مجلّة جامعة عدن للعلوم الانسانيّة والاجتماعية
EJUA-HS Vol. 4 No. 3 (2023)
https://doi.org/10.47372/ejua-hs.2023.3.282
ISSN: 2708-6275

مقالة بحثية

الفرق بين التحليل العنقودي والتحليل التمييزي

هالة محمد أحمد علي بابريش*

قسم إحصاء ومعلوماتية، كلية العلوم الإدارية، جامعة عدن، عدن، اليمن

*** الباحث الممثل: هالة محمد أحمد علي بابريش؛ البريد الالكتروني: hala199062@gmail.com**

استلم في: 28 يوليو 2023 / قبل في: 22 أغسطس 2023 / نشر في: 30 سبتمبر 2023

المُلخّص

هدفت الدراسة إلى التعريف بالتحليل العنقودي والتحليل التمييزي؛ إذ يعتبران من المواضيع المتقدمة في الإحصاء والغير متناول بكثرة من الباحثين؛ وذلك لإعطائهم فكرة عن كيفية استعمالهما في عملية تصنيف الظواهر متعددة المتغيرات، ومن أجل تحقيق هدف الدراسة استعملت الباحثة المنهج الوصفي لمناسبته لطبيعة الدراسة. وتوصلت الدراسة إلى فروق عديدة بين التحليل التمييزي والتحليل العنقودي في دراسة الظواهر، وأبرزها أن التحليل التمييزي يهتم بمسألة التمييز والفصل بين المجموعات ويتطلب المعرفة المسبقة بعدد المجاميع حيث يسعى إلى تكوين نموذج إحصائي يوضح العلاقة المتبادلة بين المتغيرات المختلفة.

الكلمات المفتاحية: التحليل العنقودي، التحليل التمييزي، التحليل متعدد المتغيرات.

المقدمة:

يعد التحليل العنقودي والتحليل التمييزي أحد أساليب التحليل متعدد المتغيرات والذي يهتم بدراسة متغيرات متعددة أو مجموعة من المتغيرات في وقت واحد ولكن الكثير من الباحثين لا يستطيع التفريق بين استعمال أسلوب التحليل العنقودي والتحليل التمييزي، حيث يهدف كل منهما إلى التصنيف، إذ تعطي الدارسة صورة مختصرة لأوجه الاختلاف بين التحليل العنقودي والتحليل التمييزي.

مشكلة الدراسة:

تتلخص مشكلة الدراسة بصعوبة بعض الباحثين التمييز في استعمال أسلوب التحليل العنقودي والتحليل التمييزي في عملية تصنيف الظواهر متعددة المتغيرات؛ لذا ستوضح الدراسة أسلوب التحليل العنقودي وأهدافه، وشروط استعماله وأنواعه، ثم ستتناول التحليل التمييزي وشروط استعماله، وأهدافه، والمراحل التي يمر بها الباحث أثناء التحليل، وأنواع الدوال التمييزية.

أهمية البحث:

من المقارنة بين أسلوب التحليل العنقودي والتمييزي؛ سيكون هناك توسع لدى الباحثين في التعرف على كيفية استعمال كلٌّ منهما في الدراسات التطبيقية؛ من حيث الطرائق والخطوات والشروط والفرضيات باعتبارهما أسلوبًا إحصائيًا متعدد المتغيرات.

أهداف الدراسة:

تهدف الدراسة إلى التعريف بالتحليل العنقودي والتحليل التمييزي؛ إذ يعتبران من المواضيع المتقدمة في الإحصاء، إذ ستعطى فكرة للباحثين عن كيفية استعمال كلٌّ منهما في عملية تصنيف الظواهر متعددة المتغيرات.

منهجية الدراسة:

اعتمدت الدراسة على المنهج الوصفي في وصف التحليل العنقودي والتحليل التمييزي.

مكونات الدراسة:

تتكون الدراسة من مبحثين:

الأول: يتضمن التحليل العنقودي.

الثاني: يتضمن التحليل التمييزي.

الدراسات السابقة:

1-دراسة (عاشور،2019م):"تصنيف المحافظات العراقية صحيا باستخدام التحليل العنقودي لعا م2016"

هدفت الدراسة إلى معرفة الاختلاف بين المحافظات العراقية، من حيث مستوى المؤشرات الصحية المقدمة للمواطن، بالإضافة إلى تحديد أي من المؤشرات التي أسهمت بدرجة كبيرة في هذا الاختلاف والتفاوت بين المحافظات.

وتوصلت الدراسة إلى أن محافظة بغداد هي الأفضل في تقديم الخدمات الصحية للمواطن؛ إذ كانت المسافة بينها وبين بقية المحافظات تتراوح من 3.875 إلى 4.841، وأن محافظتي النجف والقادسية متقاربة في تقديم هذه الخدمات للمواطن؛ إذ كانت المسافة بينهما (0.411) ولوحظ أن المحافظات تجمعت في ثلاثة عناقيد، الأول ضم المحافظات (كركوك، ديالى، بابل، كربلاء، واسط، صلاح الدين، النجف، القادسية، المثنى، ذي قار، ميسان) والثاني ضم محافظة البصرة فقط والثالث ضم محافظة بغداد فقط.

2-دراسة (بسيوني،2021م):" استخدام التحليل التمييزي في التصنيف والتنبؤ"

هدفت الدراسة إلى تحديد العوامل المؤثرة في الإصابة بمرض السكري؛ إذ أُنشئت دالة تمييزية للفصل والتمييز بين الأشخاص إلى مجموعتين (مصاب – غير مصاب).

وتوصلت الدراسة إلى تكوين دالة تمييزية للفصل بين الأشخاص؛ إذ تبين أن أكثر العوامل المؤثرة في الإصابة بمرض السكري هي الوزن، السن، ضغط الدم، التدخين، ممارسة الرياضة، الوراثة، الكوليسترول، النوع، وتم استبعاد بقية المتغيرات لعدم معنويتها.

تميزت الدراسة الحالية بدراسة أوجه الاختلاف بين أسلوبي التحليل العنقودي والتحليل التمييزي وتوضيح بعض طرق المستخدمة في كلٍّ منهما.

المبحث الأول: التحليل العنقودي

مفهوم التحليل العنقوديCluster Analysis:

عرف الشمراني(2020) التحليل العنقودي بأنه: "عبارة عن مجموعة من الأدوات لإنشاء مجموعات(عناقيد) من بيانات متعددة المتغيرات والهدف من ذلك هو تكوين مجموعات ذات خصائص متجانسة من أصل مجموعات كبيرة غير متجانسة".

أهداف التحليل العنقودي:

أوضح رزق الله (2002م) إلى أهم أهداف استعمال أسلوب التحليل العنقودي وهي:

1. وصف التصنيفات واستكشافها.
2. تبسيط البيانات واختزالها؛ ليتم التعامل معها بشكل مجاميع عوضًا عن الكميات الكبيرة من البيانات والتي يصعب التعامل معها.
3. تحديد العلاقات والتي يمكن التعرف عليها بعد تشكيل القطاعات(المجاميع).

حدود استعمال التحليل العنقودي:

أشار رزق الله(2002م) إلى حدود استعمال التحليل العنقودي:

1. يوصف التحليل العنقودي بأنه: تحليل وصفي ونظري وغير استنتاجي؛ إذ يعتبر استعماله استكشافي بطبيعته.
2. يمكن الوصول للعديد من الحلول من التغير في العوامل والإجراءات المستخدمة في التحليل.
3. يعتمد التحليل العنقودي بشكل كلي على المتغيرات التي استعملت لتحديد مقياس التشابه؛ إذ أن إضافة أو حذف أي متغير يؤدي إلى تغير جوهري في النتائج.

تطبيقات التحليل العنقودي:

يستعمل التحليل العنقودي في مجالات وتطبيقات عديدة، ويمكن تحديدها في أربعة اتجاهات كما أشار إليها هندوش(2010م):

1. توليد واختبار الفرضيات.
2. التنبؤ المعتمد على المجاميع فالأنماط غير المعروفة نستطيع تصنيفها إلى عناقيد بالاعتماد على درجة التشابه.
3. تحليل البيانات الكبيرة التي تحتاج إلى معالجة، إذ تستعمل العنقدة لتجزئة البيانات إلى عدد من العناقيد لتسهل معالجتها.

خصائص التحليل العنقودي:

أشار Banerjee et al.(2007) و Bekkerman et al.(2005) إلى أن أهم الخصائص التي تتميز بها العنقدة وهي:

1. أن عملية العنقدة تعدُّ من التعليم غير الموجَّه؛ وذلك لعدم توافر معلومات سابقة متعلقة بالصفات المميزة لكل عنقود.

2. يعتمد مفهوم العنقدة على تقسيم مجموعة من البيانات على مجاميع متشابهة فيما بينها ومختلفة عن باقي العناقيد.
3. ليس من السهل تحديد عدد العناقيد المطلوبة؛ ولذلك من الممكن أن يكون هناك العديد من الحلول الصحيحة لمسالة العنقدة الواحدة.

خطوات التعنقد (Clustering Steps):

أوضح أحمد(2015م)، والعلي(2020م)، وعاشور(2019م) أهم خطوات تطبيق التحليل العنقودي وهي:

1. حساب مصفوفة التباعد أو مصفوفة التشابه.
2. ربط العنصرين اللذين تكون المسافة بينهما أقصر المسافات ضمن المصفوفة؛ لتشكيل العناقيد الأولية.
3. حساب مصفوفة المسافة الجديدة بعد تشكيل العناقيد الأولية والاستمرار بعملية ربط العناصر؛ اعتمادًا على المسافة بينهما إلى أن يربط العنقودان الآخران في نهاية التحليل.

أنواع التحليل العنقودي:

يتفرع التحليل العنقودي عمومًا إلى نوعين كما وضحه العلي(2020م)، وأحمد(2015م)، وطه وحسين (2012م):

1-التحليل العنقودي الهرمي (Hierarchical):

عرفه أحمد (2015م) بأنه: "لا يتطلب التحليل العنقودي الهرمي المعرفة المسبقة بعدد العناقيد التي سيتم تصنيف الحالات على أساسها حيث يناسب التحليل العنقودي الهرمي العينات الصغيرة نسبيًّا".

ويقسم هذا التحليل على:

1. التحليل العنقودي الهرمي للحالات(المفردات).
2. التحليل العنقودي الهرمي للمتغيرات.

وأشار علي (2015م) إلى أن هناك أسلوبين علميين لعنقدة مفردات العينة باستعمال التحليل الهرمي وهما:

-أسلوب التجميع (The Agglomerative Technique):

يبدأ هذا الأسلوب من التحليل بعنقود واحد لكل حالة ثم تجمع العناقيد المتشابهة تدريجيا حتى نصل إلى العدد المطوب.

-أسلوب التقسيم (The Divisive Technique):

يبدأ هذا الأسلوب بافتراض أن جميع الحالات تتجمع في عنقود واحد ثم يتم تصنيف الحالات في عناقيد أصغر فأصغر.

حساب مصفوفة التباعد (Dissimilarity) للمتغيرات الكمية:

عرف العلي(2020م) أن مصفوفة التباعد تتألف من قيم المسافات بين أفراد العينة ويرمز لها بالرمز D وهي كالآتي:

$$D=\begin{bmatrix} 0 & d12 & d13 \dots\dots & d1n \\ & 0 & d23 \dots\dots & d2n \\ & & 0 \quad djk & djn \\ & & & 0 \\ & & & & 0 \end{bmatrix}$$

حيث إن العنصر d_{jk} هو المسافة بين المفردتين j , k ونلاحظ أن هذه المصفوفة هي مصفوفة مثلثية عليا وعناصر قطرها الرئيس يساوي أصفارًا لأن المسافة بين المفردة j ونفسها تساوي صفرًا.

يتم حساب المسافة باستخدام المسافة الإقليدية وفق الصيغة الآتية:

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

حساب مصفوفة التشابه أو التقارب(Similarity):

لحساب مصفوفة التقارب يتم اتباع الخطوات التالية كما وضحها العلي(2020م):

أولًا: تحويل المتغيرات النوعية (رتبيه، اسمية) إلى متغيرات ثنائية وتأخذ احدى القيمتين (1) عند التحقق و(0) عند عدم التحقق.

ثانيًا: حساب التقارب بين كل مفردتين من إنشاء جدول التوافق.

ثالثًا: حساب عناصر مصفوفة التقارب ويرمز لها S من استعمال المقياس الرياضي الآتي:

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

من استعمال المقياس تنتج مصفوفة التقارب وهي كالآتي:

$$S = \begin{vmatrix} 1 & & & \\ s_{21} & 1 & & \\ s_{31} & s_{32} & 1 & \\ s_n & s_{n2} \ldots\ldots\ldots & & 1 \end{vmatrix}$$

نلاحظ أن المصفوفة مثلثية سفلى من الرتبة n*n كما أن عناصر قطرها الرئيسي تساوي 1.

طرق التعنقد الهرمية (Hierarchical Clustering Method):

ويطلق عليها بقوانين الربط كما أشار إليها كل من أحمد(2015م)، واسمييو وحنيش(2019م)، وعاشور(2019م) وهي كالآتي:

أولًا طريقة الربط المنفرد(Linkage Single) :

وتسمى أيضًا بطريقة الجوار الأقرب (Neighbor Nearest) وتعدُّ هذه الطريقة كل مفردة تشكل عنقودًا خاصًا ومن ثم تضاف الترابطات الأقوى بين المفردات لتجميع العناصر وتشكيل العناقيد، وتعتمد على حساب مصفوفة التباعد أو مصفوفة التقارب ومن ثم تتم عملية الربط بالاعتماد على أقل مسافة بين أزواج المفردات وربطها معًا بحسب الصيغة الآتية:

$$(\mathrm{A}, \mathrm{B}) = \mathrm{Min}\left\{\mathrm{d}(y_i\ , y_j)\ , y_i \in \mathrm{A}\ , y_j\ \in B\right\}$$

إذ i,j تمثل العناصر في العناقيد ، $\mathrm{d}(y_i\ , y_j)$ هي المسافة المحسوبة من مصفوفة التباعد باستعمال أحدى المقاييس الرياضية .

ثانيًا طريقة الربط الشامل (Complete linkage):

ويطلق عليها أيضًا الجوار الأبعد (Furthest Neighbor)، تعمل هذه الطريقة بشكل معاكس للطريقة السابقة؛ إذ تحدد المسافات بين العناقيد بأكبر مسافة بين أي عنصرين ضمن العناقيد المختلفة (أبعد جوار)، ولهذا فعملية الحساب تبدأ بالبحث عن أصغر عنصر في المصفوفة D ثم يتم تحديد العنقودان الأكثر تقاربًا من الصيغة الآتية:

$$\mathrm{D}(\mathrm{A}, \mathrm{B}) = \mathrm{Max}\left\{\mathrm{d}(y_i\ , y_j)\ , y_i \in \mathrm{A}\ , y_j\ \in B\right\}$$

ثالثًا طريقة الربط المتوسط (Average Linkage):

أوضح أسمييو وحنيش(2019م) أنه: " 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 وMax" .

ويُعبر عن ذلك بالمعادلة الآتية:

$$proximity(s_1\ , s_2) = \frac{\sum_{\substack{x_1 \in s_1 \\ x_2 \in s_2}} proximity(x_1, x_2)}{size(s_1\ ) * size(s_2)}$$

رابعًا طريقة الربط المركزية (Centroid):

أشار علي(2002م) أن المسافة بين العنقودين في الطريقة المركزية تعرف على أنها المسافة الإقليدية بين متجهي الوسط الحسابي للعنقودين.

خامسًا طريقة الوسيط(Median):

تستخدم في حالة كانت عدد مفردات أحد العناقيد أكبر من الأخرى.

سادسًا طريقة وورد (Ward's method):

أشار الشمراني (2020م) إلى أنها من الطرق الأكثر استعمالًا؛ إذ تتبع سلسلة من خطوات التجميع التي تبدأ بالعناقيد ويحوي كل منها على عنصرًا واحدًا، وتنتهي بمجموعة واحدة تشمل العناصر جميعًا، ويطلق عليها في بعض الأحيان بطريقة أصغر تباين (Minimum Variance Method)؛ لأنها تستعمل أسلوب تحليل التباين ومؤشرات مجموع المربعات بين العناقيد.

إذ يكون التعبير عنها بالصيغ الآتية:

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

إذ SSE_A هي مجموع مربعات الخطأ في العنقود A ، nA هي عدد المشاهدات الكلية في العنقود.

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

إذ SSE_B هي مجموع مربعات الخطأ في العنقود B ، nB هي عدد المشاهدات الكلية في العنقود.

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

إذ SSE_{AB} هي مجموع مربعات الخطأ للعنقود الناتج من ربط العنقودينA و B ، nAB هي عدد المشاهدات الكلية في العنقود.

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

إذ I_{AB} هو مقدار الزيادة الناتج عن ربط العنقودين بحيث يقلل الزيادة في مربع المسافات.

2-التحليل العنقودي الغير هرمي(Non-Hierarchical):

أوضح العلي(2020م) أن التحليل العنقودي غير الهرمي يطبق على العينات الكبيرة، ويستعمل لتجميع بيانات المفردات الكثيرة أكثر مما يستعمل لتجميع المتغيرات ضمن عنقودا محددا؛ إذ يعتمد على أسلوب التقسيم؛ ويمكن تحديد عدد العناقيد سابقًا، شخصيًّا، أو من أسلوب العنقدة.

وتوجد طرائق عديدة للتحليل العنقودي غير الهرمي، إذ تعدُّ طريقة K متوسط أفضلها وذات فعالية عالية للتعامل مع البيانات الضخمة.

طريقة K متوسط (K-means):

أشار أحمد (2015م) إلى أنه: "تقوم هذه الطريقة على أساس تصنيف الحالات (المفردات) في مجموعات متجانسة من حيث الخصائص والصفات، وذلك باستخدام خوارزميات يمكنها معالجة عدد كبير من الحالات، وتسمى هذه الطريقة بطريقة التحليل العنقودي السريع لأنها تقوم بعملية التحليل والتصنيف في وقت قصير نسبيًّا".

وقد أشار أحمد(2018م)، والموسى وآخرون(2015م) إلى خطوات تطبيق أسلوب المتوسطات:

1. تحديد عدد العناقيد المطلوب لإجراء عملية التجميع (العنقدة).
2. حساب البعد بين كل مفردة وبين جميع المراكز باستخدام المسافة الإقليدية.
3. ضم كل مفردة إلى العنقود الأقرب اليها.
4. تحديد مراكز العناقيد من إيجاد مراكز المفردات الموجودة في كل عنقود، وتحدد المراكز من حساب متوسط المفردات التي تنتمي للعنقود.
5. تكرار الخطوات من 2 إلى 4 حتى نصل إلى الاستقرار (أي عدم توافر مفردات تنتقل ضمن العناقيد) أو عند عدد معين من التكرارات.

مشكلات التحليل العنقودي:

أشار مانلي (2001م) إلى أهم مشكلات استعمال التحليل العنقودي وهي:

1. توجد عدة خوارزميات مختلفة للتحليل العنقودي ولا تعطي نفس النتائج عند تطبيقها على نفس البيانات.
2. يجب أن تكون متغيرات الدراسة ذات صلة بالتصنيف المرغوب لكون العناقيد أكثر حساسية لاختيار المتغيرات.

المبحث الثاني: التحليل التمييزي

تعريف التحليل التمييزي:

عرف قاسم ومحمد(2018م) بأنه: أحد الطرائق الإحصائية في تحليل البيانات متعددة المتغيرات؛ إذ يهتم بمسألة تمييز K من المجموعات التي تكون متشابه في كثير من الخصائص (الصفات) اعتمادا على P من المتغيرات المستقلة من خلال استخدام الدالة المميَّزة والتي هي عبارة عن تركيب خطي للمتغيرات المستقلة.

أهداف التحليل التمييزي:

وضح جودة (2008م) أهداف التحليل التمييزي وهي:

1. تصميم التركيبات الخطية للمتغيرات الأفضل في التمييز بين فئات المتغير التابع.
2. التحقق فيما إذا كان هناك فروق ذات دلالة إحصائية بين المجموعات فيما يتعلق بالمتغيرات.
3. تصنيف المتغيرات التي تسهم بأكبر قدر من الاختلاف بين فئات المتغير التابع.
4. تقييم دقة التصنيف كنسبة مئوية.

خصائص التحليل التمييزي:

أشار طاقية وآخرون(2016م) إلى أهم الخصائص التي يتصف بها التحليل التمييزي وهي:

1. يعدُّ أحد الأساليب الإحصائية متعدد المتغيرات ويستعمل لدارسة مدى تداخل المجموعات.
2. يهدف إلى الفصل بين المجموعات محل الدراسة بناءً على عينة من المشاهدات تسحب من المجموعات.

شروط استخدام التحليل التمييزي:

أشار الصويعي وبنيني (2020م) إلى أهم الشروط اللازمة لإجراء التحليل التمييزي وهي:

1. يجب أن يكون اختيار العينة بشكل عشوائي وذات حجم كبير.
2. أن تكون المجتمعات الإحصائية موضوع الدراسة تتوزع توزيعا طبيعيا.
3. تجانس مصفوفة التغاير والتباين في المجتمعات الإحصائية موضوع الدراسة.
4. عدم وجود ارتباط قوي بين المتغيرات المستقلة.

استعمال التحليل التمييزي:

أشار برايس وعبان (2022م) إلى أن التحليل التمييزي يستعمل للتعرف على:

1. الدلالة الإحصائية للتنبؤ.
2. عدد الدوال التمييزية الدالة إحصائيًا.
3. الأهمية النسبية لمتغيرات التنبؤ ويقصد به أي من المتغيرات المستقلة هي الأكثر أهمية في التنبؤ بمجموعات المتغير التابع.
4. حجم التأثير ويقصد به مقدار الارتباط بين مجموعات المتغير التابع ومجموعة من المتغيرات المستقلة.
5. نسبة التصنيف ويقصد به نسبة الحالات التي صنفت بشكل صحيح.

مراحل التحليل التمييزي:

أشار بقريش وشترواي (2017م) إلى مراحل التحليل التمييزي وهي:

1. تحديد المتغير التابع والمتغيرات المستقلة.
2. التحقق من توافر الشروط اللازمة لإجراء التحليل.
3. تحديد النموذج والطريقة الإحصائية المقترح استعمالها.

4. تقدير معالم النموذج المقترح.
5. اختبار جودة توفيق النموذج المقترح مع تحليل البواقي.
6. اختبار معنوية مقدرة النموذج المقترح على التنبؤ باستعمال القيم المعرفة سابقًا، وتفسير النتائج.

أنواع التحليل التمييزي:

أوضح بالرايس وعبان(2022م) أنه توجد أنواع للتحليل التمييزي عديده سواء من عدد الدوال التمييزية، أو من الطرق المستعملة في إدخال المتغيرات المستقلة في التحليل.

أولًا من حيث طرق إدخال المتغيرات المستقلة في التحليل: ينقسم التحليل التمييزي من حيث الطرق المستعملة في إدخال المتغيرات المستقلة في التحليل إلى ثلاثة أنواع وهم:

- **التحليل التمييزي المباشر:**

إذ يتم إدخال جميع المتغيرات مره واحده إلى التحليل.

- **التحليل التمييزي الوهمي:**

إذ يتم إدخال المتغيرات بناءً لما يراه الباحث من أهمية للمتغيرات وبالترتيب الذي يراه مناسب.

- **التحليل التمييزي المتدرج:**

إذ يتم تحديد معيار إحصائي يحدد أولوية إدخال المتغيرات في التحليل.

ثانيًا من حيث عدد الدوال التمييزية: ينقسم التحليل التمييزي من حيث عدد الدوال التمييزية إلى نوعين وهما:

- **التحليل التمييزي الخطي:** يعتمد التحليل التمييزي الخطي على النماذج الخطية للفصل بين المجموعات، ويشترط تساوي تباينات المجتمع محل الدراسة، وينصف التمييز الخطي إلى نوعين هما:
  - التمييز الخطي لمجموعتين.
  - التمييز الخطي لأكثر من مجموعتين.
- **التحليل التمييزي الغير خطي:** يعتمد التحليل التمييزي الغير خطي على النماذج غير الخطية للفصل بين المجموعات ويستعمل في حالة عدم تساوي تباينات المجتمع المدروس، ويصنف التمييز الغير خطي إلى:
  - التحليل التمييزي التربيعي(Quadratic Discriminant).
  - الشبكات العصبية متعددة الطبقات(Multi-Layer Perceptron).
  - شجرة القرار(Decision Trees).
  - خوارزميات الغابة العشوائية(Random Forest).
  - خوارزمية الجار الأقرب(K-Nearest Neighbors).
  - وفي هذا البحث سيتم توضيح أسلوب التحليل التمييزي الخطي لمجموعتين وذلك لتوضيح خطوات، وشروط، وفرضيات تطبيق التحليل التمييزي.

التحليل التمييزي الخطي لمجموعتين:

أشار بسيوني(2021م) إلى أهم خطوات حساب الدالة التمييزية الخطية وهي:

1. حساب متوسطات المتغيرات في كل مجموعة وإيجاد الفرق بين المتوسطات:

- حساب متوسطات المتغيرات في المجموعة الأولى:

$$\bar{x}_{i(1)} = \begin{bmatrix} \bar{x}_{1(1)} \\ \bar{x}_{2(1)} \\ . \\ . \\ \bar{x}_{k(1)} \end{bmatrix}$$

- حساب متوسطات المتغيرات في المجموعة الثانية:

$$\bar{x}_{i(2)} = \begin{bmatrix} \bar{x}_{1(2)} \\ \bar{x}_{2(2)} \\ . \\ . \\ . \\ \bar{x}_{k(2)} \end{bmatrix}$$

إذ K عدد المتغيرات المستقلة.

- حساب الفرق بين متوسط المتغيرات في المجموعتين:

$$d_i = \bar{x}_{i(1)} - \bar{x}_{i(2)} = \begin{bmatrix} \bar{x}_{11} - \bar{x}_{12} \\ \bar{x}_{21} - \bar{x}_{22} \\ . - . \\ . - . \\ \bar{x}_{k(1)} - \bar{x}_{k(2)} \end{bmatrix} = \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_k \end{bmatrix}$$

حيث d هي المسافة بين متوسط المتغيرات.

2. إيجاد التباين والتغاير المشترك بين المجموعتين:

التباين المشترك:

$$v_{ii} = \frac{s_{ii(1)} + s_{ii(2)}}{n_1 + n_2 - 2}$$

التغاير المشترك:

$$v_{ij} = \frac{s_{ij(1)} + s_{ij(2)}}{n_1 + n_2 - 2}$$

إذ يمكن إيجاد s_{ii} و s_{ij} من الصيغ الآتية:

$$s_{ii} = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$$

$$s_{ij} = \sum x_i\, x_j - \frac{\sum x_i \sum x_j}{n}$$

مصفوفة التباين والتغاير المشترك بين المجموعتين:

$$v = \begin{bmatrix} v_{11} & v_{12} & v_{13} & \dots . v_{1k} \\ v_{21} & v_{22} & v_{23} & v_{2k} \\ . & . & . & . \\ . & . & . & . \\ v_{k1} & v_{k2} & v_{k3} & v_{kk} \end{bmatrix}$$

إذ يمثل التباين المشترك عناصر القطر الرئيسي للمصفوفة والعناصر المتبقية تمثل التغاير المشترك.

3. بناء الدالة التمييزية:

تعدُّ الدالة التمييزية الخطية تركيب خطي من المتغيرات بناءً على الصيغة الآتية:

$$\hat{l} = \alpha_1 x_1 + \alpha_2 x_2 + \dots + \alpha_k x_k$$

إذ أن:

α معاملات النموذج وتستخدم في عملية التصنيف.

k عدد المتغيرات.

x متجه المتغيرات.

علمًا بأن:

$$\alpha = v^{-1}\, d$$

إذ أن:

v^{-1} معكوس مصفوفة التباين والتغاير المشترك.

d مصفوفة المسافة بين متوسط المتغيرات في كل من المجموعتين.

4. الأهمية النسبية للعوامل المؤثرة (المتغيرات المستقلة):

في هذه المرحلة تحدد الأهمية النسبية للمتغيرات المستقلة المؤثرة في عملية التمييز والفصل بين المجموعات؛ إذ تعد المعاملات المعيارية ذات اعلى قيمة هي الأكثر أهمية، وتحدد نسبة مساهمة المتغير في عملية التمييز بواسطة عامل الارتباط القانوني.

5. اختبار قدرة الدالة على التمييز بين المجموعتين:

توجد مجموعة من الاختبارات لقياس قدرة الدالة على التمييز والفصل بين المجموعتين وهي:

- اختبار F (F test):

ويتم ذلك عن طريق اختبار الفرضية التالية:

الدالة ليس لديها قدرة على التمييز: H_0

ويعتمد هذا الاختبار على تكوين جدول تحليل التباين:

Source	SS	DF	MS	F
بين المجموعات	*SSB*	*K-1*	M_{SB}	$\frac{M_{SB}}{M_{SE}}$
الخطأ	*SSE*	*n-K*	M_{SE}	
الكلي	*SST*	*n-1*		

إذ أن :

$$SSE = D^2 = \widehat{\alpha_1}\, d_1 + \widehat{\alpha_2}\, d_2 + \cdots + \widehat{\alpha_k}\, d_k$$

$$SSB = \frac{n_1\, n_2}{(n_1 + n_2)(n_1 + n_2 - 2)} \times (D^2)^2$$

$$SST = SSE + SSB$$

ولحساب F المحسوبة نستعمل الصيغة التالية:

$$F = \frac{M_{SB}}{M_{SE}}$$

ولحساب القيمة الجدولية لـ F نستعمل الصيغة التالية:

$$F(k-1\,, n-k)$$

إذا كانت قيمة F المحسوبة أكبر من القيمة الجدولية؛ فإنه يتم رفض الفرض الصفري(العدمي) ويتم قبول الفرض البديل وهو أن الدالة لديها قدرة على التمييز بين المجموعتين.

- اختبار ويلكس لمدا(A) Wilk's lambda:

ويتم ذلك عن طريق اختبار الفرضية التالية:

الدالة ليس لديها قدرة على التمييز $H_0 : \mu_1 = \mu_2$

ويحسب من العلاقة:

$$\Lambda = \sum_{i=1}^{k} \frac{1}{1 + \lambda_i}$$

إذ أن:

λ_i الجذر الكامن eigenvalues لكل المتغيرات.

k عدد المتغيرات.

تنحصر قيمة Λ بين الصفر و الواحد الصحيح $0 \leq \Lambda \leq 1$

إذا كانت :

$\Lambda = 1$ ويقصد بها تساوي متوسطات المجوعتين (عدم مقدرة الدالة على التمييز).

$\Lambda = 0$ ويقصد بها عدم تساوي متوسطات المجوعتين (مقدرة الدالة على التمييز).

أما إذا اقتربت قيمة Λ من الواحد الصحيح يدل ذلك على عدم مقدرة الدالة على التمييز، وإذا اقتربت من الصفر دل ذلك على قدرة الدالة على التمييز.

- اختبار هوتلنج Hotelling- Lawely test(T^2):

ويأخذ الصيغة التالية:

$$T^2 = \sum_{i=1}^{s} \lambda_i$$

إذ أن:

S عدد المتغيرات.

ويمكن حساب قيمة F من اختبار هوتلنج باستعمال الصيغة الآتية:

$$F = \frac{n_1 + n_2 - K - 1}{(n_1 + n_2 - 2)k} \times T^2$$

والقيمة الجدولية:

$$F_\alpha = (K-1\,, n_1 + n_2 - K - 1)$$

إذا كانت F المحسوبة أكبر من F الجدولية فإن الدالة لها قدرة عالية على التمييز.

6. نقطة الفصل Cut off Point:

أشار سليمان وآخرون (2011م) إلى أن نقطة الفصل بين المجموعتين تستعمل لغرض تصنيف مفردة معينة إلى المجموعة الأقرب لها، فإذا كانت قيمة الدالة بعد تعويض قيم المفردة فيها أكبر من هذه النقطة فالمفردة تعود إلى المجموعة الأولى، أما إذا كانت قيمة الدالة أكبر من هذه النقطة فالمفردة تعود إلى المجموعة الثانية، ويتم ذلك من حساب الصيغة الآتية:

$$\bar{\bar{l}} = \frac{\bar{l}_{(1)} + \bar{l}_{(2)}}{2}$$

إذ أن:

$\bar{\bar{l}}$ نقطة الفصل.

$\bar{l}_{(1)}$ متوسط القيم التمييزية للمجموعة الأولى.

$\bar{l}_{(2)}$ متوسط القيم التمييزية للمجموعة الثانية.

يتم تصنيف المفردة إلى احدى المجموعات من اتباع قاعدة التصنيف وهي:

1) إذا كان $\bar{L}_1 > \bar{L}_2$:

- ❖ إذا كانت القيمة التمييزية للمفردة الجديدة أكبر من نقطة الفصل؛ تصنف المفردة الجديدة ضمن المجموعة الأولى.
- ❖ إذا كانت القيمة التمييزية للمفردة الجديدة أقل من نقطة الفصل؛ تصنف المفردة الجديدة ضمن المجموعة الثانية.

- إذا تساوت القيمة التمييزية للمفردة الجديدة ونقطة الفصل؛ تصنف المفردة الجديدة عشوائيًا ضمن أي مجموعة من المجموعتين.

2) إذا كان $\bar{L}_1 < \bar{L}_2$:

- إذا كانت القيمة التمييزية للمفردة الجديدة أعلى من نقطة الفصل؛ تصنف المفردة الجديدة ضمن المجموعة الثانية.
- إذا كانت القيمة التمييزية للمفردة الجديدة أقل من نقطة الفصل؛ تصنف المفردة الجديدة ضمن المجموعة الأولى.
- إذا تساوت القيمة التمييزية للمفردة الجديدة مع نقطة الفصل؛ تصنف المفردة الجديدة عشوائيًا ضمن أي مجموعة من المجموعتين.

أخطاء التصنيف:

ويقصد بأخطاء التصنيف هو أن يحدث خطأ في تصنيف البيانات في فئات خاطئة؛ وذلك بوضع المفردة في مجموعة غير المجموعة التي تنتمي إليها، ويعتبر خطأ التصنيف عامل مهم عند الحكم على كفاءة الدالة التمييزية.

أشار المخلافي (2019م) إلى أن هناك نوعان من أخطاء التصنيف وهي:

- احتمال خطأ التصنيف P12 ويقصد به احتمال تصنيف المفردة إلى المجموعة الثانية وهي تنتمي إلى المجموعة الأولى.
- احتمال خطأ التصنيف P21 ويقصد به احتمال تصنيف المفردة إلى المجموعة الأولى وهي تنتمي إلى المجموعة الثانية.

ويمكن تقدير احتمال خطأ التصنيف من الصيغة الآتية:

$$P_{12} = P_{21} = \Phi(\frac{-D}{2})$$

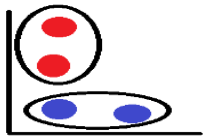
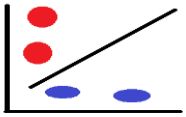
إذا أن:

Φ تمثل دالة التوزيع الطبيعي القياسي.

D يمثل مقياس مهالانوبيس (Mahalanobis Distance).

ويمكن تلخيص أهم الفروق بين التحليل العنقودي والتحليل التمييزي من الجدول الآتي:

الجدول (1): الفرق بين التحليل العنقودي والتحليل التمييزي

أوجه المقارنة	التحليل العنقودي	التحليل التمييزي
التعريف	تجميع المفردات في مجاميع بحسب الاختلاف والتشابه بين المفردات.	توصيف عناصر المجتمع المحددة مسبقًا والموزعة على مجموعات واستخلاص قاعدة محددة لتحديد انتماء أي عنصر إليها.
البيانات	يعد من التعلم غير الموجَّه؛ إذ لا يتطلب المعرفة المسبقة بالمجاميع.	يعد من التعليم الموجَّه؛ إذ يتطلب المعرفة المسبقة بانتماء المفردات إلى المجاميع.
الهدف	يهدف إلى تجميع مجموعة من الأشياء لمعرفة ما إذا كانت هناك أي علاقة بينها.	يهدف إلى العثور على الفئة التي ينتمي إليها المفردة الجديد لتشكيل مجموعة من الفئات المحددة مسبقًا.
آلية العمل	يتناول خطوة واحدة فقط وهي التجميع.	يتناول عمليتين أساسيتين وهما التمييز والتصنيف.
الشكل		

المصدر: من إعداد الباحثة بالاعتماد على المراجع ذات العلاقة.

النتائج:

يمكن تلخيص أهم النتائج التي توصلت إليها الدراسة في النقاط الآتية:

1. التحليل العنقودي يكون أكثر فائدة عندما تهدف الدراسة إلى اكتشاف أنماط أساسية وفق بيانات غير معرفة مسبقًا، ويكون التحليل التمييزي مناسبًا عند تخصيص بيانات جديدة للمجموعات المعرفة مسبقًا.

2. التحليل التمييزي يسعى للوصول إلى دالة التمايز؛ والتي تهدف إلى تعظيم الفروق بين متوسط المجموعات وتقليل التشابه في أخطاء التصنيف في نفس الوقت.
3. التحليل العنقودي أقل تعقيدًا من التحليل التمييزي؛ إذ انه يتطلب تجميع المفردات في عناقيد فقط، بعكس التحليل التمييزي والذي يتطلب تصنيف البيانات وتمييزها.
4. يساعد التحليل التمييزي في عملية التنبؤ بناءً على البيانات المعرفة مسبقًا.
5. نتيجة لتقنية التحليل العنقودي الاستكشافية؛ إذ يمكن أن يستعمل في العديد من العمليات مثل تحليل الصور وضغط البيانات والتعلم الآلي.

التوصيات:

1. التوسع في استخدام التحليل العنقودي والتحليل التمييزي في المجالات الاجتماعية والصحية والاقتصادية.
2. تزويد المكتبة العربية بالعديد من الدراسات المتعلقة بالتحليل العنقودي والتحليل التمييزي والذي يعتبر من المواضيع المتقدمة في الإحصاء.
3. إجراء المزيد من البحوث والدراسات حول أساليب التحليل الإحصائي متعدد المتغيرات.

المراجع:

[1] أحمد، رجاء إدريس. (2018). استخدام مفهوم العنقدة في تحليل أنماط المبيعات دراسة حالة (شركة مزيان) [رسالة ماجستير غير منشورة]. جامعة النيلين.

[2] أحمد، طالب. (2015). تصنيف المحافظات السورية حسب الإنفاق الاستهلاكي للأسرة باستخدام التحليل العنقودي. مجلة جامعة تشرين للبحوث والدراسات العلمية، 37(2)، 45-64.

[3] الشمراني، محمد موسى. (2020). توظيف أسلوبي التحليل العنقودي والتحليل التمييزي في تصنيف البيانات وبناء الدوال التمييزية. مجلة كلية التربية، 39(186 ج1)، 11-39.

[4] الصويعي، عبدالحليم مولود، بنيني، فاطمة خليفة. (2020). التحليل التمييزي وفعاليته في تصنيف تأثير وزن الحقيبة المدرسية على صحة التلاميذ "دراسة تطبيقية على تلاميذ مرحلة التعليم الأساسي بمدينة الزاوية". International Multilingual Journal of Science and Technology (IMJST)، 5(9)، 1659-1662.

[5] العلي، إبراهيم محمد. (2020). أسس التحليل الإحصائي متعدد المتغيرات. تم الاسترجاع من الرابط https://www.researchgate.net/publication/340825697

[6] المخلافي، فؤاد عبده.(2019). تصنيف وتمييز المحافظات اليمنية بحسب مصادر الدخل باستخدام التحليل العنقودي والتحليل التمييزي. مجلة بحوث جامعة تعز،(19)،32-57.

[7] الموسى، ياسر ، الجاسم، عبد الناصر ، ودهان، محمد.(2015). تحسين خوارزمية العنقدة K Means- باستخدام التحليل العاملي. مجلة بحوث جامعة حلب،(16)،1-21.

[8] بالرايس، نورة ، عبان ، ذكرى.(2022). فعالية استخدام أسلوب التحليل التمييزي في تيسير مخاطر الائتمان في المؤسسات الصغيرة و المتوسطة دراسة حالة فرع الوكالة الوطنية لتسيير القرض المصغر تبسة خلال الفترة 2016-2020[رسالة ماجستير منشورة] . جامعة العربي التبسي.

[9] بسيوني، عبد الرحيم. (2021). استخدام التحليل التمييزي في التصنيف والتنبؤ (دراسة تطبيقية). التجارة والتمويل، 41(3)، 297-325.

[10] بقريش, & شتراوي.(12-13 نوفمبر 2017). التحليل التمييزي كأحد الأساليب المعلمية في التنقيب عن البيانات لتسيير مخاطر القرض دراسة حالة فرع الوكالة الوطنية لتسيير القرض المصغر بالمسيلة. الملتقى العلمي الدولي حول: التحول الرقمي للمؤسسات والنماذج التنبؤية على المعطيات الكبيرة، جامعة محمد بوضياف بالمسيلة، كلية العلوم الاقتصادية والتجارية وعلوم التسيير.

[11] جودة، محفوظ. (2008) . التحليل الإحصائي الأساسي باستخدام SPSS. دار وائل للطباعة والنشر والتوزيع.

[12] حنيش، إبراهيم سليمان، وأسميو، خلود سليمان. (2019). المقارنة بين طرق التعنقد الهرمي واختيار أفضلها مع تطبيق عملي على بعض أنواع الحليب المبيع في مدينة مصراتة. مجلة العلوم الأساسية والتطبيقية، (عدد خاص)،275-284. مجلة العلوم(misuratau.edu.ly) .

[13] رزق الله، عايدة .(2002).دليل الباحثين في التحليل الإحصائي "الاختبار و التفسير". كلية التجارة، جامعة عين شمس، مصر.

[14] سليمان، مثنى صبحي، قاسم ، عمر صابر، وحسين، طلال فاضل. (2011). مقارنة بين طريقة السيطرة المضببة والدالة التمييزية في تصنيف بعض آبار محافظة نينوى. المجلة العراقية للعلوم الإحصائية، 11(2)، 315-330.

[15] طاقية، البيومي عوض، المجي، هشام محمد، والعتابي، كريم خلف. (2016). استخدام الانحدار اللوجيستي والتحليل التمييزي لدراسة حالات الإصابة بمرض الإسهال لدى الأطفال في العراق. المجلة المصرية للدراسات التجارية، 40(1)، 233-255.

[16] طه، حذيفة حازم، وحسين، محمد زيد. (2012). استخدام التحليل العنقودي لتصنيف نوعية المياه الجوفية في آبار منطقة بعشيقة في محافظة نينوى. المجلة العراقية للعلوم الإحصائية، (12)، 215-233.

[17] عاشور، وفاء عبد الصمد. (2019). تصنيف المحافظات العراقية صحياً باستخدام التحليل العنقودي لعام 2016. مجلة العلوم الطبيعية والحياتية والتطبيقية، 3(3) ،121-135.

[18] علي، كنان أحمد. (2015). فاعلية استخدام التحليل العنقودي والتحليل التمييزي في التحقق من الدلالة التمييزية لاختبارات الذكاء والشخصية (دراسة ميدانية مقارنة في محافظة دمشق)[رسالة ماجستير منشورة]. جامعة دمشق.

[19] مانلي، بريان. (2001) . الأساس في الطرق الإحصائية المتعددة المتغيرات (عبد الرحمن محمد أبو عمة، مترجم). النشر العلمي والمطابع، جامعة الملك سعود، الرياض، السعودية.

[20] هندوش، رنا وليد. (2010). استخدام العنقدة في نمذجة سلم الرواتب الشهري. المجلة العراقية للعلوم الإحصائية، (18)، 297-320.

Foreign References:

[21] Banerjee, A., Basu, S., & Merugu, S. (2007, April). Multi-way clustering on relation graphs. In Proceedings of the 2007 SIAM international conference on data mining (pp. 145-156). Society for Industrial and Applied Mathematics.

[22] Bekkerman, R., El-Yaniv, R., & McCallum, A. (2005, August). Multi-way distributional clustering via pairwise interactions. In Proceedings of the 22nd international conference on Machine learning (pp. 41-48).

RESEARCH ARTICLE

THE DIFFERENCE BETWEEN CLUSTER ANALYSIS AND DISCRIMINANT ANALYSIS

Hala Mohammed Ahmed Ali Babresh*

Dept. of Information & Statistics, Faculty of Administration Science, University of Aden, Aden, Yemen

***Corresponding author: Hala Mohammed Ahmed Ali Babresh; E-mail: hala199062@gmail.com**

Received: 28 June 2023 / Accepted 22 August 2023 / Published online: 30 September 2023

Abstract

The study aimed to introduce cluster analysis and Discriminant analysis, as they are considered advanced topics in statistics and are not accessible to many researchers, to give them an idea of how to use them in the process of classifying multivariate phenomena, and to achieve the goal of the study, the researcher used the descriptive approach for its suitability to the nature of the study. The study found many differences between Discriminant analysis and cluster analysis in the study of phenomena, most notably that Discriminant analysis is concerned with the issue of discrimination and separation between groups and requires prior knowledge of the number of totals, as it seeks to form a statistical model that illustrates the interrelationship between different variables.

Keywords: Cluster analysis, Discriminant analysis, Multivariate analysis

كيفية الاقتباس من هذا البحث:

بابريش، هـ. م. أ. ع.، (2023). الفرق بين التحليل العنقودي والتحليل التمييزي. مجلّة جامعة عدن الإلكترونيّة للعلوم الانسانيّة والاجتماعية، 4(3)، ص476-489. https://doi.org/10.47372/ejua-hs.2023.3.282