استخدام التحليل العنقودي Cluster Analysis في تحليل العوامل المؤثرة على مرض القلب

*تهاني مهدي عباس ، **سميرة مزهر حميد و *قتيبة نبيل نايف
* جامعة بغداد ، كلية الأدارة والأقتصاد ، قسم الأحصاء.
**جامعة الموصل ، كلية الفنون الجميلة.

الخلاصة

يهدف هذا البحث الى دراسة مرض القلب والعوامل التي تؤثر على الاصابة به وتصنيف هذه العوامل او المتغيرات في مجموعات مختلفة حسب درجة تشابهها باستخدام التحليل العنقودي Cluster Analysis، والذي يعتبر من الاساليب الاحصائية المهمة والتي لها العديد من الاستخدامات ومنها تصنيف البيانات (المتغيرات او المشاهدات).

تم تطبيق هذا الاسلوب على عينة متكونة من (630) مريضا بالقلب، وقد تم تطبيق اربعة من طرق التعنقد باستخدام البرنامج الجاهز Minitab.

بعد تحليل النتائج تبين ان ثلاثة من هذه الطرق اظهرت نفس النتائج بالنسبة لتعنقد المتغيرات حيث تم تجميعها في ثلاثة عناقيد حسب درجة تشابهها.

المقدمة وهدف البحث

يعد القطاع الصحي من القطاعات المهمة التي حظيت باهتمام بالغ من قبل بلدان العالم كافة من أجل بناء صرح صحي متين من شأنه أن يحقق للإنسان الاستقرار والضمان الصحي والنفسي ومن أهم أسس ودعائم بناء هذا الصرح هو درء جميع الأمراض التي تحد من فعالية الإنسان وتعيقه عن العمل ومن أهم تلك الأمراض وأخطرها التي تودي بحياة العديد من البشر في مختلف أرجاء العالم هي أمراض القلب وتصلب الشرايين التي باتت تحتل المركز الأول في لائحة أسباب الوفيات في البلدان المتقدمة والمتطورة وكذلك في البلدان النامية حيث ان المرضى المصابين بهذا المرض مهددة حياتهم نظراً لكون القلب أهم عضلة في جسم الإنسان وان توقفها عن الحركة يعني توقف حياة الإنسان. ومن هنا كان الهدف من هذا البحث هو دراسة هذا المرض والعوامل المؤثرة على الإصابة به وتصنيف هذه العوامل أو المتغيرات في مجموعات مختلفة حسب درجة تشابهها وذلك باستخدام التحليل العنقودي (Cluster Analysis) الذي يعتبر من الأساليب الإحصائية المهمة التي يمكن استخدامها في كثير من مجالات الحياة ومنها المجال الصحي.

الجانب النظري [2] [3] [4] [6]

ان التحليل العنقودي(Cluster Analysis) يستخدم في تعنقد البيانات المتعددة المتجانسة. ان مثل هذه الأساليب تستخدم لغرض تجميع العناصر (Objects) أو المتغيرات (Variables) تحت الدراسة في مجاميع متجانسة فيما بينها (داخل المجموعة الواحدة) ومختلفة عن المجاميع الأخرى وذلك اعتماداً على العديد من الصفات.

هنالك العديد من الباحثين الذين استخدموا مصطلح التحليل العنقودي للإشارة إلى الأساليب التي تبحث في تجميع المتغيرات والتي تم اقتراحها كبديل لتحليل المركبات الرئيسية (Principle Components Analysis) ومن الباحثين الذين استخدموا هذه التسمية للغرض أعلاه الباحثان (Kendall and Sturat) واللذان استخدما أيضاً التصنيف (Classification) لغرض وصف الأساليب التي تبحث في تجميع المفردات.

ان الاستخدامات المختلفة للتحليل العنقودي غالباً ما تكون غير متشابهة ولكن بصورة عامة يمكن أن نحدد أهمها كما يلي:

1- اختصار البيانات Data reduction.
2- توليد الفرضيات Hypothesis generating.
3- اختبار الفرضيات Hypothesis testing.
4- التنبؤ المبني على المجاميع Prediction based groups.
5- مطابقة النماذج Models fitting.
6- التشخيص Identification.

إن الباحث قد يواجه كميات كبيرة من المشاهدات (Observations) ويكون من الصعب دراستها ما لم يتم وصفها في مجاميع متجانسة وبالتالي تتم معاملة هذه المجاميع كوحدات (units)، وكذلك يستخدم التحليل العنقودي لعنقدة المتغيرات (Variables) إلى مجاميع متجانسة وذلك لاختصار عدد المتغيرات. حيث تستخدم طريقة التسلسل الهرمي الذي يبدأ بجميع المتغيرات بشكل منفصل أي كل متغير في مجموعة لوحده وفي الخطوة الأولى يتم تجميع متغيرين في مجموعة واحدة وفي الخطوة الثانية يندمج المتغير الثالث مع المجموعة الأولى أو يندمج متغيرين في مجموعة جديدة ثانية. وهكذا وهذه العملية تستمر حتى تندمج كل المجموعات في مجموعة واحدة ولكن يجب تحديد عدد المجموعات التي تتوقف فيها العملية.

معاملات المسافة Distance Coefficients [5] [2]:

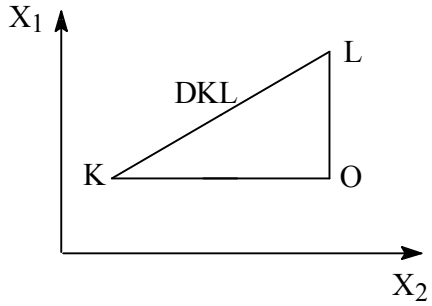
وهي معاملات مصممة لقياس الاختلاف وذلك بأنها تقيس المسافة بين المتغيرات وكلما زادت المسافة بين أي متغيرين يزداد عدم التشابه (الاختلاف) بينهما ويقل التشابه وفي هذا المجال هنالك مقاييس عدة لتقدير الاختلاف:

1. مقياس المسافة الاقليدية Euclidian Distance.
2. طريقة منكوفسكي لقياس المسافات The Minkowfski Distance M.D.
3. مقياس المسافة للمتغيرات الثنائية Distance for binary variables .
4. مقياس المسافة للبيانات الممزوجة Distance for mixed data .

حيث سنقوم في هذا البحث باستخدام مقياس المسافة الاقليدية لحساب المسافة بين المتغيرات لذا سنتناوله بشيء من التفصيل:

مقياس المسافة الاقليدية (Euclidian Distance):

ان هذا المقياس هو أكثر المقاييس شيوعاً في الاستخدام. ان كلاً من المتغيرين K, L الممثلين بالشكل رقم (1) كلاهما متصفان بالخاصيتين $X_{.1}, X_{.2}$.



شكل(1): يوضح التمثل للمسافة بين المتغيرين K, L.

فلإيجاد المسافة بين المتغيرين K, L نستخدم نظرية فيثاغورس.

$$D^2_{K,L} = (\overline{ok})^2 + (\overline{oL})^2$$
$$= (X_{k1} - X_{L1})^2 + (X_{K2} - X_{L2})^2$$
$$D_{K,L} = \sqrt{\sum_{j=1}^{2} (X_{kj} - X_{ij})^2}$$

وبصورة عامة:

$$D_{K,L} = \sqrt{\sum_{j=1}^{p} (X_{Kj} - X_{Lj})^2} \quad \ldots\ldots\ldots\ldots (1)$$

ويمكن الحصول على المسافات بشكل موجهات (Vector Notation) وكما يلي:

$$\underline{X}_{k.} = \left(X_{k1}, \ldots, X_{kj}, \ldots, X_{kp}\right)$$
$$\underline{X}_{L.} = \left(X_{L1}, \ldots, X_{Lj}, \ldots, X_{Lp}\right)$$

وعليه:

$$\left(\underline{X}_{k.}\, \underline{X}_{L.}\right) = \left(X_{L1} - X_{L1}, \ldots, X_{Kp} - X_{Lp}\right)$$

وهذه تمثل موجهات على شكل صفوف (row factors) ومبدلتها هي:

$$(\underline{X}_{K.} - X_{L.})' = \begin{bmatrix} X_{K_1} - X_{L_1} \\ \vdots \\ X_{K_j} - X_{L_j} \\ \vdots \\ X_{K_P} - X_{L_P} \end{bmatrix}$$

لذلك:

$$D_{K,L} = \sum_{j=1}^{p} (X_{Kj} - X_{Lj})^2$$
$$= \left(\underline{X}_{K.} - \underline{X}_{L.}\right)\left(\underline{X}_{K.} - \underline{X}_{L.}\right)' \quad \ldots\ldots\ldots\ldots (2)$$

طرق التعنقد (Clustering Methods) [2]:

إن عملية التعنقد تؤدي إلى وضع المتغيرات في مجاميع بطريقتين:

أ- تكون مجاميع أكبر نتيجة لدمج المتغيرات (أو المجاميع الصغيرة).

ب- تكوين مجاميع صغيرة نتيجة لتجزئة المجاميع الكبيرة وابتداءاً من أكبر مجموعة (أي التي تحوي جميع المتغيرات).

وحيث أنه تم اعتماد الطريقة الأولى في هذه الدراسة فقد تم تناولها بشكل مفصل.

إن إجراءات الاندماج الكلي (agglomerative) (اعتباراً من دمج أول متغيرين إلى انتهاء عملية التعنقد). لغرض اجراء عملية التعنقد نتبع الخطوات التالية:

1- حساب مصفوفة المسافة (Distance Matrix) أو مصفوفة معاملات الارتباط ذات الأبعاد N×N.

$$D = \|D_{ij}\|$$

من مجاميع المتغيرات الأولية ان الـ D_{ij} تمثل المسافة بين المتغيرين لذا على الأغلب تتم عملية حساب المسافة بالاعتماد على مقياس المسافة الاقليدية (Euclidian distance).

2- يتم البحث عن أقصر المسافات داخل المصفوفة اعلاه حيث يتم ربط المتغيرين الذين تكون المسافة بينهما أقصر المسافات ضمن المصفوفة ليشكلا نواة العنقود.

3- بعد أن يتم تشكيل العناقيد الأولية تأتي المرحلة الثالثة وهي مرحلة حساب مصفوفة المسافة الجديدة والتي تأخذ بنظر الاعتبار التغيرات التي حصلت في المرحلة الثانية وعلى افتراض ان المرحلة الثانية تضمنت عملية ربط واحد لعنقودين (أو متغيرين)، فإن عنقوداً جديداً يكون قد تشكل بالاعتماد على أقصر المسافات وبالتالي فإن المصفوفة الجديدة ستكون ذات أبعاد (P-1) (P-1).

4- الاستمرار بعملية الربط بالاعتماد على أقصر مسافة ممكنة إلى أن يتم ربط آخر عنقودين (أو عنقود بمتغير).

وهناك العديد من طرق التعنقد، وأهم هذه الطرق هي:

1- طريقة الربط المفرد Single Linkage [3]:

تعتبر هذه الطريقة من أبسط طرق التعنقد وأقدمها وتعتمد بالأساس على معيار الأكثر تشابه بين المتغيرات هو الذي يكون نواة العنقود ثم تضاف اليه بقية الوحدات إلى تلك النواة وبالتسلسل، أي يتم إضافة الوحدات إلى العنقود حسب درجة الشبه مع وحدات العنقود النواة والأكثر شبهاً يضاف أولاً، وهكذا بقية المتغيرات، أما في حالة ربط مجموعة عناقيد مع بعضها فيلاحظ أقرب المسافات بين عناصر العناقيد وحسب الصيغة التالية:

$$D(I,J) = \min(D_{ij}),\ i \in I,\ j \in J \quad \ldots\ldots (3)$$

حيث ان i, j تمثل المتغيرات في العناقيد I, J على التوالي.

2- طريقة الربط الشامل Complete Linkage Method [2]:

وتعرف أيضاً بطريقة الـ Maximum Method في هذه الطريقة يتشكل العنقود بطريقة معاكسة للطريقة الأولى حيث انها تعتمد على الأقل تشابه بين المتغيرات (أو بعد المسافات). وبتعبير آخر ان المتغير المرشح للدخول إلى العنقود يدخل فقط إذا كانت المسافة بينه وبين أي من متغيرات العنقود هي أكبر مسافة أما في حالة ربط مجموعة عناقيد مع بعضها فيتم اعتماد أبعد المسافات بين متغيرات العناقيد وحسب الصيغة التالية:

$$D(I,J) = \max(D_{ij}),\ i \in I,\ j \in J \quad \ldots\ldots (4)$$

حيث ان i, j تمثل المتغيرات في العناقيد I, J على التوالي.

3- الطريقة المعتمدة على الوسط الحسابي Average Method [3]:

تكون هذه الطريقة كحالة وسيطة بين طريقتي الربط المفرد والشامل وتعتمد على الوسط الحسابي وتعتبر من الطرائق المستخدمة في الوقت الحاضر وقد طورت اساساً من قبل (Sokal and Michener) اساس هذه الطريقة هو إعطاء أوزان متساوية للمتغيرات عند حساب المسافة بين أي عنقودين والقانون العام لحساب المسافة بين أي عنقودين مثل J,I هو:

$$D(I,J) = \frac{1}{P_i P_j} D(i,j),\ i \in I,\ j \in J$$

عند ربط عنقودين مثل I, J فالمسافة بين العنقود الجديد واي عنقود آخر مثل (R) هي:

$$D(R,I,+J)=\frac{P_iD(R,I)+P_jD(R,J)}{P_i+P_j} \quad \ldots\ldots\ldots\ldots(5)$$

حيث ان:

P_i: عدد المتغيرات في العنقود I.

P_j: عدد المتغيرات في العنقود J.

4- الطريقة المركزية Centroid Method [5]:

في هذه الطريقة عملية التعنقد تعتمد على مراكز المتغيرات (أو العناقيد) بمعنى آخر ان المسافة بين العناقيد تمثلها المسافة بين مراكز العناقيد. ان المشكلة التي تواجه مستخدم هذه الطريقة هي ان عملية الربط بين عنقودين بينها اختلاف كبير بالحجم (أي وجود متغيرات كثيرة في العنقود الأول ومتغيرات أقل في العنقود الثاني) تجعل عملية حساب مراكز العناقيد الجديدة يبقى قريباً جداً من مراكز المجاميع الكبيرة لذلك فإن خصائص العناقيد الصغيرة تكون قد ألغيت ولتجنب هذه الحالة يجب أن تبنى هذه الطريقة بصورة مستقلة عن الحجم وذلك بافتراض ان المجاميع المدمجة متساوية الحجم وبالتالي فإن موقع (أي مركز) العنقود الجديد الناتج من عملية الاندماج يكون بين عنقودين، ان الطريقة المركزية تكون كما يلي:

أ- طريقة مركزية موزونة weighted pair-group centroid method.

ب-طريقة مركزية غير موزونة -unweighted pair group centroid method

في الطريقة غير الموزونة يكون هنالك وزن أثقل للفروع قبل النهاية التي تربط بالعنقود النهائي، أما في الطريقة الموزونة فإن العناقيد التي تربط اخيراً تعطى لها نفس الوزن المعطى للعناقيد الموجودة أصلاً في العنقود الرئيس. ان الطريقة المركزية الموزونة تسمى ايضا بالطريقة الوسطية (Medium Method).

5- الطريقة الهرمية لـ Ward [5] :

هذه الطريقة مبنية على أساس أقل فقدان في المعلومات لعمل العنقدة. لقد اعتمد مقياس المصاحبة لعمل العنقدة وهو مقياس مشابه لتباين العينة ذات البعد الواحد فإذا كانت قيمة المتغير (j) للعنصر (i) في العنقود (k) يقال لها (X_{ijK}) فإن مركز العنقود (k) سيكون:

حيث ان $[\bar{X}_{.1K},\bar{X}_{.2K},\ldots,\bar{X}_{.nK}]$:

$$\bar{X}_{.jK}=\frac{1}{mk}\sum_{j=1}^{mk}X_{ijK}$$

mk: تمثل عدد المتغيرات في العنقود (k).

وعليه فإن مجموع مربعات الخطأ الخاص بالعنقود (k) سيكون:

$$E_k=\sum_{i=1}^{P}\sum_{j=1}^{mk}(X_{ijK}-\bar{X}_{.jK})^2 \quad \ldots\ldots\ldots\ldots(6)$$

ومجموع التشتت الكلي:

$$E\ total=\sum_{K=1}^{n}E_K \quad \ldots\ldots\ldots\ldots(7)$$

حيث أن:

E total (Et): تمثل قياس لمقدار المعلومات المفقودة في حالة الاستعاضة عن المتغيرات بمراكز العناقيد عند إجراء عملية الربط في بداية عملية التعنقد يكون مجموع الخطأ يساوي (صفراً) (لأن الربط يتم بين المتغيرات) ولكن عند ربط عنقودين سنلاحظ أن مجموع مربعات الخطأ سيزداد.

في طريقة ward في أي مرحلة من مراحل التعنقد نجد أن العنقودين اللذين سيكونان مرشحين للارتباط هما العنقودان اللذان يحققان مجموع مربعات خطأ أقل مايمكن.

6- طريقة الاتحاد الخطي Linear Combination Method [1] [5]:

هذه الطريقة مطورة من قبل (Lance & William) وقد أوضح هذان الباحثان أن مختلف طرائق التعنقد مشتقة بصورة مباشرة من المعادلة التالية:

$$D_{H,K}=\alpha_iD_{H,I}+\alpha_j D_{H,J}+\beta D_{I,J}+\gamma|D_{H,J}-D_{H,I}| \quad \ldots\ldots(8)$$

حيث أن:

γ, β, α: تمثل معالم تحدد طبيعة التعنقد.

$D_{H,K}$: تمثل المسافة بين العنقود (H) والعنقود (K).

(العنقود K هو العنقود الناتج من اندماج العنقودين (J,J))

ويمكن أن نعبر بهذه المعادلة عن الطرائق السابقة وذلك من خلال تغير قيم المعالم γ,β,α، وجدول رقم (1) يوضح قيم معالم الطرائق المستخدمة:

جدول رقم (1).

الطريقة	αI	αJ	β	γ
طريقة الربط المفرد	0.5	0.5	0.0	−0.5
طريقة الربط الشامل	0.5	0.5	−0.0	0.5
طريقة الوسط الحسابي	$\frac{n_I}{n_K}$	$\frac{n_J}{n_K}$	0.0	0.0
الطريقة المركزية غير الموزونة	$\frac{n_i}{n_K}$	$\frac{n_j}{n_K}$	$\frac{-n_i \cdot n_j}{-n^2_K}$	0.0
الطريقة المركزية الموزونة	0.5	0.5	-0.25	0.0
طريقة ward	$\frac{P_{hi}}{P_{hK}}$	$\frac{P_{hj}}{P_{hK}}$	$-\frac{n_h}{n_K}$	0.0

حيث أن:

P_i, P_j, P_k: تمثل عدد المتغيرات في العناقيد I, J, K على التوالي.

الجانب التطبيقي:

تم جمع البيانات الخاصة بالبحث من خلال استمارة استبيان*، وبالاعتماد على الإضبارة الخاصة بكل مريض حيث كانت العينة المستخدمة تتألف من (630) مريضاً من كلا الجنسين من المرضى الراقدين في الردهات الباطنية والقلبية في المستشفيات التالية:

1-مستشفى ابن النفيس للقلب والأوعية الدموية.

2-مستشفى اليرموك التعليمي.

3-مستشفى مدينة الطب التعليمي.

حيث تم اختيار 310 مريضاً من مستشفى ابن النفيس و 110 مريضا من مستشفى اليرموك التعليمي و 210 مريضاً من مستشفى مدينة الطب التعليمي. وقد تضمنت استمارة الاستبيان أهم العوامل المؤثرة على زيادة نسبة الإصابة بأمراض القلب وهي كالتالي:

1-**العمر**: تراوحت الفئات العمرية بين (80-15) سنة وتمثل المتغير X_1.

2-**الوزن**: تم قياس الزيادة عن الوزن الطبيعي** لكلا الجنسين وتمثل المتغير X_2.

3-**ضغط الدم**: تم قياس مقدار ضغط الدم بالنسبة لارتفاعه عن المقياس الطبيعي لضغط الدم للإنسان العادي حيث ان المقدار الطبيعي لضغط الدم هو (120/80) وتمثل المتغير X_3.

4-**السكر**: تم قياس الارتفاع عن المقدار الطبيعي للسكر في الدم والذي يتراوح بين (140-80) ملغم/ملليتر من الدم وتمثل المتغير X_4.

5-**الكولسترول**: تم قياس مقدار الكولسترول عن طريق حساب الزيادات عن المقادير الطبيعية والتي تتراوح بين (200-180) ملغم لكل مئة ملليتير من الدم ويمثل المتغير X5.

6-**نسبة الحمض البولي**: تم قياس الحمض البولي في الدم بحساب الزيادة عن النسبة الطبيعية والتي تتراوح بين %(7-5) ملغم في الليتر الواحد ويمثل المتغير X6.

7-**نسبة يوريا الدم**: تم قياس نسبة اليوريا في الدم بحساب الزيادة عن النسبة الطبيعية والتي تبلغ (40%) ويمثل المتغير X7.

8-**التدخين**: تم قياس هذا العامل وفق عدد السيكارات التي يدخنها يومياً حيث تراوح عددها بين (60-3) سيكارة وتمثل المتغير X8.

9-**حجم الاسرة**: لقد تم اعتماد عدد افراد الاسرة التي يمثل المريض رب او ربة الاسرة وقد تراوحت حجم الاسر لجميع المرضى بين (15-0) فرد، حيث ان القيمة الصفرية ترمز الى ان المريض غير متزوج أي غير مسؤول عن عائلة وتمثل المتغير X9.

10- **الوراثة**: صنف المرضى وفقاً لعامل الوراثة حسب الحالات المرضية الموجودة في العائلة وتمثل المتغير X10. حيث تم التصنيف بالشكل التالي:

* نظمت استمارة الاستبيان من قبل الباحثين والاستعانة بالأطباء ذوي الاختصاص.

** الوزن الطبيعي هو الوزن المثالي للشخص حسب الطول والجنس.

الأب والأم 2 الاب أو الأم 1 أحد الأقارب 2/1 لايوجد 0.

11- **عدد ساعات العمل:** تم تحديد عدد ساعات العمل اليومية بين (3-16) ساعة ويمثل المتغير X11.

ولقد تم تحليل البيانات بواسطة البرنامج الجاهز Minitab حيث تم اعتماد مقياس المسافة الاقليدية (Euclidian Distance) لحساب الاختلاف، حيث تم اعتماد بعض طرق التعنقد التي تم التطرق اليها في الجزء النظري وكانت النتائج كالتالي:

1-**طريقة الربط المفرد Single linkage**: عند استخدام هذه الطريقة كانت النتائج موضحةفي الجدول رقم (2):

جدول رقم (2).

Step	Number of clusters	Similarity level	Distance level	Clusters joined		New clusters	Number of variables
1	10	86.7763	0.26447	4	11	4	2
2	9	79.4106	0.4117	3	6	3	3
3	8	78.8470	0.4786	1	7	1	4
4	7	76.0682	0.4651	5	10	5	5
5	6	71.5459	0.5890	2	9	2	6
6	5	65.3391	0.6732	4	8	4	7
7	4	61.5958	0.7680	1	2	1	8
8	3	56.7422	0.8773	3	5	3	9
9	2	55.4390	0.8912	1	4	1	10
10	1	54.1042	0.9211	1	3	1	11

وبالاستناد الى الجدول اعلاه تم تكوين ثلاثة عناقيد للمتغيرات قيد البحث:

العنقود الاول: يتضمن المتغيرات: العمر، الوزن، يوريا الدم، حجم الاسرة.

العنقود الثاني: يتضمن المتغيرات: ضغط الدم، الكولسترول، الحمض البولي، الوراثة.

العنقود الثالث: يتضمن المتغيرات التالية: السكر، التدخين، عدد ساعات العمل.

2-**الطريقة المركزية Central Method**: عند استخدام هذه الطريقة كانت النتائج كالاتي:

جدول رقم (3).

Step	Number of clusters	Similarity level	Distance level	Clusters joined		New clusters	Number of variables
1	10	87.7631	0.1421	4	11	4	2
2	9	81.4060	0.2641	4	7	4	3
3	8	79.6210	0.4160	4	9	4	4
4	7	78.4107	0.4217	4	8	4	5
5	6	76.2820	0.4711	2	4	2	6
6	5	71.1141	0.5656	1	2	1	7
7	4	65.4590	0.6842	1	3	1	8
8	3	61.3119	0.7891	1	6	1	9
9	2	56.4139	0.8640	1	5	1	10
10	1	55.3334	0.8911	1	10	1	11

وبالاستناد الى الجدول اعلاه تم تكوين ثلاثة عناقيد للمتغيرات قيد البحث:

العنقود الاول: يتضمن المتغيرات: العمر، الوزن، ضغط الدم، السكر، الحمض البولي،يوريا الدم، التدخين، حجم الاسرة، عدد ساعات العمل.

العنقود الثاني: يتضمن المتغير: الكولسترول.

العنقود الثالث: يتضمن المتغير: الوراثة.

3- **الطريقة المعتمدة على الوسط الحسابي Average Method:** عند استخدام هذه الطريقة كانت النتائج كالاتي:

جدول رقم (4).

Step	Number of clusters	Similarity level	Distance level	Clusters joined		New clusters	Number of variables
1	10	86.7763	0.26447	4	11	4	2
2	9	79.4106	0.4117	3	6	3	3
3	8	78.8470	0.4786	1	7	1	4
4	7	76.0682	0.4651	5	10	5	5
5	6	71.5459	0.5890	2	9	2	6
6	5	65.3391	0.6732	4	8	4	7
7	4	61.5958	0.7680	1	2	1	8
8	3	56.7422	0.8773	3	5	3	9
9	2	55.3187	0.9158	1	4	1	10
10	1	53.9954	0.9355	1	3	1	11

وبالاستناد الى الجدول اعلاه تم تكوين ثلاثة عناقيد للمتغيرات قيد البحث:

العنقود الاول: يتضمن المتغيرات: العمر، الوزن، يوريا الدم، حجم الاسرة.

العنقود الثاني: يتضمن المتغيرات: ضغط الدم، الكولسترول، الحمض البولي، الوراثة.

العنقود الثالث: يتضمن المتغيرات التالية: السكر، التدخين، عدد ساعات العمل.

4- **طريقة Ward:** عند استخدام هذه الطريقة كانت النتائج كالاتي:

جدول رقم (5).

Step	Number of clusters	Similarity level	Distance level	Clusters joined		New clusters	Number of variables
1	10	86.7763	0.26447	4	11	4	2
2	9	79.4106	0.4117	3	6	3	3
3	8	78.8470	0.4786	1	7	1	4
4	7	76.0682	0.4651	5	10	5	5
5	6	71.5459	0.5890	2	9	2	6
6	5	65.1478	0.6955	4	8	4	7
7	4	60.5958	0.8158	1	2	1	8
8	3	56.5581	0.8874	3	5	3	9
9	2	54.8547	0.9048	1	4	1	10
10	1	55.1254	0.9987	1	3	1	11

وبالاستناد الى الجدول اعلاه تم تكوين ثلاثة عناقيد للمتغيرات قيد البحث:

العنقود الاول: يتضمن المتغيرات: العمر، الوزن، يوريا الدم، حجم الاسرة.

العنقود الثاني: يتضمن المتغيرات: ضغط الدم، الكولسترول، الحمض البولي، الوراثة.

العنقود الثالث: يتضمن المتغيرات التالية: السكر، التدخين، عدد ساعات العمل.

الاستنتاجات

1. عند استخدام كل من طريقة الربط المفرد والطريقة المعتمدة على الوسط الحسابي وطريقة Ward تم تكوين ثلاثة عناقيد الاول يحتوي على المتغيرات (العمر، الوزن، يوريا الدم، حجم الاسرة)، اما العنقود الثاني فيحتوي المتغيرات (ضغط الدم، الكولسترول، الحمض البولي،وجود عامل الوراثة)، اما العنقود الثالث يحتوي على المتغيرات (السكر، التدخين، عدد ساعات العمل). وهذا يدل على وجود تشابه كبير عند استخدام الطرق الثلاثة اعلاه.
2. عند استخدام الطريقة المركزية جاءت النتائج مختلفة تماما عن الطرق الثلاث اعلاه حيث ان العنقود الاول قد احتوى على اكثر المتغيرات وهذه احدى عيوب هذه الطريقة (العمر، الوزن، ضغط الدم، السكر، الحمض البولي، يوريا الدم، التدخين، حجم الاسرة، عدد ساعات العمل)، اما العنقود الثاني فيتكون من متغير واحد هو مستوى الكولسترول، اما العنقود الثالث ايضا احتوى على متغير واحد هو وجود عامل الوراثة.
3. يمكن دمج المتغيرات او العوامل التي تؤثر على الاصابة بمرض القلب الى ثلاثة عناقيد بالاعتماد على الطرق الثلاثة المذكورة في الفقرة (1). حيث ان المتغيرات التي دمجت في العنقود الاول هي (العمر، الوزن، يوريا الدم، حجم الاسرة)، اما المتغيرات التي دمجت في العنقود الثاني هي (ضغط الدم، الكولسترول، الحمض البولي، وجود عامل الوراثة)، اما المتغيرات التي دمجت في العنقود الثالث فهي (السكر، التدخين، عدد ساعات العمل).عليه يمكن القول ان المتغيرات في العنقود الواحد قد دمجت حسب درجة تشابهها وارتباطها مع بعض.

المصادر

[1] Afifi,A. & Clark, V., (1996), "Computer-Aided Multivariate Analysis", 3rd ed. Chapman & Hall, New York, pp.361-391.

[2] Aldenderfer, Mark S. & Blast, Roger K., (1984), "Cluster Analysis", 1st ed., Sage Puplications, Newbury, Park, Call, pp.9-59.

[3] Everitt, B. S., Landau, S. and Leese, M. (2001), "Cluster Analysis", 4th Edition, Edward Arnold, pp.450-462.

[4] Rolf, F. James, (2004), "Clustering Variables". Available at: http://www.life.bio.sonysb.edu/ee/rohlf.

[5] Stanley, L. Selove, (2001), "Notes on Cluster Analysis".Available at: http://www.statsoftince.com/textbook/stcluan.html.

[6] Lawley, D. N. & Maxwel, A. E., (1971), "Factor Analysis as Statistical Method", 2nd ed., John Wiley & Sons, Inc., New York, pp.66.

Abstruct

This research aims to study the heart disease and the factors which are cause it, and classify these factors into different groups according to its similarity degree by using cluster analysis which is consider the important statistical ways which have many uses, e.g. data classification (variables or observations).

This way was applied on sample contain 630 heart patients; four ways of cauterizing were applied by using the program (Minitab).

After results analysis we find that three of these ways show the same results according clustering variables which are collecting in three clusters according to similarity degree.