* تعريف خوارزمية k_means :

هي خوارزمية تقسيم البيانات إلى مجموعات k-means خوارزمية متجانسة بحيث يكون كل مجموعة تحتوي على نقاطمتشابهة. (clusters)
الهدف من هذه الخوارزمية هو تقليل المسافة بين نقاط البيانات في نفس المجموعة وزيادة المسافة بين نقاط المجموعات المختلفة. يتم تحديد عدد المجموعات المطلوبة مسبقاً وتقوم الخوارزمية بتحديد مراكز هذه المجموعات بناءً على البيانات المدخلة

* طرق حساب k في k_means

1. Elbow Method: هذه الطريقة تتضمن تشغيل خوارزمية k-means عدة مرات مع قيم مختلفة لـ k وحساب متوسط المربعات البيضاء (Sum of Squared Errors) لكل قيمة من k. يتم رسم هذه القيم على الرسم البياني على الرسم البياني وهي (elbow) "التي تظهر "كوع k ويتم اختيار قيمة القيمة التي يعتبر بها انخفاض معدل الخطأ أبطأ

* مثال Elbow Method :

في k يتم تحديد قيمة مناسبة لـ ،(Elbow Method) طبقاً للطريقة الكوعية .
(Sum of Squared Errors - SSE) عن طريق رسم خطأ التجميع k-means خوارزمية (clusters). يتم اختيار مقابل عدد مختلف من المجموعات
مما يشبه شكل كوع ،حيث يحدث تغيير حاد في الخطأ التجميعي k قيمة.
لنفترض أن لدينا مجموعة بيانات تحتوي على نقاط مثل التالي:

نقطة	x	y
1	2	3

2	3	4
3	2	5
4	6	7
5	7	6
6	8	9

باستخدام الطريقة الكوعية. يمكننا تطبيق k نريد تحديد قيمة مناسبة لـ k-means خوارزمية وحساب SSE، (مثلاً من 1 إلى 6) k لعدة قيم مختلفة لـ k (k). مقابل عدد المجموعات (SSE) لكل قيمة. ثم يتم رسم خطأ التجميع

نبحث عن النقطة التي تشبه كوعًا في الرسم البياني. هذه النقطة تشير، بعد ذلك إلى أنه يجب اختيار قيمة k في هذا الموضع تقع.

K = 1: Inertia = 2560
K = 2: Inertia = 1800
K = 3: Inertia = 1200
K = 4: Inertia = 900
K = 5: Inertia = 800
K = 6: Inertia = 750

أفضل قيمة عندما تكون k = 4

2. Silhouette Method: تستخدم هذه الطريقة مقياس الـ Silhouette لتقييم جودة التجزئة. يتم تشغيل خوارزمية k-means لعدة قيم مختلفة من k. يتم اختيار قيمة k التي تعطي أعلى قيمة Silhouette. ويتم حساب قيمة Silhouette لكل قيمة من k.

3. Gap Statistics: هذه الطريقة تقارن متوسط الـ SS (Sum of Squared Errors) للبيانات الفعلية بالقيم المتوقعة إذا كانت البيانات توزعت عشوائياً. يتم اختيار قيمة k التي تعطي أكبر فجوة بين متوسط الـ SS للبيانات الفعلية والبيانات المتوقعة.

4. BIC (Bayesian Information Criterion) أو AIC (Akaike Information Criterion): تستخدم هذه الطرق معاقبة إضافية لعدد المجموعات (k) وتحاول اختيار قيمة k التي تحقق أفضل توازن بين دقة النموذج وبساطته.

-