I- التحليل العنقودي:

التحليل العنقودي (Cluster Analysis) عبارة عن أدوات تهدف إلى تصنيف مجموعة حالات cases أو متغيرات variables بطرق معينة وترتيبها داخل عناقيد clusters. بحيث تكون الحالات المصنفة داخل عنقود معين متجانسة فيما يتعلق بخصائص محددة وتختلف عن حالات أخرى موجودة في عنقود آخر.

وإذا أجريت التحليل العنقودي لتصنيف الحالات، حيث تشارك الخصائص لفريق ما صفاتهم فإنه يصبح من السهل أن تتوقع إيمان أو المدير سلوك الأفراد والمجموعات اعتمادا على عضوية المجموعة التي تشترك أعضاؤها في خصائص متشابهة.

وهناك نوعان من التحليل لتكوين العناقيد:

أ- التحليل العنقودي التقسيمي: يبدأ هذا النوع من التحليل العنقودي باعتبار أن جميع الحالات تجمع في عنقود واحد، وبعد ذلك يتم تصنيف الحالات تدريجيا في عناقيد أصغر فأصغر.

ب- التحليل العنقودي التجميعي: يبدأ التحليل التجميعي بعنقود واحد لكل حالة، ثم يتم تجميع العناقيد المتشابهة تدريجيا في نهاية المطاف إلى العدد المطلوب من العناقيد.

1- التحليل العنقودي (K-Means):

تقوم هذه الطريقة على أساس تصنيف الحالات في مجموعات متجانسة من حيث خصائص أو صفات معينة وذلك باستخدام خوارزميات يمكن أن تعالج عدد كبير من الحالات. وتسمى هذه الطريقة أحيانا بطريقة التحليل العنقودي السريع (Quick clustering) وذلك بسبب أنها تقوم بعملية التحليل والتصنيف في وقت قصير نسبيا، ويمكن إيجاز خطوات هذه الطريقة بما يلي:

① * تحويل البيانات الموجودة في المتغيرات إلى قيم معيارية إذا كانت متغيرات متصلة بوحدات مختلفة (المتغير الأول معبر عنه بالسنوات والثاني معبر عنه بالدنانير والثالث معبر عنه بالدرجات ... الخ)، ويمكن إجراء ذلك من خلال اختيار القائمة الرئيسية Analyse ثم القائمة الفرعية Descriptive statistics ثم الضغط على Descriptives والتأشير على المربع الصغير أمام Save standardized values as variables داخل صندوق الحوار Descriptives.

② * تحديد عدد العناقيد المطلوب أن يجرى التصنيف على أساسها.

3 - تحديد متوسط قيم بيانات (Centroids) بشكل مبدئي، مراكز المتوسطات
4 - حساب المسافات بين نقطة التقاء كل زوج من البيانات
٥ - تصنيف كل نقطة التقاء ببيانات للعنقود الذي يعتبر متوسطه
الأقرب إليها وعادة يتم استخدام مقياس (Euclidean Distance)
6 - إعادة حساب متوسطات قيم بيانات.
7 - استمرار تكرار نفس الخطوات (3-6) حتى نتوصل إلى أنه ليس هناك أية
نقطة بيانات تحركها إلى متوسط قيم بيانات بشكل أكثر قربا من موقع حالي.

خصائص التحليل العنقودي:

من أهم الخصائص التي يتميز بها هذا التحليل عن غيره:

- أن عملية التحليل العنقودي تعد من التعليم غير موجه، وذلك لعدم توفر معلومات سابقة متعلقة بالفئات المميزة لكل عنقود.
- يعتمد على تقسيم مجموعة من البيانات إلى مجاميع متشابهة فيما بينها ومختلفة فيما بينها.
- العناقيد: هي مجموعة من البيانات المتشابهة فيما بينها والمختلفة عن البيانات في العناقيد الأخرى.

خطوات التحليل العنقودي:

أهم خطوات تطبيق التحليل العنقودي هي:

- حساب مصفوفة التشابه أو مصفوفة التباعد: أقصر المسافات ضمن المصفوفة تشكل العناقيد الأولية.
- ربط العناصر التي تكون المسافة بينها أقصر المسافات ضمن المصفوفة لتشكيل العناقيد الأولية.
- حساب مصفوفة المسافة من جديد بعد تشكيل العناقيد الأولية، والاستمرار بعملية ربط العناصر والعناقيد.
- حساب مصفوفة المسافة من جديد بعد تشكيل العناقيد الأولية، والاستمرار بعملية ربط العناصر والعناقيد.
- الاستمرار في هذه العملية حتى نصل إلى عدد العناقيد المطلوب.

أنواع التحليل العنقودي:

تتفرع التحليل العنقودي عموما إلى نوعين:

أ- التحليل العنقودي الهرمي (Hierarchical):

التحليل العنقودي الهرمي المعروف بطبيعته بعدد العناقيد التي يتم تصنيف الحالات فيها، حيث يتطلب التحليل العنقودي الهرمي عينات صغيرة نسبيا، على أساسها يتم تحديد التحليل العنقودي الهرمي المناسب.

وهناك أسلوبين لتكوين العنقود مفردات العينة باستعمال التحليل الهرمي وهما:

- أسلوب التجميع: يبدأ هذا الأسلوب من التحليل بعنقود واحد لكل حالة ثم يتجمع العناقيد المتشابهة تدريجيا حتى نصل إلى العدد المطلوب.
- أسلوب التقسيم: يبدأ هذا الأسلوب بافتراض أن جميع الحالات تتجمع في عنقود واحد ثم يتم تصنيف الحالات في عناقيد أصغر فأصغر.

حساب مصفوفة التباعد:

يرمز لها بالرمز D:

$$D = \begin{pmatrix} 0 & d_{12} & d_{13} & \cdots & d_{1n} \\ & 0 & d_{23} & \cdots & d_{2n} \\ & & 0 & d_{jk} & d_{jn} \\ & & & 0 & \vdots \\ & & & & 0 \end{pmatrix}$$

d_{jk} : المسافة بين المتغيرتين k و j وبما أن هذه المصفوفة هي مثلثية عليا وعناصر قطرها الرئيسي تساوي الصفر لأن المسافة بين المتغيرة ونفسها تساوي صفر، يتم حساب المسافة من خلال العلاقة :

$$D_e = \sqrt{\sum_{i=1}^{d} (x_{ij} - x_{ik})^2}$$

الاقليدية

حساب مصفوفة التشابه أو التقارب :

لحساب مصفوفة التقارب يتم اتباع الخطوات التالية :

- تحويل المتغيرات النوعية (رئيسية، اسمية) إلى متغيرات ثنائية وتأخذ احدى القيمتين (1) عند تحقق المتغير و (0) عند عدم تحقق
- حساب التقارب بين كل متغيرتين من خلال الأشياء دون التوافق .
- حساب عناصر مصفوفة التقارب ويرمز لها بالرمز S_{ik} :

$$S_{ik} = \frac{a+d}{p} = \frac{\text{عدد المزاوجات المتشابهة}}{\text{عدد المتغيرات المؤثرة}}$$

- استعمال المعطيات لتشكيل مصفوفة التقارب كما يلي :

$$S = \begin{pmatrix} 1 & & & \\ s_{21} & 1 & & \\ s_{31} & s_{32} & 1 & \\ \vdots & \vdots & \vdots & \\ s_{n1} & s_{n2} & \cdots & 1 \end{pmatrix}$$

نلاحظ أن المصفوفة مثلثية سفلى من الرتبة $(n \times n)$ كما أن عناصر قطرها الرئيسي تساوي 1 .

ب - التحليل العنقودي غير الهرمي:

يطبق التحليل العنقودي غير الهرمي على العينات الكبيرة، ويستعمل لتجميع بيانات المفردات الكبيرة أكثر مما يستعمل لتجميع المتغيرات ضمن عنقود محدد. إذ يعتمد على أسلوب التقسيم، وتوجد عدة طرق للتحليل العنقودي غير الهرمي، إذ تعد طريقة K متوسط (K-means) أفضلها وذات فعالية عالية للتعامل مع البيانات الضخمة.