المقارنة بين طرق العنقدة غير الهرمية وتحديد أفضلها (دراسة تطبيقية على بعض أنواع الحليب المبيع بمدينة مصراته)

أ. خلود سليمان عمر إسميو

كلية العلوم، جامعة مصراته

المستخلص:

إجريت دراسة تحليلية على بعض مكونات الحليب المبيع في الأسواق المحلية بمدينة مصراتة لعام (2016) بغية الوصول إلى مجموعات متجانسة التي تجمعها صفات مشتركة وذلك بالاعتماد على بعض المعادن الداخلة في تركيب الحليب. من هذا المنطلق قمنا في هذه الورقة بتطبيق بعض طرق التحليل العنقودي الغير هرمي. أظهرت نتائج الدراسة أن أفضل الطرق الغير الهرمية والمتمثلة في كلا من طريقة k-means, k-mediods, كانت طريقة k-means حيث كان لها أفضل أداء من طريقة k-medoids والتي بدورها قسمت أنواع الحليب إلى 4 مجموعات حيث تضمنت المجموعة الأولى أنواع الحليب التالية حليب جهينة كامل الدسم(U5), حليب كاديا كامل الدسم(U6), حليب الريحان خالي الدسم (U7), حليب مبستر مصراتة(P1), حليب مبستر طرابلس(P2), و حليب أبقار خام منطقة السكت, واحتوت المجموعة الثانية على الأنواع التالية حليب السهول كامل الدسم(U2), حليب الزهرات كامل الدسم(U3),و حليب الربيع(U4), أما بالنسبة للمجموعة الثالثة فكانت تحتوي على نوعين فقط وهما حليب الريحان كامل الدسم(U1)، وحليب جهينة خالي الدسم(U8), واحتوت المجموعة الرابعة على أنواع الحليب الخام من المناطق الدافنية الغيران قصر أحمد كرزاز وطمينة، ومن ثم تم تقييم النتائج باستخدام بعض المؤشرات الخاصة بصلاحية العناقيد لها، كذلك من خلال نتائج الدراسة اتفقت طريقة k-means في تقسيم أنواع الحليب إلى أربع مجموعات متجانسة.

الكلمات المفتاحية: التحليل العنقودي، العنقدة الغير هرمية، مؤشر دن، مؤشر الصورة الظلية.

المقدمة Introduction

تعرف العنقدة أو التحليل العنقودي بأنها طريقة نموذجية لتجميع نقاط البيانات (العناصر) ضمن محيط التصنيف, حيث يتم تقسيم مجموعة من البيانات إلى عدد من المجاميع الجزئية أو العناقيد, وذلك اعتماداً على تشابه العناصر داخل العنقود الواحد درجة بدرجة نسبية من التشابه, بينما تملك العناصر المنتمية إلى عناقيد أخرى درجة عالية في الاختلاف أو عدم التشابه , حيث تتم عملية تصنيف (تقسيم العناصر) إلى عناقيد على أساس المقاييس الموضوعة على هذه العناصر, فمثلا إذا توفرت لدينا مجموعة بيانات فإن الهدف هو التجزئة أو تحديد المجاميع الجزئية أو العناقيد لعناصر متشابهة على أساس تقسيم المجتمع إلى مجموعات تحتوي على مجموعة من العناصر المنتمية إلى المجتمع, والهدف من استعمال هذا الأسلوب هو عملية تجزئة وتصنيف البيانات بطريقة علمية بحثه

2.1 الدراسات السابقة Literature Review

في هذا البند سنقوم بعرض لبعض الدراسات السابقة ذات الصلة بموضوع التحليل العنقودي بشكل عام محاولين وضع دراسات تكون ذات صلة بموضوع أنواع الحليب ومنها :

1- في عام 2016 تم نشر ورقة علمية من قِبل (Sharifuddin M.zain&et.al) تم من خلالها دراسة المكونات الرئيسية لأنواع من الحليب مُختلف المنشأ مستخدمين في ذلك طريقة Ward`s حيث أظهرت النتائج تقسيم عينة الدراسة إلى عنقودين .

2- قام (إبراهيم, وآخرين,2005) باستخدام التحليلين العاملي والعنقودي في أنواع مختلفة من الحليب المجفف والتمييز بين الأسلوبين, حيث تم استخدام طريقة التعنقد الهرمي, و أظهرت نتائج البحث وجود ثلاث مجموعات نتيجة عملية العنقدة حيث احتوت كل من المجموعتين الأولى والثانية على نماذج للحليب كامل الدسم, واحتوت المجموعة الثالثة على نماذج للحليب نصف الدسم.

3- أجريت دراسة في أيرلندا في عام 2017 من قبل (EL Feeney,et.al) على أنماط تناول الألبان الغذائية وأثرها في تكوين الجسم وعلامات صحة التمثيل الغذائي حيث هدفت الدراسة إلى دراسة الارتباطات بين تناول الألبان وصحة الأيض, وتم استخدام التحليل العنقودي بالخصوص وأظهرت النتائج ثلاثة أنماط (مجموعات) من

استهلاك الألبان (لبن كامل الدسم),(لبن قليل الدسم وزبادي),(زبدة كريمية), وأهتم الباحثون بدراسة نسبة الدهون المشبعة والثلاثية في المجموعات الثلاث .

4- قام (Simone S ,et.al, 2010) بدراسة لمراقبة جودة الحليب UHT البرازيلي عالي الحرارة من الناحية الكيميائية, المعالج في المصانع الصناعية الواقعة في مناطق مختلفة من البلاد, باستخدام تحليل المكونات الرئيسية والتحليل العنقودي حيث أتاحت هذه التحاليل التحقق من حدوث بعض الغش في مناطق محددة.

5- في عام 2012 قام (السباح, الأعرجي, 2012) باستخدام التحليل العنقودي السريع لفرز نوع الرضاعة حسب خصائص الأمهات, حيث تم تطبيق أسلوب التحليل العنقودي السريع على عينة عشوائية من الأمهات وتمكن من فرزها حسب نوع الرضاعة إلى خمسة مجموعات.

6- قدم (Arbin.N&et.al,2015) دراسة لمقارنة كل من خوارزميتي:

k-means,k-medoids لمجموعة من البيانات المختلفة لوضع نقاط القوة والضعف لكليهما.

7- قـّدم (رشيد, مهدي ,2011) ، دراسة مقارنة باستخدام طرائق التحليل العنقودي, ومن أهم النتائج التي توصلوا إليها أن محافظة بغداد كانت الأفضل في تقديم الخدمات الخاصة بالتربية والتعليم, وأن محافظة نينوى كانت متأخرة في هذا المجال, وكانت الطريقة الهرمية (طريقة المتوسطات بين المجاميع) أفضل الطرائق المستخدمة لامتلاكها أقل قيمة للمقياس وفق مقياس العلاقة المقلص.

مشكلة البحث Problem Statement

بالرغم من أهمية تحليل التباين أحادي المتغير وثنائي المتغير إلا أن قصور هذه الطرق فيما يتعلق الأمر بعدد كبير من المتغيرات أدى إلى اللجوء إلى نوع آخر من التحليل وهو التحليل متعدد المتغيرات. حيث يهتم هذا الأسلوب بتحديد أوجه التشابه بين متغيرات متعددة أو مجموعة من المتغيرات في وقت واحد.

أهمية البحث Research Importance

- الدراسات المتعلقة بهذا المجال قليلة من الناحيتين النظرية والتطبيقية حسب علم الباحث.

- تتمثل أهميته في كيفية تحليل وتحديد أوجه التشابه من متغيرات متعددة باستخدام التحليل العنقودي، ولفت أنظار الباحثين المتخصصين إلى الاهتمام بهذا الموضوع ذو الأهمية بمكان في تحليل البيانات عديدة المتغيرات.
- أهمية استيعاب برامج التحليل الإحصائي وخاصة التحليل العنقودي ضمن مقررات العملي في مؤسسات التعليم العالي والبحث العلمي.

أهداف البحث **Research Objectives**

يهدف البحث إلى تطبيق التحليل الإحصائي، اعتماداً على تحليل متعدد المتغيرات وعلى وجه الخصوص أسلوب التحليل العنقودي والتعرف على أهميته في التصنيف ومقارنة بعض طرق العنقدة، كما يهدف البحث إلى معرفة أوجه التشابه لعينات الحليب المختلفة.

تساؤلات البحث **Research Assumptions**

- المقارنة بين الخوارزميات الغير هرمية لتحديد أفضلها في عملية العنقدة.
- المقارنة بين الطرق المختلفة لتحديد العدد الأمثل للعناقيد.
- المقارنة بين الطرق المختلفة لتحديد صلاحية العناقيد.

منهجية البحث **Research Methodology**

تتمثل منهجية البحث في الأسلوب الوصفي والتحليلي الذي يعتمد على تطبيق أهم أساليب متعددة المتغيرات ومنها التحليل العنقودي للوصول إلى أهم المتغيرات التي تؤثر في اختلاف بيانات الدراسة. واعتمدت الباحثة على بيانات مأخوذة من رسالة ماجستير بقسم الكيمياء كلية العلوم جامعة مصراته لعام 2016(بن حليم, 2016).

وكذلك يمكننا تطبيق أسلوب التحليل العنقودي للوصول إلى العناقيد المتجانسة، باستخدام البرمجة بلغة R.

المصطلحات المستخدمة في التحليل العنقودي **Terminology Used in Cluster Analysis**

نظراً لأهمية المصطلحات المستخدمة في التحليل العنقودي لابد من إعطاء فكرة مبسطة عن كلا منها:

- العنقود (The Cluster): ويقصد به مجموعة من العناصر التي ترتبط مع بعضها البعض طبقاً لمجموعة من الخواص بحيث أن العنقود يحوي العناصر الأكثر تجانساً وفي الوقت نفسه تكون العناصر بين العناقيد مختلفة أكثر ما يمكن.

- العنصر (The Object): ويقصد به عنصر الدراسة ويستعمل بمعناه الواسع في تمثيل الأشياء من أبسطها حتى أعقدها.
- المسافة (The Distance): هي تصور للتماثل بين العناصر والتعامل معها، ولها خواص من أهمها المسافة لا تكون سالبة وتبادلية. وتعرف المسافة أيضاً على أنها الحيز أو الفضاء الفاصل بين عنصرين والعلاقة بين التشابه والمسافة هي علاقة عكسية ويمكن إجراء التحليل العنقودي بالاستناد على أي منهما.
- الشجرة البيانية (The Diagram or Dendogram) : هي الشكل الهرمي الناتج بعد أجراء عملية العنقدة ويمكن الوصول إليها وفق الطريقتين التاليتين:
  - طريقة التجميع Agglomerative
  - طريقة التجزئة Divisive

ومن الملاحظ أن النتائج التي تظهرها الطريقتين هي شجرة تسلسل هرمي حيث يطلق على بداية الشعبة بالجدر ونقاط التفرعات بالعقد، والعقدة النهائية أو الأخيرة على الشجرة ليس لها تفرعات يطلق عليها الأوراق وهي تمثل العناصر التي اجتمعت مع بعضها وكل واحدة من العقد في الشجرة ويضمنها الجذر تمثل مجموعة نوعية لكافة الأشياء التي يمكن الوصول إليها في تلك العقدة باتجاه المقدمة ومن خلال الشجرة (عبد الله, 1995). والعنصر الذي لا يتمكن من الدخول إلى أي عنقود يطلق عليه القزم runt (كاظم ,2006).

- **مصفوفة البيانات Data Matrix**

في عملية التحليل العنقودي تستعمل بيانات غالباً ما تكون بيانات كمية (أعداد)، وبيانات نوعية أو بيانات مختلطة من كلا النوعين، تمثل البيانات بصورة أساسية مشاهدات لبعض العمليات الطبيعية في فضاء متعدد الأبعاد، إذ أن كل بعد من هذه الأبعاد يمثل خاصية معينة لوصف مشاهدات العناصر.

إن مجموعة العناصر يمكن تمثيلها بمصفوفة ذات بعد $(n*p)$ ، إذ أن هناك n من الصفوف تمثل كل واحدة منها مشاهدات العناصر $[x_j / j = 1,2,...,n]$ و p من الأعمدة تمثل كل واحدة منها خاصية من خواص العناصر $x_j \in R^p ; \left[x_j = (x_{j1}, x_{j2}, ..., x_{jp})' \right]$.

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdot & \cdot & \cdot & x_{1p} \\ x_{21} & x_{22} & \cdot & \cdot & \cdot & x_{2p} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ x_{n1} & x_{n2} & \cdot & \cdot & \cdot & x_{np} \end{bmatrix}$$

عند مجال تمييز النموذج يطلق على صف المصفوفة X بالنماذج أو العناصر، ويطلق على الأعمدة بالخواص أو الصفات المميزة، أما المصفوفة X فتدعى بمصفوفة النموذج pattern matrix، وغالباً ما يطلق على مصفوفة X في التطبيقات العملية بمصفوفة البيانات، ويعتمد تعريف صفوف وأعمدة المصفوفة على وفق المجالات المستعملة في عملية العنقدة (جودة، 2008).

- **التصنيف Classification**

وهو ترتيب الأشياء اعتماداً على ما بينها من تشابه أو اختلاف أو قد ترتب هذه الأشياء وفق أكثر من أسلوب، أي أنه يمكن وضع أكثر من ترتيب للعناصر أو الأشياء، وذلك حسب صفة التشابه أو الاختلاف قيد الاهتمام.

- **مصفوفة القرابة The Proximities Matrix**

وهي المصفوفة التي عادة ما يبدأ التحليل العنقودي بتكوينها ، وتكون متماثلة وهي عبارة عن جدول للتشابه النسبي بين العناصر قيد الدراسة, وكل قيمة من قيم هذه المصفوفة تمثل المسافة بين عنصرين " معامل الارتباط في حالة دراسة المتغيرات " ويكون الشكل العام لها :

$$D = \begin{pmatrix} d_{11} & d_{12} \cdots & d_{1m} \\ d_{21} & d_{22} \cdots & d_{2m} \\ \vdots & \vdots & \vdots \\ d_{m1} & d_{m2} & d_{mm} \end{pmatrix}$$

حيث m عدد العناصر(العزاوي,2013).

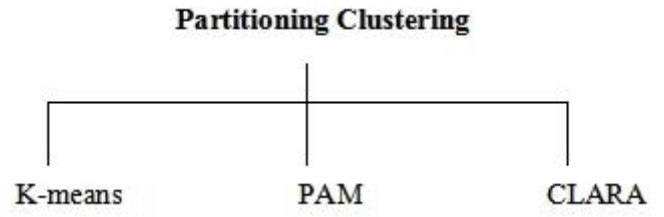
طرائق التحليل العنقودي Cluster Analysis Methods

هنالك نوعان رئيسيان من أنواع التحليل العنقودي الحاد وهما التحليل الهرمي والتحليل اللاهرمي حيث يرتكز هذا البحث على الجزء الغير هرمي فقط.

طريقة التعنقد غير الهرمية Nonhierarchical Clustering Method

هي أسلوب يعد مفيداً جداً في العديد من التطبيقات التي تتضمن مجموعة كبيرة من البيانات بحيث يصبح من الصعب بناء الشكل الهرمي لمثل هذه البيانات، تهدف العنقدة اللاهرمية إلى تجزئة مجموعة من البيانات إلى عدد K من المجاميع العنقودية المتجانسة، وذلك بالاعتماد على فكرة النزعة المركزية التي هي نقطة الوسيط لمجموعة من النقاط (بإمكان أي نقطة مركزية واحدة أن تمثل عنقود) . ومن أهم طرق العنقدة اللاهرمية هي (طريقة العنقدة باستخدام K من المتوسطات (k-Mean Clustering)، يعتمد التحليل العنقودي بالدرجة الأساسية على التعامل المباشر مع البيانات لذلك فان الأمر يتطلب تحويل البيانات إلى حالة المتغير الواحد (القيمة القياسية)، وبعدها يتم استعمال أساليب التحليل العنقودي لاكتشاف التقارب بين المتغيرات وذلك بالاعتماد على مقاييس خاصة تدعى بمقاييس المسافة التي يتم بواسطتها قياس مدى التشابه أو الاختلاف بين العناصر اعتماداً على الصفات المقرونة بكل عنصر(العزاوي، 2013).

حيث يمكن تقسيم طرق التعنقد كما في الشكل(1):



شكل (1) يوضح طرق التعنقد غير الهرمي

1- طريقة المتوسطات (K-means)

من أهم طرق العنقدة غير الهرمية طريقة المتوسطات k-means، إذ تعتمد الفكرة الأساسية لـــ k-means على تعريف المجموعات (العناقيد-clusters) بحيث يتم تقليل التغير داخل العنقود والمعروف باسم التباين الكلي داخل العنقود (total within-cluster variation) (Kassambara,2017).

حيث هنالك العديد من الخوارزميات حول k-means, وتعتبر الخوارزمية القياسية هي خوارزمية (Hartigan-Wong,1979), والتي تحدد التباين الكلي داخل العنقود كمجموع لمربع المسافة الإقليدية بين العناصر والقيمة الوسطى كما يلي :

$$W(C_K) = \sum_{x_i \in c_k} (x_i - \mu_k)^2$$

حيث أن :

x_i : نقطة من البيانات المنتمية للعنقود C_k .

μ_k : القيمة المتوسطة (المراكز العنقودية) للنقاط المنتمية للعنقود C_k .

يتم توزيع المشاهدات x_i الى عناقيد معينة (الخاصة بها), بحيث يكون مجموع مربع المسافة SS من المشاهدات x_i إلى المراكز العنقودية الخاصة μ_k بها أقل ما يمكن.

ويمكن حساب التباين الكلي داخل العنقود (total within-cluster variation) من خلال العلاقة التالية (Hartigan&Wong,1979):

$$wss = \sum_{k=1}^{k} W(C_K) = \sum_{k=1}^{k} \sum_{x_i \in c_k} (x_i - \mu_k)^2$$

وبالتالي يمكن القول بان وظيفة هذه الطريقة تتلخص في تقسيم البيانات إلى مجموعات متشابهة، وتكمن الفكرة الأساسية في البحث حول K من المتوسطات والتي تُعتمد في عملية العنقدة للبيانات موضوع البحث، وتكون البيانات قريبة جداً من متوسطها إذا كان التباين لمجموعة من البيانات ذو قيمة صغيرة، وفي التحليل العنقودي تهدف إلى قياس التقارب في البيانات حول المتوسط، هنا نقسم البيانات إلى عناقيد كل منها لها متوسط خاص بها وعليه سوف يتم الاعتماد على تباين العنقود Cluster Variance (Johnson&Wichern,2007).

i. خوارزمية k-means

يمكن تلخيص خطوات عمل طريقة k-means كما يلي :

1- تحديد عدد العناقيد K التي سيتم إنشاءها بواسطة المُحلل .
2- حدد عناصر K عشوائياً من مجموعة البيانات كمراكز ابتدائية للعنقود.
3- تُعين كل مشاهدة إلى أقرب نقطة وسطى، استنادا إلى المسافة الإقليدية بين العنصر والمركز.
4- لكل مجموعة من مجموعات العناقيد K تحديث لمركز العنقود عن طريق حساب القيمة المتوسطة الجديدة لكافة قيم البيانات في العنقود . مركز العنقود ذو الرتبة

K_m يمثل متجه الطول P والذي يحتوي على متوسطات جميع المتغيرات لكل المشاهدات في العنقود K_m , P عدد المتغيرات.

5- تكرارياً قم بتقليل التباين الكلي داخل العنقود (total within sum of square (total.withinss)), بمعنى تكرار الخطوات 3و4 حتى يتوقف تعينات العنقود أو الحصول على العدد الأقصى لعدد التكرارات المحدد .

ii. مزايا وعيوب k-means

k-means clustering advantages and disadvantages

تعتبر خوارزمية k-means من الخوارزميات البسيطة جداً والسريعة، وبالرغم من فعاليتها في التعامل مع مجموعات البيانات الكبيرة إلا أنه هنالك نقاط ضعف في هذه الخوارزمية ومنها:

1- يفترض معرفة مسبقة للبيانات ويتطلب من المحلل (analyst) اختيار العدد المناسب من العناقيد K مسبقاً.
2- النتائج النهائية التي يتم الحصول عليها حساسة للاختيار العشوائي الأولي لمراكز العنقود، ويرجع السبب في ذلك لأنه لكل دورة مختلفة من الخوارزمية على نفس مجموعة البيانات يمكن اختيار مجموعة مختلفة من المراكز الأولية، مما يؤدي إلى ظهور نتائج مختلفة للتجميع (clustering) خلال دورات مختلفة من الخوارزمية.
3- حساسة للقيم المتطرفة (outliers).
4- عند إعادة ترتيب البيانات، فمن المحتمل الحصول على حل مختلف في كل مرة تغير فيها الترتيب.

ويمكن التغلب على نقاط الضعف السابقة من النقاط التالية:

1- حساب k-means لمجموعة مختلفة من قيم K , مثلا تغيير قيم K بين 2 و10 ومن ثم اختيار أفضل k من خلال مقارنة نتائج التجميع المتحصل عليها عند قيم مختلفة لـ K
2- حساب خوارزمية k-means عدة مرات مع تغيير المراكز الأولية للعنقود، واختيار خوارزمية الحل التي تعطي أدنى حد للتباين داخل العنقود.

3- لتجنب التغيرات الناجمة عن القيم المتطرفة يمكن استخدام خوارزمية PAM والتي تعتبر أقل حساسية للقيم المتطرفة (Kassambara, 2017)

2- **طريقة K-Mediods**

تعتبر خوارزمية K-Mediods طريقة من طرق التجميع المتعلقة بطريقة k-means لتقسيم مجموعة البيانات إلى K مجموعة أو عنقود, حيث يتم تمثيل كل عنقود في k-mediods بواسطة نقطة واحدة في بيانات العنقود, ويطلق على هذه النقاط اسم cluster medoids, ويشير مصطلح mediods إلى عنصر داخل العنقود بحيث لا يختلف فيه المتوسط بينه وبين باقي عناصر العنقود ويكون هو الحد الأدنى, بحيث يقابل النقطة الأكثر مركزية في العنقود.تعتبر طريقة K-Mediods بديلاً قوياً لطريقة k-means وذلك لان خوارزميتها أقل تأثراً بالقيم المتطرفة بالمقارنة مع k-means, وذلك لأنها تستخدم mediods كمراكز عنقودية بدلاً عن المتوسطات المستخدمة في k-means .

خوارزمية K-Mediods تتطلب من المُحلل تحديد عدد العناقيد K مسبقاً كما هو الحال في طريقة k-means, وتعتبر طريقة silhouette الطريقة الأمثل لتحديد عدد العناقيد والتي سيتم التطرق إليها لاحقاً (Kassambara, 2017).

ومن أكثر طرق K-Mediods الشائعة والمستخدمة بكثرة خوارزمية

PAM(Partitioning Around Mediods) , (Kaufman & Rousseeuw , 1990)

- **خوارزمية PAM** **PAM algorithm**
- تعتبر خوارزمية PAM أكثر قوة من خوارزمية k-means التي غالباً ما تتأثر بالقيم المتطرفة مما يؤثر بشدة على تخصيص المشاهدات للعناقيد .

ويمكن حساب الخوارزمية كما يلي:

1- اختر عناصر K لتكون mediods .
2- قم بحساب مصفوفة الاختلاف إذا لم يتم توفيرها.
3- تعيين كل عنصر إلى أقرب mediods .
4- لكل عنقود ابحث عن عنصر يقلل متوسط معامل الاختلاف وإن وجد حدد العنصر الذي يقلل هذا المعامل على أنه mediods لهذا العنقود.

5- إذا تغير mediods مرة على الأقل كرر الخطوة 3 , وإلا قم بإنهاء الخوارزمية.

ومن الملاحظ أن خوارزمية PAM تعتمد على مصفوفة الاختلاف والتي يمكن حسابها باستخدام أحد المقياسين التاليين:

- المسافة الإقليدية Euclidean distances (مجموع مربعات الاختلافات).
- Manhattan distances (مجموع المسافات المطلق).

3- طريقة CLARA

تعتبر هذه الطريقة امتداد لطريقة k-medoids حيث تتعامل مع البيانات التي تحتوي على عدد كبير، وسنتطرق في بحثنا هذا إلى الطريقة k-medoids لأنها تتناسب مع حجم بيانات البحث.

تحديد العدد المثالي للعناقيد Determining the Optimal Number of Cluster

يعتبر تحديد العدد الأمثل للعناقيد لمجموعة من البيانات القابلة للعنقدة أمر أساسي لتقسيم هذه البيانات إلى عناقيد كما هو الحال في طريقة k-means حيث يطلب من المُحلل تحديد عدد العناقيد K , ولكن تحديد هذا العدد يعتبر ذاتي وغير موضوعي ويعتمد على الطريقة المستخدمة لقياس أوجه التشابه والمعايير المستخدمة في التقسيم, ويوجد عدة طرق مختلفة يمكن استخدامها لتحديد العدد المثالي للعناقيد سواء كان ذلك لطريقة k-means أو k-medoids(PAM) وتتضمن هذه الطرق :

1- الطريقة المباشرة Direct methods : وتتضمن هذه الطريقة تحسين المعيار مثل مجموع المربعات sums of squares, والصورة الظلية المتوسطة average silhouette للعنقود, وتكون التسمية المقابلة لهذه الطرق : طريقة الكوع elbow method والصورة الظلية silhouette method على التوالي.
2- طرق الاختبار الإحصائي:statistical testing methods تعتمد هذه الطريقة على مقارنة الأدلة ضد فرض العدم ومن الأمثلة على هذه الطريقة طريقة الفجوة الإحصائية the gap statistic .

والآن سيتم التطرق إلى الطرق السابقة وإعطاء نبذة عليها والتعرف على خوارزمية عملها:

- **طريقة الكوع Elbow method**

تعتمد أساليب التقسيم المتطرق إليها سابقاً مثل k-means إلى تعريف العناقيد بحيث

يتم تقليل التغير الكلي بين المجموعات إلى الحد الأدنى بمعنى أن يكون wss التباين الكلي داخل العنقود صغيراً قدر الإمكان .

$$wss = \sum_{k=1}^{k} W(C_k) = \sum_{k=1}^{k} \sum_{x_i \in C_k} (x_i - \mu_k)^2$$

ويمكن التعرف على العدد الأمثل للعناقيد كما يلي :

1- حساب خوارزمية التجميع مثلا لـ k-means عند قيم مختلفة لـ K وذلك من خلال تغيير قيم K من 2 إلى 10 عناقيد .
2- لكل K نقوم بحساب المجموع الإجمالي للمربعات (wss) داخل العنقود .
3- يتم رسم منحنى wss وفقاً لعدد العناقيد K .
4- يتم اعتبار موقع الانحناء (knee) في الشكل بشكل عام مؤشراً على العدد المناسب للعناقيد (Kassambara,2017). ولكن هذه الطريقة تعتبر غامضة ambiguous في بعض الأحيان والبديل لها هو أسلوب متوسط الظل Average Silhouette الذي يمكن استخدامه مع أي طريقة تجميع (Kaufman&Rousseeuw,1990).

- **طريقة الصورة الظلية المتوسطة (متوسط الظل) Average Silhouette Method**

تعتبر طريقة الصورة الظلية من الطرق التي تعتمد على قياس جودة العنقود، حيث تُحدد كيفية انضمام كل عنصر إلى العنقود الخاص به، بمعنى أنها تعمل على قياس جودة التعنقد(Kassambara,2017) .

حيث تعمل هذه الطريقة على حساب متوسط الظل average silhouette للمشاهدات عند قيم مختلفة لـ k, ومن تم يتم تحديد العدد المثالي للعناقيد من بين مجموعة القيم المحتملة لـ k على أنه يحقق الحد الأقصى للصورة الظلية بمعنى أنه يعمل على زيادة متوسط الصورة الظلية(Kaufman&Rousseeuw,1990) .

- **خوارزمية متوسط الظل**

حساب العدد الأمثل في هذه الطريقة يشبه حسابه في طريقة الكوع ويتم حساب الخوارزمية كما يلي:

1- حساب خوارزمية التجميع مثلاً خوارزمية k-means , وذلك عند قيم مختلفة لـ K .
2- حساب متوسط الصورة الظلية (avg.sil) عند كل قيمة من قيم K .
3- ارسم منحنى ((avg.sil وفقاً لعدد العناقيد K .
4- يتم اعتبار موقع الحد الأقصى على أنه العدد المناسب للعناقيد .

- **طريقة إحصاءه الفجوة Gap Statistic Method**

تم نشر هذه الطريقة من قبل[Tibshirain,et.al ,2001], يمكن تطبيق هذه الطريقة على أي طريقة تجميع, وتقارن إحصاءه الفجوة بين التباين الكلي داخل العنقود wss عند قيم مختلفة لـ K مع قيمها المتوقعة expected values في ظل التوزيع المرجعي للبيانات reference distribution, ويتم اعتبار القيمة التي تعمل على تعظيم إحصاءه الفجوة (تنتج أكبر إحصاءه فجوة - largest gap statistic) على أنها العدد المثالي للعناقيد, بمعنى أن بنية التجميع بعيدة عن التوزيع العشوائي للنقاط random uniform distribution of point.

- **خوارزمية إحصاءه الفجوة**

تعمل الخوارزمية على النحو التالي:

1- تجميع البيانات، وتغيير عدد العناقيد $k = 1,2,\ldots,k_{\max}$, ومن تم القيام بحساب القيم المقابلة لإجمالي الاختلاف الكلي بين العناقيد wss :

$$wss = \sum_{k=1}^{k} W(C_k) = \sum_{k=1}^{k} \sum_{x_i \in C_k} (x_i - \mu_k)^2$$

2- إنشاء B من مجموعات البيانات المرجعية reference data sets في ظل توزيع عشوائي، ومن تم يتم حساب إجمالي الاختلاف لكل عنقود من مجموعة البيانات المرجعية وذلك عند قيم مختلفة لـ k , $k = 1,2,\ldots,k_{\max}$.

3- حساب القيمة التقديرية لاحصاءة الفجوة كانحراف لقيم w_k عن قيمها المتوقعة تحت فرض العدم

$$H_0 : Gap(k) = \frac{1}{B}\sum_{b=1}^{B} \log(W_{kb}^{*}) - \log(w_k)$$

كذلك حساب الانحراف المعياري للإحصاءات.

4- يتم اختيار عدد العناقيد كأصغر قيمة لـ k بحيث تكون إحصاءه الفجوة مع انحرافها المعياري عند $k+1$:

$$Gap(k) \geq Gap(k+1) - S_{k+1}$$

ومن المؤشرات التي سيتم استخدامها لقياس صحة التعنقد معامل الخيال ومؤشر دن. ومن الممكن استخدام هذين المؤشرين لتحديد العدد الأمثل للعناقيد.

- **معامل الظل Silhouette Coefficient**

يقيس تحليل الظل مدى جودة تجميع المشاهدات داخل العناقيد وتقدير متوسط المسافة بين العناقيد، لرسم البياني لـ Silhouette يقيس تقارب كل نقطة في عنقود واحد إلى نقاط أخرى في عناقيد مجاورة، ويتم حساب Silhouette Width (عرضه) لكل مشاهدة i ويرمز له بالرمز S_i كما يلي :

1- لكل مشاهدة i احسب متوسط الاختلاف a_i بين i وكل النقاط الأخرى في العنقود الذي تنتمي له i .

2- لكل العناقيد الأخرى C التي لا ينتمي لها i قم بحساب متوسط الاختلاف $d(i,C)$ من i إلى كل المشاهدات من C ، ويتم تعريف أصغر قيمة من $d(i,C)$ كما يلي :

$$b_i = \min_C d(i,C)$$

حيث يمكن اعتبار قيمة b_i كمقياس للاختلاف بين i والعنقود المجاور لها . أي أقرب عنقود لا تنتمي إليه .

3- أخيراً يتم حساب Silhouette Width للمشاهدة i كما يلي :

$$S_i = (b_i - a_i) / \max(a_i, b_i)$$

يمكن تفسير عرض Silhouette كما يلي:

- المشاهدات التي لها قيمة S_i كبيرة (تقريباً 1) تجميعها ممتاز .
- قيمة S_i صغيرة (حول الصفر) يعني المشاهدة تقع بين عنقودين .
- S_i سالبة يعني أنه من الممكن أنه تم وضع المشاهدات في العنقود الخاطئ .

2.9.3 مؤشر دن Dunn index

هو عبارة عن تدقيق آخر للتجميع الداخلي ويمكن حسابه كما يلي :

1- لكل عنقود قُم بحساب المسافة بين كل عنصرين في العنقود و العناصر الموجودة في عنقود آخر.
2- استخدم أقل قيمة من المسافات الزوجية كفاصل بين العناقيد (min.separation) .
3- لكل عنقود احسب المسافة بين عناصر نفس العنقود.
4- استخدم المسافة القصوى (القطر الأقصى maximum diameter) كاكتناز داخل العنقود .
5- احسب مؤشر دن D كما في الصيغة التالية:

$$D = \frac{\min .separation}{\max .diameter}$$

ملاحظة: إذا كانت مجموعة البيانات مقسمة بشكل جيد فإن قطر العنقود من المتوقع أن يكون صغيراً والمسافة بين العناقيد من المتوقع أن تكون كبيرة وبالتالي يجب تكبير مؤشر دن (Kassambara,2017).

3. بيانات البحث Search Data

تتضمن بيانات البحث قراءات لبعض المعادن الثقيلة الداخلة في تركيب الحليب كما هو موضح في الجدول (1)، وهذه المعادن هي (الصوديوم Na، البوتاسيوم K، الكالسيوم Ca، الماغنيسيوم Mg، الحديد Fe، الزنك Zn، النحاس Cu) ، والتي تعتبر من المعادن الرئيسية للحليب، والتي سيتم من خلالها دراسة تعنقد (ترابط) هذه العناصر مع بعضها البعض.

جدول (1): يبين أنواع العينات المدروسة

ت	اسم العينة	رمز العينة
1	حليب الريحان كامل الدسم	U1
2	حليب السهول كامل الدسم	U2
3	حليب الزهرات كامل الدسم	U3
4	حليب الربيع	U4
5	حليب جهينة كامل الدسم	U5
6	حليب كانديا كامل الدسم	U6

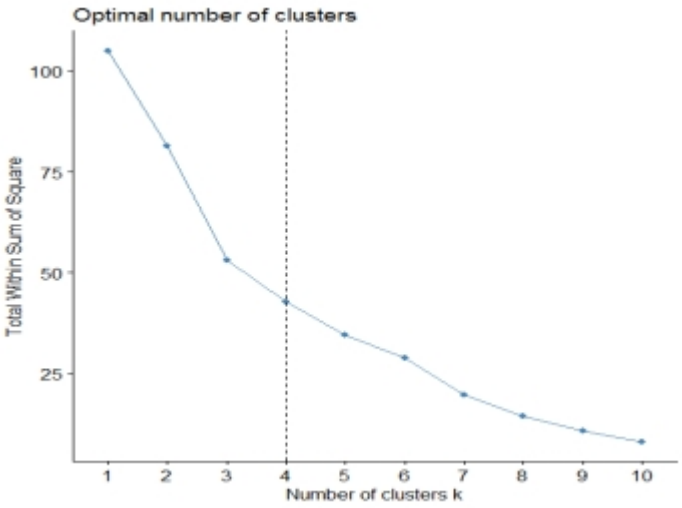
U7	حليب الريحان خالي الدسم	7
U8	حليب جهينة خالي الدسم	8
P1	حليب مبستر (طرابلس)	9
P2	حليب مبستر (مصراتة)	10
R1	حليب أبقار خام(منطقة الدافنية)	11
R2	حليب أبقار خام(منطقة الغيران)	12
R3	حليب أبقار خام(منطقة قصر أحمد)	13
R4	حليب أبقار خام(منطقة كرزاز)	14
R5	حليب أبقار خام(منطقة طمينة)	15
R6	حليب أبقار خام(منطقة السكت)	16

جدول(2) محتوى بعض العناصر المعدنية (جزء/مليون) لعينات الحليب (بن حليم,2016)

sample	Zn	Mg	Cu	Fe	Ca	K	Na
U1	2.53	61.96	0.147	1.23	562.7	798.7	462.9
U2	1.54	60.95	0.155	1.26	518.4	793.8	560.3
U3	1.24	72.93	0.165	2.18	632.7	861.7	624.7
U4	1.86	71.41	0.174	1.12	526.9	769.6	571.4
U5	2.34	80.44	0.148	1.87	583.8	898	489.5
U6	2.31	80.19	0.188	1.11	544.6	894.7	512.5
U7	2.11	80.23	0.164	4.33	563.7	901.6	514.3
U8	2.23	73.76	0.163	2.17	638.9	827.8	350.4
P1	2.03	73.64	0.185	1.68	597.3	852.9	556.5
P2	1.98	80.27	0.165	1.57	611.6	893.7	554.3
R1	1.98	75.92	0.305	2.32	682.7	885.7	439.4
R2	1.57	65.9	0.354	2.46	675.3	865.2	435.5
R3	1.96	72.61	0.321	2.75	691.3	893.3	493.3
R4	1.76	69.3	0.421	3.1	701.8	834.8	434.8
R5	2.1	71.8	0.636	2.81	683.8	799.8	469.8
R6	1.97	78.5	0.045	2.6	674.6	879.6	449.6

طريقة k-means .

- **تحديد العدد المثالي للعناقيد .**عند إجراء التحليل باستخدام طريقة k-means يستوجب الأمر تحويل البيانات إلى الصورة المعيارية كخطوة أولى, ومن تم يتم استخدام الطريقة البيانية لتحديد العدد المثالي للعناقيد, وذلك من خلال استخدام الشكل البياني بطريقة الكوع Elbow والذي يحدد بمكان الانحناء keen كما هو موضح بالشكل (2) .



شكل (2) يوضح العدد المثالي للعناقيد لطريقة K-means

من خلال الشكل (2) يتضح أن العدد المثالي للعناقيد هو4 عناقيد، بعد ما تم تحديد العدد المثالي للعناقيد الخاص بطريقة k-means الموضحة سابقاً يمكن إجراء خطوات التعنقد الخاصة بها وذلك من خلال إيجاد مراكز العناقيد الموضحة بالجدول التالي:

جدول (3) يوضح مراكز العناقيد لطريقة k-means

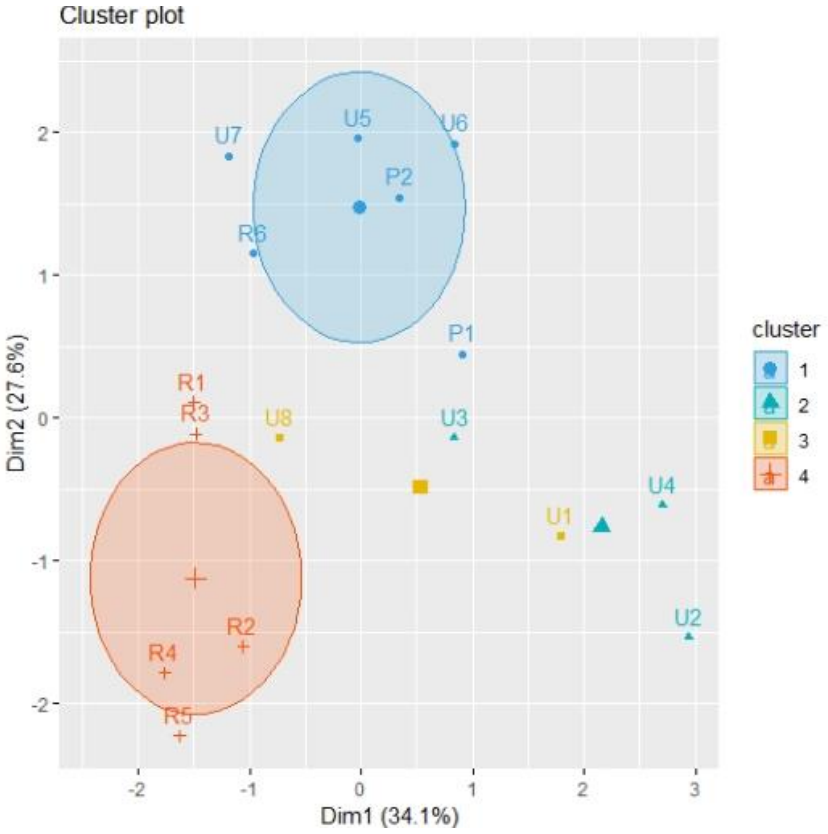
	Zn	Mg	Cu	Fe	Ca	K	Na
1	0.4702114	0.9215639	-0.5884491	0.03858549	-0.3530360	0.76860439	0.2613799
2	-1.2910134	-0.7485938	-0.4802954	-0.7408413	-0.9315230	-1.0260937	1.3266864
3	1.2541091	-0.8397078	-0.5477461	-0.5324796	-0.2756365	-0.9142829	-1.2941971
4	-0.2912893	-0.3208373	1.2134146	0.61119408	1.0949716	0.05904416	-0.59198

من خلال نتائج الجدول رقم (3) يتضح لنا مراكز العناقيد، والتي تمثل بمصفوفة فيها الصفوف عدد العناقيد من 1 إلى 4 , والأعمدة المتغيرات الرئيسية, وكذلك نجد أن متجه العناقيد كما يلي :

جدول (4) يوضح متجه العناقيد لطريقة k-means

البيانات	R6	R5	R4	R3	R2	R1	P2	P1	U8	U7	U6	U5	U4	U3	U2	U1
رقم العنقود	1	4	4	4	4	4	1	1	3	1	1	1	2	2	2	3

حيث يحتوي المتجه على أعداد صحيحة من 1 إلى k تشير مواقع البيانات في العناقيد، حيث يتضح أن العنقود الأول به 6 عناصر والثاني به 3 عناصر والثالث به عنصران أما العنقود الرابع والأخير فبه 5 عناصر، والتي يمكن تمثيلها من خلال أشكال بيضاوية حول نقطة تمركز العنقود، أما فيما يخص النقاط التي تقع خارج الشكل فإنها تمثل أقرب نقطة مسافة من تمركز العنقود، حيث يمكن تميز كل عنقود بلون كما في الشكل (3):

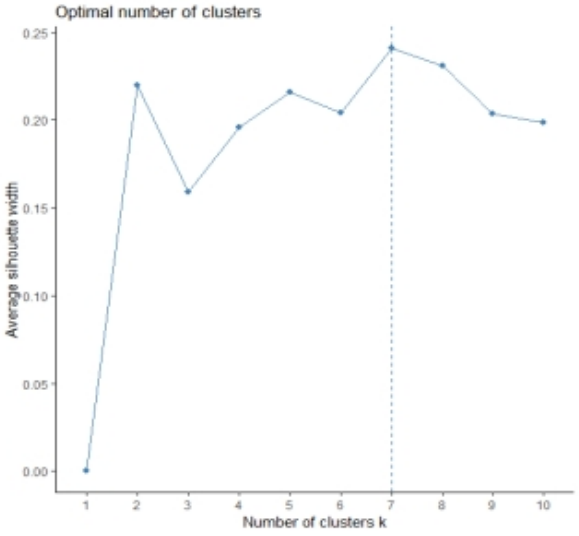


شكل (3) يوضح تمركزات العناقيد لطريقة k-means

من الشكل (3) يتضح كيفية تصنيف البيانات في 4 عناقيد حيث احتوى العنقود الأول على أنواع الحليب التالية (U5, U6, U7, P1, P2, R6), وكذلك الأنواع U6, P2, (U5) واقعة داخل الشكل البيضاوي, بينما كانت أنواع الحليب التالية (U7, P1, R6) خارجه حيث تعتبر المسافة بينها وبين مركز العنقود أقرب ما يمكن, واحتوى العنقود الثاني على الأنواع (U2, U3, U4), أما بالنسبة للعنقود الثالث فكان يحتوي على نوعين فقط وهما (U1, U8)، واحتوى العنقود الرابـع الأخيـر علـى أنـواع الحليـب (R1, R2, R3,R4, R5), حيث كـانت فيه أنـــواع الحليـب (R1, R3, R5) خارج الشكل البيضاوي, وهذا يعني أن الأنواع الموجودة في كل عنقود أكثر تجانساً لبعض من ناحية التركيب والمعادن وفقاً للمسافة الإقليدية.

طريقة PAM. تعتبر طريقة K-Mediods بديلاً قوياً لطريقة k-means وذلك لان خوارزميتها أقل تأثراً بالقيم المتطرفة بالمقارنة مع k-means، وذلك لأنها تستخدم mediods كمراكز عنقودية بدلاً عن المتوسطات المستخدمة في k-means، حيث تعتبر طريقة PAM من أكثر الطرق الشائعة والأكثر استخداماً، والتي تعتمد على نقطة تمركز البيانات.

- **تحديد العدد المثالي للعناقيد:** تتطلب هذه الطريقة تحديد العدد المثالي للعناقيد كما هو الحال في طريقة k-means حيث يتم استخدام الطريقة البيانية (طريقة المتوسط الظلي) لتحديد العدد المثالي للعناقيد والذي يمكن تحديده كما هو موضح بالشكل (4):



شكل (4) يوضح العدد المثالي للعناقيد لطريقة PAM

من خلال الرسم يتضح أن العدد المثالي للعناقيد لهذه الطريقة هو 7 عناقيد، بعد ما تم تحديد العدد المثالي للعناقيد الخاص بطريقة PAM الموضحة سابقاً يمكن إجراء خطوات التعنقد الخاصة بها وذلك من خلال إيجاد نقطة تمركز البيانات الموضحة بالجدول (5):

جدول (5) يوضح نقاط تمركز البيانات لطريقة PAM

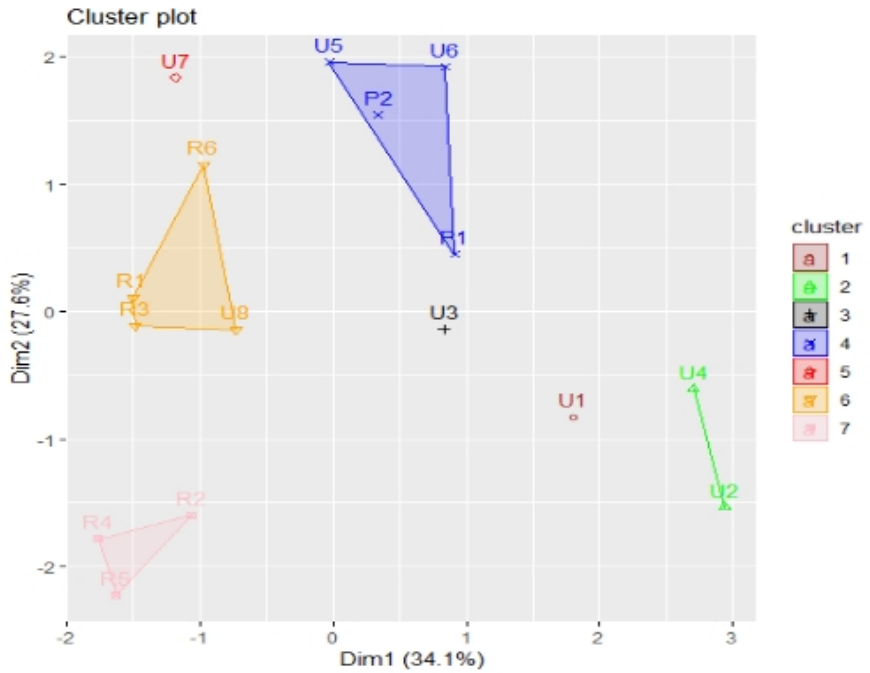
	Zn	Mg	Cu	Fe	Ca	K	Na
U1	1.71223121	-1.7828181	-0.6035674	-1.0765350	-0.8815795	-1.2474259	-0.4697510
U4	-0.3340473	-0.2722431	-0.4151706	-1.2038671	-1.4509433	-1.9137119	1.1205138
U3	-2.2276185	-0.0292723	-0.4779695	0.02315129	0.2317016	0.1950489	1.9017223
P2	0.03245031	1.14402075	-0.4779695	-0.6829630	-0.1038731	0.9277346	0.8698822
U7	0.42948943	1.13762679	-0.4849472	2.51191508	-0.8656755	1.1086163	0.2836094
R1	0.03245031	0.44867673	0.4989025	0.18521033	1.0269024	0.7445631	-0.8141863
R4	-0.6394620	-0.6095249	1.3083108	1.08811068	1.3306691	-0.4200864	-0.8816077

من خلال نتائج الجدول (5) يتضح لنا العناقيد بطريقة mediods, والتي تمثل بمصفوفة فيها الصفوف نقطة تمركز البياناتmedoids، والأعمدة المتغيرات الرئيسية, وكذلك نجد أن متجه العناقيد كما يلي :

جدول (6) يوضح متجه العناقيد لطريقة PAM

البيانات	**R6**	**R5**	**R4**	**R3**	**R2**	**R1**	**P2**	**P1**	**U8**	**U7**	**U6**	**U5**	**U4**	**U3**	**U2**	**U1**
رقم العنقود	6	7	7	6	7	6	4	4	6	5	4	4	2	3	2	1

حيث يحتوي المتجه على أعداد صحيحة من 1 إلى k تشير مواقع البيانات في العناقيد، حيث يمكن ربط البيانات داخل كل عنقود برسم إطاري كما في الشكل (5)



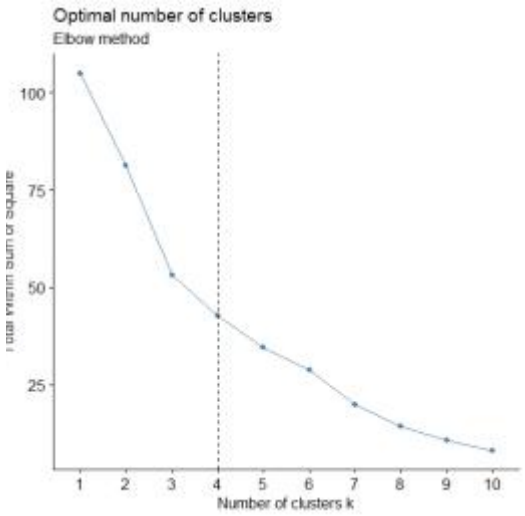
شكل (5) يوضح ربط البيانات في عناقيد لطريقة PAM

من الشكل (5) يتضح كيفية تصنيف البيانات في 7 عناقيد, حيث أحتوى العنقود الأول على نوع واحد من الحليب U1, والعنقود الثاني أحتوى على نوعين حليب (U2, U4), واحتوى العنقود الثالث على نوع الحليب U3, بينما ضم العنقود الرابع أكبر عدد من البيانات والمتمثلة في أنواع الحليب التالية (U5, U6, P1, P2), والعنقود الخامس به النوع U7 فقط , بينما ضم العنقود السادس الأنواع التالية (R1, R3, R6, U8), وكان العنقود السابع الأخير يحتوي على أنواع الحليب (R2, R4, R5), حيث تم استخدام الشكل الإطاري لأنه يقوم بحساب نقاط التواصل بين البيانات بناءً على الإطار, ومن تم تقسيمها إلى مجموعات, وبمقارنة النتائج التي تحصلنا عليها مع ما سبق من دراسات في نفس المجال نلاحظ أن نتائجنا المتحصل عليها في تفضيل k-means على k-mediods تتفق مع النتائج التي توصل إليها كلا من (Batra,2011, Velmurugan,2012) عند إجراءهم دراسة للمقارنة بين خوارزمية k-means و خوارزمية k-mediods حيث كانت خوارزمية k-means أفضل أداءً من k-mediods. **التأكد من العدد المثالي للعناقيد:** بعد إجراء التحليل ومعرفة أفضل طرق الربط يمكن الآن التأكد من العدد الأمثل للعناقيد , حيث يعتبر تحديد العدد الأمثل للعناقيد

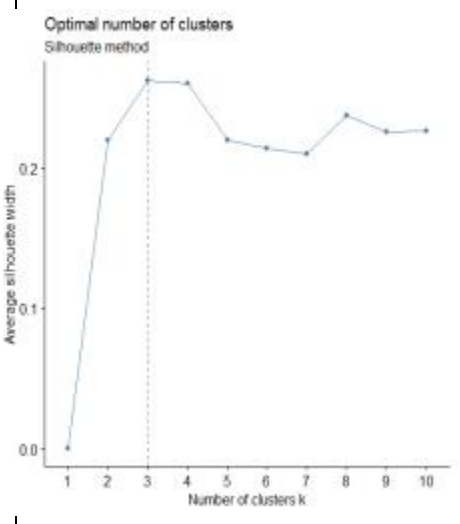
لمجموعة من البيانات القابلة للعنقدة أمر أساسي لتقسيم هذه البيانات إلى عناقيد لأي طريقة عنقدة.

تحديد العدد المثالي للعناقيد بطريقة k-means .

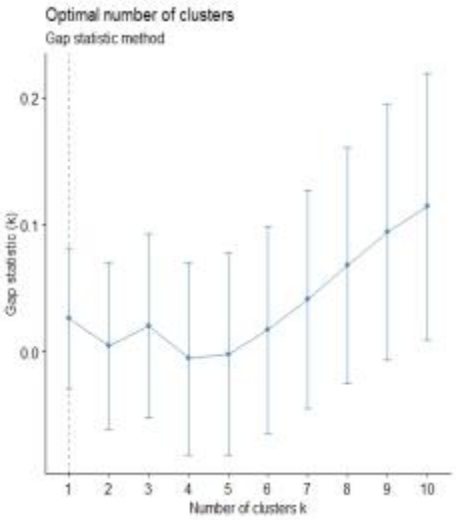
تم استخدام كل من الطرق التالية (Elbow و Silhouette و Gap Statistic) لتحديد العدد المثالي للعناقيد لطريقة k-means, حيث كانت النتائج على النحو التالي من خلال الأشكال التالية:



شكل (6)يوضح العدد المثالي للعناقيد بطريقة الكوع



شكل (7)يوضح العدد المثالي للعناقيد بطريقة الصورة الظلية



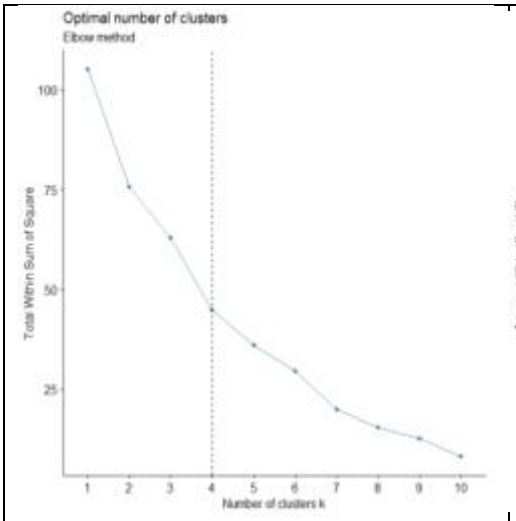
شكل (8)يوضح العدد المثالي للعناقيد بطريقة إحصاءه الفجوة

حيث أظهرت النتائج أن طريقة Elbow قامت بتجميع البيانات في 4 عناقيد، بينما قامت طريقة

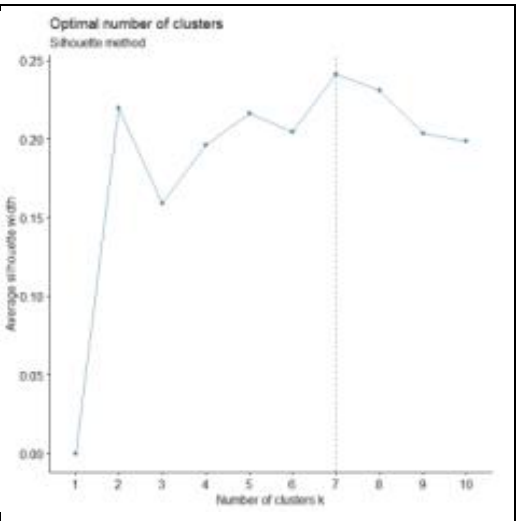
Silhouette بتجميع البيانات في 3 عناقيد، وقامت Gap statistic بتجميع البيانات في عنقود واحد.

- تحديد العدد المثالي للعناقيد بطريقة pam

تم استخدام كل من الطرق التالية (Elbow و Silhouette) لتحديد العدد المثالي للعناقيد لطريقة pam, حيث كانت النتائج على النحو التالي من خلال الأشكال التالية:

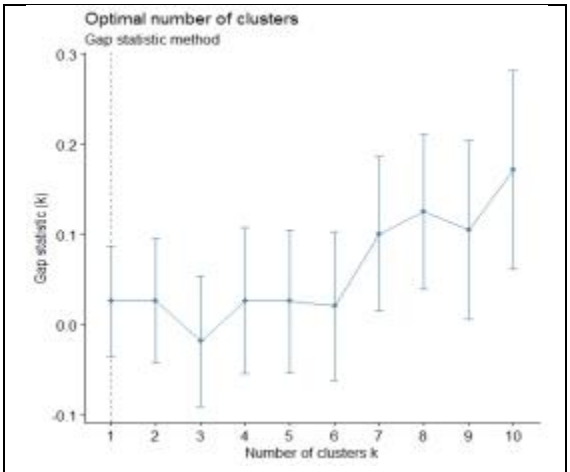


شكل (9)يوضح العدد المثالي للعناقيد بطريقة الظلية



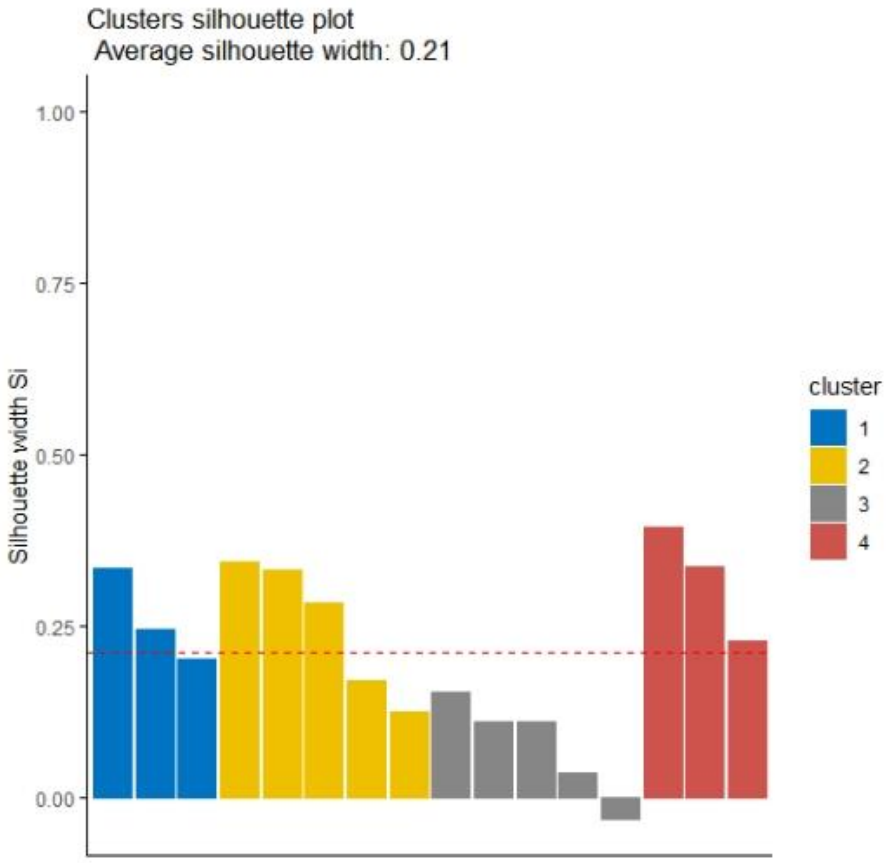
شكل (10)يوضح العدد المثالي للعناقيد بطريقة الكوع

حيث أظهرت النتائج أن طريقة Elbow قامت بتجميع البيانات في 4 عناقيد، بينما قامت طريقة Silhouette بتجميع البيانات في 7 عناقيد, بينما كانت نتيجة طريقة الفجوة الإحصائية تجميع كل أنواع الحليب في عنقود واحد.



شكل (11)يوضح العدد المثالي للعناقيد بطريقة الفجوة

التأكد من صحة العناقيد: من المعروف أن صحة التأكد من العناقيد يستخدم في العادة لتصميم إجراءات تقييم صلاحية نتائج خوارزمية التجميع , حيث تنقسم إحصاءات التأكد إلى ثلاث أقسام وهي: التحقق من العناقيد داخلياً، التحقق من العناقيد خارجياً، التحقق من العناقيد نسبياً، حيث يقتصر هذا البحث على التحقق من العناقيد داخلياً فقط, وذلك لأنها تعتمد على استخدام المعلومات الداخلية لمعالم العنقود لتقييم أفضل تركيب (بُنية) للعناقيد ويتم ذلك دون الرجوع إلى معلومات خارجية، حيث يمكن من خلالها تقدير عدد العناقيد و ملائمة خوارزمية التعنقد دون أي معلومات خارجية، بمعنى آخر يتم تجميع نتائج الخوارزمية بناءً على المعلومات الواردة في البيانات نفسها من خلال مجموع مربعات الخطأ [Redon,et.al,2011], وقبل القيام باختبار المعاملات الخاصة بهذه الطريقة يستوجب الأمر تحويل البيانات إلى الصورة المعيارية ومن تم استخدام إحدى طرق التجميع طريقة kmeans، وبعد القيام بخطوات التحليل كانت النتائج كما يلي:



شكل (12) يوضح العرض البياني للصورة الظلية

حيث يمكن حساب معامل silhouette لكل المتغيرات الرئيسية (أنواع الحليب) وعرضها في ثلاث أعمدة كما يلي:

جدول (7) يوضح معامل silhouette لكل أنواع الحليب

	Cluster	Neighbor	sil_width
U2	1	2	0.33456905
U4	1	2	0.24464758
U1	1	2	0.20100700
P2	2	3	0.34413916
U6	2	3	0.33192848
P1	2	1	0.28230486
U5	2	3	0.16997740
U3	2	3	0.12601035
R6	3	2	0.15315075
U8	3	4	0.11141637
R1	3	4	0.11038781
U7	3	2	0.03715911
R3	3	4	-0.03204234
R4	4	3	0.39307551
R5	4	3	0.33539274
R2	4	3	0.22906409

من خلال ما سبق يمكن تحديد القيمة السالبة ومكان تواجدها , وتظهر النتائج قيمة واحدة سالبة:

جدول (8) يوضح قيم معامل silhouette السالبة

	Cluster	Neighbor	sil_width
R3	3	4	-0.03204234

يلاحظ أن إحدى قيم $S_i = -0.0320423$ سالبة والتي تقابل R3 والتي تم وضعها في العنقود الثالث ويمكن اعتبارها أقرب للعنقود الرابع, وهذا يفسر أن R3 من الممكن أنها وضعت في العنقود الخاطئ, أما بالنسبة لباقي القيم فكانت صغيرة (حول الصفر) وهذا يفسر أن المشاهدات تقع بين عنقودين, فمثلاً U2 تم تصنيفها ضمن العنقود الأول وقيمة $S_i = 0.3345690$ وكانت قريبة من العنقود الثاني بمعنى أن U2 وقعت بين العنقودين الأول والثاني , وعند النظر إلى U4 تم تصنيفها ضمن العنقود الأول وقيمة $S_i = 0.2446475$ وكانت قريبة من العنقود الثاني بمعنى أن U4 وقعت بين العنقودين الأول والثاني, وهكذا لباقي العينات, وعند استخدام معامل دن Dunn index للتأكد من صحة العناقيد أظهرت النتائج أن قيمة مؤشر دن هي : 0.4528993, وكان عدد العناقيد 4, وهي نفس النتائج المتحصل عليها من معامل silhouette, وهذه النتيجة تتفق مع ما وجده (Redon E,2011) حيث وجد أن المؤشرات الداخلية ذات دقة عالية.

مقارنة خوارزميات التجميع: عند مقارنة خوارزميات التجميع يتم اعتبار مصفوفة البيانات فيها الصفوف هي العناصر المراد تجميعها والأعمدة هي العينات، وعند مقارنة خوارزميات التجميع وتحديد طرق التجميع واستخدام المؤشرات الداخلية تظهر النتائج ما يلي:

جدول (9) يوضح مقارنة خوارزميات التجميع

الطريقة	عدد العناقيد	2	3	4
	المعامل			
K-means	Connectivity	6.6381	16.7841	18.8948
	Dunn	0.4700	0.4337	0.4809
	Silhouette	0.2672	0.2625	0.2409

وبمقارنة قيمة معامل Silhouette في لطريقة K-means نلاحظ أن أفضل تقسيم لأنواع الحليب يكون بتجميعها في 4 عناقيد حيث كانت قيمته 0.2409

الاستنتـــاجـــــات Conclusions

من خلال دراسة التحليل العنقودي باستخدام أسلوبي التحليل الهرمي التجميعي والأسلوب غير الهرمي وتطبيقهم على أنواع مختلفة من الحليب المبستر والحليب طويل الأمد نستنتج ما يلي:

1- أظهرت النتائج أن خوارزمية k-means فعالة في عملية تصنيف أنواع الحليب مقارنة مع أداء خوارزمية k-medoids.
2- بتطبيق الطريقة المثلى للعنقدة والمتمثلة في k-means تم تقسيم أنواع الحليب إلى 4 مجموعات حيث تضمنت المجموعة الأولى أنواع الحليب التالية حليب جهينة كامل الدسم(U5), حليب كاديا كامل الدسم(U6)، حليب الريحان خالي الدسم (U7)، حليب مبستر مصراتة(P1)، حليب مبستر طرابلس(P2)، وحليب أبقار خام منطقة السكت، واحتوت المجموعة الثانية على الأنواع التالية حليب السهول كامل الدسم(U2), حليب الزهرات كامل الدسم(U3), وحليب الربيع(U4)، أما بالنسبة للمجموعة الثالثة فكانت تحتوي على نوعين فقط وهما حليب الريحان كامل الدسم(U1), وحليب جهينة خالي الدسم(U8), واحتوت المجموعة الرابعة على أنواع الحليب الخام من المناطق الدافنية الغيران قصر أحمد كرزاز وطمينة.
3- تقسيم أنواع الحليب إلى أربع عناقيد متجانسة وفقا طريقة k-means,pam, عند استخدام طريقة الكوع وطريقة الصورة الظلية وطريقة احصاءة الفجوة.
4- عند استخدام المؤشرات الداخلية وبالأخص معامل الصورة الظلية ومؤشر دن نتحصل على نفس النتائج التي تحصلنا عليها عند تطبيق طريقة k-means، حيث تم تقسيم أنواع الحليب إلى أربع عناقيد متجانسة.

التوصيـات Recommendations

من خلال الاستنتاجات التي تم التوصل إليها فإن الباحثة تتقدم بالتوصيات التالية:

1- استخدام خوارزمية k-means في تصنيف أنواع الحليب المختلفة في الأسواق المحلية.
2- إجراء الدراسة على عينة أكبر لتشمل كافة أنواع الحليب المبيعة بالأسواق المحلية بمدينة مصراتة، وعدداً أكبر من العناصر الداخلة في تركيب الحليب للاستفادة من النتائج للتعرف على نوعية الحليب.

3- الاهتمام بأساليب التنقيب عن البيانات بجانبيها المعلمي واللامعلمي وإدراجهما ضمن المقررات الدراسية لأقسام الإحصاء.

4- ضرورة التوسع في استخدام التحليل العنقودي ليشمل دراسات اقتصادية واجتماعية وتربوية، وذلك لنجاحه في تنظيم المشاهدات وتقسيمها إلى مجاميع متجانسة تتمتع بصفات مشتركة.

أولاً / قائمة المراجع العربية:

أ) الكتب:

1- جودة، محفوظ، (2008)، التحليل الإحصائي المتقدم باستخدام spss، عمان، الأولى، دار وائل.

ب) الرسائل العلمية:

2- عبد الله، وليد، (1995)، استخدام تحليل الانحدار والتحليل العنقودي في تشخيص العوامل المسببة لتصلب الشرايين، رسالة ماجستير، كلية الإدارة والاقتصاد، جامعة المستنصرية.

3- كاظم، فريال،(2006)، تحليل إحصائي للأهداف الإنمائية للألفية باستخدام أسلوب التحليل العنقودي، رسالة دبلوم عالي, المعهد العربي للبحوث الإحصائية، بغداد.

10 – العزاوي، إخلاص عبد الأمير، (2013) , تحليل إحصائي للعوامل المؤثرة على الوضع الاقتصادي للنساء ضمن نتائج مسح شبكة معرفة العراق, رسالة دبلوم عالي، بغداد.

11 -- بن حليم، انتصار، التقييم الكمي والنوعي للحليب المعقم والحليب المبستر التجاري في ليبيا، (2016)، رسالة ماجستير في الكيمياء، كلية العلوم، جامعة مصراتة.

ج) المجلات العلمية:

12 – السباح، شروق & الأعرجي، إمتناء، استخدام التحليل العنقودي السريع لفرز نوع الرضاعة حسب خصائص الأمهات، مجلة كربلاء العلمية- المجلد العاشر- العدد الرابع/علمي/2012.

13 - إبراهيم، علي& حميد، ماجد& حسن، صباح، (2005)، أثر التحليلين العاملي والعنقودي في نماذج من الحليب المجفف باستخدام نظام التطبيقات الإحصائية spss على الحاسبة الالكترونية، المجلة العراقية للعلوم الإحصائية العدد 8، ص ص(96-119)

14 - رشيد، أسيل عبد الرازق & مهدي، نبأ، (2011)، تحليل واقع التربية والتعليم باستخدام طرائق التحليل العنقودي، مجلة القادسية للعلوم الإدارية والاقتصادية المحور الإحصائي، كلية الإدارة والاقتصاد جامعة المستنصرية.

ثانياً / قائمة المراجع الإنجليزية:

1- Redon E., Abundez I., Arizmendi &Elvia M.,(2011) ,Internal versus External Cluster validation indexes , International Journal-Issue 1,volume 5.

2- Arbin N., Suhaimi N., Mokhtar N., Othman Z., (2015) .Comparative Analysis between k-means and k-mediods for statistical clustering, third international conference on Artificial Intelligence .

3- E L Feeney, A O'Sullivan, A P Nugent, B McNulty, J Walton, A Flynn & E R Gibney (2017), Patterns of dairy food intake, body composition and markers of metabolic health in Ireland: results from the National Adult Nutrition Survey, *Nutrition & Diabetes* volume 7, page e243 (2017)

4- Simone S.Souza[a]Adriano G.Cruz[b]Eduardo H.M.Walter[b]Jose A.F.Faria[b]Renata M.S.Celeghini[b]Márcia M.C.Ferreira[c]DanielGranato[d]Anderson de S.Sant'Ana[d], Monitoring the authenticity of Brazilian UHT milk: A chemometric approach,https://doi.org/10.1016/j.foodchem.2010.06.074Get rights and content

5- Zain M., Behkami S.,& Bakirdere S.,(2016) Milk authentication and discrimination via metal content clustering A case of comparing milk from Malaysia and selected countries of the world, Food control 66(2016) 306-314

Journal homepage: www.elsevier.com/locte/foodcont.

6- Kassambara A.,(2017) Multivariate Analysis I ,Practical Guide To Cluster Analysis in R , Edition 1

http://www.sthda.com, alboukadel.kassambara@gmail.com

7- Hartigan J. A., and Wong M. A., (1979). A K-means clustering algorithm.
8- Johnson R., and Wichern D.,(2007) Pearson Education ,6ed, applied multivariate statistical analysis, London.
9- Kaufman L., and Rousseeuw P.J., (1990). Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York.
10- Tibshirani R., Walther G., and Hastie T., (2001). Estimating the number of data clusters via the Gap statistic. Journal of the Royal Statistical Society B, 63, 411–423. PDF
11- Batra A.,(2011) Analysis and Approach: K-Means and K-Medoids Data Mining Algorithms, *Asst Professor, L. J. Institute of Computer Applications, Ahmedabad, India.,E–mail: batra.aishwarya@gmail.com*, 5th IEEE International Conference on Advanced Computing & Communication Technologies [ICACCT-2011].
12- Velmurugan T.,(2012) Efficiency of k-Means and K-Medoids Algorithms for Clustering Arbitrary Data Points, Vol 3 (5), 1758-1764 IJCTA | Sept-Oct 2012 A , E-Mail: velmurugan_dgvc@yahoo.co.in