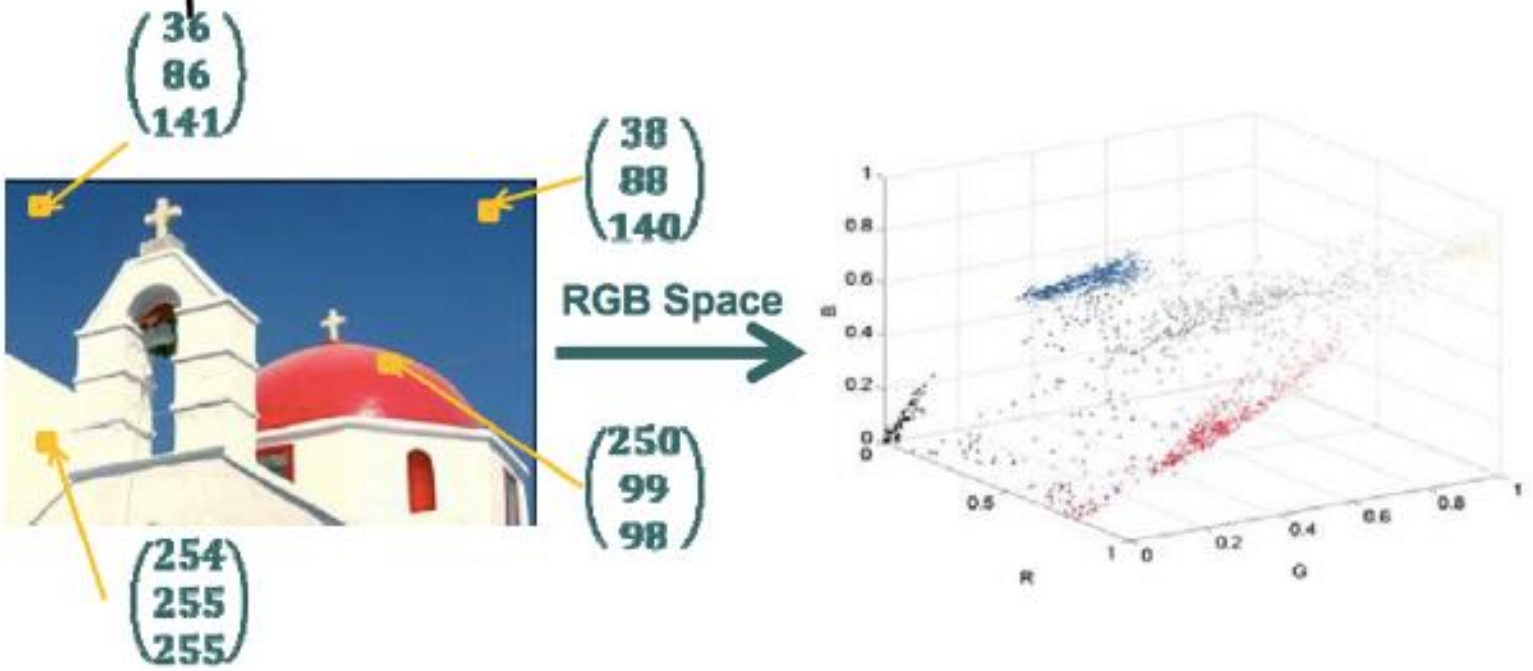
$$v_i = \begin{pmatrix} x_i \\ y_i \\ R_i \\ G_i \\ B_i \\ f_1(x_i, y_i) \\ \vdots \\ f_k(x_i, y_i) \end{pmatrix}$$

- Select a set of image features; position $\{x, y\}$, color $\{R, G, B\}$, a set of filter responses $\{f_1(x, y) \ldots f_k(x, y)\}$
- For each pixel p_i form a feature vector v_i

Features in Feature Space

- The vector represents a point in an N dimensional feature space.
- The feature vectors for similar pixels should occupy nearby locations in this feature space
- Thus, homogeneous image regions become dense clouds of feature vectors in feature space.

Example



- The feature space has 3 dimensions (RGB).
- Principal image regions generate dense clusters in feature space.

Spatial + RGB Space

- The feature space has 5 dimensions (XY-RGB).

K-Means Clustering

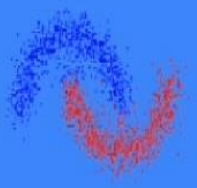
- Compute the feature space vectors
- Randomly select K cluster centers in feature space
- Iterate until convergence
  - Assign feature vectors to the closest cluster center
  - Re-compute the cluster centers as a (weighted) mean of the feature vectors assigned to each cluster
- Label pixels according to the cluster their feature vectors belong to

K-Means Clustering Sample

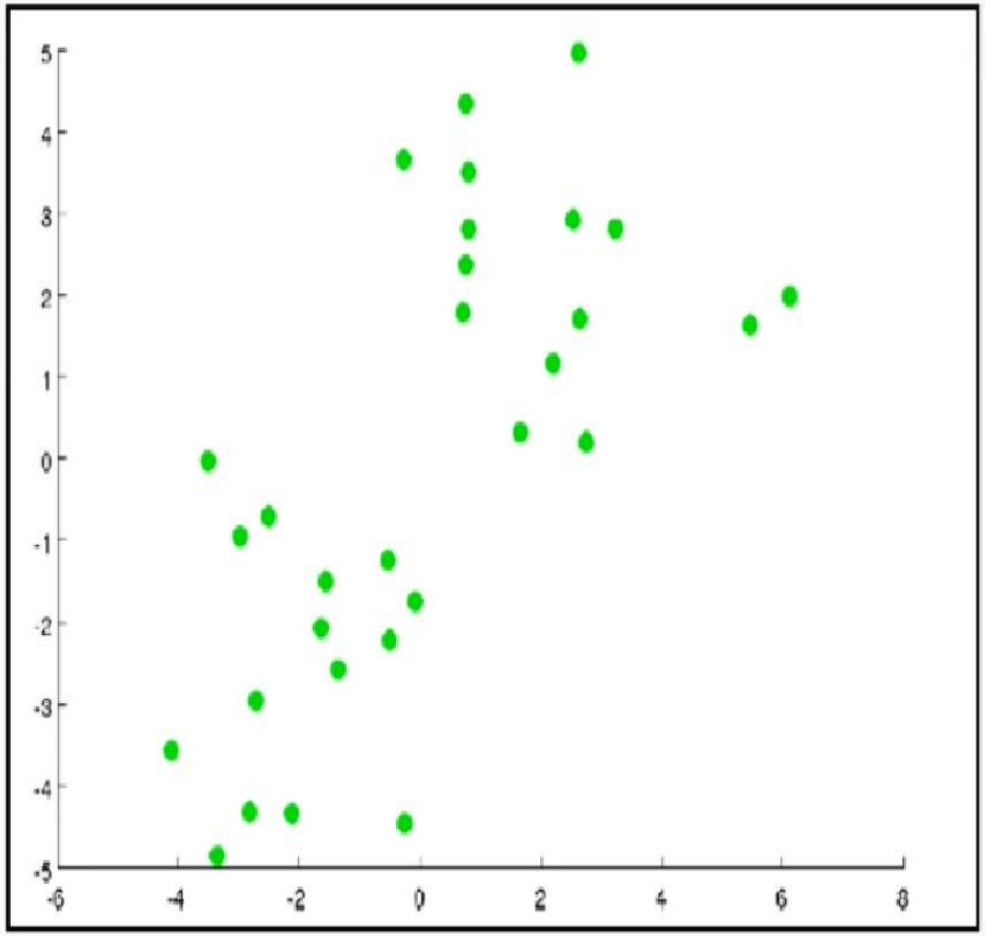
خوارزم K-means

المفهوم :

- هي طريقة لعمل تقسيم للبيانات الغير معنونة , unlabeled data
- يتم أولا تحديد عدد المجموعات المطلوب Clusters
- يقوم الخوارزم بتحديد عدد من النقاط العشوائية وسط النقاط تسمي cluster centroid , ويكون عددا هو نفس عدد المجموعات المطلوب
- ثم يقوم بتقسيم نقاط العينة عبر المراكز
- يقوم بعمل تغيير في اماكن المراكز , و يعيد الخطوة , حتي يصل للشكل الأمثل

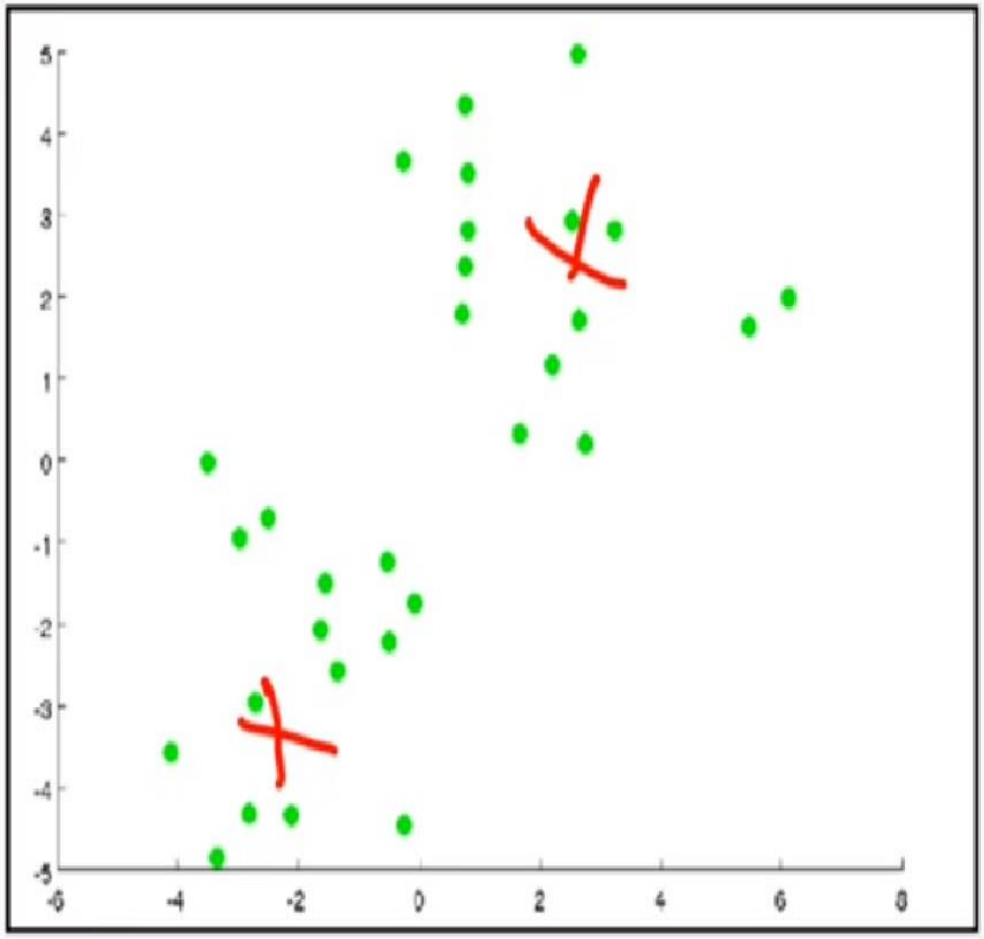


خوارزم K-means



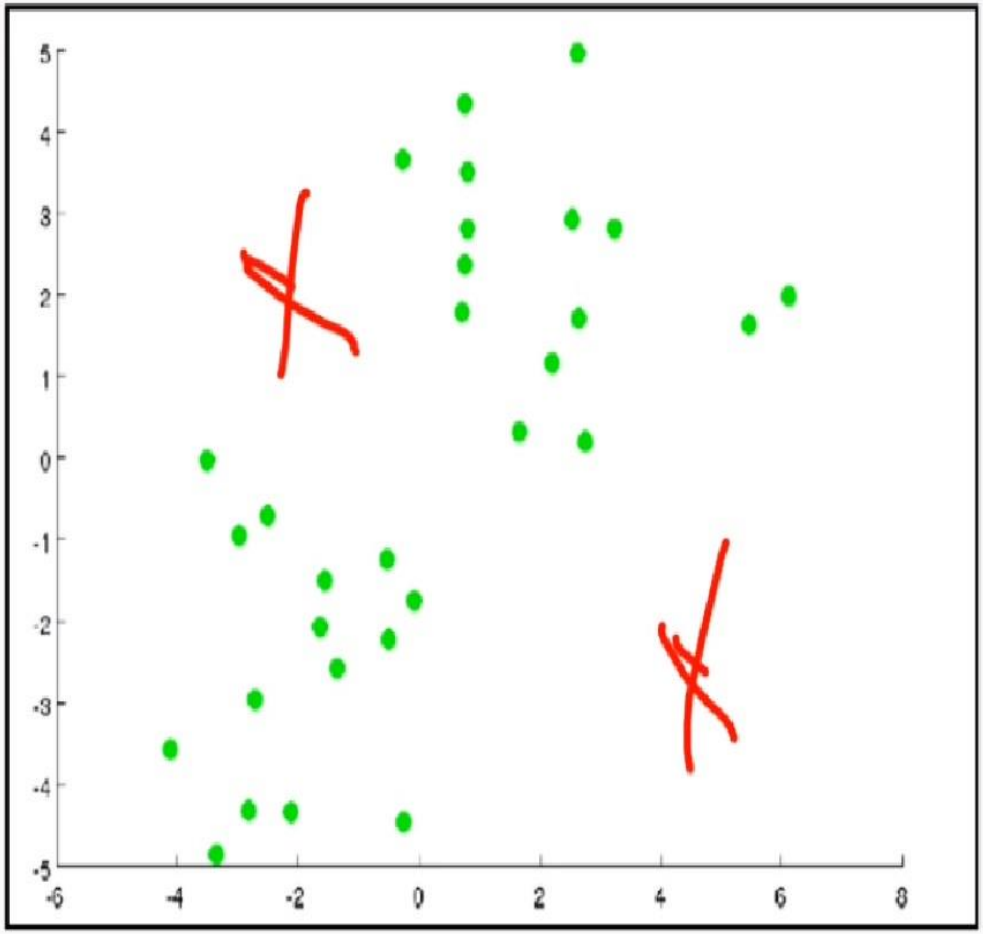
خطوات الـ K-means :

1 – عرض البيانات الغير معنونة , و تحديد عدد المجموعات , وليكن 2



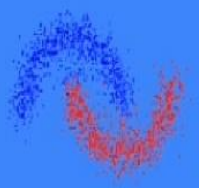
خطوات الـ K-means :

1 – عرض البيانات الغير معنونة , و تحديد عدد المجموعات , وليكن 2



خطوات الـ K-means :

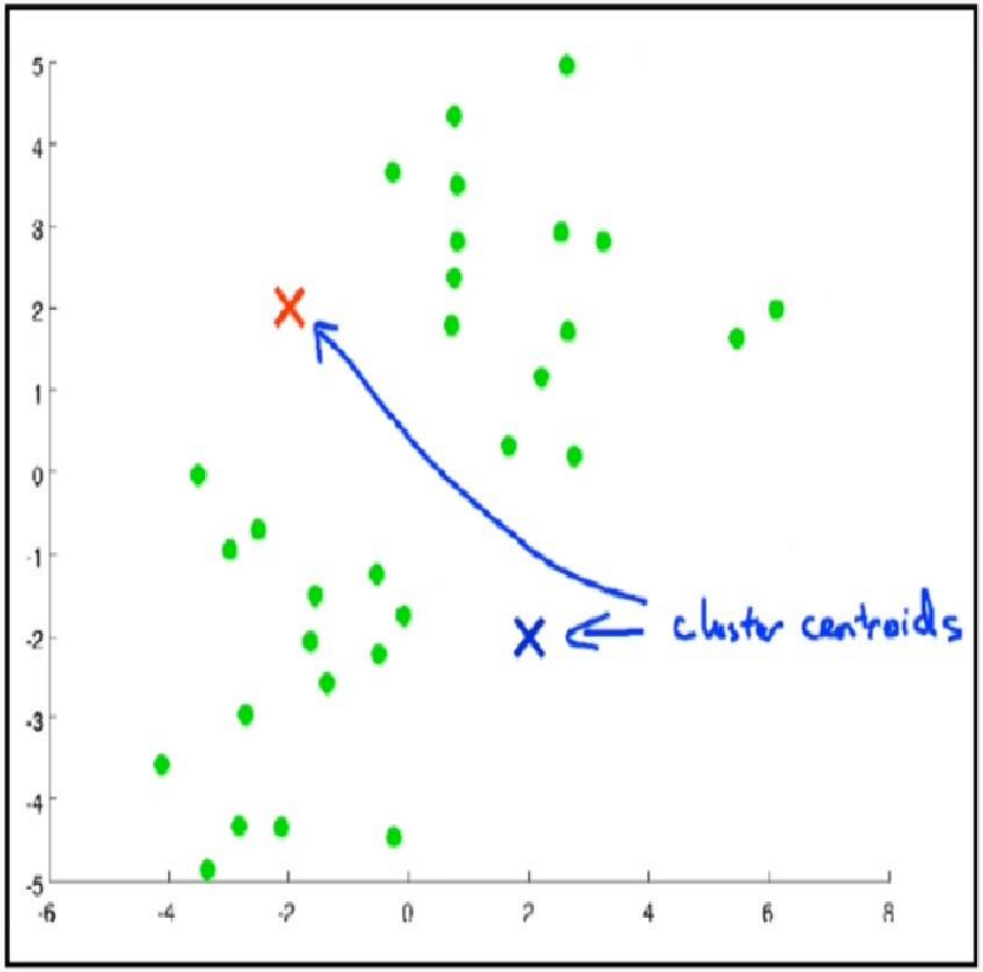
1 – عرض البيانات الغير معنونة , و تحديد عدد المجموعات , وليكن 2

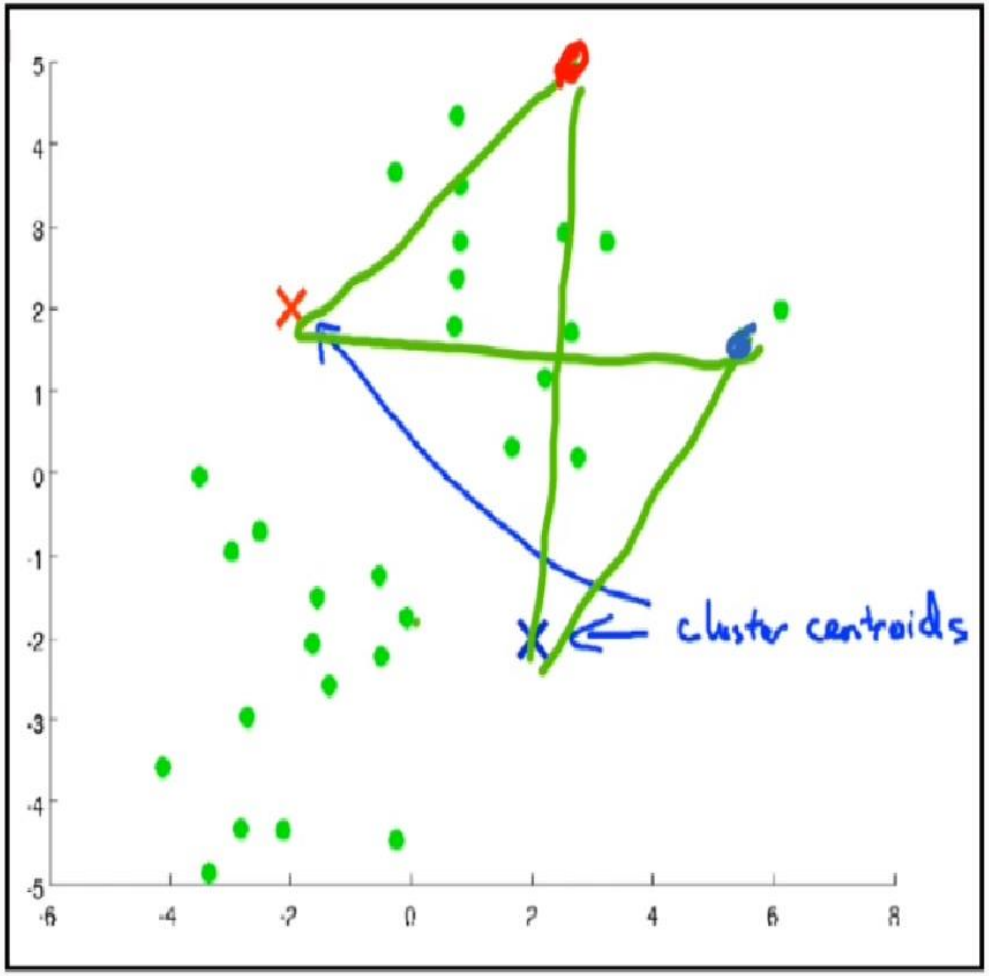


خوارزم K-means

خطوات الـ K-means :

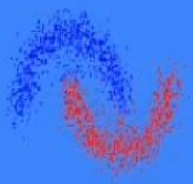
2 – تحديد نقطتين عشوائيتين كمركزين للمجمعتين , cluster centroids





خطوات الـ K-means :

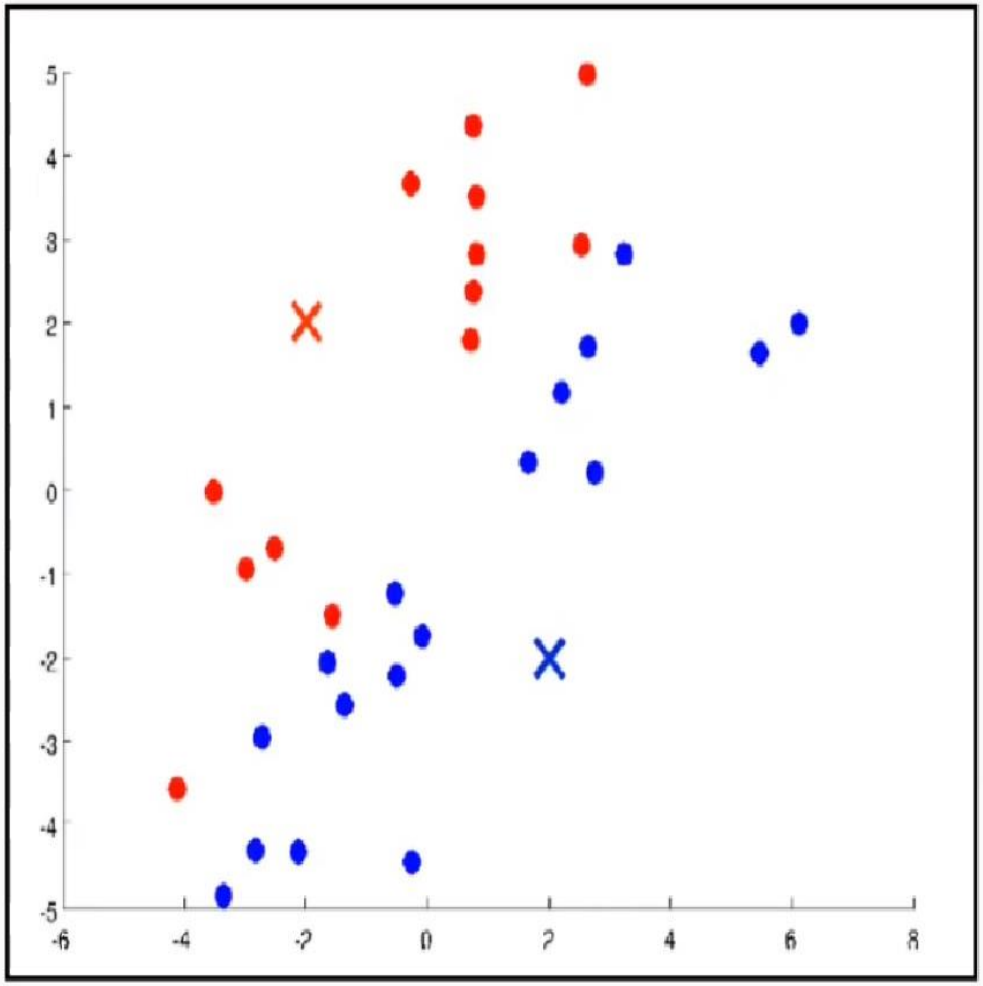
2 – تحديد نقطتين عشوائيتين كمركزين للمجمعتين , cluster centroids

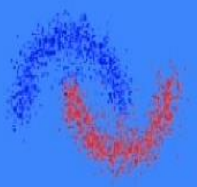


خوارزم K-means

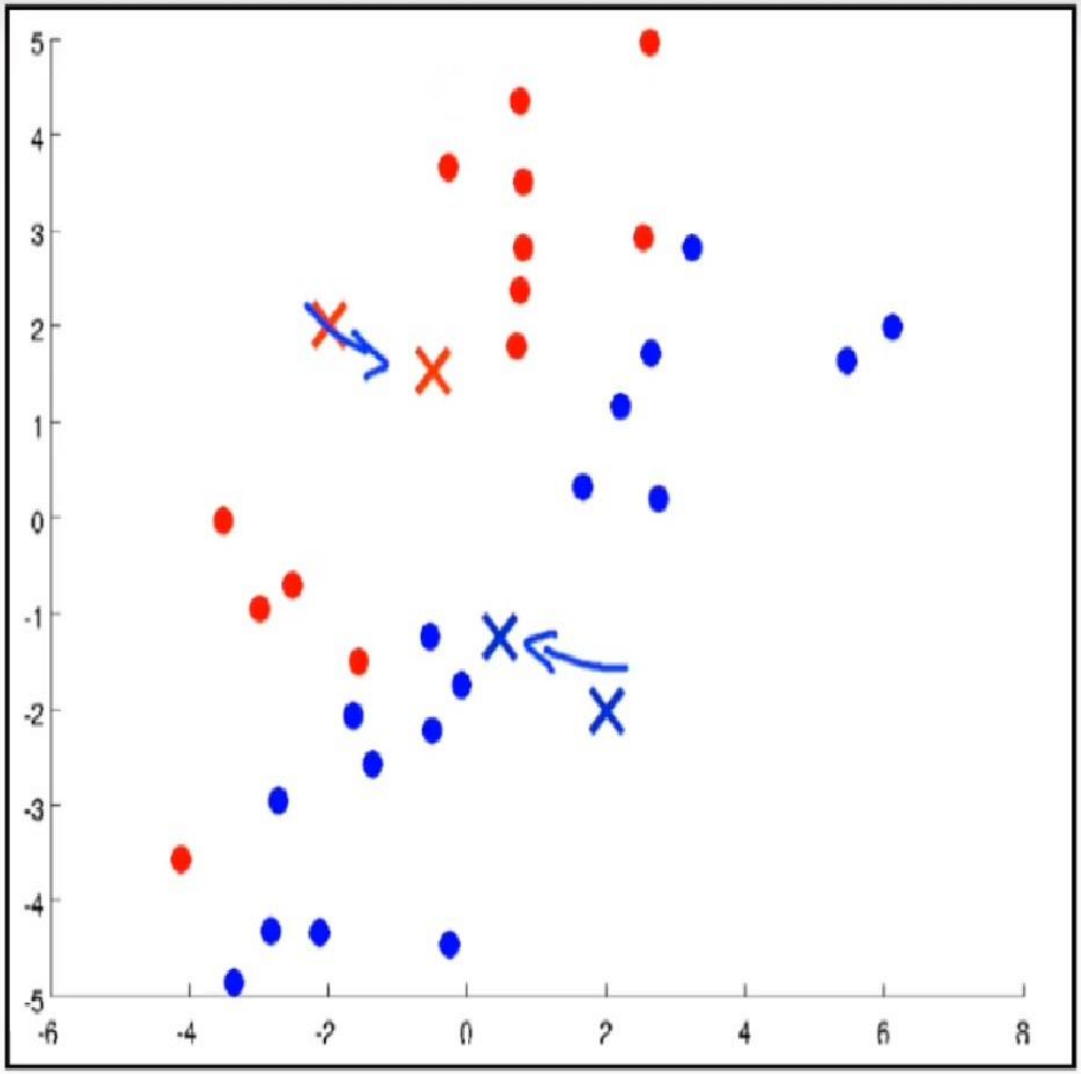
خطوات الـ K-means :

3 – قياس المسافة بين كل نقطة , وبين المركزين , وتحديد كل نقطة من نقاط العينة تابعة لأقرب مركز لها



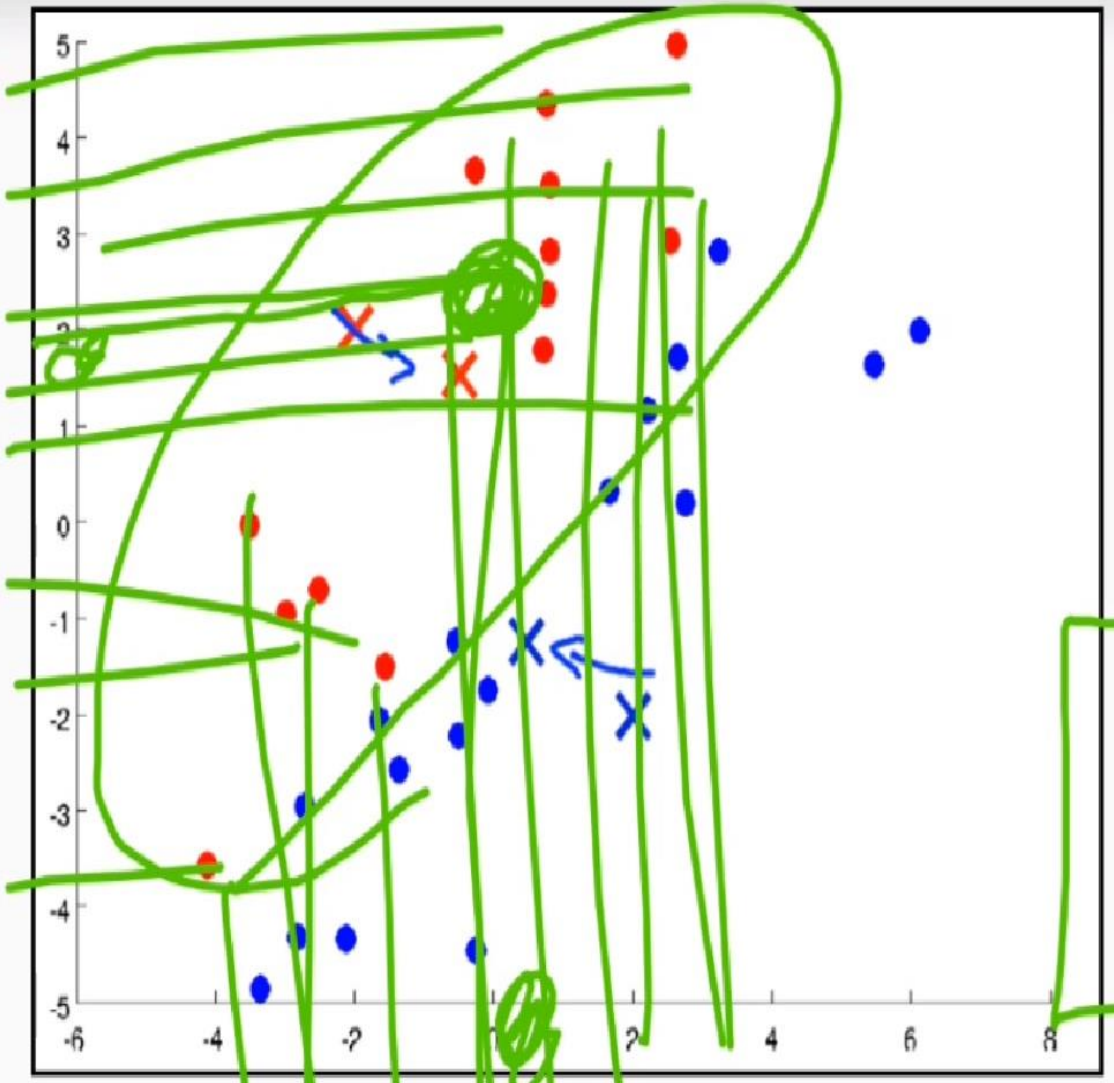


خوارزم K-means



خطوات الـ K-means :

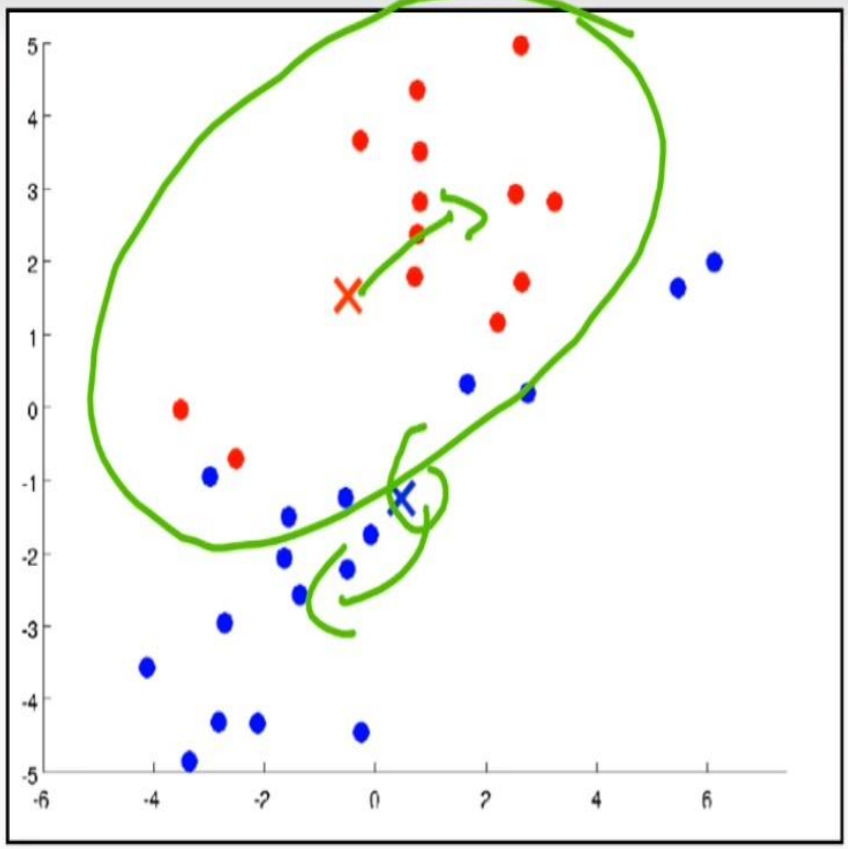
4 – تحريك كل مركز فيهم , إلي قلب نقاط مجموعته



خطوات الـ K-means :

4 – تحريك كل مركز فيهم , إلي قلب نقاط مجموعته

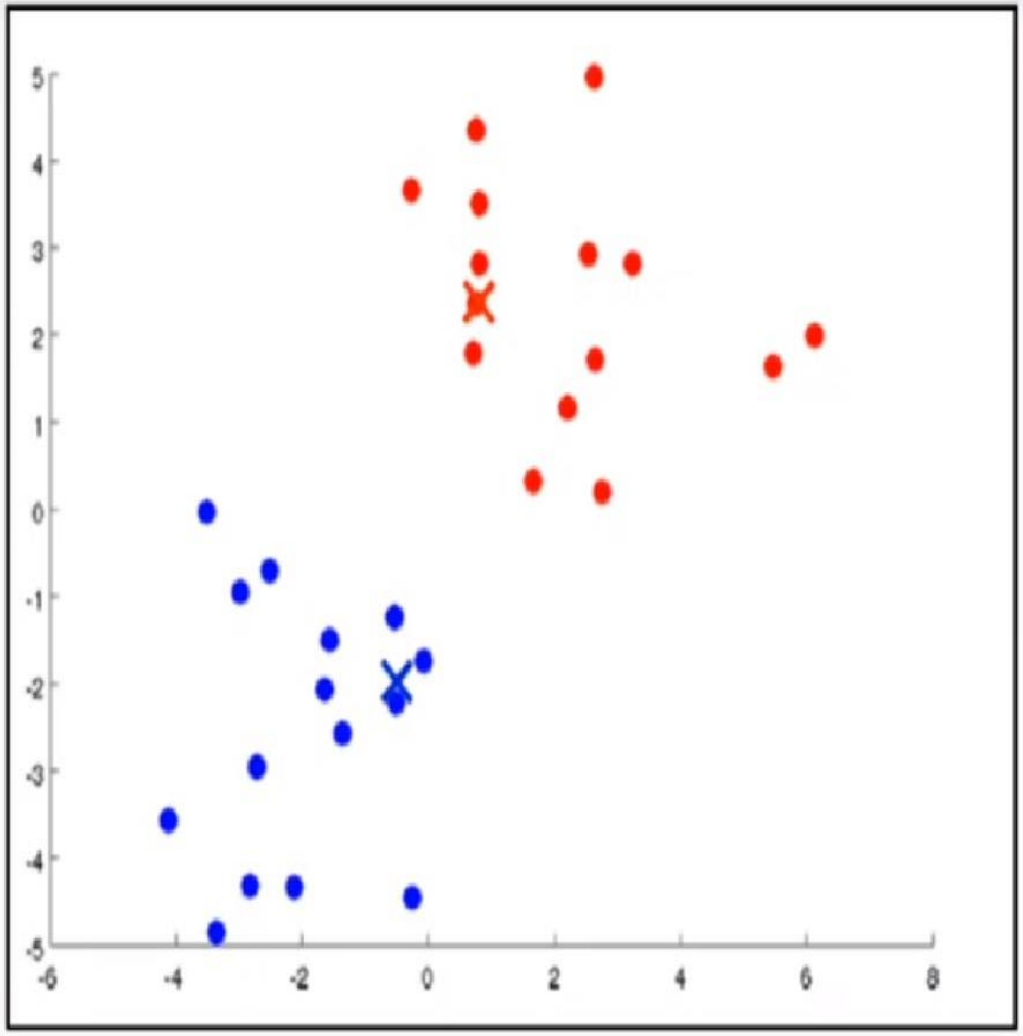
خوارزم K-means



خطوات الـ K-means :

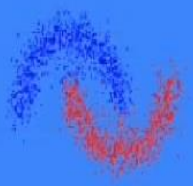
5 – إعادة الخطوة الثالثة , في جعل كل نقطة تابعة للمركز الأقرب لها

خوارزم K-means



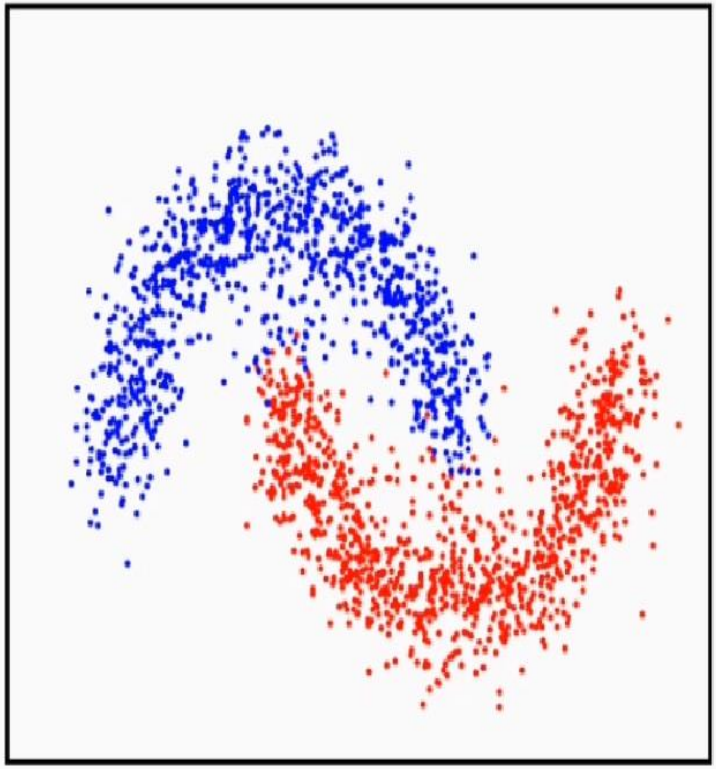
خطوات الـ K-means :

7 – تكرار خطوتي (ازاحة المركز + تقسيم النقاط) عدد من المرات , حتي نصل للتقسيم المثالي



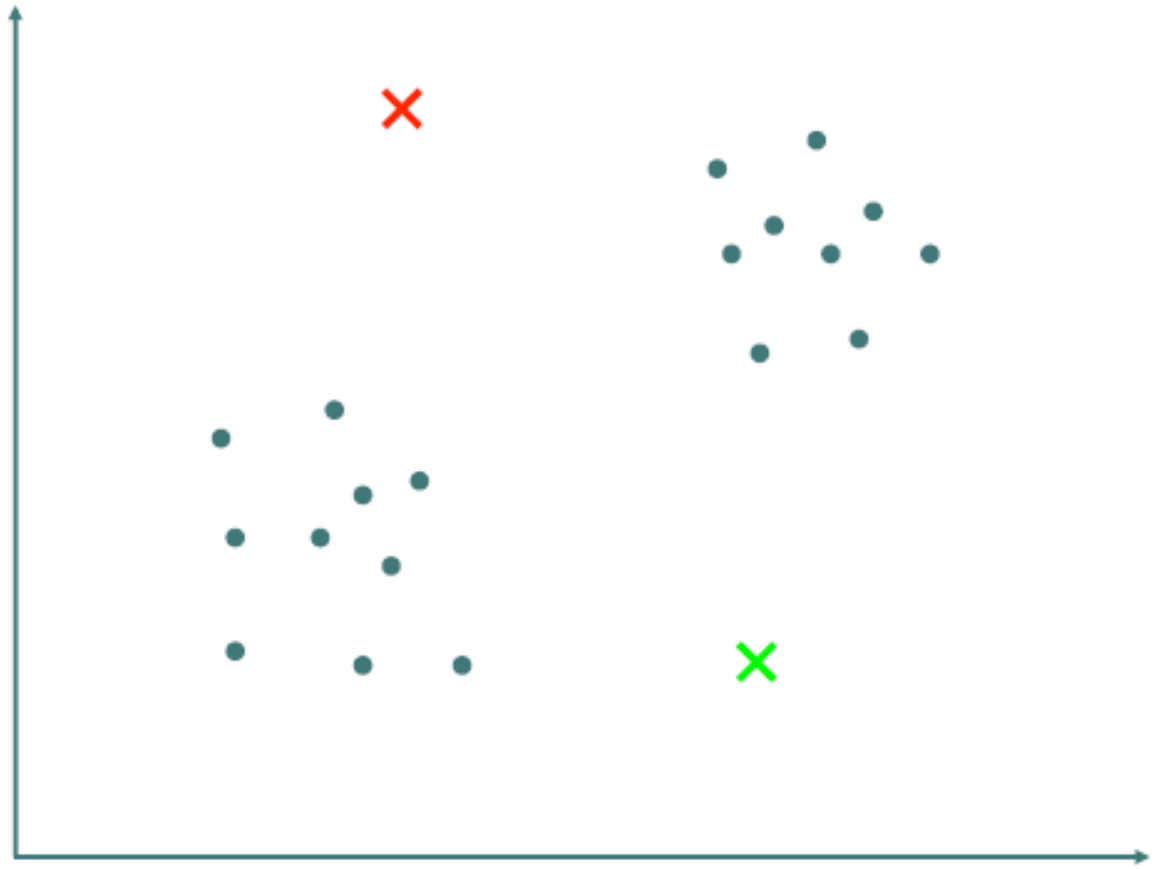
خوارزم K-means

مع العلم أن :

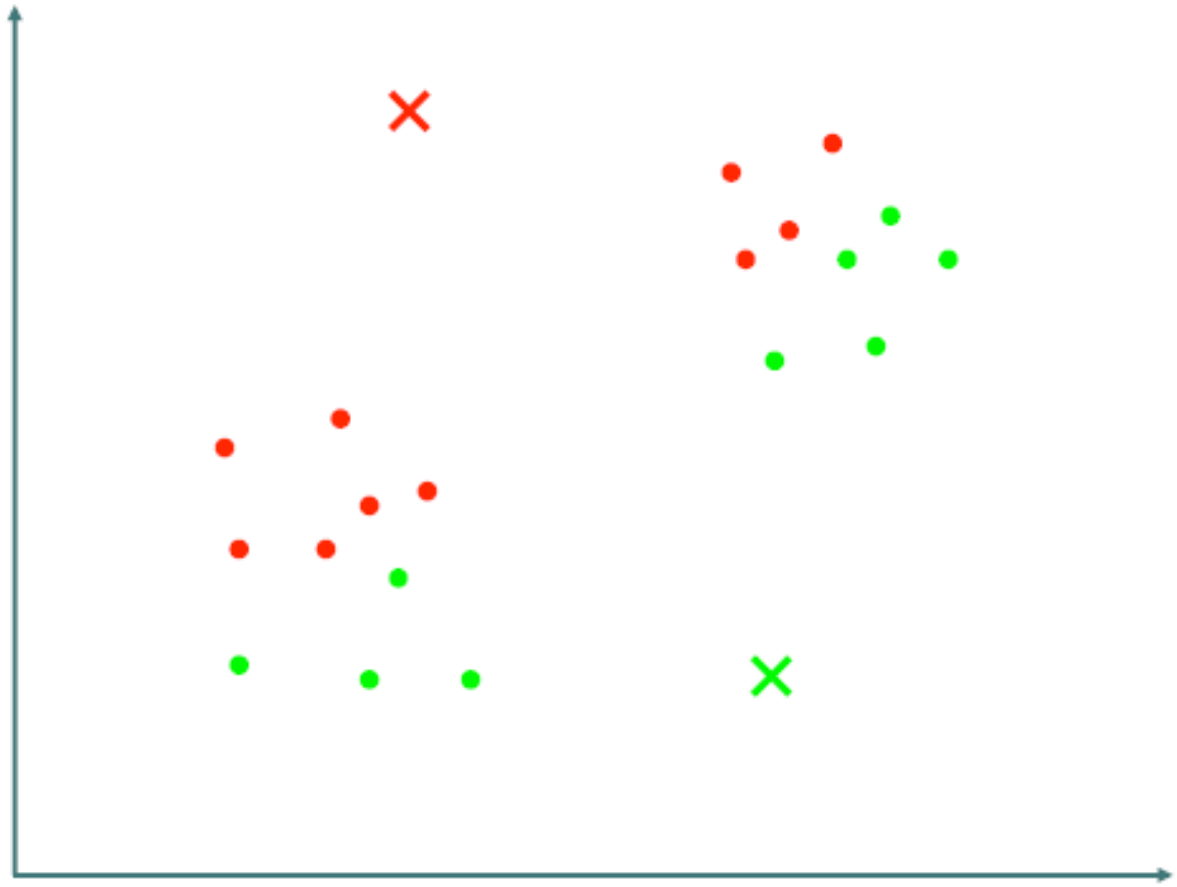


- يمكن أن ينطبق هذا الأمر علي أكثر من عنقودين , ممكن أي عدد
- يمكن أن أقوم أنا بتحديد عدد العناقيد , أو أن أجعل الخوارزم هو الذي يقوم بتحديد العدد المناسب حسب مدي تجانس البيانات
- يتم الأمر عبر طبيق خطوتين تباعا هي : الإختيار + الإزاحة

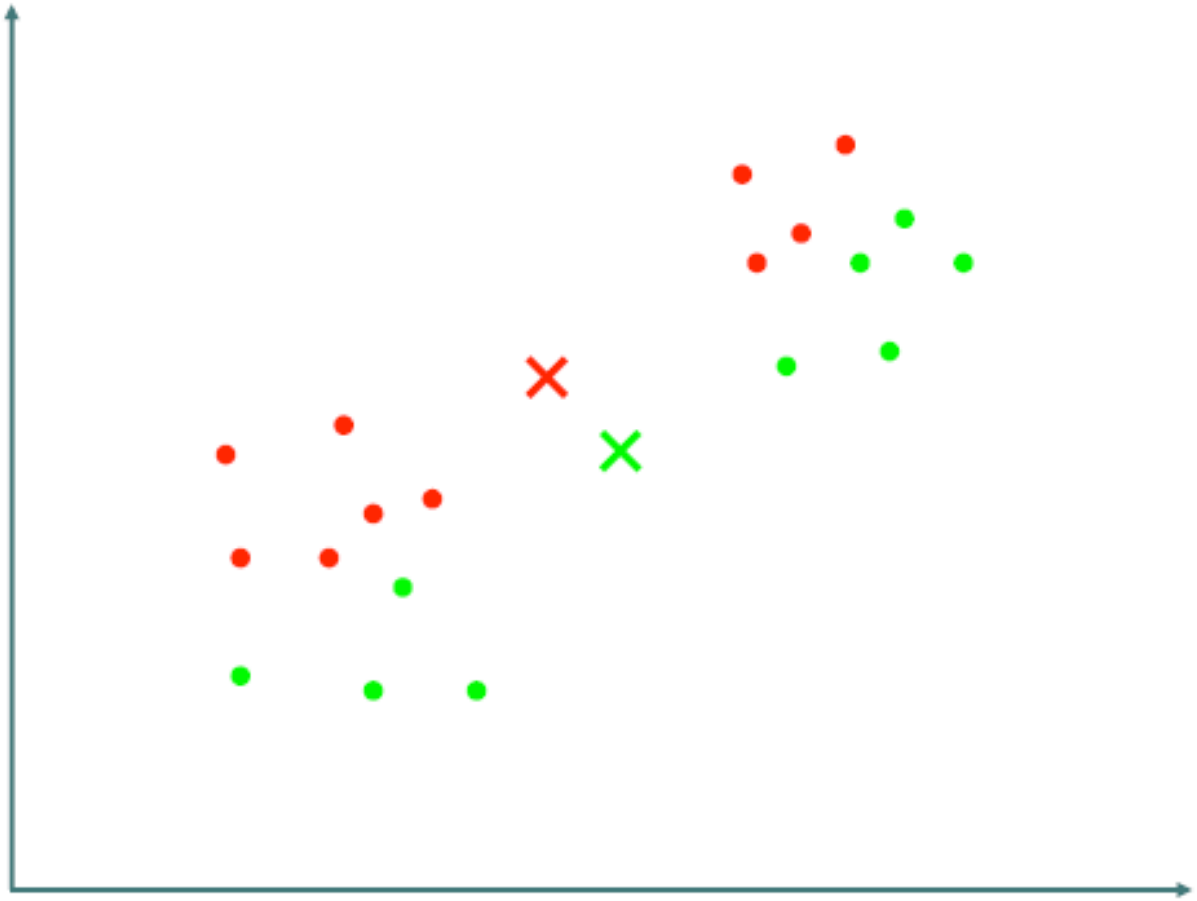
K-Means Clustering



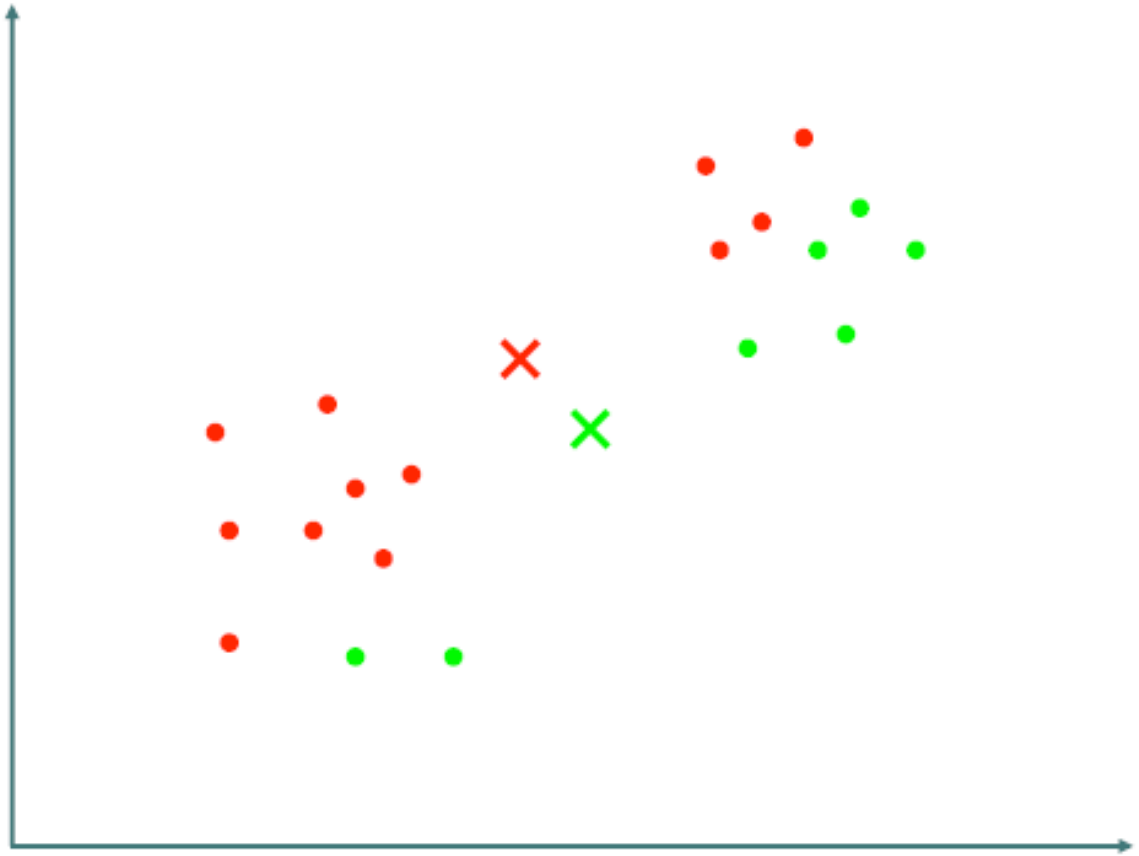
K-Means Clustering



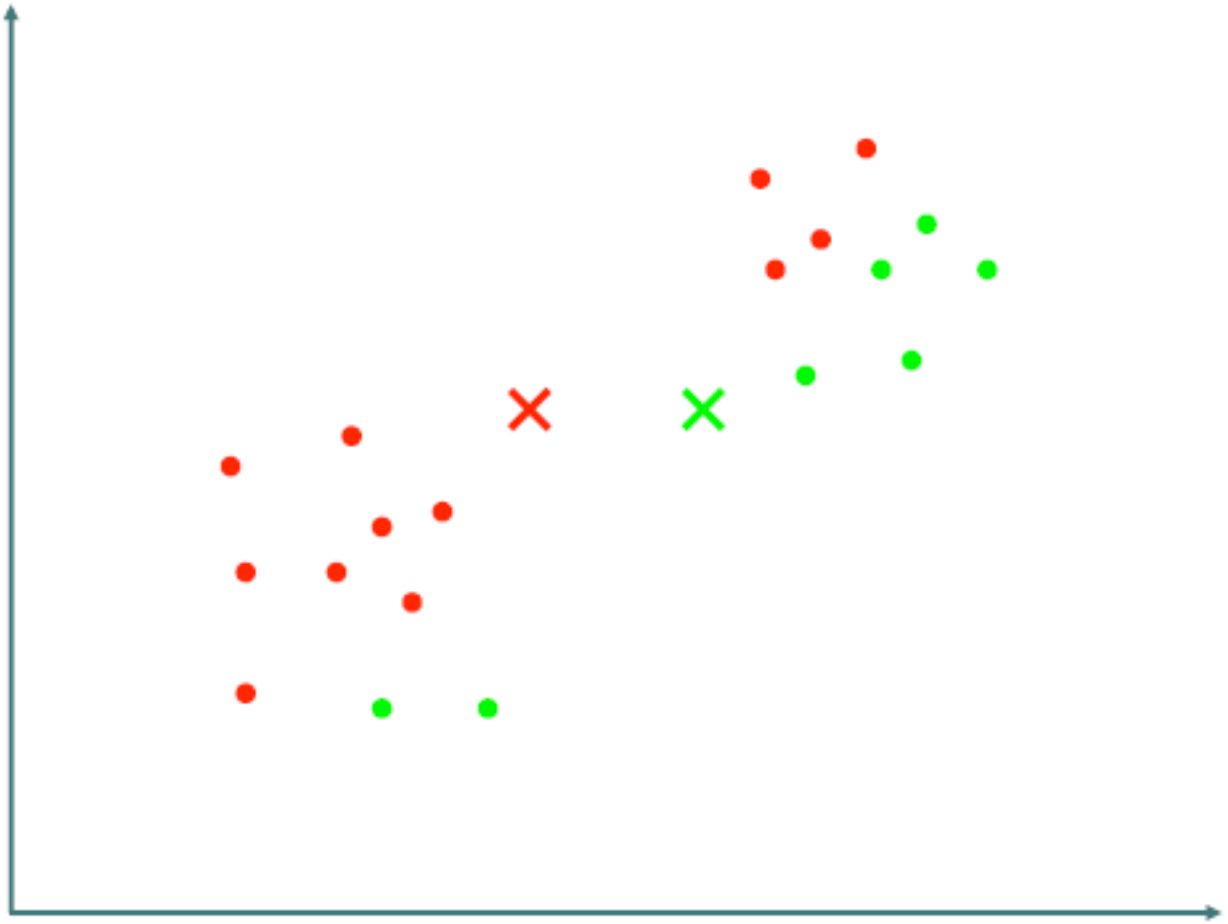
K-Means Clustering



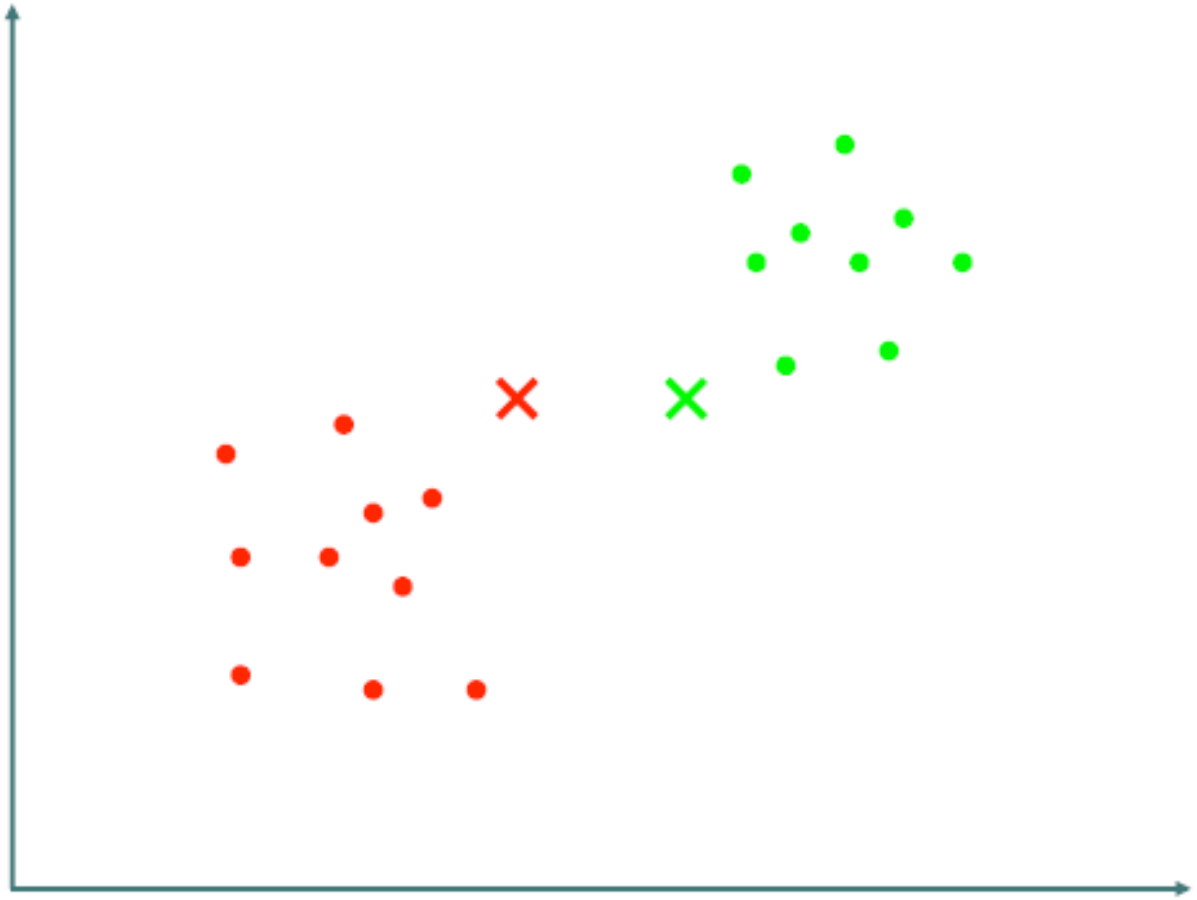
K-Means Clustering



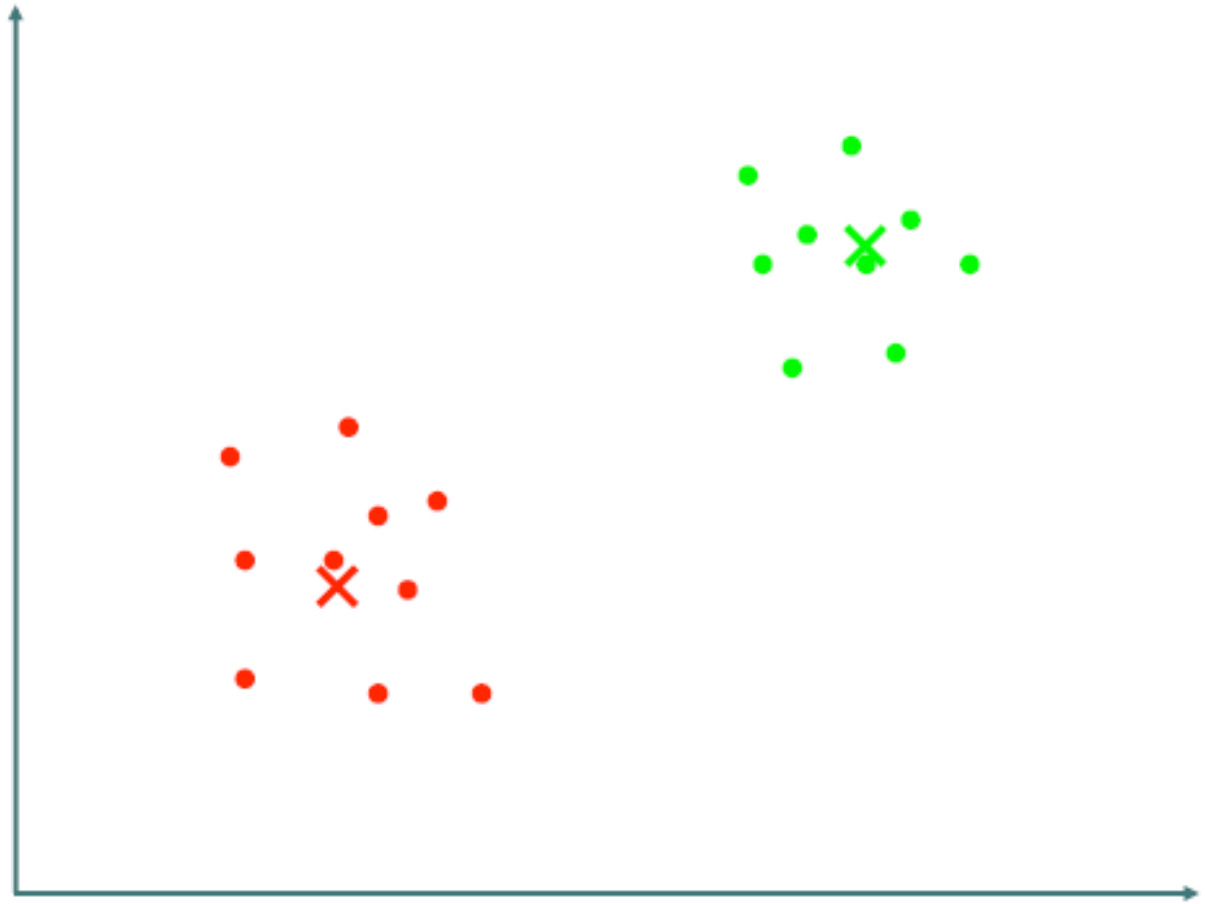
K-Means Clustering



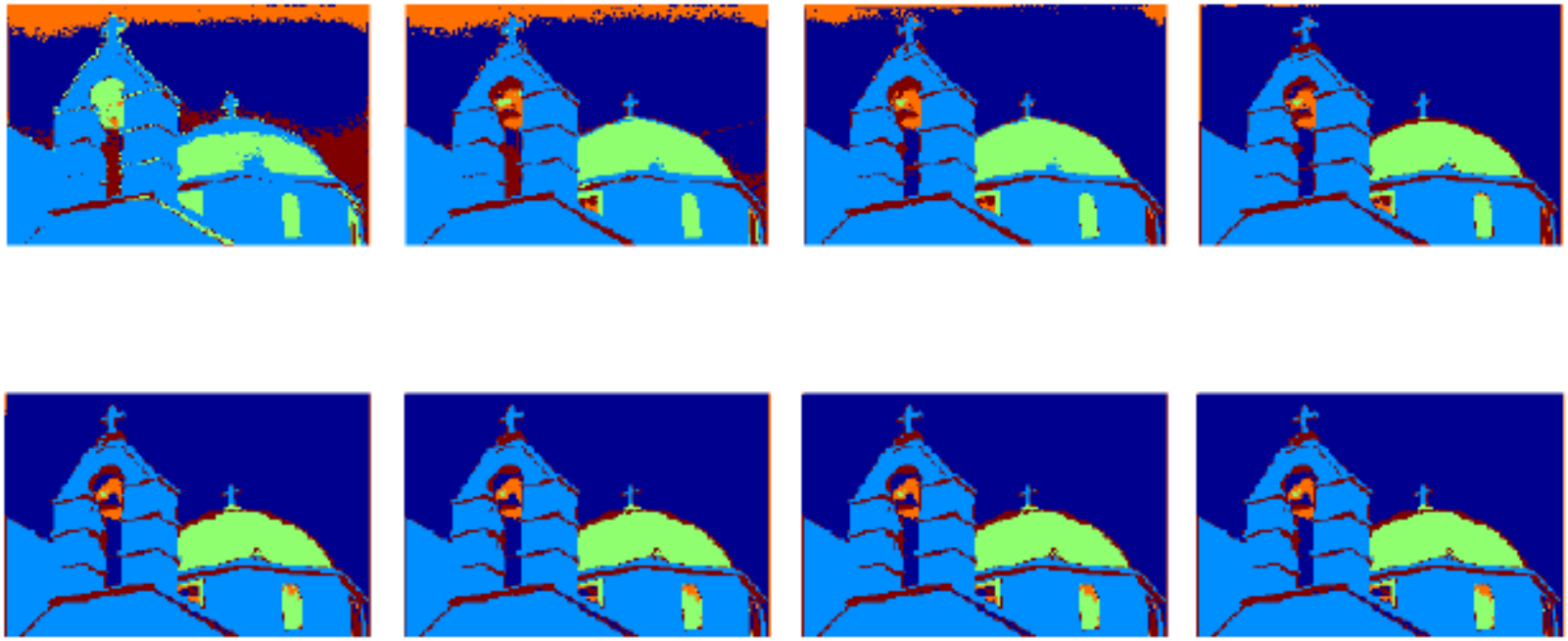
K-Means Clustering



K-Means Clustering



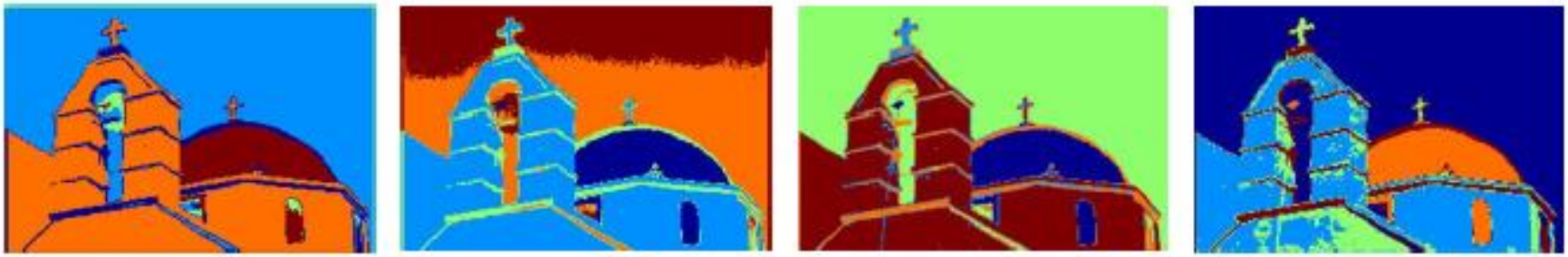
Sample Iterations



8 iterations of the K means procedure, K=5

Parameter Selection

Effect of random initialization, K=5



Effect of the choice K

K=3 K=5 K=8 K=15

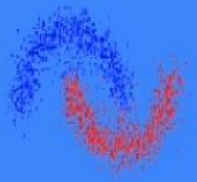
K-Means pros and cons

Pros

- Simple and fast
- Converges to a local minimum of the error function

Cons

- Need to pick K
- Sensitive to initialization
- Sensitive to outliers
- Only finds “spherical” clusters



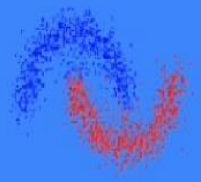
خوارزم K-means

الإختيار :

- يقصد بها تحديد كل نقطة هي تتبع أي مركز فيهم
- يتم هذا عبر إيجاد قيمة الـ norm لكل نقطة , مع كل مركز فيهم , واختيار الـ norm الأقل
- الـ norm يعبر عنها بالمسافة بين النقطة و المركز
- يرمز لها بالرمز ميو وتأخذ قيمة k وهي عدد العناقيد , فتكون ميو1 , ميو 2 , وهكذا , إشارة لمسافة النقطة مع μ_kز الأول و الثاني و هكذا

الإختيار :

- يشار لكل مركز لرقم k سمول , بينما عدد المراكز (عدد العناقيد) K كابيتال
- يرمز للـ norm بالرمز C و يكون له رقم , أي .. C1 , C2
- المعادلة المستخدمة هي : $C^i = ||X^i - \mu_k||$
- أحيانا يتم عمل تربيع للمعادلة , وبنفس التأثير



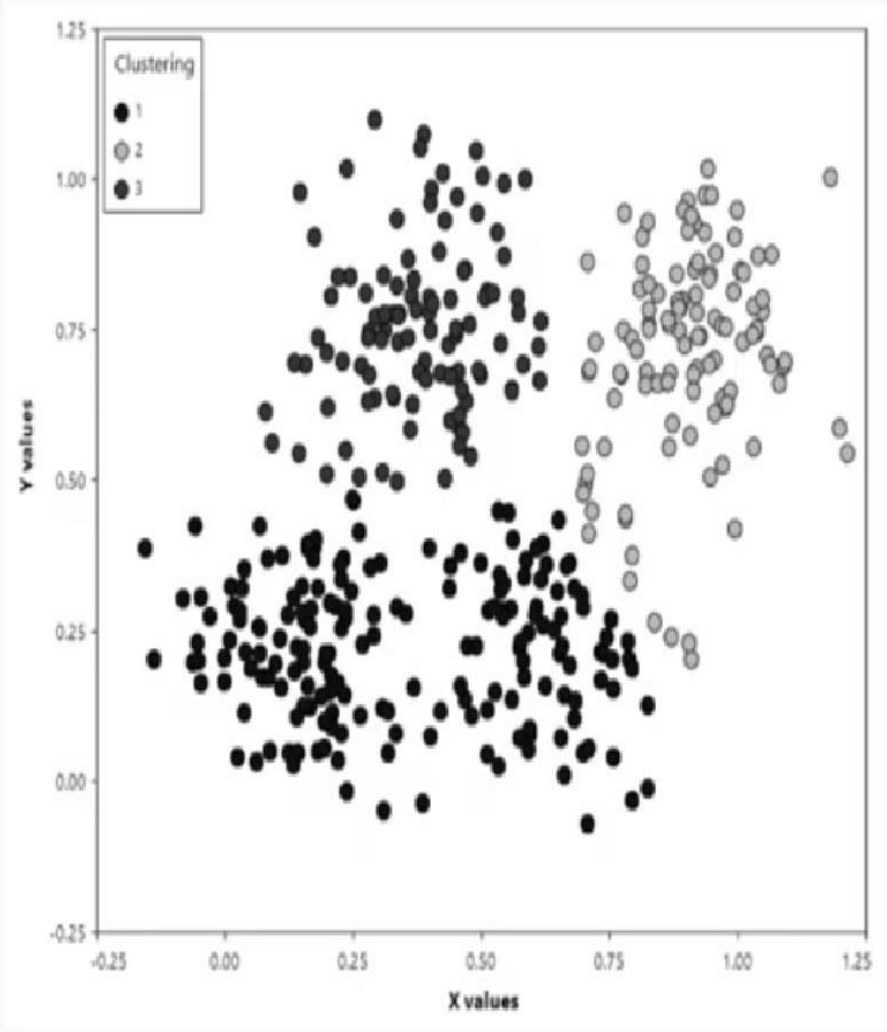
خوارزم K-means

الإزاحة :

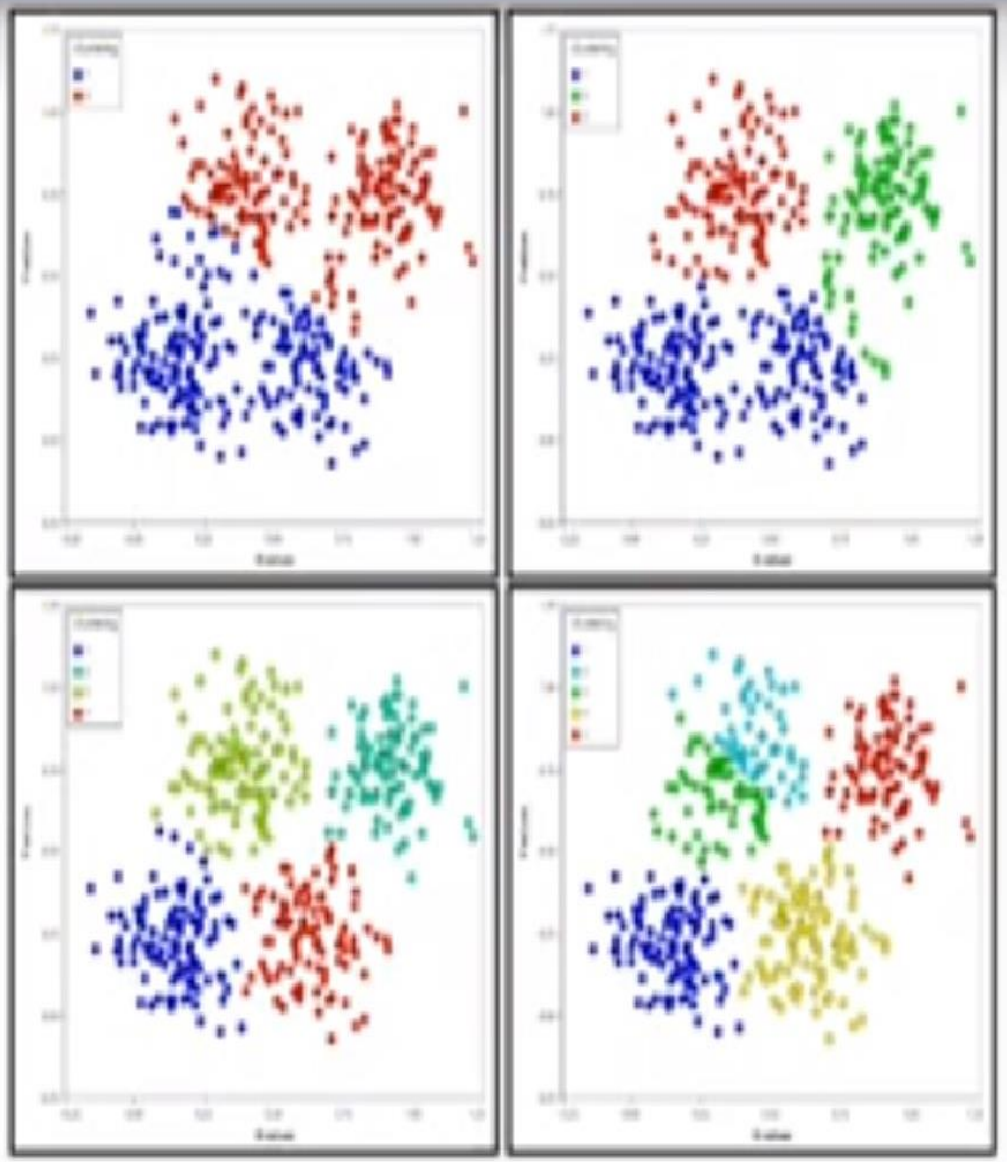
- تتم عبر إيجاد المتوسط الحسابي arithmetic mean لجميع نقاط العنقود
- يتم اختيار المركز بناء علي المتوسط الحسابي , ويكون هو المركز الجديد
- يتم تكرار خطوة الإختيار مرة اخري مع كل النقاط من البداية
- عبر تكرار الخطوتين معا , يتم عمل التقسيم بشكل سليم

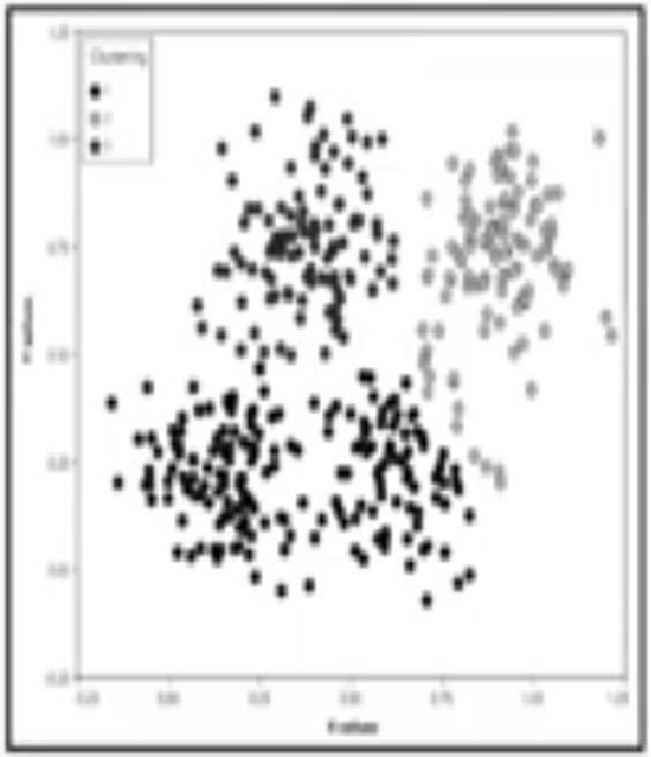
خوارزم K-means

كيفية التقسيم :



خوارزم K-means

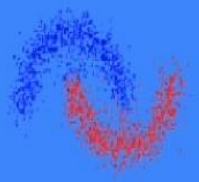




ضبط العناصر

عدد من العناصر الهامة الواجب ضبطها :

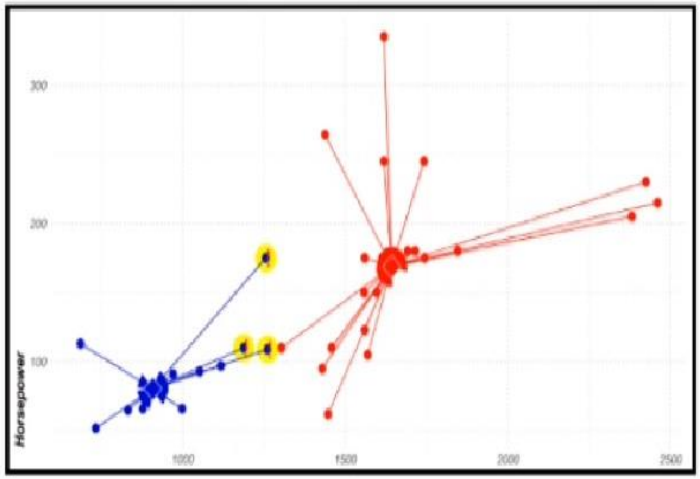
- دالة الـ Optimization Objective
- اختيار مكان المركز
- القيم الدنيا المحلية و العامة Local minimum Vs Global minimum

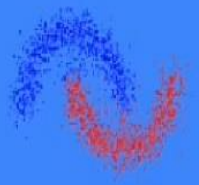


دالة الـ Optimization Objective

المعني :

- هي الدالة المطلوب تقليلها ,للوصول للشكل الأمثل من التقسيم
- كلما قلت قيمة الـ Optimization Objective كلما كانت كل نقطة أقرب للمركز الخاص بها
- هي تشبه دالة الخطأ cost function المستخدمة في التوقع و التصنيف





دالة الـ Optimization Objective

المعني :

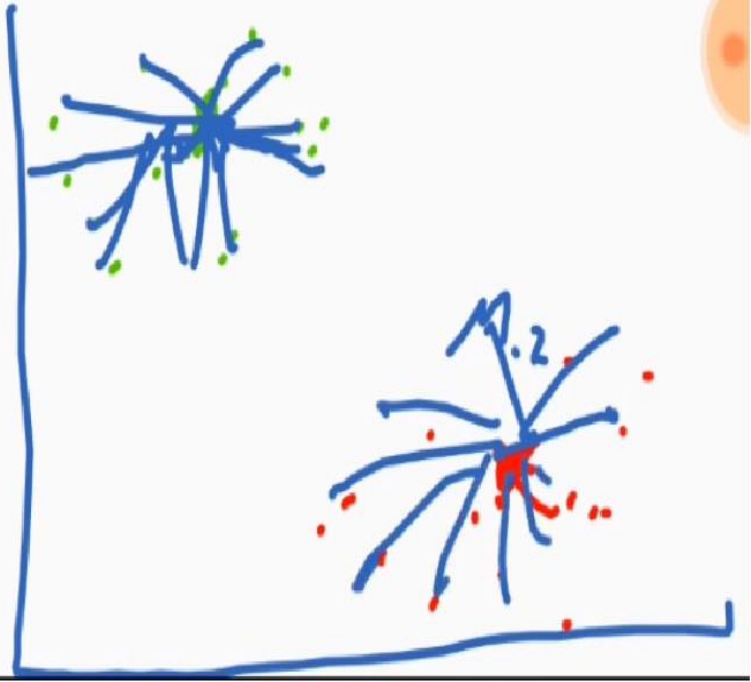
- يرمز لها بنفس الرمز J
- قيمتها , هي مجموع مربعات فارق المسافة بين كل نقطة و المركز التابع لها , مقسوم علي عدد النقاط
- زيادة القيمة معناها تباعد النقاط عن المركز , وتقليل القيمة معناه تراكم النقاط معا بالقرب من المركز

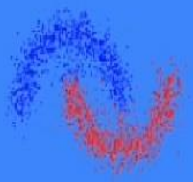
$$\frac{1}{m}\sum_{i=1}^{m}||x^{(i)}-\mu_{c^{(i)}}||^2$$

المعني :

- يرمز لها بنفس الرمز μ_i
- قيمتها , هي مجموع مربعات فارق المسافة بين كل نقطة و المركز التابع لها , مقسوم علي عدد النقاط
- زيادة القيمة معناها تباعد النقاط عن المركز , وتقليل القيمة معناه تراكم النقاط معا بالقرب من المركز

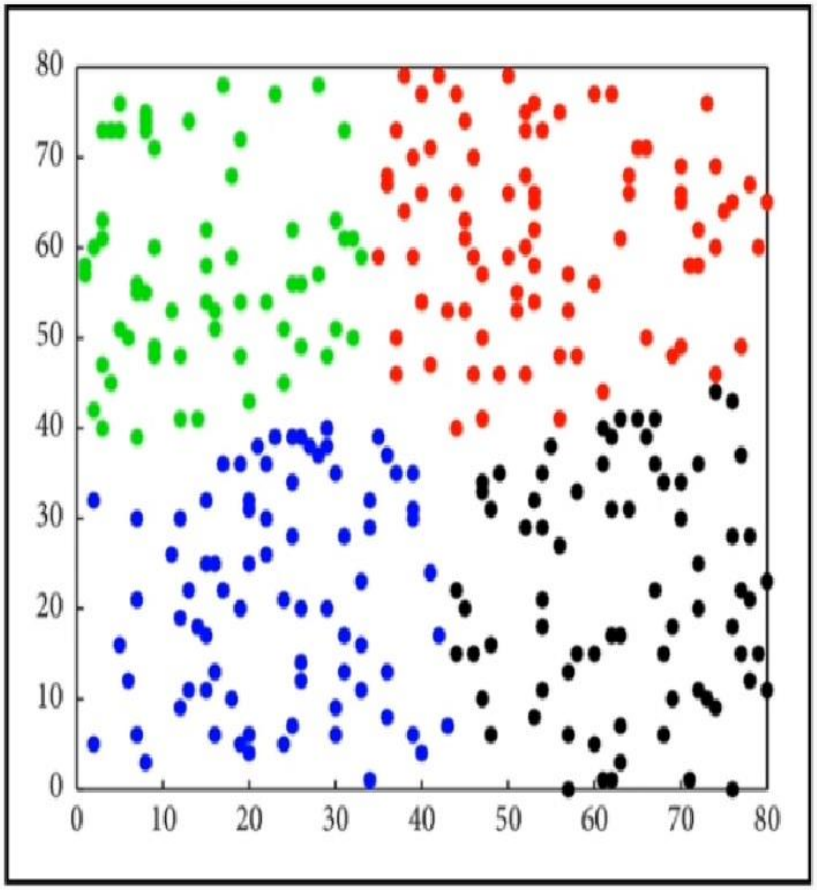
$$\frac{1}{m}\sum_{i=1}^{m}||x^{(i)} - \mu_{c^{(i)}}||^2$$

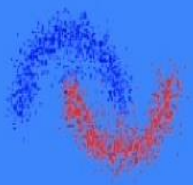




Optimization Objective دالة الـ

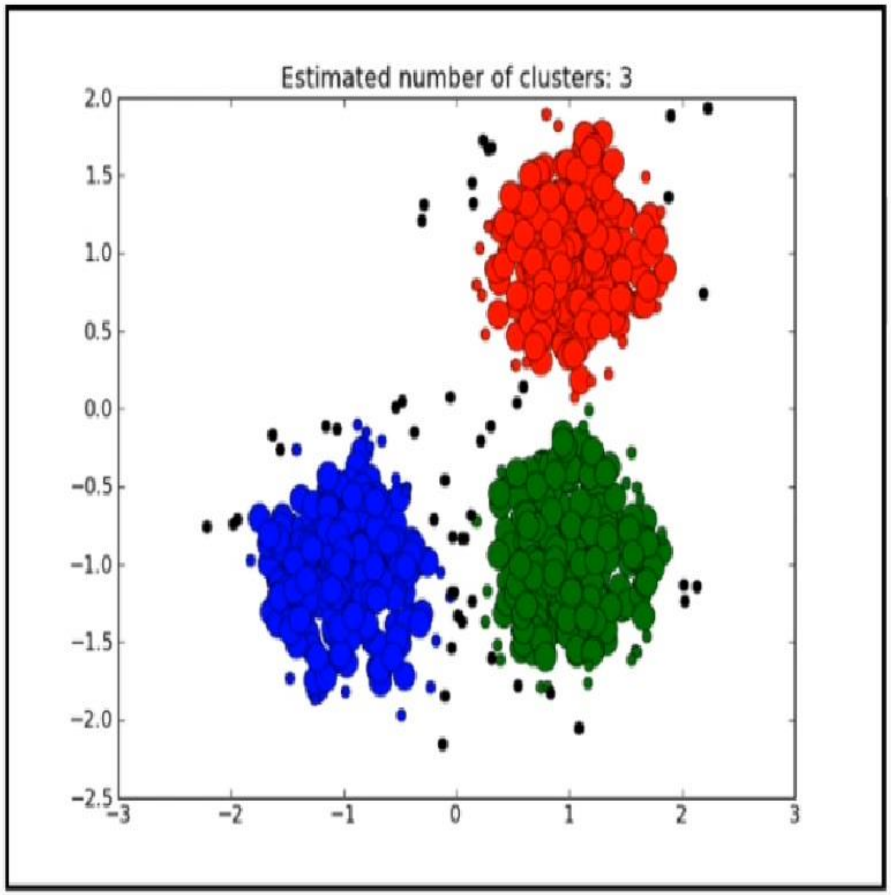
قيمة ر عالية



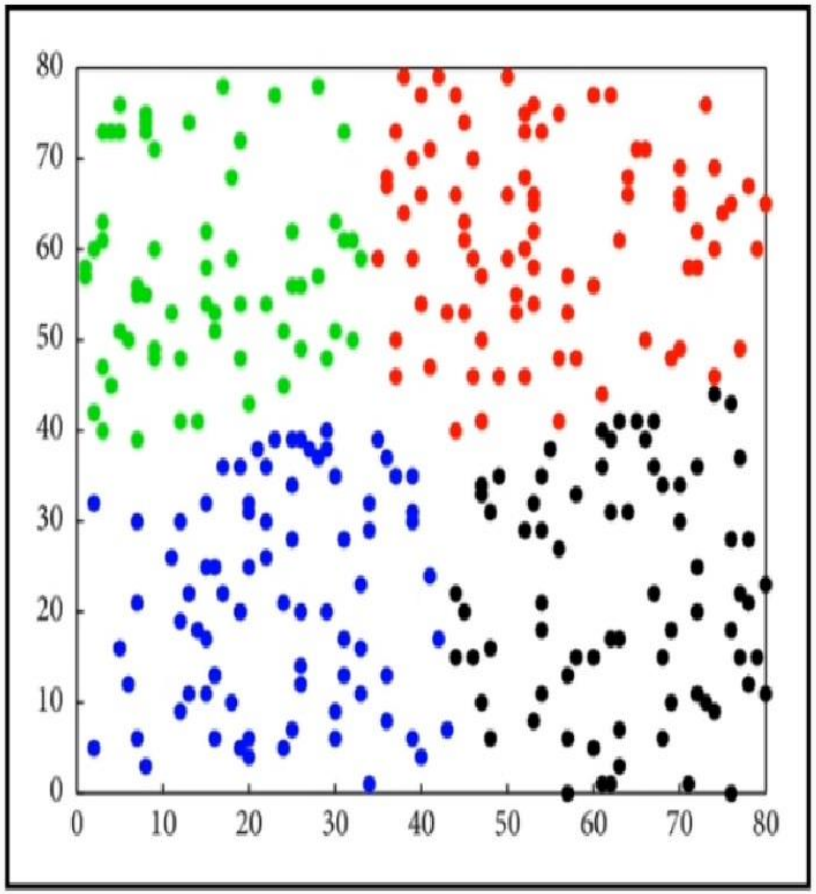


Optimization Objective دالة الـ

قيمة ر قليلة



قيمة ر عالية



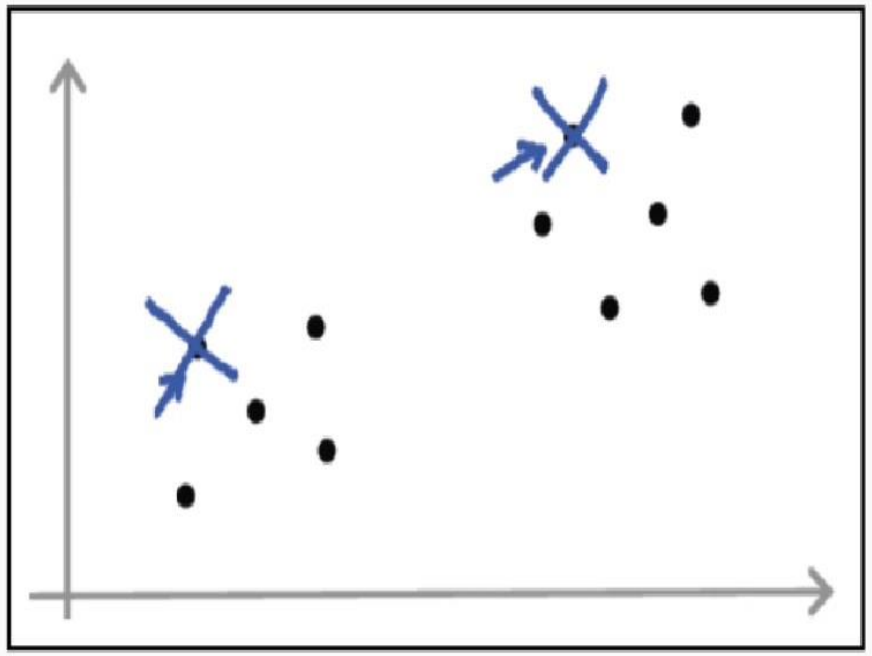
اختيار مكان المركز

الفكرة :

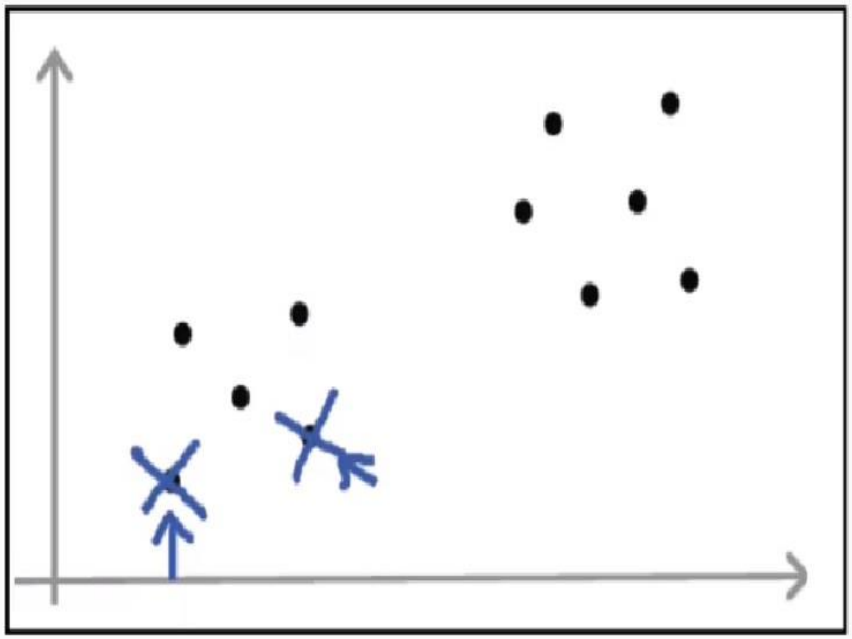
- الفكرة في اختيار المراكز , ان تكون قريبة من التجمعات الموجودة , ليسهل علي الخوارزم الوصول للتقسيم السليم
- لأن مع اقتراب كل مركز من مجموعة متراكمة , تقل قيمة J و يسهل عمل خطوتي القياس و التحريك

اختيار مكان المركز

اختيار مناسب

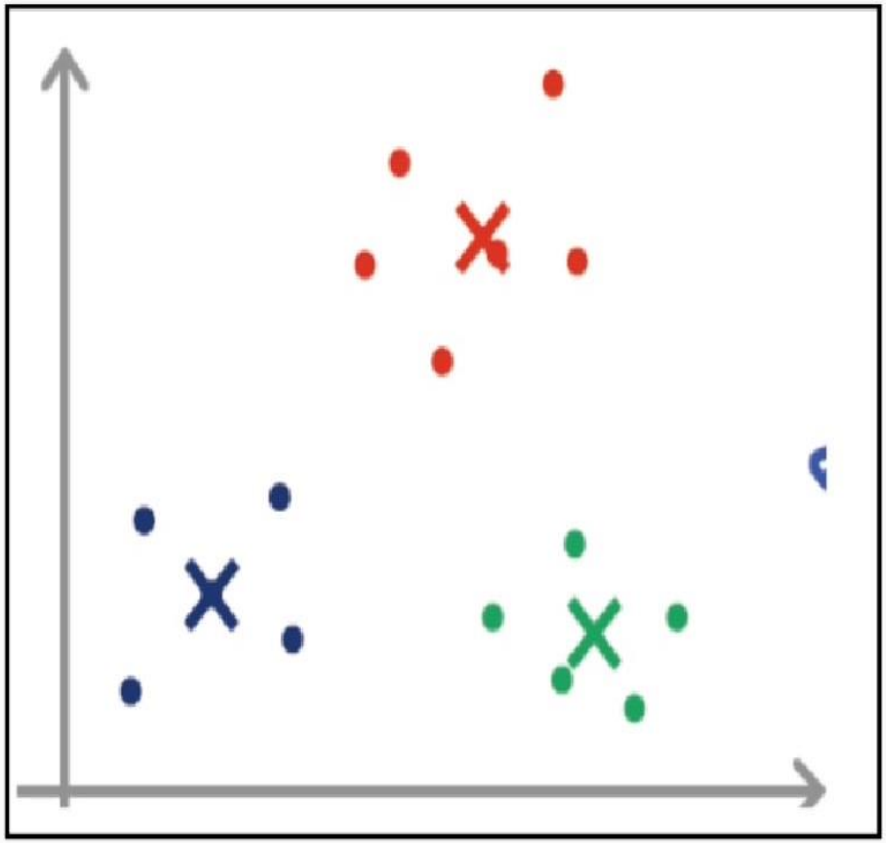


اختيار غير مناسب

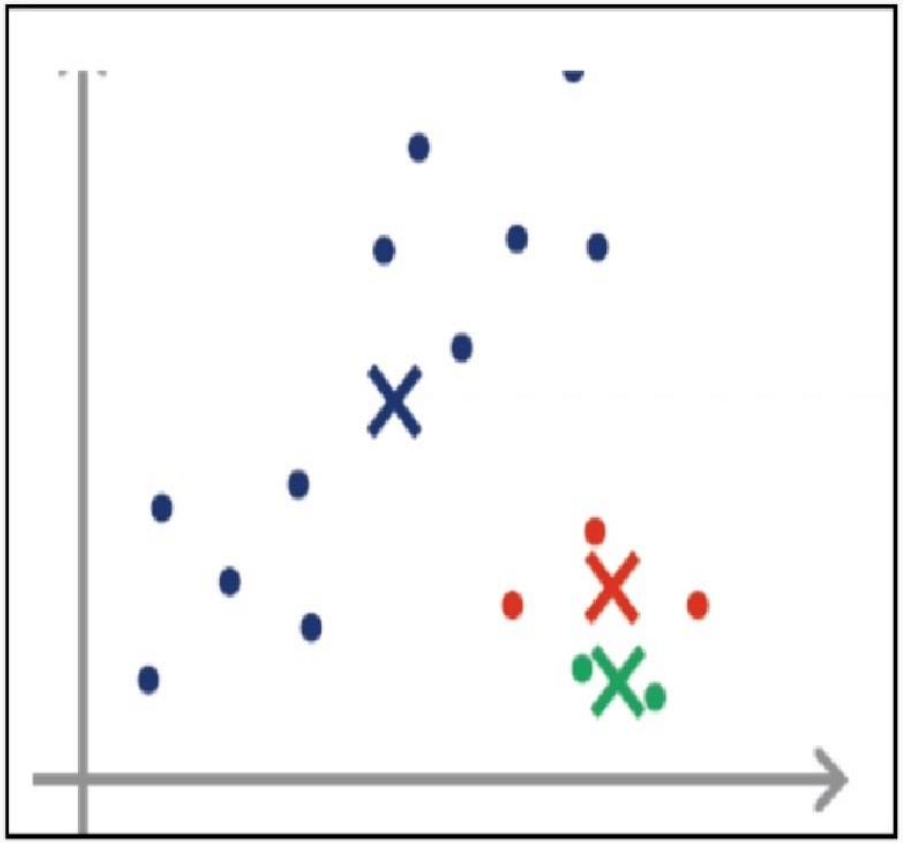


اختيار مكان المركز

اختيار مناسب



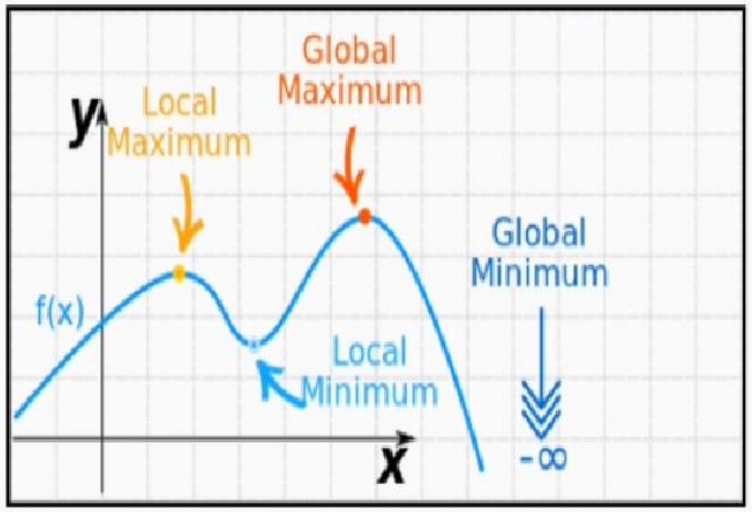
اختيار غير مناسب

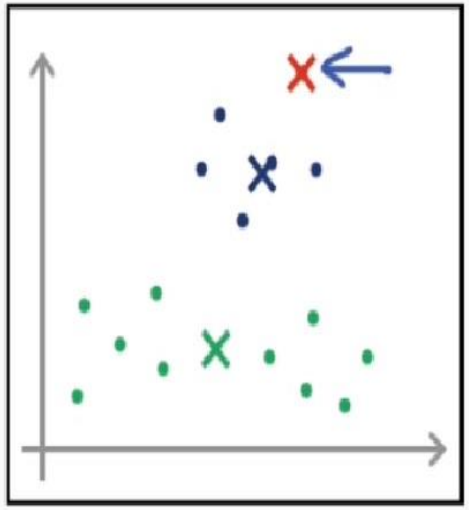


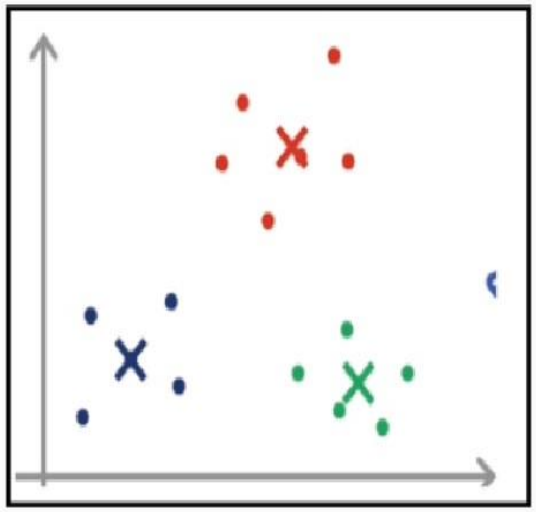
القيم المحلية و العامة Local vs Global min.

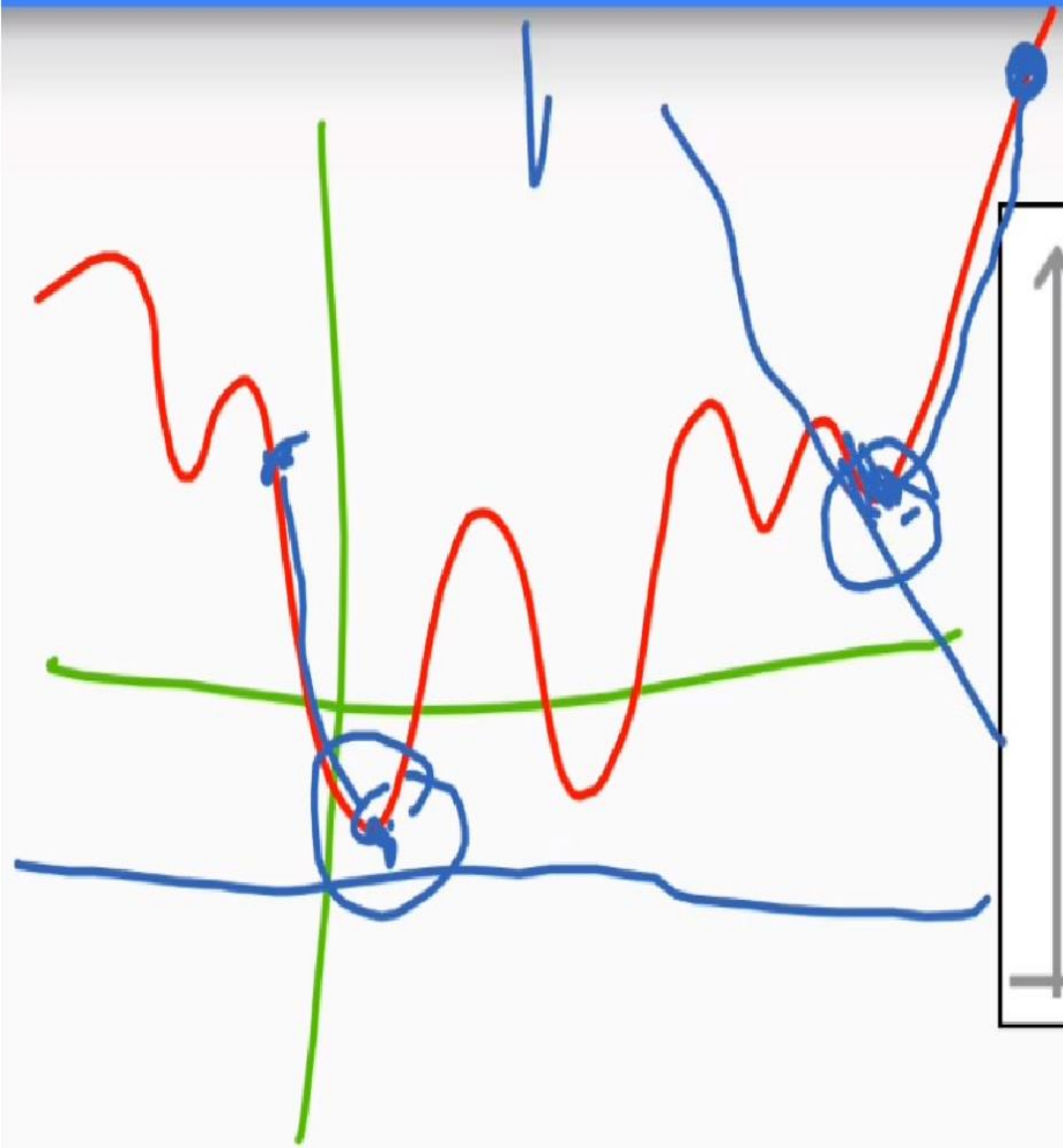
المعني :

- القيمة المحلية local minimum هي قيمة قليلة للـ (ل) لكن ليست مثالية
- القيمة العامة global minimum هي أقل قيمة ممكن للـ (ل)
- الاختيار الغير مناسب للمراكز يؤدي غالبا للقيم المحلية , بعكس الاختيار المناسب للمراكز

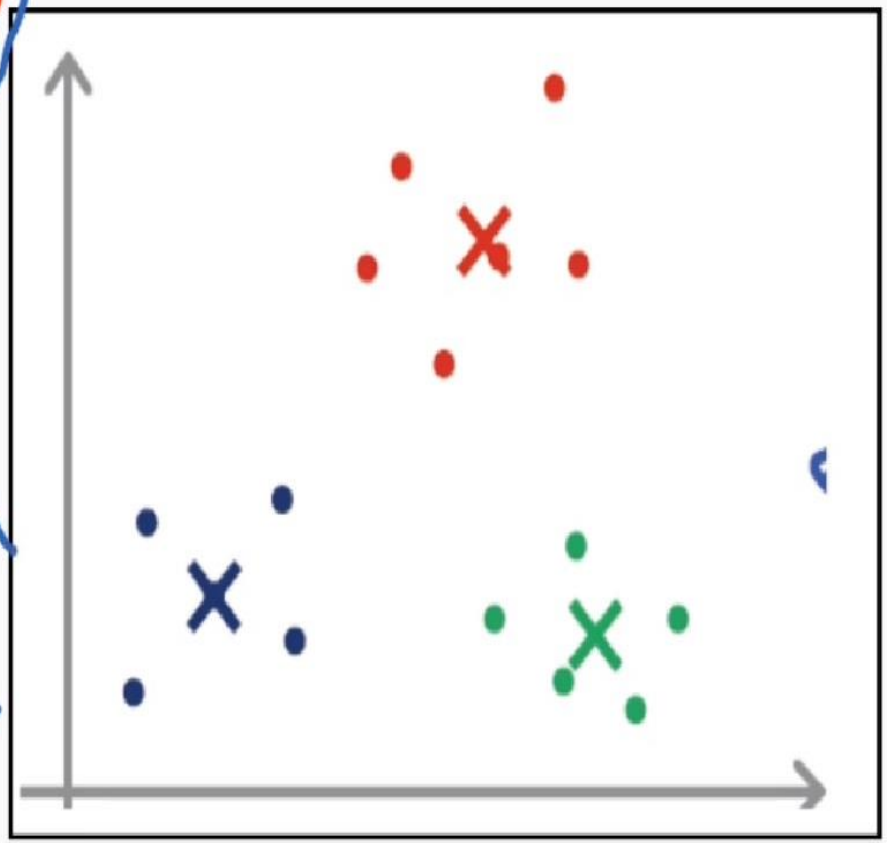


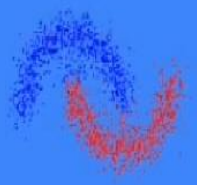






اختيار مناسب





القيم المحلية و العامة Local vs Global min.

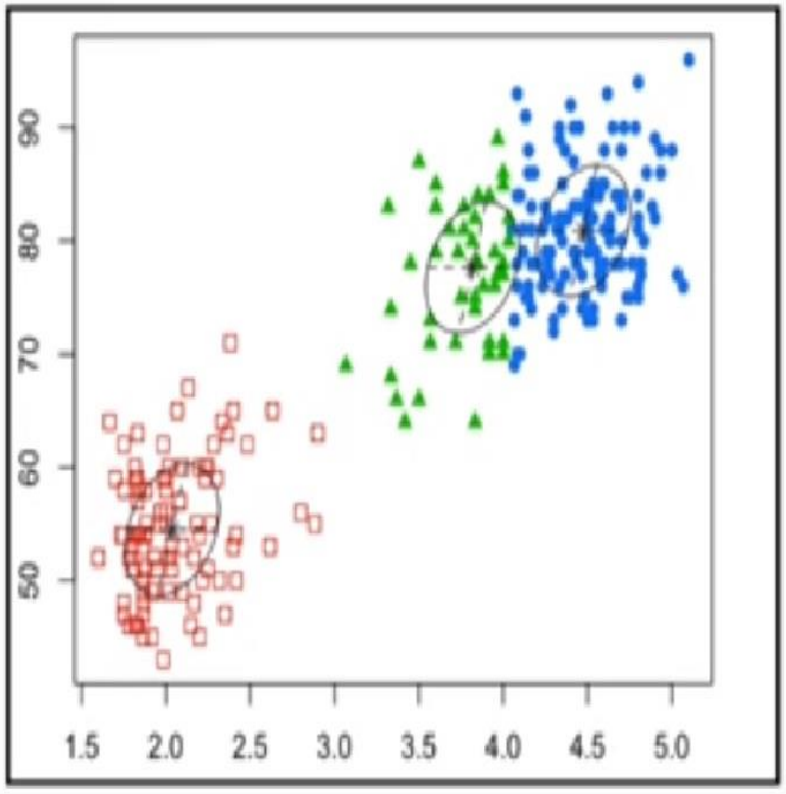
المعني :

- إذا كان عدد العنايد قليل , فجيب عمل عدد محاولات iterations كبير , حتي يتأكد من الوصول للـ global
- لا يمكن الإعتماد علي الفحص البصري , لتشابك البيانات و تعقيدها

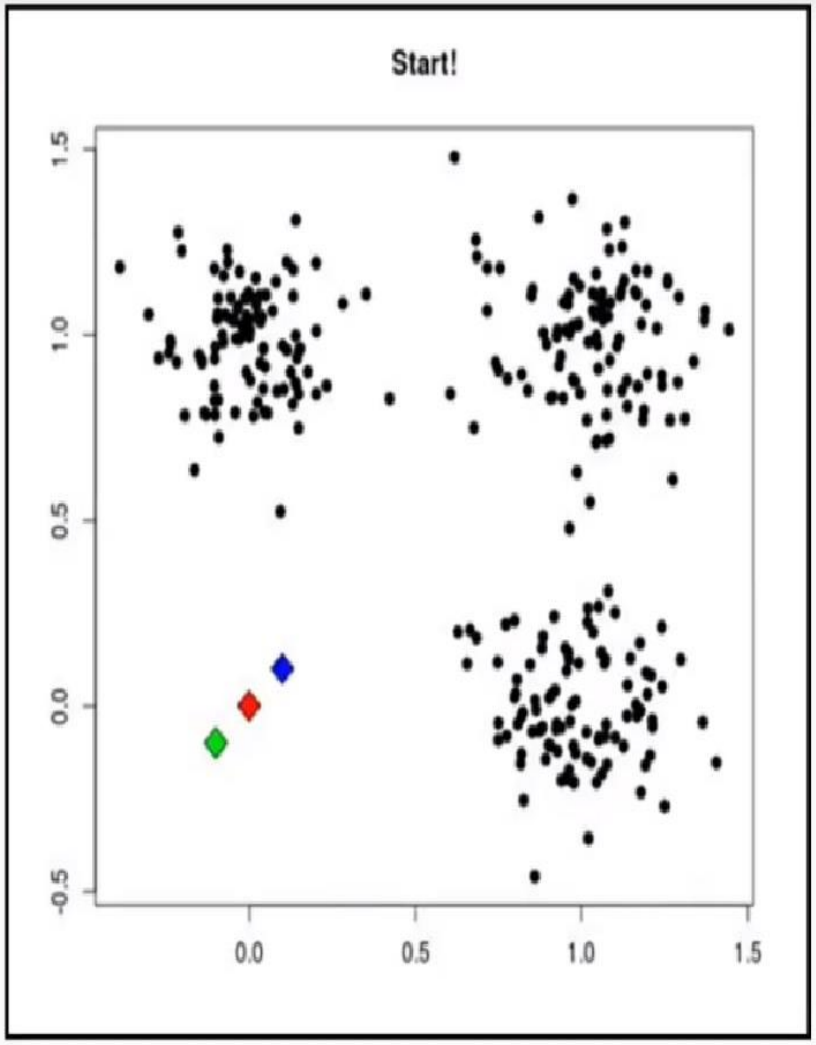
عدد العناقيد

المعني :

- وهو عدد المجموعات او العناقيد التي سنقوم بتحديدها للخوارزم
- أحيانا نقوم بتحديد العدد , وأحيانا نجعل الخوارزم نفسه هو من يحدد العدد المناسب
- في حالة قيامك باختيار العدد , فهناك طريقتين : الطريقة البصرية , وطريقة الكوع elbow method

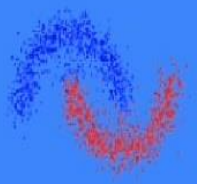


عدد العناقيد



الطريقة البصرية :

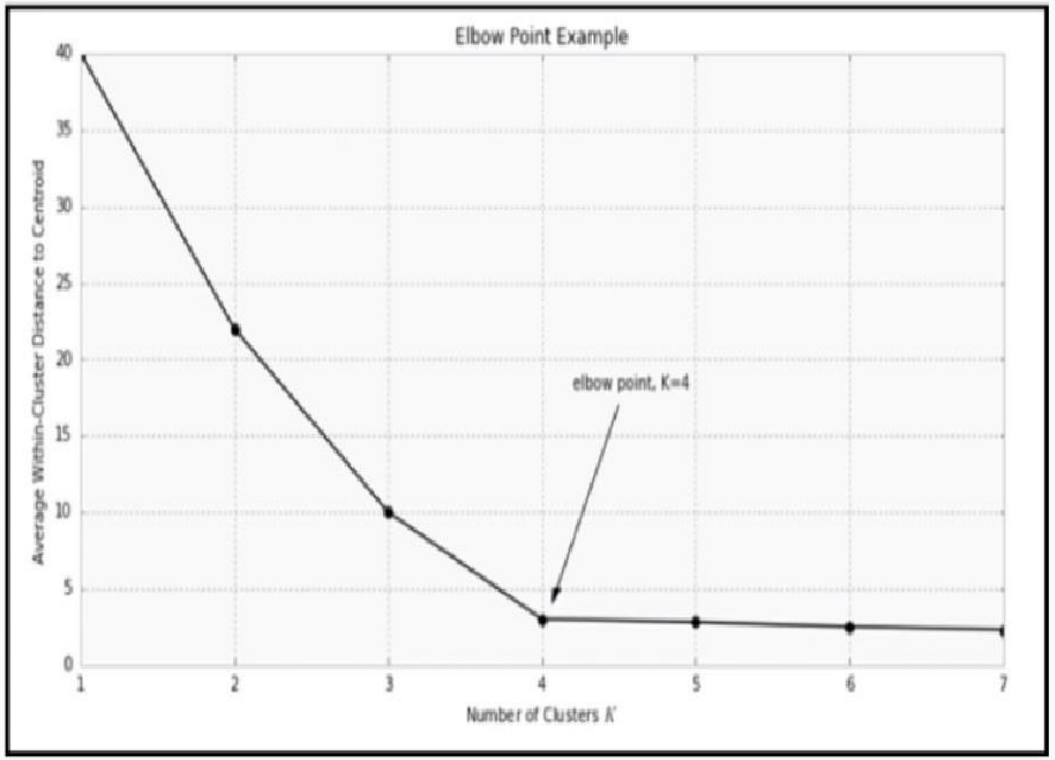
- وهي عبر عمل جراف للنقاط , ثم تحديد رقم تقريبي للعناقيد عبر رؤية تكتلات النقط
- تصلح فقط مع الأرقام المعقولة
- قد لا تكون دقيقة بشكل عام
- أقل في الوقت والتكلفة



عدد العناقيد

طريقة الكوع Elbow method :

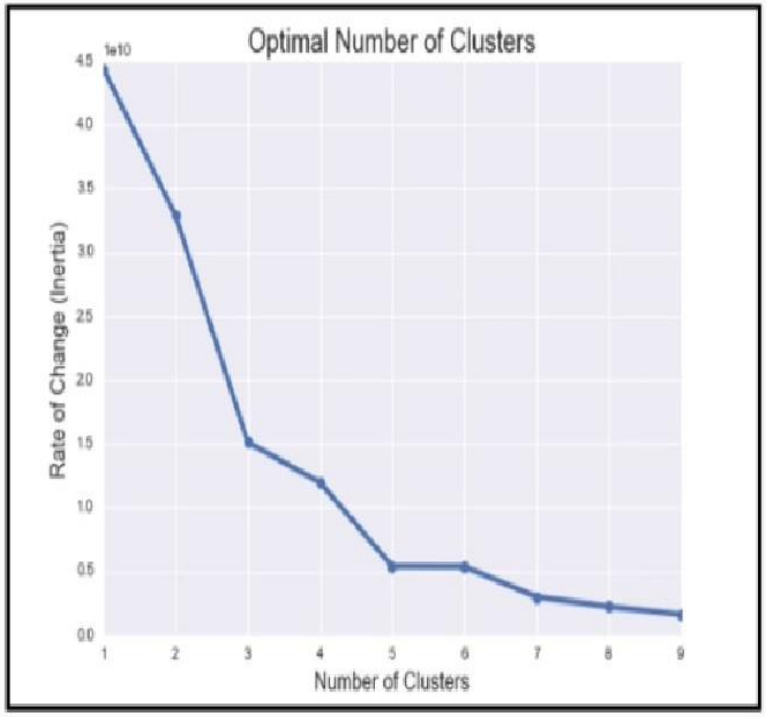
- والمقصود بها عمل جراف فيه عدد المحاولات k في محور إكس , وقيمة J في محور واي ومنها اختيار رقم مناسب



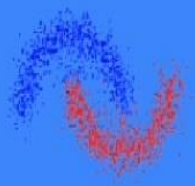
عدد العناقيد

طريقة الكوع Elbow method :

- علي الرقم الذي يتم اختياره أن يصاحبه عدد عناقيد مناسب , وأن يكون له قيمة ل قليلة لتجنب أي نقص في الكفاءة



عدد العناقيد

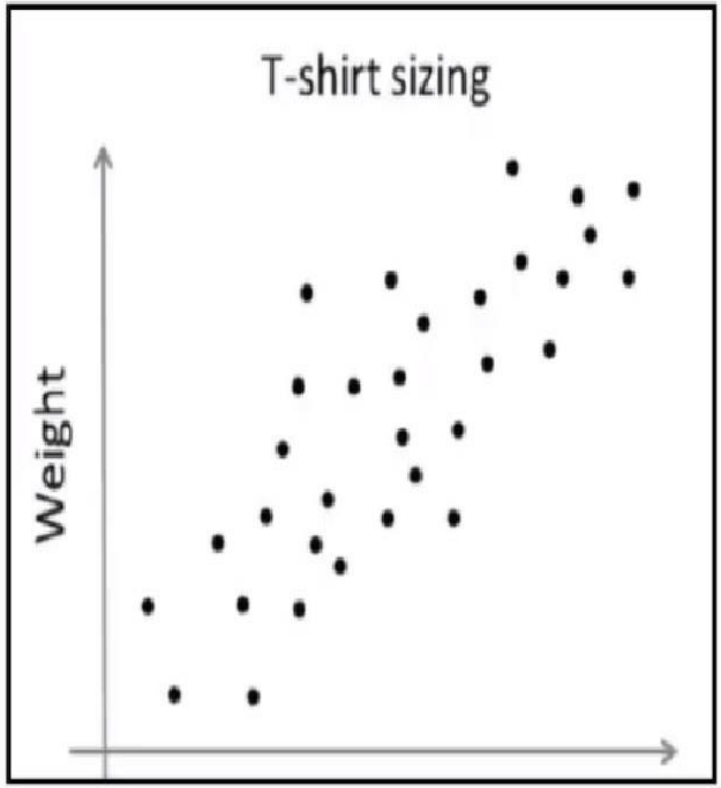


العوامل الأخري :

- تحديد عدد المجموعات ليس فقط علي أساس قيمة الـ (J) لكن هناك عوامل حياتية أخري تتدخل و قد تكون أقوي
- فلو سيتم تقسيم العملاء لشركة معينة لعدد من الأقسام , فلو كان الرقم الأمثل هو 6 أقسام , لكن انا عندي بس اربع موظفين مختصين بخدمة العملاء , فممكن اختار 4 بدل 6
- لو هاعمل تقسيم لطلاب الجامعة بعدد من الأقسام , وانا عايز اعمل بس 3 شرائح تعمليمية لأن المنهج لا يحتمل إلا التقسيم ده , فهيتم التجاوز عن قاعدة الكوع

العوامل الأخري :

- لو عندي بيانات طول ووزن العملاء اللي بيشترو تيشيرت معين بالشكل ده



عدد العناقيد

العوامل الأخري :

- فممكن الخوارزم يعمل 3 اقسام او 5 اقسام