

K-means

التجميع هو خوارزمية غير خاضعة للرقابة، والهدف هو اكتشاف أنماط مثيرة للاهتمام في البيانات. تستخدم تقنيات التجميع البيانات الأولية لتكوين مجموعات بناءً على عوامل مشتركة بين نقاط البيانات المختلفة. مثال: يعد تقسيم العملاء للتسويق المستهدف أحد أهم تطبيقات خوارزمية التجميع.

العوامل المهمة التي يجب علينا مراعاتها أثناء استخدام خوارزمية K-means :

- **عدد المجموعات (K) :**

يجب تحديد عدد المجموعات التي تريد تجميع نقاط البيانات فيها مسبقًا.

- **القيم الأولية/البذور :**

يمكن أن يكون لاختيار مراكز المجموعة الأولية تأثير على تكوين المجموعة النهائي. خوارزمية K-means غير حتمية. وهذا يعني أن نتائج التجميع يمكن أن تكون مختلفة في كل مرة يتم فيها تشغيل الخوارزمية حتى على نفس مجموعة البيانات.

- **القيم المتطرفة :**

تكوين المجموعة حساس جدًا لوجود القيم المتطرفة. تقوم القيم المتطرفة بسحب الكتلة نحو نفسها، مما يؤثر على تكوين الكتلة الأمثل.

- **مقاييس المسافة :**

قد يؤدي استخدام مقاييس مسافة مختلفة) تُستخدم لحساب المسافة بين نقطة البيانات ومركز المجموعة (إلى مجموعات مختلفة.

- لا تعمل خوارزمية K-Means مع البيانات الفئوية.

- قد لا تتقارب العملية في العدد المحدد من التكرارات. يجب عليك دائما التحقق من التقارب.

طريقة منحني الكوع:

تعمل طريقة الكوع على استخدام k يعني التجميع في مجموعة البيانات لمجموعة من قيم k على سبيل المثال من 1 إلى 10 :

- قم بإجراء تجميع يعني K مع كل هذه القيم المختلفة لـ K . لكل قيمة من قيم K ، نحسب متوسط المسافات إلى النقطة الوسطى عبر جميع نقاط البيانات.
- ارسم هذه النقاط وبحث عن النقطة التي يقع فيها متوسط المسافة من النقطة الوسطى فجأة ("الكوع").
- بصريًا، يمكننا أن نرى أن العدد الأمثل للمجموعات يجب أن يكون حوالي 3. لكن تصور البيانات وحده لا يمكن أن يعطي دائمًا الإجابة الصحيحة.