المقارنة بين طرق التعنقد الهرمي واختيار أفضلها مع تطبيق عملي على بعض أنواع الحليب المبيع في مدينة مصراتة

، د.إبراهيم سليمان حنيش[12]خلود سليمان إسميو*
[1]قسم الإحصاء، كلية العلوم، جامعة مصراتة، ليبيا
[2]قسم االإحصاء، كلية العلوم، جامعة مصراتة، مصراتة، ليبيا
E-mail: khuloudbll@gmail.com

الملخص:
من الطرق الإحصائية المهمة والمستخدمة في التصنيف هي أسلوب التحليل العنقودي والتي تعتمد بدورها على تحليل متغيرات محددة تعتمد على نقاط التشابه والاختلاف بين البيانات حيث يندرج هذا الأسلوب ضمن أساليب التنقيب اللامعلمي للبيانات والذي يعتبر من المجالات المهمة والحديثة في علم الإحصاء.
من هذا المنطلق وكتطبيق عملي لهذا التحليل تم إجراء 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 الهرمي والمقارنة بينها ومعرفة أفضل طريقة عنقدة للبيانات.
أظهرت نتائج الدراسة أن أفضل الطرق الهرمية هي طريقة المتوسطات والتي قامت بتقسيم أنواع الحليب إلى أربع مجموعات متنافية، حيث تضمنت المجموعة الأولى ثلاث أنواع وهي حليب الريحان كامل الدسم (U1)، حليب السهول كامل الدسم (U2)،و حليب الربيع (U4)، ، بينما احتوت المجموعة الثانية على حليب أبقار خام (R5)، وضمت المجموعة الثالثة أحد عشر نوع وهم حليب أبقارخام(R1,R2,R3,R4,R6)، حليب جهينة كامل الدسم (U5) ، حليب كاديا كامل الدسم (U6)، حليب الريحان خالي الدسم (U7)، حليب جهينة خالي الدسم(U8)، حليب مبستر مصراتة (P1)، و حليب مبستر طرابلس(P2)، بينما ضمت المجموعة الرابعة حليب الزهرات كامل الدسم(U3) فقط.

الكلمات المفتاحية: التحليل العنقودي،طرق التعنقد الهرمي،مقياس العلاقة المقلص.

1- المقدمة

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 [1].
يهدف البحث إلى التعريف بأسلوب التحليل العنقودي و أهميته في التصنيف، و كيفية استخدام هذا الأسلوب لتصنيف أنواع الحليب (الحالات) في المدينة إلى مجموعات (عناقيد)، بحيث تكون الأنواع المجتمعة في نفس العنقود متقاربة من حيث نسبة العناصر الداخلة في التركيب، ومتباعدة عن أنواع الحليب المجتمعة في العناقيد الأخر .
تم الاعتماد على بيانات متوفرة من دراسة تطبيقية على بعض أنواع الحليب في مدينة مصراتة، (رسالة ماجستير في قسم علم الكيمياء).

2- طرق التعنقد الهرمي أو قوانين الربط Hierarchical Clustering Method.

حيث تبدأ طرائق العنقدة الهرمية التجميعية بفرض انتماء كل مشاهدة إلى عنقود مفرد (n من المجاميع ذات حجم واحد) ، وعند كل خطوة يتم إدماج أقرب زوج من العناقيد حتى الحصول على عنقود واحد *فقط يحوي جميع عناصر البيانات ، وهناك عدة طرائق معتمدة في عملية وضع العناصر في مجاميع بالاعتماد على مصفوفة التشابه ومنها :*

أولاً : طريقة الربط المفرد Single Linkage Method

وتسمى أيضا بطريقة الجوار الأقرب Nearest Neighbor ، وتعتمد هذه الطريقة على تحديد المسافة بين العناقيد بأصغر مسافة (أعظم تشابه) بين أي عنصرين من العناقيد (أقرب جوار) في العناقيد المختلفة ، ويطلق على هذا الأسلوب بطريقة الربط المفرد لأنها تبدأ مع كل النقاط كعناقيد مفردة ، ومن ثم يتم إضافة الترابطات الأقوى بين النقاط لتجميع العناصر وتشكيل العناقيد ، وتقود العناقيد الناتجة إلى سلسلة طويلة من الترابطات[5]

بمعنى أنه يتم الربط بين عنقودين بالاعتماد على أقل مسافة محسوبة ، وعندها يتم تقليص عدد العناقيد بمقدار واحد وتعاد العملية بحساب المسافات بين أزواج العناقيد واختيار أقل المسافات وهكذا نحصل على الشكل الهرمي للعناقيد والذي يمكن تمثيله بشكل الشجرة [3].

وتمثل صيغة الربط المفرد بالشكل الآتي :

$$d_{\min}(S_i, S_j) = \min_{\substack{x_i \in S_i \\ x_j \in S_j}} \left\| x_i - x_j \right\| \qquad (1)$$

إذ أن $S_i = \{x_1, x_2, ..., x_n\}$ يمثل العنقود الأول و $S_j = \{x_1, x_2, ..., x_m\}$ يمثل العنقود الثاني.

ثانياً : طريقة الربط الشامل Complete Linkage Method

ويطلق عليها أيضاً بطريقة الجوار الأبعد Furthest Neighbor ، حيث يتم تحديد المسافات بين العناقيد بأكبر مسافة بين أي عنصرين ضمن العناقيد المختلفة (أبعد جوار) . يطلق على هذا الأسلوب بطريقة الربط التام لأنها تبدأ مع كل العناصر كعناقيد مفردة ومن ثم يتم إضافة أقوى ارتباطات بين العناصر، و لا تمثل مجموعة العناصر عنقود إلا بربط جميع العناصر ربطاً تاماً وتشكيل الكتل clumps الواضحة ، . وتمثل صيغة الربط التام بالشكل الآتي :

$$d_{\max}(S_i, S_j) = \max_{\substack{x_i \in S_i \\ x_j \in S_j}} \left\| x_i - x_j \right\| \qquad (2)$$

إذ أن $S_i = \{x_1, x_2, ..., x_n\}$ ويمثل العنقود الأول و $S_j = \{x_1, x_2, ..., x_m\}$ يمثل العنقود الثاني[4].

ثالثاً : طريقة معدل الربط أو معدل زوج المجموعة Average Linkage or Pair-Group Average Method

في هذه الطريقة يتم تحديد المسافة بين عنقودين باستعمال معدل التقاربات (المسافة) الزوجية بين كل أزواج العناصر في العناقيد المختلفة . ويمثل هذا أسلوب متوسط بين طريقتي ربط Min و Max ويتم التعبير عن ذلك بالمعادلة الآتية :

$$\Pr oximity(s_1, s_2) = \frac{\sum proxinity(x_1, x_2)}{Size(s_1) * Size(s_2)} \qquad (3)$$

حيث أن s_1 و s_2 يمثلان العنقودين الأول والثاني ، وتقسم طريقة معدل الربط إلى نوعين أساسيين:

- **طريقة معدل زوج المجموعة غير الموزون Unweighted Pair-Group Average Method**

في هذه الطريقة تحسب المسافة بين عنقودين كمعدل مسافة بين كل أزواج العناصر ضمن عنقودين مختلفين ، وتكون هذه الطريقة ذات كفاءة عالية عندما تشكل العناصر كتل واضحة طبيعية ، وتمثل صيغة المعدل غير الموزون بالآتي :

$$d_{ave.}(S_i, S_j) = \frac{1}{n_i n_j} \sum_{x_i \in S_i} \sum_{x_j \in S_j} \left| x_i - x_j \right| \qquad (4)$$

إذ أن S_i و S_j تمثلان العنقودين i_{th} (و j_{th}) ، وتمثل(n_i و n_j) عدد العناصر في العنقودين S_i و S_j

- **طريقة معدل زوج المجموعة الموزون Weighted Pair-Group Average Method**

في طريقة معدل الربط الموزون تعطى المجموعتين أوزان متساوية بغض النظر عن عدد المشاهدات في كل مجموعة عند تحديد المجموعة المجمعة لذلك يتم استخدام الطريقة هذه عندما تصبح أحجام العناقيد غير متساوية [6].

رابعاً : طريقة العنقدة المركزية أو مركز متوسط زوج المجموعة Centroid Clustering or Pair-Group Centroid Method .

في هذه الطريقة تحسب المسافة بين عنقودين بالاعتماد على مراكز العناقيد وهناك أسلوبين مستخدمين لطريقة العنقدة المركزية هما :

- **طريقة مركز متوسط زوج المجموعة غير الموزونة Unweighted Pair-Group Centroid Method**

يمثل المركز المتوسط Centeroid للعنقود نقطة المعدل في فضاء متعدد الأبعاد، إذ يمثل مركز الثقل للعنقود الخاص ، وفي هذه الطريقة تحدد المسافة بين عنقودين كاختلاف بين مركزين كما في الصيغة الآتية :

$$d_{mean}(S_i,S_j)=|m_i-m_j| \qquad (5)$$

إذ أن $m_i=\frac{1}{n}\sum x_i$ و $m_j=\frac{1}{m}\sum x_j$ ويمثلان متوسطي العنقودين S_i و S_j بالتعاقب.

في حالة استعمال الوسيط بدلاً من مراكز العنقود فإن الصيغة تمثل :

$$d_{med}(S_i,S_j)=|med_i-med_j| \qquad (6)$$

إذ أن (med_i و med_j) يمثلان وسيطي العنقودين (S_i و S_j) بالتعاقب .

- **طريقة مركز متوسط زوج المجموعة الموزونة Weighted Pair-Group Centroid Method**

تتشابه هذه الطريقة مع الطريقة السابقة ماعدا أنه هناك أوزان تؤخذ بنظر الاعتبار عند اختلاف أحجام العناقيد (بمعنى عدد العناصر المحتواة فيها) .

خامساً : طريقة وورد Ward's Method

حيث يتم تفضيل هذه الطريقة على بقية الطرائق الهرمية التجميعية السابقة التي اقترحت من Ward عام 1963.

والتي يطلق عليها في بعض الأحيان بطريقة أصغر تباين Minimum Variance Method لأنها تستعمل أسلوب تحليل التباين لحساب المسافات بين العناقيد، والمعطاة حسب الصيغة الآتية:

$$d_{ward}(S_i,S_j)=n.m.d_{ij}^2/(n+m) \qquad (7)$$

إذ أن d_{ij}^2 تمثل المسافة بين العنقود i و العنقود j المعرفة في طريقة العنقدة المركزية غير الموزونة و m,n يمثلان عدد العناصر في العنقودين i j , بالتعاقب وتحاول هذه الطريقة العمل على حساب مجموع مربعات الخطأ بين كل زوج من العناصر ومن ثم العمل على ربط الزوجين الذي يعطي أصغر مجموع مربعات خطأ (SSE) ، ويتم حساب (SSE) لأي عنقود حسب الصيغة الآتية :

$$SSE=\sum_{j=1}^{n}\left[\sum_{i=1}^{n_j}x_{ij}^2-\frac{1}{n_j}\left(\sum_{i=1}^{n_j}x_{ij}\right)^2\right] \qquad (8)$$

إذ أن n تمثل عدد العناصر الكلية وتمثل n_j عدد العناصر في العنقود j_{th} .

ويعرف التقارب بين عنقودين بطريقة Ward على أنه الزيادة في مربع الخطأ Square Error الناتجة عند إدماج عنقودين ، لذلك تستعمل هذه الطريقة نفس دالة الهدف المستعملة في عنقدة k-Means . وبإجراء بعض الخطوات البسيطة نلاحظ أن هذه الطريقة تشابه طريقة معدل زوج المجموعة عندما يكون التقارب بين نقطتين محدد ليمثل مربع المسافة بينهما[7].

وتُعرف أيضاً بطريقة مجموع المربعات المضافة وتعتمد على استخدام مربع المسافات داخل كل عنقود ومربع المسافات بين العناقيد والمُعبر عنها بالصيغ التالية على اعتبار AB هو العنقود الناتج من ربط العنقودين A وB:

$$SSE_A = \sum_{i=1}^{n_A} (y_i - \bar{y}_A)'(y_i - \bar{y}_A)$$

$$SSE_B = \sum_{i=1}^{n_B} (y_i - \bar{y}_B)'(y_i - \bar{y}_B)$$

$$SSE_{AB} = \sum_{i=1}^{n_{AB}} (y_i - \bar{y}_{AB})'(y_i - \bar{y}_{AB})$$

ويتم ربط أي عنقودين بحيث يقلل الزيادة في مربع المسافات SSE ويعبر عن مقدار تلك الزيادة كما يلي[3] :

$$T_{AB} = SSE_{AB} - (SSE_A + SSE_B)$$

3- اختيار أفضل طرائق العنقدة

وسوف يتم الاعتماد على طريقة حساب مقياس العلاقة المُقلص Cophenetic Correlation Coefficient حسب الصيغة التالية :

$$D_m = \frac{[\sum |d_{ij} - d_{ij}^*|]^M}{[\sum_{i \le j}^{m} d_{ij}^M]^{1/M}}$$

حيث:

M : درجة المقياس $M = 1,2,...$

d_{ij} : قيم مصفوفة المسافات الأصلية .

d_{ij}^* : قيم مصفوفة المسافات المُقلصة .

n : عدد المتغيرات .

وأن القيمة العالية لهذا المقياس تعني حصول تشويه أكبر لمصفوفة المعاملات وأن معامل التشويه يمكن قياسه بالصيغة الآتية [2]:

$$C.C.R = \sqrt{\frac{\sum_{i \le j}^{n} (d_{ij} - d_{ij}^*)^2}{\sum_{i \le j}^{n} d_{ij}^2}} \qquad (9)$$

الجزء العملي Eexperimental Part

وصف عينة البحث

أجريت هذه الدراسة في مدينة مصراتة ، وكانت البيانات مأخوذة من رسالة ماجستير بقسم الكيمياء جامعة مصراتة لعام 2016، حيث تم جمع عينات الحليب المعقم والحليب المبستر (محلية الصنع ومستوردة من عدة دول) من محلات في مدينة مصراتة الليبية بمعدل 16 عينة لكل منها ، حفظت العينات وأجريت عليها التحاليل المطلوبة من تحاليل كيميائية وتحاليل ميكروبية ، وتم تقدير بعض العناصر المعدنية ، وشملت العينة بعض أنواع الحليب بالمدينة والمتمثلة في (حليب الريحان كامل الدسم – حليب السهول كامل الدسم – حليب الزهرات كامل الدسم –حليب الربيع – حليب جهينة كامل الدسم –

حليب كانديا كامل الدسم – حليب الريحان خالي الدسم –حليب جهينة خالي الدسم –حليب مبستر (طرابلس)- حليب مبستر (مصراتة) – حليب أبقار خام)

بيانات البحث

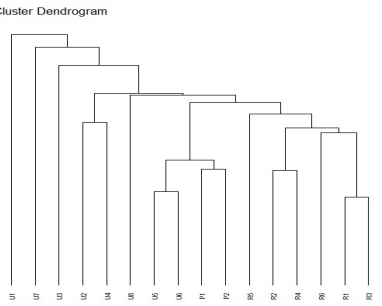
تتضمن بيانات البحث قراءات لبعض المعادن الثقيلة الداخلة في تركيب الحليب ، وهذه المعادن هي (الصوديوم Na ، البوتاسيوم K ، الكالسيوم Ca ، الماغنيسيوم Mg ، الحديد Fe ، الزنك Zn ، النحاس Cu) ، والتي تعتبر من المعادن الرئيسية للحليب ، والتي سيتم من خلالها دراسة تعنقد (ترابط) هذه العناصر مع بعضها البعض.

جدول (1) يبين أنواع العينات المدروسة

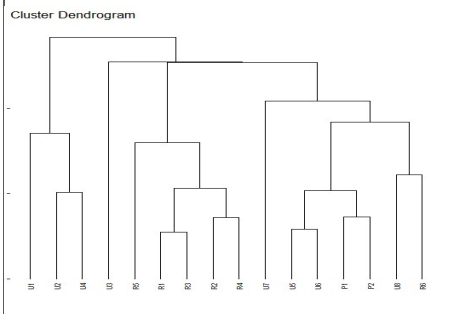
ت	اسم العينة	رمز العينة
1	حليب الريحان كامل الدسم	U1
2	حليب السهول كامل الدسم	U2
3	حليب الزهرات كامل الدسم	U3
4	حليب الربيع	U4
5	حليب جهينة كامل الدسم	U5
6	حليب كانديا كامل الدسم	U6
7	حليب الريحان خالي الدسم	U7
8	حليب جهينة خالي الدسم	U8
9	حليب مبستر (طرابلس)	P1
10	حليب مبستر (مصراتة)	P2
11	حليب خام	R1
12	حليب خام	R2
13	حليب خام	R3
14	حليب خام	R4
15	حليب خام	R5
16	حليب خام	R6

التحليل الإحصائي:

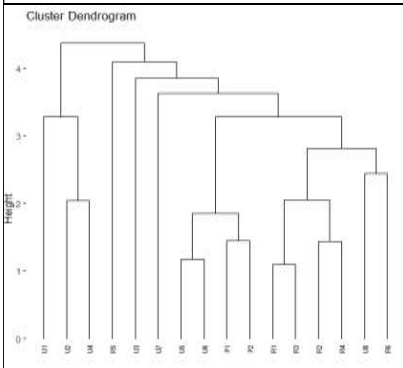
الأشكال التالية تبين نتائج التعنقد للطرق الهرمية بطريقة التجميع :



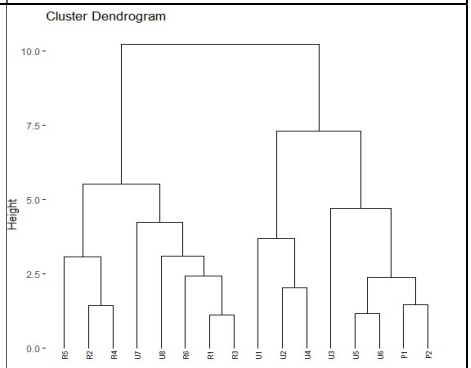
شكل (1) يوضح المخطط الهرمي لطريقة الربط المفرد



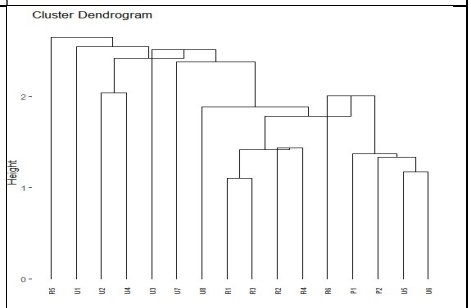
شكل (2) يوضح المخطط الهرمي لطريقة الربط التام



شكل (3) يوضح المخطط الهرمي لطريقة الربط المتوسط



شكل(4)يوضح المخطط الهرمي لطريقة الربط ward.



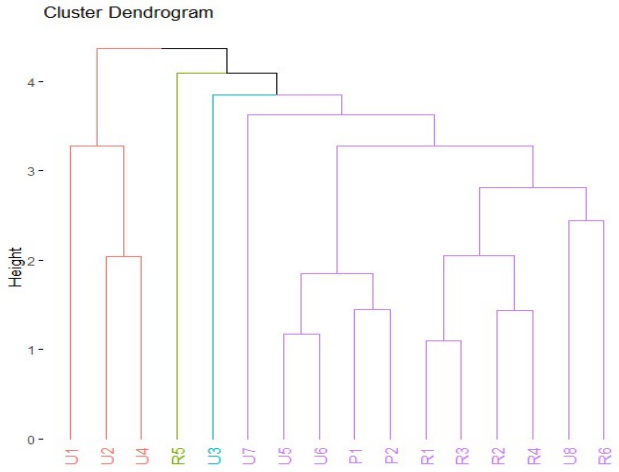
شكل (5)يوضح المخطط الهرمي لطريقة الربط Centroid

لتحديد أفضل طريقة هرمية بطريقة التجميع بناءً على الأشكال السابقة تم استخدام معامل الارتباط المقلص Cophenetic correlation coefficient في اختيار أفضل الطرق الهرمية المستخدمة في هذا البحث ، حيث يبين الجدول (2) قيم هذا المعامل .

جدول(2) يوضح قيم مقياس العلاقة المقلص

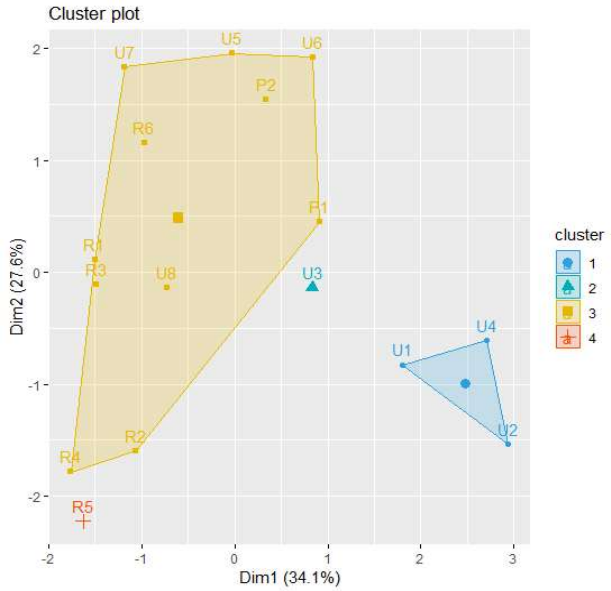
الطريقة	قيمة المقياس
طريقة الربط المفرد	0.58583
طريقة الربط الشامل	0.7360259
طريقة الربط المتوسط	0.7830906
طريقة Centeroid	0.6976552
طريقة Ward.D	0.5887712

وبمقارنة قيم معامل الارتباط في الخمس طرق السابقة يتضح أن أفضل طريقة ربط هي طريقة الربط المتوسط وكانت قيمة معامل التجميع Agglomerative Coefficient لهذه الطريقة 0.7830906 ، وبالتالي سيتم العمل على هذه الطريقة بالتحديد، من الممكن عرض الشجرة البيانية الخاصة بطريقة الربط بالمتوسطات average على ارتفاع معين من أجل تقسيم البيانات إلى مجموعات، حيث يتم التقسيم عن طريق تحديد العدد المطلوب من المجموعات أو عند ارتفاع معين من المسافة الإقليدية، فمثلاً إذا تم قطع الشجرة عند 4 عناقيد k=4 فإنه يمكن إيجاد عدد العناصر في كل عنقود من خلال رسم الشجرة بألوان مختلفة لكل مجموعة وذلك من خلال الشكل (6).



شكل رقم (6)يوضح المخطط الهرمي لتقسيم البيانات إلى 4 مجموعات لطريقة average

حيث تم تمثيل كل عنقود بلون محدد ، حيث ضم العنقود الأول ثلاث عناصر وهي (U1,U2,U4) ، بينما احتوى العنقود الثاني على R5 ، وضم العنقود الثالث أحد عشر عنصر وهم (U7,U5,U6,P1,P2,R1,R3,R2,R4,U8,R6) ، بينما ضم آخر عنقودU3 فقط .
يمكن ضم عناصر كل عنقود مع بعض بطريقة أخرى تعرف بالشكل الإطاري ، والذي يعمل على تمثيلها كنقاط ويتم الوصل بينها لكل عنقود كما في الشكل (7) :



شكل رقم (7)يوضح الشكل الإطاري لتقسيم البيانات إلى 4 مجموعات لطريقة average

يمكن تفسير ما سبق بأن الأنواع الموجودة في كل عنقود أكثر تجانساً لبعض من ناحية التركيب، إذ تم تمثيلها كنقاط ويتم الوصل بينها لكل عنقود .
بعد تطبيق أسلوب التحليل العنقودي على أنواع الحليب في مدينة مصراتة، و تصنيفها بالإعتماد على بعض الطرق الهرمية، و وفقاً للبيانات المتوفرة، توصلت الدراسة إلى بعض الاستنتاجات و التوصيات.

الاستنتاجاتConclusions

من خلال الدراسة التحليل العنقودي باستخدام أسلوب التحليل الهرمي التجميعي وتطبيقه على أنواع مختلفة من الحليب المبستر والحليب طويل الأمد نستنتج ما يلي:

1- أظهر اختبار أفضل طرق العنقدة أن طريقة ربط المتوسطات average هي أفضل الطريقة الهرمية لعنقدة مكونات الحليب.
2- وفقا لهذه الطريقة تضمنت المجموعة الأولى ثلاث عناصر وهي (U1,U2,U4) ، بينما احتوت المجموعة الثانية على R5 ، وضمت المجموعة الثالثة أحد عشر عنصر وهم (U7,U5,U6,P1,P2,R1,R3,R2,R4,U8,R6) ، بينما ضمت المجموعة الرابعة U3 فقط .

المراجعReferences

1- الحمامي ، علاء حسين ، (2008) ، تنقيب البيانات data mining العنقدة وتحليل العنقود ، عمان ، إثراء للنشر والتوزيع .
2- العزاوي ، إخلاص عبد الأمير ، (2013) ، تحليل إحصائي للعوامل المؤثرة على الوضع الاقتصادي للنساء ضمن نتائج مسح شبكة معرفة العراق ، بغداد .
3- رشيد ، أسيل عبد الرازق & مهدي ، نبأ ، تحليل واقع التربية والتعليم باستخدام طرائق التحليل العنقودي ، مجلة القادسية للعلوم الإدارية والاقتصادية المحور الإحصائي ، كلية الإدارة والاقتصاد جامعة المستنصرية .
4-Goodman,L. and kruskal,w.;((measures of association for cross validations)).jour.amer.stat.assoc. -1954.

5-Webb, A.R.; ((Statistical pattern Recognition)). John Wiley 1 Sons, LTD, (2002).
6- McQuitty, L.L.; ((Hierarchical linkage analysis for the Isolation of types)). Educ.Psychol. Measurements, 20(1), 1960.
7- Ward, J.H.; ((Hierarchical grouping to optimize an objective function)). Jour. Ofthe Amer. Stat. Assoc., 58:236-244,1963.
8- Kassambara ,A , Multivariate Analysis I ,Practical Guide To Cluster Analysis in R ,(2017) , Edition1
http://www.sthda.com, alboukadel.kassambara@gmail.com

Comparison of methods of hierarchical selection and selection of the best with practical application to some types of milk sold in the city of Misurata

Khuloud S. Esmewo[1]and Ibrahim S. Henaish[2]
[1] Statistics Department, Faculty of Sciences, Misurata University, Misurata, Libya
[2] Statistics Department, Faculty of Sciences, Misurata University, Misurata, Libya
E-mail: khuloudbll@gmail.com

Abstract:
One of the important statistical methods used in the classification is the cluster analysis method, which depends on the analysis of specific variables based on similarities and differences between the data, which falls within the methods of exploration of the non-learning data, which is one of the important areas and modern statistics.
As a practical application of this analysis, an analytical study was conducted on some of the components of milk sold in the local markets in Misrata (2016) in order to reach homogeneous groups that combine common characteristics, depending on some minerals involved in milk synthesis. The message applies some methods of hierarchical cluster analysis, comparing them to the best and most complex method of data.
The results of the study showed that the best hierarchical method is the method of averages (U1, U2, U4), while the second group consisted of R5, and the third group included eleven species (U7, U5, U6, P1, P2, R1, R3, R2, R4, U8, and R6), while the fourth group only included U3
Keywords: Cluster analysis**,** Hierarchical Clustering Method, Cophenetic Correlation Coefficien