جامعة السودان للعلوم والتكنلوجيا

كلية الدراسات العليا

دراسة مقارنة لتحليل وتصنيف متغيرات جواز السفر الالكتروني

باستخدام طريقتي الخوارزمية الجينية والانحدار اللوجيستي

(دراسة حالة جواز السفر الالكتروني السوداني)

A comparative Study for Analysis and Classification of the Electronic Passport using Genetic Algorithm and Logistic Regression

(Case study of the Sudanese Electronic Passport)

رسالة مقدمة لنيل درجة الدكتوراه في الإحصاء

اعداد الباحث/

فائز حامد سلمان

اشراف/

الدكتور/ ا.م. احمد محمد حمدي

الدكتورة/ ا.م. امل السير خضر

الخرطوم

1440 ه - 2019م

بسم الله الرحمن الرحيم

الاستهلال

قال الله تعالى:

(إِنَّا نَحْنُ نُحْيِي الْمَوْتَى وَنَكْتُبُ مَا قَدَّمُوا وَآثَارَهُمْ وَكُلَّ شَيْءٍ أَحْصَيْنَاهُ فِي إِمَامٍ مُبِينٍ)

صدق الله العظيم

سورة يس ، آية: (12)

الإهداء

أهدي هذا العمل المتواضع إلى روح والدي الطاهر... رحمة وغفراناً

وإلى أمي التي زودتني بالحنان والمحبة
والى رفيقة الدرب.... زوجتي واولادي حباً ووفاءاً

أقول لهم: أنتم وهبتموني الحياة والأمل والنشأة على شغف الاطلاع والمعرفة

وإلى إخوتي وأسرتي جميعاً

ثم إلى كل من علمني حرفاً أصبح سنا برقه يضيء الطريق أمامي

الشكر والتقدير

الشكر لله سبحانه وتعالى الذي من علي بهذا الفعل العظيم واساله سبحانه وتعالى ان يتم فضله ويكمل نعمنا برضاه في الاخرة.

لا يسعني الا ان أتقدم بوافر الشكر والعرفان لجامعة السودان للعلوم والتكنلوجيا التي منحتني فرصة الالتحاق ببرنامج الدكتوراه.

والتقدير كل التقدير الى استاذي ومعلمي الكبير المرحوم الأستاذ احمد محمد حمدي واكن له كل الوفاء والاحترام، والى استاذتي المشرفة الدكتورة امل السر لجهودهما في إتمام وإنجاز رسالتي.

كما اخص بالشكر والتقدير مع خالص احترامي وتقديري للدكتور صباح منفي التدريسي في جامعة بغداد والمدرس هشام الدليمي التدريسي في جامعة ديالى لجهودهم القيمة.

كذلك ارسل صوت شكري وتقديري الى أساتذة وموظفي الدراسات العليا في جامعة السودان للعلوم والتكنلوجيا وفي مقدمتهم الست اماني لمساعدتهم لي في تسهيل مهمتي.

والى جميع الذين وقفوا من خلفي وشدوا من ازري اساتذتي وزملائي واصدقائي.

المستخلص

لغرض المقارنة للطرق الإحصائية التقليدية المستعملة في تحليل وتصنيف البيانات ذات الكم الهائل من المعلومات والطرق الإحصائية الحديثة المحوسبة للحصول على علاقات ونماذج تساعد في استخلاص المعلومات المفيدة والساندة لاتخاذ القرار ومن هنا تبرز مشكلة الدراسة في كيفية توظيف هذا المجال في تنظيم وتحليل وتصنيف مثل هذا الكم الهائل من البيانات، ووضع قاعدة عامة تخص الجواز السفر الالكتروني (السودان انموذجا) باستعمال افضل الطرق الإحصائية، تم دراسة وتحليل النماذج والطرق الاحصائية التقليدية والحديثة المحوسبة والمقارنة بينها وربط الخوارزمية الجينية ببعض الأساليب الإحصائية التقليدية تم ذلك من خلال استعمال الجانب التجريبي لهذه الطرق باستعمال المحاكاة أولا لتحويلها الى بيانات افتراضية تحاكي الواقع الحقيقي من اجل التوصل الى اختيار افضل طريقة لحل هذه المشاكل باقل كلفة واقصر وقت ممكن، تم ذلك بتصميم برنامج محاكاة باستعمال الخوارزمية الجينية ومن خلال برنامج ماتلاب الإصدار (R2017B) وبعد تحليل البيانات تم تصنيفها باستعمال برنامج الحزمة الإحصائية للعلوم الاجتماعية (SPSS 20) وقد تناولت الدراسة جانبين الأول تمثل بالجانب النظري والتجريبي، وتم استعراض الطرق التقليدية في التحليل وبيان مدى أهمية تطبيقها في الخوارزمية الجينية لاختيار الأفضل، وهذا ما اثبته الجانب الثاني العملي الذي تم تصميمه وتطبيقه على البيانات الاصلية للدراسة من خلال تطبيق نموذج برمجي يلائم واقع البيانات، تم التوصل الى ان افضل طريقة هي طريقة المربعات الصغرى الموزونة (WLS) من نتائج الطرق الكلاسيكية ولمختلف حجوم العينات في تقدير الانحدار اللوجيستي في الجانب التجريبي، وظهرت طريقة الإمكان الأعظم (MLE) المعتمدة على الخوارزمية الجينية كأفضل طريقة ولجميع حجوم العينات، اذ تفوقت جميع الطرق المعتمدة على الخوارزمية الجينية عن الطرق الكلاسيكية.

وتوصلت الدراسة الى التوصيات باستعمال الطرق المعتمدة على الخوارزمية الجينية في تقدير معلمات الانحدار اللوجيستي، ويمكن اعتماد نتائج خوارزميات العنقدة في التنقيب عن البيانات خاصة عند التعامل مع كم هائل من البيانات، اذ يمكن ان تتعامل مع بيانات الحوكمة الالكترونية الذكية وباستعمال خوارزميات الذكاء الاصطناعي، ومنها الخوارزمية الجينية مع خوارزميات التحليل العنقودي في تنقيب البيانات.

Abstract

For the purpose of comparison of the traditional statistical methods used in analysing and classifying data with a vast amount of information and modern statistical methods computerized to obtain relationships and models that help in extracting useful information and support for decision making and hence the problem of study in how Employing this field in the organization, analysis and classification of such a large amount of data, and the establishment of a general rule concerning the electronic passport (Sudan model) using the best statistical methods, studied and analyzed the traditional and modern statistical models and methods and the comparison between them and linking The genetic algorithm with some traditional statistical methods was done by using the experimental side of these methods using simulations first to convert them into hypothetical data that simulates real reality in order to arrive at the best way to solve these problems at the lowest cost and shortest possible time, designed A simulation program using the genetic algorithm and through the MATLAB version program (R2017B) and after analyzing the data was compiled using the program of the statistical Package of social Sciences (SPSS 20) The study examined two aspects of the first represented by the theoretical and experimental side, and reviewed the methods Traditional analysis and demonstrate how important it is to apply it in the genetic algorithm to choose the best, this is demonstrated by the second practical aspect that has been designed and applied to the original data of the study through the application of a programmatic model that fits the reality of the data,

it was found that the best way is the way of squares Weighted Micro (WLS) results from classical methods and for different sample sizes in estimating the logistic regression on the experimental side, and the greatest possible method (MLE) based on the genetic algorithm emerged as the best method and for all sample sizes, as all methods based on the algorithm exceeded Genetic for classical methods.

The study found recommendations using methods based on the genetic algorithm to estimate the parameters of the logistic regression, and the results of the algorithms for data mining can be adopted especially when dealing with a large amount of data, as it can deal with e-governance data Smart and using artificial intelligence algorithms, including the genetic algorithm with cluster analysis algorithms in data mining.

المحتويات

الموضوع	الصفحة
الاستهلال	أ
الاهداء	ب
الشكر والتقدير	ج
المستخلص باللغة العربية	د، هـ
Abstract	و، ز
المحتويات	ح، ط، ي
قائمة الجداول	ك
قائمة الاشكال	ل
قائمة الملاحق	ن
الفصل الأول: المقدمة	19-2
1-1: المقدمة	2
2-1 مشكلة الدراسة	4
3-1 أهمية الدراسة	5
4-1 اهداف الدراسة	5
5-1 فرضيات الدراسة	6
6-1 منهجية الدراسة	6
7-1 مصادر جمع المعلومات	6

8-1 الدراسات السابقة	7
9-1 هيكل الدراسة	19
الفصل الثاني: مفاهيم اساسية	34-21
1-2 جواز السفر " مفاهيم عامة"	21
2-2 تنقيب البيانات	24
3-2 المحاكاة	34-33
الفصل الثالث: الجانب النظري الطرق الإحصائية المستخدمة	65-36
1-3 الاطار النظري للتحليل العنقودي	36
2-3 الاطار النظري للمركبات الرئيسة	46
1-2-3 مفهوم تحليل المركبات الرئيسية	46
2-2-3 خصائص المركبات الرئيسية المحسوبة من المصفوفة (s)	48
3-2-3 اختبار المعنوية للمركبات الرئيسية	49
3-3 طريقة الإمكان الأعظم في التقدير	50
1-3-3 مفهوم طريقة الإمكان الأعظم	50
4-3 مفهوم طريقة بيز في التقدير	51
5-3 مفهوم الخوارزمية الجينية	54
1-5-3 الخوارزميات الجينية	55
2-5-3 عناصر الخوارزميات الجينية	56
3-5-3 ترميز الصبغيات	57
4-5-3 طرائق الترميز	57

3-5-5 دالة الكفاءة	57
3-5-6 الية عمل الخوارزمية الجينية	57
3-5-7 معايير الاختبار والعبور والطفرة في الخوارزميات الجنية	58
3-6 نموذج الانحدار اللوجيستي	60
3-6-1 طبيعة النموذج اللوجيستي	61
3-6-2 التحويل الخطي للنموذج اللوجيستي واسبابه	65-63
الفصل الرابع: الجانب التطبيقي	85-67
4-1 الجانب التجريبي	67
4-1-1 المحاكاة وتقدير معادلات الانحدار في الخوارزمية الجينية	67
4-1-2 المقارنة بين نتائج الطرق الكلاسيكية والخوارزمية الجينية	71
4-2 الجانب العملي	74
4-2-1 تطبيق افضل الطرق التي تم التوصل اليها في الجانب التجربي	77
4-2-2 تطبيق التحليل العنقودي في تصنيف المتغيرات للبيانات الحقيقية	85-80
الفصل الخامس: النتائج والتوصيات	88-87
أولا: النتائج	87
ثانيا: التوصيات	88
قائمة المصادر والمراجع	93-90
قائمة الملاحق	ا-ع

قائمة الجداول

رقم الجدول	عنوان الجدول	رقم الصفحة
1-4	احجام العينات حسب النموذج	68
2-4	القيم الافتراضية	68
3-4	قيم mse للطرق الكلاسيكية للنموذج الاول	69
4-4	قيم mse للطرق الكلاسيكية للنموذج الثاني	69
5-4	قيم mse للطرق الكلاسيكية في الخوارزمية الجينية للنموذج الاول	70
6-4	قيم mse للطرق الكلاسيكية في الخوارزمية الجينية للنموذج الثاني	70
7-4	قيم mse للطرق الكلاسيكية و الخوارزمية الجينية ولحجوم عينات مختلفة وبمعالم مختلفة للنموذج الاول	71
8-4	قيم mse للطرق الكلاسيكية و الخوارزمية الجينية ولحجوم عينات مختلفة وبمعالم مختلفة للنموذج الثاني	72
9-4	نتائج الانحدار اللوجيستي بطريقة (WLS)	75
10-4	نتائج الانحدار اللوجيستي بطريقة الخوارزمية الجينية	78
11-4	جدول تحليل التباين بالنسبة لمتغيرات البحث	80
12-4	المسافة بين مراكز المجموعات بالنسبة لمتغيرات الدرسة	81
13-4	متوسطات المتغيرات في المجموعات المختلفة بالنسبة لمتغيرات الدراسة	82
14-4	مصفوفة القرابة باستخدام الطريقة الهرمية للمفردات (الأقاليم)	83
15-4	خطوات التجميع بالطريقة الهرمية	84
16-4	عضوية كل متغير بالمجموعة التي تنتمي اليها	85

قائمة الاشكال

رقم الشكل	عنوان الشكل	رقم الصفحة
1-2	معمارية التنقيب في البيانات	24
2-2	علاقة التنقيب في البيانات مع مجالات اخرى	32
3-2	علاقة تنقيب البيانات بعلم الاحصاء	33
1-3	طريقة الجار الأقرب (الربط الأحادي)	39
2-3	طريقة الجار الابعد	40
3-3	طريقة الربط بالاعتماد على المعدل	40
4-3	طريقة وورلد لربط العنقدة لمجموعة من العناصر	42
5-3	محاور المركبات الرئيسية	47
6-3	توزيع P	62
7-3	المنحنى اللوجيستي	62
8-3	العلاقة الخطية بين المتغير التوضيحي ونسبة الاستجابة المحولة Z	64
1-4	الالواح الجليدية باستخدام الطريقة الهرمية	84
2-4	المخطط الهرمي لمتغيرات الدراسة	85

قائمة الملاحق

رقم الملحق		
1	البرنامج الأول: برنامج المحاكاة (البرنامج التجريبي)	أ
2	البرنامج التطبيقي للبيانات الاصلية	ب

الفصل الاول

المقدمة

(الاطار المنهجي والدراسات السابقة)

1.1 المقدمة

يتميز عصرنا الراهن (عصر الانترنيت و الاقتصاد الرقمي) بالسيل العظيم والانتشار الواسع النطاق للبيانات حتى أضحى من المستحيل على المحللين استخلاص معلومات ذات معنى باللجوء فقط إلى المداخل التقليدية للتحليل التمهيدي للبيانات.

ومع وجود كميات كبيرة من البيانات المخزنة في قواعد البيانات و مخازن البيانات ازدادت الحاجة إلى تطوير أدوات تمتاز بالقوة لتحليل البيانات و استخراج المعلومات والمعارف منها، من هنا ظهر ما يسمى بالتنقيب في البيانات كتقنية تهدف إلى استخراج المعرفة من كميات هائلة من البيانات. وهي تقنية حديثة فرضت نفسها بقوة في عصر المعلوماتية، واستخدامها يوفر للشركات والمنظمات في جميع المجالات القدرة على استكشاف والتركيز على أهم المعلومات في قواعد البيانات، كما تركز تقنيات التنقيب على بناء التنبؤات المستقبلية واستكشاف السلوك والاتجاهات مما يسمح باتخاذ القرارات الصحيحة واتخاذها في الوقت المناسب. والتي تعتبر بدورها مرحلة من مراحل عملية أكثر تعقيدا هي استكشاف المعرفة في قواعد البيانات، والمرتبطة إلى حد بعيد بعملية تطوير أخرى مهمة جدا هي مستودعات البيانات. حيث أن الكثير من الشركات و المنظمات الرائدة اليوم تستخدم عملية استكشاف المعرفة في قواعد البيانات بشكل منهجي و منظم بوصفها تشكل جوهر العمل الذي يعتمد عليه في تفعيل النشاط و تحقيق الميزة التنافسية.

اذ تمر البشرية في مرحلة تفوق العلوم والتقنيات، ولا يخفى على أحد الأثر الذي تركه هذا التفوق في حياة الإنسان، ومما لا شك فيه أن المجتمعات التي استملكت المعلومة والمعرفة هي الآن التي تحكم العالم وتتصدر مشهد الدول المتقدمة والعصرية.

أحد أشكال التقدم والتقنية وأحد أبرز آثارها هو تضخم المعلومات الهائل، والذي يحدث باضطراد مستمر، حيث أصبحت مخازن البيانات وقواعد البيانات التي تعتمد عليها أغلب النظم

والتطبيقات، أصبحت ممتلئة بكم كبير وضخم من البيانات العشوائية والتاريخية، ونستطيع القول فعلاً أن البيانات لا تنام ولا تتوقف عن النمو. حسب آخر الإحصائيات زاد عدد مستخدمي شبكة الانترنت في عام 2017 بنسبة 7.5% عن عام 2016 ليصل إلى 3.7 مليار مستخدم. الموقع الاجتماعي تويتر على سبيل المثال، يضم كماً كبيراً من المستخدمين الفاعلين على مدار الساعة، ويتولد عنهم في كل دقيقة ما يقرب 456,000 تغريدة تحتوي على نصوص، صور، فيديوهات، روابط وهشتاقات، ولك أن تتخيل كم البيانات الناتج عن ذلك خلال يوم واحد فقط.

مع هذا النمو في البيانات، أصبحنا أمام تحدي كبير يتمثل في كيفية الاستفادة من هذه البيانات المُخزنة في مستودعاتها، وكيفية القيام بعملية استخراج المعلومة المفيدة منها، وهذا التحدي أصبح لا يمكن حله بواسطة الطرق التقليدية لتحليل البيانات، لذلك فإننا بحاجة إلى تقنيات غير تقليدية تستوعب التعامل مع الكميات الضخمة للبيانات بالإضافة للقدرة على التعامل مع الأنواع والاشكال المختلفة للبيانات، وهذا يعني تقنيات على مستوى عالي من الكفاءة والذكاء.

ان اتجاه العالم الى حوسبة المعلومات له مزايا عديدة منها سرعة الوصول الى المعلومات، استخدام اوفر للتخزين، ومن أهمها هي دراسة البيانات وتحليلها وتصنيفها احصائيا وتعتبر الحكومة الالكترونية هي اكبر مصدر للبيانات حول العالم، حيث يتراوح كم البيانات الموجودة الى ترليون من البيانات والتي ترتبط فيما بينها، ولها أيضا قيود دولية تسهل عمل تبادل البيانات، وفي هذه الدراسة تم التطرق الى الجواز السفر الالكتروني كوثيقة مهمة وتخضع للقيود الدولية، ويمكن تعريف الجواز الالكتروني بانه وثيقة رسمية للسفر، تقر من خلال الدولة التي يتبعها مواطن ما، والتي تعرف حاملها من حيث جنسيته وهويته طبقا للدولة التي ينتمي اليها، وتسمح هذه الوثيقة لحاملها بالدخول والمرور خلال الدول الأخرى.

ولغرض تحليل وتصنيف هذه البيانات بالطرق الإحصائية الكلاسيكية والحديثة تم استعمال حجوم مختلفة من العينات لغرض التعرف على حجم التغيرات التي تحصل في النتائج عند زيادة حجوم العينات للمشكلة المدروسة، تم ذلك باستعمال أسلوب المحاكاة الذي يعتبر من الأساليب المهمة في حل المشاكل التي يصعب حلها بالطرائق التحليلية او العددية والذي من خلاله تم اختيار افضل الطرق الكلاسيكية للتحليل والتصنيف بتطبيق الخوارزمية الجينية عليها واستعمال معيار المقارنة متوسط مربعات الخطأ (MSE) وتم تصنيف البيانات باستعمال التحليل العنقودي.

وكانت من اهم النتائج التي تم التوصل اليها هي:

1) من نتائج الجانب التجريبي ظهرت افضل طريقة هي طريقة المربعات الصغرى الموزونة من الطرق الكلاسيكية في تقدير الانحدار اللوجيستي.
2) ظهرت طريقة الإمكان الأعظم المعتمدة على الخوارزمية الجينية كأفضل طريقة ولجميع حجوم العينات في تقدير الانحدار اللوجيستي.
3) تفوقت جميع الطرق المعتمدة على الخوارزمية الجينية من الطرق الكلاسيكية.
4) في الجانب التطبيقي ظهرت المتغيرات لون العين، لون البشرة ذات تأثير معنوي على نوع الجنس بالطرق الكلاسيكية.

2.1 مشكلة الدراسة:

ان جواز السفر هو وثيقة رسمية، تقر من خلال الدولة التي يتبعها مواطن ما، التي تعرف حاملها من حيث جنسيته وهويته طبقا للدولة التي ينتمي اليها، وتسمح هذه الوثيقة لحاملها بدخول والمرور خلال الدول الأخرى، وان جوازات السفر مرتبطة بصفة الحماية للشخص الحامل لها من قبل الدولة المنتمي لها ذلك الشخص، والحق بدخول أي دولة بصفته وجنسته.

ان دخول المجال الحوسبي في الكثير من مجالات الحياة ومنها مجال التحقق من الهوية أصبح من أكثر التقنيات المستخدمة، وبالأخص في مجال المعلومات الالكترونية حيث تتوفر ملايين البيانات التي يجب التعامل معها وتصنيفها حسب الطرق العلمية والخوارزميات اللازمة، ومن هنا تبرز مشكلة الدراسة في كيفية توظيف هذا المجال في تنظيم وتصنيف مثل هذا الكم الهائل من البيانات ووضع قاعدة بيانات عامة تخص الجواز الالكتروني (السودان انموذجا).

3.1 اهمية الدراسة:

تنبثق أهمية الدراسة من حيث استخدامها لبعض الطرق الإحصائية في تحليل جواز السفر الالكتروني والمتمثلة بتحليل المركبات الأساسية، التحليل البيزي، تحليل الإمكان الأعظم والخوارزمية الجينية، فضلا عن تصنيف المعلومات الخاصة بجواز السفر الالكتروني باستخدام التحليل العنقودي بطريقتي التحليل الهرمي والمتوسطات.

كذلك فان ندرة الدراسات المحلية والعربية والأجنبية التي استخدمت كل من التحليل العنقودي في التصنيف والخوارزمية الجينية في تحليل قاعدة البيانات الخاصة بجواز السفر الالكتروني، رغم الضرورة الماسة لها في هذا المجال.

4.1 اهداف الدراسة:

هدفت الدراسة الى تسليط الضوء على النقاط التالية:

1. دراسة وتحليل النماذج الإحصائية التقليدية والحديثة والمحوسبة والمقارنة بينها لتحديد الأفضل بالاعتماد على مؤشر اقل متوسط مربعات الخطأ.
2. ربط الخوارزمية الجينية ببعض الأساليب الإحصائية والمقارنة مع الطرق الكلاسيكية في التحليل والتصنيف.
3. التعرف على الطرق التقليدية والخوارزمية الجينية في تحليل وتصنيف البيانات.

5.1 فرضيات الدراسة:

تضمنت الدراسة الفرضيات التالية:

1- الطرق الاحصائية الحديثة افضل من الطرق الكلاسيكية في تحليل وتصنيف الانحدار اللوجستي.

2- الخوارزمية الجينية افضل طريقة من الطرق الكلاسيكية لجميع حجوم العينات في التقدير.

3- تضم العينات متغيرات متجانسة وغير متجانسة.

6.1 منهجية الدراسة:

تضمنت المنهجية المتبعة في هذه الدراسة على:

1. العمل الأساسي يعتمد على الخوارزميات و النماذج وذلك بسبب الحاجة الى تحليل وتصنيف البيانات.
2. دراسة وتحليل للأدبيات والدراسات السابقة من اجل استنباط النماذج الرياضية والتي يمكن نمذجتها باستخدام برامج النمذجة الحديثة.
3. نمذجة ومحاكاة الطرق الإحصائية وربطها بالخوارزمية الجينية
4. عمل مقارنة بين الطرق الكلاسيكية والحديثة وتحديد معيار المفاضلة.
5. اختيار نوع الخوارزمية المناسبة وعمل برنامج تطبيقي يتم من خلاله التوصل الى النتائج، وذلك عن طريق جمع المعلومات عبر الاستعانة بعينات حقيقية من قاعدة البيانات الخاصة بالجواز الالكتروني السوداني.

7.1 مصادر جمع المعلومات:

أولا: الجانب النظري: تم الاعتماد في هذا الجانب على:

1. المصادر العربية والأجنبية والكتب.
2. الاطاريح والرسائل التي تخص موضوع الدراسة.
3. المقالات المنشورة في المجلات العلمية والدوريات والبحوث
4. الانترنت

ثانيا: الجانب العملي: تم الاعتماد في هذا الجانب على:

1. جمع المعلومات عبر الاستعانة بعينات حقيقية من قاعدة البيانات الخاصة بالجواز الالكتروني السوداني.

8.1: الدراسات السابقة

تناول هذا المبحث الدراسات التي لها صلة بموضوع البحث والتي تناولت المحاور التالية:

1. الدراسات التي تخص طرق التحليل الكلاسيكية والحديثة المستخدمة في الدراسة.
2. الدراسات التي تناولت تحليل جواز السفر

1:- الدراسات التي تخص طرق التحليل الكلاسيكية والحديثة المستخدمة في الدراسة.

- **الدراسة الأولى (الاسدي، خروفة، شهلة 2007)**

قامت الباحثة بأجراء دراسة حول استخدام الخوارزمية الجينية المهجنة لتصنيف صور الاقمار الصناعية وكان الهدف منها استخدام الخوارزمية الجينية المهجنة لتصنيف صور الاقمار الصناعية" وقد تم اقتراح تهجين طريقة (k means) مع الخوارزمية الجينية وتبينت هذه الطريقة المقترحة بانها تنتج الحل الامثل او القريب من الحل الامثل ضمن مجموعة حلول ومساحة بحث معقدة، وقد ثبت ذلك عمليا اذ تم استخدام اسلوب (K means) مع الخوارزمية الجينية (الطريقة الهجينة المقترحة" وحققت نتائج جيدة في تصنيف الصور ذات التدرجات

الرمادية مقارنة مع طريقة (k means). وقد توصلت هذه الدراسة الى مجموعة من الاستنتاجات أهمها:

- ان شجرة القرار واحدة من الأساليب سهلة الاستخدام في التصنيف مقارنة مع خوارزمية المتوسطات.
- الاعتماد على الطريقة المهجنة المقترحة في التصنيف.

- **الدراسة الثانية (متراس، 2007)**

قامت الباحثة بأجراء دراسة حول استخدام الخوارزمية الجينية المهجنة لتصنيف صور الاقمار الصناعية، وكان الهدف منها اجراء دراسة حول الخوارزمية الجينية التي تعد احد أساليب الذكاء الاصطناعي الحديثة وتطبيقها على مسالة النقل، حيث تم اقتراح خوارزمية نقل جينية لحل مسالة النقل وادى تطبيقها على مسائل النقل المتوازنة الى إيجاد عدد من الحلول بعدد مرات التوليد من ضمنها الحل الأمثل " اقل كلفة كلية ممكنة " وتم أيضا اقتراح طفرة جديدة من خلال خوارزمية النقل الجينية أدى استخدامها الى تحسين كبير في النتائج " تقليل قيمة الكلفة الكلية". وقد توصلت هذه الدراسة الى مجموعة من الاستنتاجات أهمها:

- ان شجرة القرار واحدة من الأساليب سهلة الاستخدام في التصنيف مقارنة مع خوارزمية المتوسطات.

وكانت اهم التوصيات هي الاعتماد على الطريقة المهجنة المقترحة في التصنيف.

- **الدراسة الثالثة (بحبوح، ريشة، 2007)**

قام الباحث بأجراء دراسة حول تأثير متغيرات الخوارزميات الجينية في مسائل إيجاد الحل الأمثل، وكان الهدف منها التعريف بأحد فروع الذكاء الاصطناعي وهي الخوارزميات الجينية، نظرا لقدرتها على حل كثير من المسائل المعقدة، وفي مجالات علمية متعددة سواء في

علم الحاسبات او بحوث العمليات ومعالجة الصور او حتى في علم الاجتماع، حيث تم استخدام الخوارزميات الجينية لإيجاد النهاية الحدية العظمى لتابع مستمر ضمن مجال محدود وذلك بهدف دراسة تأثير بعض اهم متغيرات الخوارزميات الجينية في الأداء ودقة النتائج، وقد توصل الباحث الى ان تأثير احتمال الطفرة وحجم الجيل وعدد مرات التكرار في دقة النتيجة وزمن التنفيذ باختيار إجرائية العجلة المتدحرجة في عملية الانتخاب، بعد ذلك تم مقارنة أداء العجلة المتدحرجة مع أداء إجرائية انتخاب أخرى هي حكم النخبة. واما من اهم التوصيات فكانت استخدام اجراءية حكم النخبة لما ابدته من ثبات واستقرار اعلى في النتائج التي نحصل عليها من العجلة المتدحرجة.

- **الدراسة الرابعة (البدراني، 2007)**

قام الباحث بأجراء دراسة حول استخدام الخوارزمية الجينية في تطابق أنماط الحرف الإنكليزي، وكان الهدف منها تركز الاهتمام فيه على مسالة تطابق الأنماط للحرف الإنكليزي باستخدام خوارزمية بحث احتمالية تدعى الخوارزمية الجينية، حيث يوضح البحث إمكانية الخوارزمية في تطابق الأنماط ما بين النمط الأصيل (المثالي) للحرف والانماط المشوهة، لكون الخوارزمية الجينية جيدة لعدد من الاعمال التي تتطلب الأمثلة، فهي تطبق على المسائل التي تمتلك مساحة واسعة ومتغيرات كبيرة وفي الإمكان حلها بسهولة وسرعة، كما انها تعطي حلا قريبا جدا من الحل المثالي للأنماط المستخدمة، وقد توصل الباحث الى قدرة الخوارزمية على اجراء عملية تطابق الأنماط للحروف الإنكليزية مقارنة بخوارزميات أخرى. اما من اهم التوصيات التي تم التوصل اليها فهي تطبيق البرنامج لأكثر من أربعة نماذج ي قواعد البيانات

- **الدراسة الخامسة (الكلاك، شعبان، 2008)**

قام الباحثان بأجراء دراسة حول الخوارزمية الجينية في جدولة العمليات مع عدم إمكانية القطع، وقد كان الهدف منها استخدام مفهوم الخوارزمية الجينية في جدولة المعالجات المتعددة، من خلال دمج خوارزميتين من خوارزميات الجدولة مع عدم إمكانية قطع العمليات والمتمثلة بالبيان الحلقي المباشر، فالخوارزمية الأولى تمثل جدولة المستوى الأعلى أولا مع تقدير الزمن، والخوارزمية الثانية تمثل جدولة المستوى الأصغر أولا مع تقدير الزمن، وتم التوصل الى نتائج إيجابية من خلال ابراز الحل الأفضل من بين عدد كبير من الحلول في تقليل طول الجدولة (تقليل زمن انهاء العمل) من خلال الجدولة المثلى والمتمثلة بالخوارزمية الجينية المقترحة اذ أظهرت كفاءة عالية وسهولة مرنة مقارنة مع استخدام خوارزميات الجدولة لوحدها. اما اهم التوصيات فهي اقتراح دالة لياقة جديدة واضفاء بعض التحويرات والاضافات الى روتين عجلة الروليت، من اجل تطبيق مفهوم الخوارزمية الجينية والوصول الى الحل الأفضل.

- **الدراسة السادسة (خروفه، 2009)**

قامت الباحثة بأجراء دراسة حول استخدام الخوارزمية الجينية لتحديد خصائص صور الأبنية، وقد كان الهدف منها الاستفادة من خواص الخوارزمية الجينية لتحديد خصائص صور الأبنية من خلال التعامل مع عدد من الصور المنتخبة لواجهات مجموعة من الأبنية وتطبيق الخوارزمية الجينية بعناصرها المختلفة من انتقاء وتقاطع وطفرة، اذ جاء التعامل مع الصور المنتخبة على أساس اعتماد خصائص معينه لواجهات الأبنية يمكن من خلالها استقراء الوظائف التي تضمنتها هذه الأبنية، فضلا عن وجود اليات لتحليل هذه المباني بالرجوع الى عدة اعتبارات سياقية وتقنية ووظيفية تحكم عمل المصمم المعماري، وتوصلت الباحثة الى أظهرت النتائج أهمية الخوارزمية الجينية في الأداء من حيث التعامل مع واجهات الأبنية، من جانب اخر فقد تم تحديد عدة ميزات يمكن من خلالها التعامل مع الأبنية وفق صيغ تحليلية تدخل النمط والسياق

ومواد البناء فضلا عن تأثير البيئة وعواملها المختلفة على المبنى، كذلك فقد أظهرت النتائج بعد تطبيق الخوارزمية المقترحة ورسم المخططات البيانية للصورة ان التدرجات اللونية التي تمتلك عددا قليلا من التكرار في بعض الأحيان تم تجاوزها باستخدام الخوارزمية الجينية الامر الذي يؤشر إمكانية التحكم والتصرف بحرية اكثر عند التحليل للوصول الى الحلول المثلى التي تعطي افضل النتائج. ومن اهم التوصيات فقد كانت الاهتمام بطرق تكنلوجيا المعلومات لما توفره من دقة عالية في التصنيف.

- **الدراسة السابعة (الجمل، 2009)**

قام الباحث بأجراء دراسة حول الثروة الحيوانية والمراعي في بعض قرى شمال محافظة نابلس، وقد كان الهدف منها دراسة واقع الإنتاج الحيواني في بعض قرى شمال محافظة نابلس، وذلك في محاولة منه لتطوير إمكانية الثروة الحيوانية وقدراتها، وذلك من خلال دراسة شاملة متخصصة تدرس وضع الثروة الحيوانية والظروف المحيطة بها بالمزارعين القائمين عليها، وظروف الإنتاج، والقاء الضوء على أهم المشاكل والمعوقات التي يعاني منها قطاع الثروة الحيوانية في منطقة الدراسة، وتم التوصل الى مجموعة من الاستنتاجات أهمها تواجه الثروة الحيوانية والمراعي الطبيعية في منطقة الدراسة الى مشاكل عديدة، تعمل على تقليص المساحات الرعوية وانخفاض الإنتاج الحيواني، مثل الزحف العمراني والرعي الجائر والتحطيب غير المسؤول. كما أظهرت النتائج ان العوامل الطبيعية تؤثر على الثروة الحيوانية والمراعي، ومن اهم التوصيات هي تحسين الأنواع والاصناف العلفية المحلية، والاهتمام باستنباط أصناف جديدة ووفيرة تتلاءم مع البيئة وقد تم استخدام الطرق الاحصائية التقليدية.

- **الدراسة الثامنة (عودة، 2009)**

قامت الباحثة بأجراء دراسة حول دراسة تحليلية للمشكلات الإنتاجية والمالية والإدارية والتسويقية لمشاريع تربية فروج اللحم في محافظة الديوانية وقد كان الهدف منها التعرف على

أهم المشكلات التي تواجه مشاريع تربية فروج اللحم في محافظة الديوانية، والتي تم تصنيفها الى مشكلات إنتاجية (فنية)، إدارية، مالية، تسويقية، كذلك التعرف على مدى وجود فروق في المشكلات التي تواجه المنتجين تبعاً لبعض خصائصهم (التحصيل الدراسي، سنوات الخبرة في تربية دجاج اللحم، عدد الوجبات خلال السنة، عدد الطيور في الوجبة الواحدة) فضلاً عن تحديد العلاقة بين خصائص المنتجين والمشكلات التي تواجههم. وتم التوصل الى ان أكثر المشكلات تأثيراً على المنتجين هي عدم تشغيل الحقول بطاقتها القصوى، كذلك عدم الانتظام أو الانقطاع المستمر للتيار الكهربائي، ارتفاع تكاليف الأعلاف، ومنافسة الدجاج المستورد، وارتفاع تكاليف الأفراخ، عدم ضمان نوعية الأفراخ، ونقص الوزن عن الوزن المثالي عند التسويق، عدم كفاءة الأدوية واللقاحات البيطرية، ارتفاع أسعار المحروقات (الوقود)، ارتفاع نسبة الفاقد من العلف، استغلال الوسطاء، استخدام العمالة غير الكفوءة (غير المدربة)، عدم معرفة كميات العلف المقدمة لمراحل التربية المختلفة، رداءة نوعية الأعلاف، ارتفاع تكاليف الأدوية واللقاحات البيطرية. وقد كانت من اهم التوصيات هي تنظيم برامج ارشادية لمنتجي دجاج اللحم بالتنسيق مع دوائر الارشاد وجمعيات منتجي الدواجن. تم استخدام الطرق الاحصائية التقليدية.

- **الدراسة التاسعة (يوسف واخرون، 2011)**

قام الباحثون بأجراء دراسة حول استخدام التحليل العنقودي لتقييم ابعاد دليل التنمية البشرية في العراق لعام 2009 وقد كان الهدف منها استخدام التحليل العنقودي لتقييم أبعاد دليل التنمية البشرية في العراق لعام 2006 والمتمثلة بالمتغيرات (دليل الالمام بالقراءة والكتابة، دليل نسبة الالتحاق الاجمالية، دليل التعليم، دليل حصة الفرد من الناتج المحلي الإجمالي) مصنفة حسب المحافظات العراقية، وذلك باستخدام التحليل العنقودي بالطريقة الهرمية (Hierarchical cluster analysis) وطريقة المتوسطات (K- means) بالنسبة للمشاهدات (المحافظات) والمتغيرات (عناصر أبعاد دليل التنمية البشرية) وتم التوصل الى ان تم تصنيف

المحافظات الى أربعة مجاميع، ضمت المجموعة الأولى (دهوك، سليمانية، أربيل) أما المجموعة الثانية فقد كانت (البصرة، ذي قار، صلاح الدين، واسط، كربلاء، بابل، كركوك، نينوى) وبالنسبة للمجموعة الثالثة فقد كانت (ميسان، المثنى، القادسية، النجف) في حين كانت المجموعة الرابعة (بغداد، الأنبار، ديالى). كما تم تصنيف أدلة التنمية البشرية في العراق الى أربعة مجاميع وذلك بالاعتماد على قيمة معامل الارتباط البسيط (Simple correlation coefficient) وقد تبين ان أعلى معامل ارتباط بسيط كان بين دليل التعليم ودليل الالمام بالقراءة والكتابة، وهذه علامة ارتباط طردية واقعية، حيث ان زيادة دليل التعليم ستؤدي الى زيادة الالمام بالقراءة والكتابة. وقد كانت من اهم التوصيات هي ضرورة رفع مستوى التنمية البشرية في المحافظات التي تعاني ضعف في هذا المجال والارتقاء بها.

- **الدراسة العاشرة (الكريم واخرون، 2012)**

قام الباحث بأجراء دراسة حول تصميم وتطبيق نظام كشف لاختراقات الشبكة مبني على العميل الذكي، وقد كان الهدف منها تصميم وتطبيق نظام كشف لاختراقات الشبكة مبني على العميل الذكي، اذ تعاني أنظمة كشف اختراقات الشبكة في الوقت الحاضر من مشكلة أساسية تتمثل في عدم قدرتها على كشف الهجمات غير المعرفة (Unknown Attacks) مسبقاً في قواعد بياناتها، مما يؤدي بدوره إلى ارتفاع معدل الإنذارات الخاطئة. تقدم هذه الورقة تطويراً لنظام فعال لكشف اختراقات الشبكة مبني على فكرة العميل الذكي (Intelligent Agent)، بحيث تم استخدام مفهوم منظومة العميل مع الخوارزمية الجينية (Genetic Algorithm) والتي تعتبر إحدى اليات التحسين (Enhancement) والبحث (Search) والتي تكسب العميل خاصية الذكاء. بعد تطبيق النظام على نوع من الديدان الحاسوبية وهى دودة الرمز الأحمر (Code Red Worm)، وتم التوصل الى أن نظم كشف اختراقات الشبكة المبنية على نظام العميل الذكي أكثر كفاءة من نظم كشف الاختراقات التقليدية من حيث تقليل معدل الإنذارات

الخاطئة (السلبية منها والايجابية)، فضلاً عن المقدرة على اكتشاف بعض الهجمات الغير معرفة مسبقاً في النظام. ومن اهم التوصيات فقد كانت استخدام نظام اكثر من عميل ذكي عند وجود بيانات كبيرة، بذلك يكون النظام اكثر فعالية.

- **الدراسة الحادي عشر (الشمرتي واخرون، 2012)**

قام الباحثان بأجراء دراسة حول استخدام التحليل العنقودي لتقييم مؤشرات البيئة والتنمية المستدامة في محافظات العراق للسنوات 2007-2011 وقد كان الهدف منها الكشف عن إمكانية تطبيق التحليل العنقودي في تحليل مشاكل مؤشرات التنمية المستدامة والبيئة، مصنفة حسب المحافظات العراقية، وتم التوصل الى ان محافظة بغداد هي أفضل محافظة من حيث مؤشرات التنمية المستدامة، وذلك يعود الى المشاريع والخطط التنموية في السنوات الأخيرة.

جاءت بالمرتبة الثانية كل من محافظات إقليم كوردستان والنجف ونينوى من حيث مؤشرات التنمية المستدامة. ومن اهم التوصيات وضع خطط تنموية سريعة في المحافظات التي ظهرت مؤشراتها التنموية بمستوى ضعيف للارتقاء بالمحافظات نحو تقدم التنمية المستدامة.

- **الدراسة الثاني عشر (طه واخرون، 2012)**

قام الباحث بأجراء دراسة حول استخدام التحليل العنقودي لتصنيف نوعية المياه الجوفية في ابار منطقة بعشيقة في محافظة نينوى، وقد هدفت الدراسة الى تصنيف عدد من آبار المياه الجوفية للمواقع المختارة في آبار منطقة بعشيقة في محافظة نينوى، وذلك طبقاً لنوعية المياه فيها، وتم التوصل الى ان أغلب مياه الآبار الجوفية لمنطقة بعشيقة صالحة للاستخدام البشري، والقسم الآخر منها صالح للاستخدام الحيواني، عدا البئر الذي يحمل الرقم (7) فهو غير صالح للاستخدام البشري وكذلك الحيواني، وذلك لارتفاع نسبة الملوحة فيه، الأمر الذي يشكل خطورة على حياة الحيوانات الصغيرة والتي هي في دور الحضانة. وقد كانت من اهم التوصيات هي

استخدام النتائج المستحصلة في هذا التصنيف في التوجيه الميداني لمستخدمي مياه الابار للحصول على الفائدة العظمى من خلال توظيف ما يمتاز به البئر من صفات.

- **الدراسة الثالث عشر (حجوز، 2014)**

قام الباحث بأجراء دراسة حول تحسين خوارزميات K-MEANS كان الهدف منها ايجاد افضل طريقة لتحديد المراكز الاولية وايجاد عملية ناجعة لتحديد طريقة انتماء الكائن لأي عنقود، وقد توصل الباحث الى ان الخوارزمية المحسنة التي تستخدم البيانات المقسمة على طول محور البيانات وفقا لأعلى اختلاف من اجل تعيين المراكز الأولية للخوارزمية.

- تم تقسيم الخوارزمية المحسنة باستخدام مجموعة من البيانات المختلفة، وهي مجموعة من السلع التي تباع في احدى المتاجر كالسكر والعصير والخبز، وتم استخدام عدد مرات التكرار والوقت المنقضي كمعايير للمقارنة، اذ بينت النتائج افضلية الخوارزمية المحسنة من خلال كونها تحتاج الى عدد اقل من التكرارات الذي بدوره يقلل من وقت التجميع.

اما التوصيات فهي ضرورة العمل على تحسين خوارزمية التصنيف k من اجل زيادة الدقة والتقليل من الوقت.

- **الدراسة الرابع عشر،(محمد بكري، 2015)**

قام الباحث بأجراء دراسة حول تحديد العوامل المؤثرة في مرض السكري باستخدام طرائق متعدد المتغيرات، وقد كان الهدف منها استخدام التحليل التمييزي والتحليل العاملي والتحليل العنقودي بغية الوصول الى مؤشرات احصائية تساعد المعنيين في ايجاد تفسيرات ملائمة لها واسبابها وبالتالي الخروج باستنتاجات وتوصيات تساعد في اتخاذ القرار الملائم بشأنها من قبل اصحاب القرار. وتم التوصل الى تم تصنيف عوامل الإصابة بمرض السكري بولاية شمال كردفان اعتمادا على عشرة متغيرات، وتم التوصل الى أربعة عناقيد حيث ان

العنقود او المجموعة الأولى تضم المهنة فقط، وتضم المجموعة الثانية العمر فقط وتضم المجموعة الثالثة القرابة والتدخين والكحول والرياضة وصنف الدم والمستوى التعليمي بينما المجموعة الرابعة تضم تنسيب كتلة الجسم (bmi) فقط.

وفيما يتعلق بمضاعفات مرض السكري بولاية شمال كردفان فقد تم تصنيفها الى اربعة عناقيد فقد ضمت المجموعة الأولى الإصابة بأمراض أخرى وامراض العيون والاصابة بأمراض وراثية، اما المجموعة الثانية فقد ضمت امراض القلب وامراض الكلى وامراض الجهاز العصبي وجروح باليد وجروح بالرجل وقد ضمت المجموعة الثالثة المضاعفات فقط، اما المجموعة الرابعة فقد ضمت بتر الأطراف فقط. اما اهم التوصيات فقد كانت عمل برامج تثقيفية لمرضى السكري بخصوص كيفية التغذية الصحية، فضلا عن عمل برامج تتضمن مسوحات للاكتشاف المبكر لهذا المرض.

- **الدراسة خامس عشر (أيوب، 2017)**

قام الباحث بأجراء دراسة حول دالة صلاحية جديدة لاستخدام الخوارزمية الجينية في كسر شفرات النصوص العربية والإنكليزية المشفرة بطريقة نابساك – ميركل هيلمن، وقد كان الهدف منها كسر شفرة نابساك باستخدام الخوارزمية الجينية بأسلوب يخرج عن كلاسيكيات البحوث التي نشرت في هذا المجال وتم التوصل الى ان النظام المقترح عن جدارة في كسر الشفرات المتولدة بطريقة نابساك والفضل في ذلك يعود الى استخدام دالة صلاحية جديدة إضافة الى رؤية جديدة في استخدام الخوارزمية الجينية ككل. وكان من اهم التوصيات استخدام طرائق أخرى لتنفيذ عمليات التزاوج والانتقاء والطفرة وقياس وقت البرنامج قبل التعديل وبعده لدراسة مدى تأثير هذه الطرائق على النظام لغرض اعتماد الخيارات الأفضل.

2. الدراسات التي تناولت تحليل جواز السفر

- **الدراسة السادس عشر (الجبور، 1987)**

قام الباحث بأجراء دراسة حول مكافحة تزوير الجوازات ووثائق السفر، وقد هدفت الى تحديد مفهوم التزوير في جواز السفر وذلك بالوقوف على محل التزوير في جوازات السفر، ووسائل تزوير جوازات السفر بوجه عام. وقد توصلت هذه الدراسة الى مجموعة من الاستنتاجات أهمها:

- اختيار مادة ورق جواز السفر وغلافه ضمن مواصفات امنية عالية ويمكن ان يتم ذلك بالاستعانة برأي التخصصات العلمية في حقل الكيمياء والتراكيب قبل تقرير تركيب المادة.
- تعبئة بيانات تحقيق الشخصية في جواز السفر الكترونيا وذلك للحصول على قاعدة بيانات يمكن الرجوع اليها مستقبلا.

اما من اهم التوصيات فقد كانت اتخاذ سبل الرقابة الإدارية التي تضمن عدم اصدار جوازات السفر بالسند لأوراق ثبوتية مزورة، فضلا عن أهمية التعامل مع التدقيق الالكتروني للمعلومات الشخصية والمتعلقة بجواز السفر.

- **الدراسة الثامن عشر (أسماء، 2017)**

قامت الباحثة بأجراء دراسة حول استثمار تكنلوجيا المعلومات والاتصالات في مجال الخدمة العمومية (ضعف الأداء المؤسسي كدافع للاستثمار) دراسة استكشافية حول مشروع جواز السفر البيوميتري في الجزائر. وقد هدفت الى تسليط الضوء عن التأثير الذي تلعبه تكنلوجيا المعلومات والاتصالات في تحسين الأداء المؤسسي في المؤسسات العمومية، والقاء الضوء على

اهم المفاهيم المرتبطة بتكنلوجيا المعلومات وبالأداء المؤسسي العمومي. وتوصلت هذه الدراسة الى مجموعة من الاستنتاجات أهمها:

من ناحية البنية التحتية للمؤسسة محل الدراسة، ان مستوى التجهيز بالمعلوماتية يعد في المستوى المطلوب.

ومن اهم التوصيات هي الحرص على تهيئة المناخ المناسب للاستفادة من تطبيقات تكنلوجيا المعلومات والاتصالات الجديدة بالدقة والسرعة المطلوبة.

- **الدراسة التاسع عشر (قارطي، مداوي 2017)**

قام الباحث بأجراء دراسة حول دراسة اثر استخدام الإدارة الالكترونية من طرف مصالح الخدمة العمومية في تحسين جودة الخدمة العمومية في الجزائر: دراسة استطلاعية حول خدمة استخراج جواز السفر البيوميتري، وقد هدفت هذه الدراسة الى التعرف على مدى تأثير استخدام الإدارة الالكترونية من مصالح الجماعات المحلية (البلدية والدائرة) في الجزائر في تحسين جودة خدمة استخراج جواز السفر البيوميتري بأبعادها الخمسة (الاعتمادية، الاستجابة، التوكيد، الملموسية، التعاطف) ولتحقيق هذا الهدف قامت الباحثتين بجمع المعلومات اللازمة عن طريق توزيع استبيان على عينة ميسرة من مجتمع الدراسة، وقد توصلت الدراسة إلى مجموعة من النتائج أهمها توجد علاقة طردية قوية بين كل استخدام الإدارة الالكترونية من طرف مصالح الإدارة العمومية وكل من بعد الاستجابة والملموسية والتعاطف لجودة خدمة استخراج جواز السفر البيومتري، أما بالنسبة لكل من بعدي الاعتمادية والتوكيد فهناك علاقة طردية ضعيفة بينهما وبين استخدام الإدارة الالكترونية من طرف مصالح الجماعات المحلية، ومن اهم التوصيات ضرورة رفع مستوى الإدارة الالكترونية لغرض تحسين جودة خدمة الجواز البيوميتري.

من خلال الدراسات السابقة التي تناولناها تبين ان اغلب الدراسات السابقة استخدمت الطرق الاحصائية التقليدية والكلاسيكية في عملية التجليل والتصنيف لبيانات مشكلة الدراسة ، اما الدراسة الحالية تعتمد على طرق احصائية حديثة محوسبة تعتمد على نمذجة البيانات واستخدام اسلوب محاكاة لتوظيف الخوارزمية الجينية والتي تعتبر اسلوب من اساليب الذكاء الصناعي للطرق الكلاسيكية.

1-9 هيكل الدراسة:

اشتملت الدراسة ما يلي:

1. الفصل الأول: (مقدمة) ويتضمن المقدمة و الاطار المنهجي والدراسات السابقة
2. الفصل الثاني: (مفاهيم اساسية) والمتمثل بمفهوم جواز السفر الالكتروني وتنقيب البيانات وعلاقته بعلم الإحصاء واسلوب المحاكاة.
3. الفصل الثالث: (الجانب النظري) تضمن الاطار النظري للطرق الاحصائية الكلاسيكية المستخدمة والانحدار اللوجستي والخوارزمية الجينية.
4. الفصل الرابع: (اسلوب المحاكاة) يتضمن نتائج الجانب التجريبي لطرق التحليل الاحصائي المستخدمة في الدراسة وتحديد افضل هذه الطرق بالاعتماد على معيار المفاضلة MSE وايضا تضمن نتائج الجانب العملي في التحليل والتصنيف.
5. الفصل الخامس: تضمن النتائج والتوصيات.

الفصل الثاني

مفاهيم أساسية

أولا: جواز السفر (مفاهيم عامة)

ثانيا: تنقيب البيانات

ثالثا: المحاكاة

2-1جواز السفر " مفاهيم عامة"

ان جواز السفر قد جاء ترجمة عن الكلمة الإنجليزية " Passport" وهي ذات اصل وتطور تاريخي، حيث جاءت نتيجة اتحاد كلمتين فرنسيتين هما "Passer" بمعنى يمر، و"Port" بمعنى ميناء او بوابة، وكان لهما استخدام في العصور القديمة عند المرور بالمدن الاوربية المحاطة بالأسوار والتي كانت تمر بها اعداد قليلة(Daniel C Turack 1972) وقد عرف الفقه الإنجليزي جواز السفر بانه " الطلب لدولة اجنبية بتقديم تسهيلات محددة لحاملها" Going Abroad P (5) ومثل هذا التعريف لا يتضمن المقومات الأساسية لجواز السفر، بل تناول جانبا من جوانب مجالات استخدامه.

وفي عام (1905) عرف اللورد الفرستون جواز السفر البريطاني" بانه وثيقة تصدر باسم السلطة وعلى مسؤولية وزير التاج الى شخص مسمى ليتقدم الى حكومات الدول الأجنبية ولتستخدم كحماية فردية كموضوع او اهتمام بريطاني في البلاد الأجنبية وتستند مشروعيتها الى الحقيقة ان المكتب الخارجي بالنسبة للوثيقة الرسمية يتحمل مسؤولية الشخص المسمى"، وقد بدأ هذا التعريف ملائما بعد أربعين سنة في قضية (Rex V Joyce) الا ان بعض الفقه شكك في ملائمته لمفهوم جواز السفر في الوقت الحاضر.

ومن اجل استظهار العناصر الجوهرية لجواز السفر نورد بعض التعريفات في التشريع والقضاء المقارن والتي من شأنها ان تبرز عنصرا او اكثر من عناصر جواز السفر او طبيعته او وظائفه، ففي كندا عرفت دائرة الشئون الخارجية (Canadas department of external affairs) جواز السفر بأنه" وثيقة تحقيق شخصية تصدر من قبل الدولة عادة لمواطنيها، والتي تطلب فيها من حكومة اجنبية منح حاملها ضمن اختصاصها السلامة والمرور الحر وتحقيق جميع الأهداف والغايات المشروعة وحمايته خلال وجوده ضمن اختصاصها، وبضمانة الدولة مصدرة الجواز لتمكين حامله من العودة اليها".

كما عرفت المحكمة العليا في الهند بانه" وثيقة مواطنة واثبات شخصية، تصدر من قبل الدولة لمن ينوون السفر الى دول أخرى، تطلب بموجبها من دولة أخرى منح حاملها كل مساعدة كموضع يهم الدولة مصدرة الجواز"، اما محكمة الاستئناف الدائمة في جنوب افريقيا فقد عرفت جواز السفر بأنه" وثيقة رسمية تختم وتوقع من قبل المعنيين في الدولة، ومن شأنها اثبات شخصية حاملها وتشهد بجنسيته ومواطنته، ويطلب بموجبها ممن يهمه الامر ان يسمح له بالمرور بدون عوائق وان يقدم له المساعدة والحماية التي قد يحتاج اليها"

وقد اضافت إدارة الجوازات في الولايات المتحدة عنصرا اخر لمقومات جواز السفر وهو عنصر الولاء للولايات المتحدة، حينما عرفت جواز السفر بانه" وثيقة لإثبات الشخصية والجنسية تصدر لشخص يحمل مشاعر الولاء للولايات المتحدة وينوي السفر او الإقامة المؤقتة في بلد اجنبي، ومن حق حامل هذه الوثيقة ان يتلقى الحماية والمعاملة الحسنه من الدبلوماسية الامريكية(Turack ;OP.cit,.p 17)

يتضح من التعريفات السابقة ان جواز السفر وثيقة رسمية مكتوبة تصدر عن الدولة للأغراض التالية:

- تمكين حاملها من مغادرة البلاد والعودة اليها.
- تتضمن اثبات جنسية وشخصية حاملها
- تتضمن بعض البيانات والمعلومات التي تتمتع بقوة الاثبات لتوثيقها من قبل الموظف العام المختص.

فالأصل فيمن يحمل جنسية دولة ان يتمتع بحق الإقامة الحرة داخل بلده وان يملك الحق في العودة اليها حين يغادرها الى أراضي دولة أخرى، وهذا الحق بالإضافة الى كونه حقا مقررا للأفراد فانه يختلط بحق الدولة الأخرى التي يدخلها ذلك الفرد للإقامة المؤقتة كأجنبي، الا ان بعض الدول تحضر على من تمنحهم جوازات سفرها الوطنية العودة اليها، فبريطانيا مثلا التي تمنح جوازات سفرها لبعض مواطني دول الكومنولث لا تسمح لهم بالإقامة في البلاد ولا بدخولها فيما اذا ابعدوا من دولة أخرى.

ومن جانب اخر يمكن تعريف جواز السفر الالكتروني (*Cancillería* 2005 2018 Council Regulation) بانه عبارة جواز سفر إلكتروني ورقي يحتوي على معلومات حيوية التي يمكن استخدامها لمصادقة هوية المسافرين. جواز السفر الإلكتروني يستخدم تقنية بطاقة ذكية، يشمل شريحة معالج دقيق وهوائي) لكل من الطاقة للشريحة والاتصالات) تكون في الغلاف الأمامي أو الخلفي أو صفحة في وسط جواز السفر الإلكتروني.

ان اتجاه العالم الى حوسبة المعلومات له مزايا عديدة منها سرعة الوصول الى المعلومات، استخدام اوفر للتخزين، ومن أهمها هي دراسة البيانات وتصنيفها احصائيا وتعتبر الحكومة الالكترونية هي اكبر مصدر للبيانات حول العالم حيث يتراوح كم البيانات الموجودة الى ترليون من البيانات والتي ترتبط فيما بينها، ولها أيضا قيود دولية تسهل عملية تبادل البيانات، وفي هذه الاطروحة تم التطرق الى الجواز السفر الالكتروني كوثيقة مهمة وتخضع للقيود الدولية، ويمكن تعريف الجواز الالكتروني بانه وثيقة رسمية للسفر، تقر من خلال الدولة التي

يتبعها مواطن ما، التي تعرف حاملها من حيث جنسيته وهويته طبقا للدولة التي ينتمي اليها، وتسمح هذه الوثيقة لحاملها بدخول والمرور خلال الدول الأخرى.

ان جواز السفر السوداني هي وثائق رسمية تصدر لرعايا جمهورية السودان لغرض السفر الدولي بجانب كونها دليلاً لإثبات الجنسية السودانية، تستخدم أيضاً لتسهيل الحصول على مساعدة من مسؤولو القنصليات السودانية في الخارج. تصدر الجوازات بشكل عام بمدة صلاحية 10 سنوات، ويجدد كل سنتين جميع الجوازات بعد تلك الفترة لا تكون قابلة للتجديد. ولذلك، يجب على المواطن التقدم بطلب للحصول على جواز سفر جديد مرة أخرى.

وقد أصدرت الحكومة السودانية في عام 2010 إصدارا إلكترونياً جديداً من الجواز وذلك للتماشي مع المعايير العالمية الجديدة ومتطلبات منظمة الطيران المدني الدولي. توفر جوازات السفر الجديدة اماناً أفضل وفعالية كبيرة في مكافحة التزوير، ويعمل على حد سواء مع الجواز التقليدي .

ان جوازات السفر السودانية ذات لون أخضر، عليها صقر الجديان مذهب في وسط الغلاف الأمامي. اسفل الشعار، توجد كلمة "جواز سفر" وبالإنجليزية .(Passport) يحتوي الجواز السوداني على 62 صفحة، ولأن اللغة العربية هي لغة مكتوبة من اليمين إلى اليسار، يبدأ فتح الجواز من اليمين إلى اليسار. اما الجوازات الالكترونية الجديدة، فهي ذات لون الأزرق داكن، بنفس التصميم الجواز الأخضر.

ومن أنواع جوازات السفر السودانية:

1. (جوازات سفر دبلوماسية،
2. (جوازات سفر خاصة،
3. (جوازات سفر لمهمة،
4. (جوازات سفر عادية،
5. (جوازات سفر تجارية.
6. (جوازات السفر للشئون الخارجية.

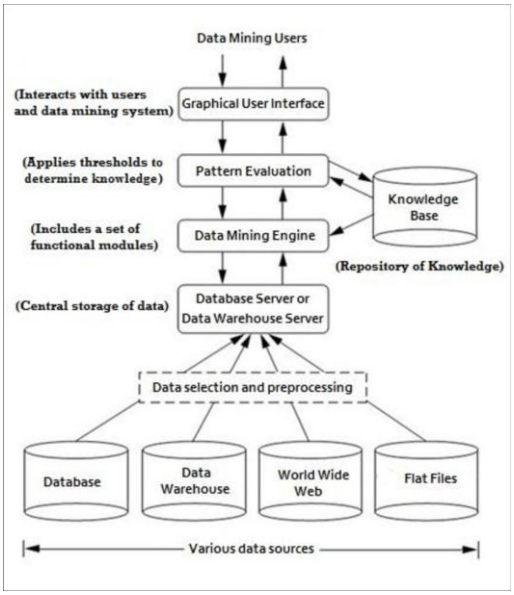
تصدر الجوازات العادية عن وزارة الداخلية (إدارة الجوازات والهجرة والجنسية)، أو من القنصليات، والسفارات السودانية خارج السودان .أما الجوازات الإلكترونية الجديدة، فهي ما يتم حاليًا إصدارها. ولا تتوافر معلومات حتى الآن عن مدى توافر الجوازات الجديدة خارج السودان.

تحتوي قاعدة البيانات للجواز السفر الالكتروني على 18 حقلا وتحتوي الحقول على قيود ادخال معينة باللغة العربية او الإنجليزية، ترجع الى اتفاقيات دولية بين كل بلدان العالم والتي تجيز استخدام الجواز الالكتروني.

2-2 تنقيب البيانات Data mining

التنقيب في البيانات هو عملية الكشف والعثور على معلومات ذات فائدة من خلال استعمال مجموعة من الأدوات المعقدة. بعض من هذه الادوات تشمل أدوات الإحصاء الاعتيادية والرسوم البيانية الحاسوبية .

فالتنقيب في البيانات منهجية تجمع بين نتائج الأبحاث في الذكاء الاصطناعي، الفهم الآلي، التعرف على الأشكال، قواعد المعلومات، الرياضيات الإحصائية، واجهات الاستعمال واللغة. وبحسب مراكز الأبحاث يمكن الفصل بين نوعين من استخراج المعلومات فتسمية تستعمل عند الباحثين في الذكاء (KDD) تستعمل عند الباحثين في الرياضيات (Data Mining) الاصطناعي والفهم الآلي وتسمية الإحصائية أو خبراء المعلومات . والشكل (الشكل (2-1) معمارية التنقيب في البيانات



الشكل (2-1) معمارية التنقيب في البيانات مصطفى، فؤاد عبيد (1995)**:**

2-2-1: مفاهيم عامة حول تنقيب البيانات

يتضمن هذا المبحث التعريف ببعض المصطلحات العامة حول تنقيب البيانات:-

1. **قاعدة البيانات (Data Base)) :** احسن، طيار وشلابي، عمار (2007)

هي مجموعة من الملفات أو الحقول أو السجلات تحتوي على مجموعة من البيانات أو المعلومات المرتبطة مع بعضها بعلاقات معينة .

2. تنقيب البيانات (Data Mining) مصطفى، فؤاد عبيد (1995)**:**

هي عملية استعمال أساليب وطرق إحصائية وتقنية جديدة لاستخراج المعلومات من كمية كبيرة من البيانات وبأساليب متعددة ومن هذه الأساليب شجرة القرار (Decision Tree) والانحدار (Regression) وقواعد الربط (Association Rules) وغيرها .

3. خوارزمية تنقيب البيانات (Data Mining Algorithm) :

هي طرق تقنية رياضية وإحصائية تحول الحالات (Cases) من البيانات الأصلية إلى نموذج تنقيب البيانات(Data Mining Model) بعدد من الخطوات المتسلسلة .

4. العناصر (Object): Ramachandran M, Pushpa . (2001)

يستعمل بمعناه الواسع في تمثيل الأشياء ، ويمكن أن يمثل كنقطة أو عنصر في الفضاء (Space) ورياضية يمثل العنصر بمتجه مثل x اذ أن:

$$X = (x_{i1},x_{i2},...,x_{in}). \quad (1\text{-}2)$$

وان X_{ij} : تمثل قيم عددية لمقاييس المسافة .

5. التنبؤ (Prediction)) الدوري، زكريا مطلب (2004)**:**

أي تقدير القيمة المستقبلية للبيانات ويعرف أحيانا تصنيف البيانات على أساس سلوكها المستقبلي ، وبما أن المتغيرات كمية فأن أدوات التنبؤ هي الانحدار بأنواعها وقواعد الربط والشبكات العصبية (Neural Networks).

6. مستودعات البيانات (Data Warehouse)) رحيمه، وليد عبد الله (1995)**:**

وهي مجموعة ضخمة من البيانات والتي جاءت من مصادر متنوعة متشابهة وغير متشابهة تم توحيدها وحفظها كي تساعد في اكتشاف المعرفة وبخطوات عدة.

7. التعليم الآلي 36(Machine Learning)) : Ramachandran M, Pushpa . (2001)"

هي القاعدة الأساسية التقنية في تنقيب البيانات. وتستعمل لانتزاع المعلومات الخام من قاعدة البيانات وتحويلها إلى صيغ مفهومة تعمل للأغراض المختلفة .

8. خادم لغة الاستفسار الهيكلية:

هو برنامج حاسوبي صمم للمساعدة في إدارة البيانات فهو يقدم طرائق لمسح وتحديث وإدخال البيانات التي يمكن من خلالها الحفاظ على قاعدة بيانات متجددة .

9. القيم المفقودة (Missing Values): Ramachandran M, Pushpa . (2001)"

هي تلك القيم التي لم يجر إدخالها في قاعدة البيانات أو غير معرفة .

10.دورة حياة البيانات (Life role of Data) : العلي، عبد القادر وقنديل، عامر إبراهيم والطمري، غسان 2006

هي المدة الزمنية التي تولد فيها البيانات ولحين الافادة منها ، بعد مرورها بمعالجات ، وتقسم إلى مراحل وخطوات متعددة .

11.العنقدة (Clustering): M.Halkidi & Y.Batistakis & M. Vazoirgiannis . (2001)

عبارة عن مجموعة من العناصر، ترتبط مع بعضها البعض طبقا لمجموعة من الخواص ، بحيث أن العنقود يحتوي العناصر الأكثر تجانسة (Homogenous) وفي نفس الوقت تكون العناصر بين العناقيد اقل تجانسة اي اكثر تباعدة .

12.المعالجة التحليلية الآنية (Wares Analytical Processing):
Ramachandran M, Pushpa . (2001)"

وهي برامج تستعمل لخزن المعلومات وتسمح للمستعملين باستكشاف المعلومات من البيانات من عدد من المصادر المختلفة من خلال التحليل الاتي والسريع للبيانات .

1. اكتشاف المعرفة في قواعد البيانات KDD: رحيمه، وليد عبد الله (1995)

وهي عملية استخلاص معلومات مخفية غير معروفة مسبقة ومفيدة من البيانات في قواعد البيانات (Data Base) .

2-2-2 مفهوم تنقيب البيانات

لابد هنا من التميز بين ثلاثة مصطلحات هامة هي البيانات Data والمعلومات Information والمعرفة Knowledge ، حيث يقصد بالأولى تلك البيانات المقترنة بالعمليات الوظيفية للمنظمة أي ما يتعلق بالأفراد والمواقع والصفقات وغيرها Wu, Jonathan 2000 ، اما المعلومات

فهي بيانات متماسكة تعطي رسالة ما (Seiner, Robert S 2002) في حين أن المعرفة تضم العامل البشري فهي استنتاج يرسم من المعلومات بعد ربطها مع معلومات اخرى ومقارنتها مع ما هو معروف سابقا"(Daft, Richard L,2001)

تجمع المنظمات مقدارا كبيرا من البيانات، لكن الذي تريده حقا هو المعلومات، لذلك فأن التقنية الأحدث والأسرع للتعامل مع كم البيانات هذا والإجابة على أسئلة العمل هي تنقيب البيانات

(DM) التي تشبه عملية تنقيب الذهب (Gold Mining) فتنقيب الذهب عملية غربلة مابين كميات من المعدن الخام لإيجاد كتلة من معدن ثمين ذي قيمة، وكذلك عملية تنقيب البيانات التي هي عملية غربلة ما بين كميات من البيانات لإيجاد المعلومات المفيدة لاتخاذ القرار، ويرجع سبب زيادة الاهتمام بتنقيب البيانات الى اهتمامها بالاكتشاف والتحليل من قواعد بيانات كبيرة جدا تحوي قيما" خفية فضلا عن مساعدتها في الكشف عن نماذج وقواعد ذات مغزى تساهم في تحسين فهم الزبون وحل العديد من مشاكل العمل.

ومن أجل فهم أفضل لتنقيب البيانات (Data Mining) لا بد من التمييز بينها وبين اكتشاف المعرفة في قواعد البيانات (KDD) فعرف (Fayyad) الأخيرة بأنها عملية غير عادية لتحديد نماذج صحيحة وجديدة وذات فائدة محتملة ويمكن فهمها بالنهاية اما تنقيب البيانات فهي أحد خطوات عملية اكتشاف المعرفة وتتكون من تطبيق خوارزميات تحليل واكتشاف البيانات التي تقدم عدد خاص من النماذج عبر البيانات.

ان مصطلح تنقيب البيانات ظهر في منتصف الستينات في الولايات المتحدة الأمريكية (احسن، طيار وشلابي، عمار (2007)) وهو مصطلح يجمع بين الإحصاء وتكنولوجيا المعلومات وقواعد البيانات (Data Bases) والذكاء الصناعي (Neural Intelligent) والتعليم الآلي (Machine Learning) ، وظهرت عدة تعريفات لمفهوم تنقيب البيانات (DM) منها ((الاستكشاف الآلي أو التقني لأنماط شائقة وغير جلية او مخفية في قاعدة البيانات)(العلاق، بشير عباس (2005)) .

وتسمى أحيانا (اكتشاف المعرفة) وهي عملية تحليل البيانات من منظورات مختلفة واستخلاص العلاقات فيما بينها وتلخيصها إلى معلومات مفيدة من خلال استكشاف قواعد جديدة من قواعد البيانات الكبيرة وكذلك اكتشاف نماذج (Model) جديدة وصولا إلى معلومات ذات قيمة وذلك باستعمال مجموعة من الأدوات والتقنيات الحديثة أو الخوارزميات وغيرها من أدوات الإحصاء الاعتيادية والرسوم البيانية . ومن خلال التعريفات السابقة يمكن القول بأن تنقيب البيانات (Data Mining) هي عملية الوصول إلى المعلومات المخبأة من قواعد البيانات (Data Base) ، وتسمى أحيانا اكتشاف المعرفة (Knowledge discovering from (Data - KDD

فضلا عن ذلك يمكن تعريفها بانها عملية بحث محوسب ويدوي عن معرفة من البيانات دون فرضيات مسبقة عما يمكن أن تكون هذه المعرفة. كما ويعرف التنقيب في البيانات على أنه عملية تحليل كمية بيانات (عادة ما تكون كمية كبيرة)، لإيجاد علاقة منطقية تلخص البيانات بطريقة جديدة تكون مفهومة ومفيدة لصاحب البيانات. يطلق اسم "نماذج models "على العلاقات والبيانات الملخصة التي يتم الحصول عليها من التنقيب في البيانات. يتعامل تنقيب البيانات عادة مع بيانات يكون قد تم الحصول عليها بغرض غير غرض التنقيب في البيانات (مثلاً قاعدة بيانات التعاملات في مصرف ما) مما يعني أن طريقة التنقيب في البيانات لاتؤثر مطلقاً على طريقة تجميع البيانات ذاتها. هذه هي أحد النواحي التي يختلف فيها التنقيب في البيانات عن الإحصاء، ولهذا يشار إلى عملية التنقيب في البيانات على أنها عملية إحصائية ثانوية. يشير التعريف أيضاً إلى أن كمية البيانات تكون عادة كبيرة، أما في حال كون كمية البيانات صغيرة فيفضل استخدام الطرق الإحصائية العادية في تحليلها.

عند التعامل مع حجم كبير من البيانات تظهر مسائل جديدة مثل كيفية تحديد النقاط المميزة في البيانات، وكيفية تحليل البيانات في فترة زمنية معقولة وكيفية قرار ما إذا كانت أي علاقة ظاهرية تعكس حقيقة في طبيعة البيانات. عادة يتم التنقيب في بيانات تكون جزءاً من كامل البيانات حيث يكون الغرض عادة تعميم النتائج على كامل البيانات (مثلاً تحليل البيانات الحالية لمستهلكي منتج ما بغرض توقع طلبات المستهلكين المستقبلية). من أحد أهداف تنقيب البيانات أيضاً هو اختزال كميات البيانات الكبيرة أو ضغطها بحيث تعبر بشكل بسيط عن كامل البيانات بدون تعميم.

3-2-2 اهمية تنقيب البيانات

أن تنقيب البيانات (DM) أسلوب تمكن من خلاله الوصول إلى المعلومات المخزونة في مستودع البيانات (Data Warehouse) ويتضمن استعمال التحليل الإحصائي (StatisticalAnalysis) لاكتشاف العلاقات الخفية في البيانات Romny,Marshall B(2000)"Accounting information system"9th USA ، كما يعتبر تنقيب البيانات أحد تكنولوجيا الذكاء الاصطناعي (Artificial Intelligent) فضلا عن الأنظمة الخبيرة والشبكات العصبية، ويهدف إلى تمكين النشاط أو المنظمة من الاستغلال الأمثل لبياناتها فهي تحاول ايجاد المعلومات في مجاميع البيانات الكبيرة التي لا تعلم المنظمة أو النشاط بوجودها ، كذلك ايجاد العلاقات وعمل التنبؤات .

ويتضح من ذلك ان اهمية تنقيب البيانات ناجمة عن عملية استكشاف وتحليل كميات كبيرة من البيانات لغرض الحصول على علاقات ونماذج خفية تساعد في استخلاص المعلومات المفيدة والساندة لاتخاذ قرارات عمل استراتيجية كفيلة بزيادة اداء المنظمة او النشاط.

4-2-2 اهداف تنقيب البيانات

هناك ثلاثة أهداف للتنقيب في البيانات:

1. من أجل تعليل بعض الظواهر المرئية.
2. من أجل التثبت من نظرية ما. مثال: التثبت من النظرية التي تقول بأن الأسر الكبيرة تهتم بالضمان الصحي أكثر من الأسر الصغيرة عددا.
3. من أجل تحليل البيانات للحصول على علاقات جديدة وغير متوقعة. مثال: كيف سيكون الانفاق العام إن كان ملازما لعمليات خداع واسعة من قبل البطاقات الائتمانية.

5-2-2 وسائل التنقيب في البيانات

هناك عدة وسائل مختلفة من أجل التنقيب في البيانات. اختيار الوسيلة المناسبة يعتمد على طبيعة البيانات تحت الدراسة وعلى حجمها. يمكن اجراء عملية التنقيب في البيانات بالمقارنة مع سوق البيانات ومخزن البيانات.

بعض من هذه الوسائل هي:

1- الاستدلال المبني على حالات سابقة :Case-Based reasoning التفكير واستخلاص النتائج والقوانين من أمثلة حية و قضايا تم حلها سابقا.
2- الكشف عن قانون :Rule Discovery البحث عن منوال معين أو علاقة معينة في جزئية كبيرة من البيانات.
3- معالجة الإشارات :Signal Processing ايجاد الظواهر المتشابه مع بعضها البعض
4- شبكات عصبونية :Neural Nets تطوير نماذج قابلة لتنبؤ النتائج. هذه النماذج تم تطويرها بناء على أسس تم استنباطها من عقل الإنسان.
5- منحنيات غير ثابتة :Fractals تصغير البيانات الكبيرة من دون ضياع المعلومات

6-2-2 مراحل عملية التنقيب في البيانات

يمكن تلخيص مراحل و خطوات عملية التنقيب في البيانات كما يلي: عبد الستار العلي، عامر إبراهيم (قنديلجي، غسان العمري، 2006، ص.)

1. فهم طبيعة الأعمال (Business Understanding): يعتبر المطلب الأول لاكتشاف المعرفة هو فهم المشاكل و المسائل التي التي تواجهها الأعمال. و بمعنى أخر، كيف يمكن تحقيق المنفعة الأعظم من التنقيب في البيانات، مما يتطلب وجود صيغة واضحة و محددة لأهداف الأعمال.
2. فهم البيانات (Data Understanding) : تعتبر مسالة معرفة ماهية وطبيعة البيانات عامل مهم في نجاح عملية التنقيب في البيانات و اكتشاف المعرفة. حيث أن معرفة البيانات بصورة جيدة تعني مساعدة المصممين على استخدام الخوارزميات أو الأدوات المستخدمة للمسائل المحددة بدقة عالية. و هذا يقود إلى تعظيم فرص النجاح بالإضافة إلى رفع الفعالية و الكفاءة لنظام اكتشاف المعرفة.

و لا تحتاج عملية التنقيب في البيانات إلى تجميع البيانات في مستودع البيانات، أما إذا كان مستودع البيانات موجود في المنظمة، فمن الأفضل عدم احتكار المستودع بشكل مباشر لغرض التنقيب في البيانات.

و يمكن تلخيص الخطوات الضرورية لعملية فهم البيانات كالآتي:

- تجميع البيانات (Data Collection): و هي الخطوة الموجهة نحو تحديد مصدر البيانات في الدراسة بما في ذلك استخدام البيانات العامة الخارجية مثل الضرائب و غيرها.
- توصيف البيانات (Data Description): و هي الخطوة التي تركز على توصيف محتويات الملف الواحد من الملفات أو الجداول.
- جودة البيانات و تحقيقها (Data Quality and Verification): هذه الخطوة تحدد ما إذا كان تقليل أو إهمال بعض البيانات غير الضرورية أو كونها رديئة الجودة و قد لا تنفع في الدراسة. لأن النموذج الجيد يحتاج إلى بيانات جيدة مما يتوجب أن تكون البيانات صحيحة و ذات مضمون دقيق.
- التحليل الاسترشادي للبيانات (Exploratory Analysis of Data): تستخدم الأساليب مثل الإظهار المرئي أو التصور أو عملية التحليل المباشر(OLAP) التي تؤدي إلى إجراء التحليل الأولي للبيانات. و تعتبر هذه الخطوة مهمة و ضرورية لأنها تركز على تطوير الفرضيات المتعلقة بالمشكلة قيد الدراسة.

3. تهيئة البيانات (Data Preparation): و تشمل الخطوات التالية:

- الاختيار (Selection) و تعني اختيار المتغيرات المتوقعة و حجم العينة.
- صياغة المتغيرات و تحويلها(Construction and Transformation Variables) حيث يجب دائما أن تصاغ المتغيرات الجديدة لبناء النماذج الفعالة.
- تكامل البيانات (Data Integration): حيث أن مجاميع البيانات في دراسة التنقيب عن البيانات من الممكن خزنها في قواعد بيانات متعددة الأغراض التي تكون بحاجة إلى توحيدها في قاعدة بيانية واحدة.
- تصميم و تنسيق البيانات (Data Formatting) حيث تتعلق هذه الخطوة في إعادة ترتيب حقول البيانات كما يتطلب في نموذج التنقيب في البيانات.

4. صياغة نماذج الحل و ثبوتها(Model Building and Validation): إن بناء وصياغة نموذج الحل السليم و الدقيق يتم من خلال عملية الخطأ و الصواب، حيث كثيرا ما تحتاج مثل هذه العملية إلى مساعدة المختصين في التنقيب عن البيانات بهدف اختبار و فحص مختلف البدائل للحصول على أفضل نموذج لحل المشكلة قيد الدراسة.

5. التقييم و تعليل نتائج النموذج (Evaluation and Interpretation): حالما يتم صياغة النموذج و التحقق من ثباته و صدقه، تجري مباشرة عملية التحقق من ثبات حزمة البيانات التي يتم تغذيتها بواسطة النموذج. وبما أن نتاجات هذه البيانات معروفة، لذا فان النتائج المتوقعة تقارن مع النتائج الفعلية في ثبات حزمة البيانات قيد التشغيل. و تؤدي هذه المقارنة أو المفاضلة إلى التحقق من دقة النموذج.

6. نشر و توزيع النموذج (Model Deployment): حيث تشتمل هذه الخطوة على نشر و توزيع النموذج داخل المنظمة لمساعدة عملية صنع القرار. و أن النموذج الصالح يجب أيضا أن يحقق الرضا لدى المستفيدين طالما أن اختيار النموذج لابد أن يتم من خلال الدراسة الاسترشادية أو نموذج مصغر من الدراسة الشاملة.

7-2-2 تطبيقات التنقيب في البيانات

وسائل التنقيب في البيانات تُستعمل وبنجاح في الكثير من التطبيقات الحقيقة حول العالم.

التطبيقات التالية تشمل بعضا من الأمثلة: (يتبعها مثال لكل تطبيق)

-كتابة تقرير مختصر عن فئة معينة :Profiling Populations تطوير وإنشاء تقارير موجزة عن الزبائن المهمين وعن بطاقات الائتمان.

-تحليل النزعة التجارية (Analysis of Business Trend) ايجاد الأسواق ذات قدرات النمو القوية أو الضعيفة.

-التسويق لفئة معين (Target Marketing) ايجاد الزبائن من أجل منح التخفيضات لهم لسبب معين.

-تحليل الاستعمال :Usage Analysis ايجاد منوال معين لاستعمال الخدمات والسلع

-فعالية الحملة :Campaign Effectiveness مقارنة استراتيجيات الحملات مع بعضها البعض من أجل ايجاد أكثرها فعالية وتأثيرا.

-جاذبية السلعة: ايجاد السلع التي تباع مع بعضها البعض.

تطبيقات التنقيب في البيانات بدأت تنمو بصورة كبيرة للأسباب التالية:

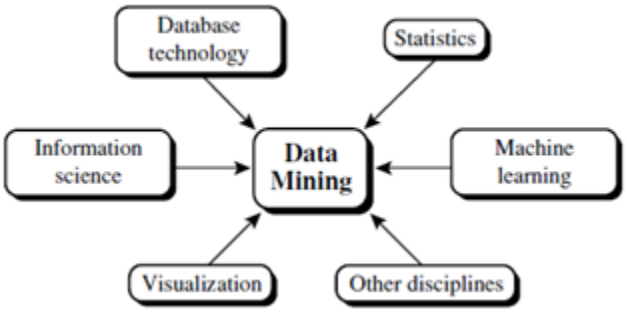
1- كمية البيانات الموجودة في مخزن البيانات وسوق البيانات تنمو بصورة أسية (exponential).

ومن أجل ذلك، فإن المستخدم يحتاج إلى أدوات متطورة من مثل التنقيب في البيانات من اجل استخلاص الفائدة والمعرفة من هذه البيانات.

2- الكثير من أدوات التنقيب عن البيانات بدأت تظهر مؤخرا، وكل أداة أفضل من الأخرى.

3- المنافسة الشديدة الموجودة في السوق تدفع الشركات إلى الاستفادة القصوى من البيانات التي بيدها. عمليات التنقيب في البيانات تفعل ذلك تماما.

الشكل (2-2) يوضح عددا من المجالات التي تستخدم التنقيب في البيانات.



الشكل (2-2) علاقة التنقيب في البيانات مع مجالات أخرى مصطفى، فؤاد عبيد (1995)**:**

وتطبيقات التنقيب في البيانات بدأت تنمو بصورة كبيرة للأسباب الآتية: 1

١) كمية البيانات الموجودة في مخزن البيانات وسوق البيانات تنمو بصورة كبيرة. ومن أجل ذلك فإن المستخدم يحتاج إلى أدوات متطورة من مثل التنقيب في البيانات من اجل استخلاص الفائدة والمعرفة من هذه البيانات.

٢) الكثير من أدوات التنقيب عن البيانات بدأ يظهر مؤخرا، وكل أداة أفضل من الأخرى.

٣) المنافسة الشديدة الموجودة في السوق تدفع الشركات إلى الاستفادة القصوى من البيانات التي بيدها، وعمليات التنقيب في البيانات توفر ذلك تماما.

2-2-8 علاقة تنقيب البيانات (DM بعلم الإحصاء

نشأ التنقيب في البيانات (DM) بمساهمة عدة تخصصات من بينها الإحصاء ، قواعد البيانات، الذكاء الصناعي وقد أثبت وجوده كأحد التخصصات الجديدة مع علوم الحاسوب التحليل الكميات الكبيرة والمعقدة من البيانات والتي لم تعد وسائل التحليل التقليدية قادرة على التعامل معها ، ومن الشكل (2-3) يتبين دور الاحصاء بتنقيب البيانات وعلاقته بالعلوم الأخرى.

الشكل (2-3) علاقة تنقيب البيانات بعلم الإحصاء رحيمه، وليد عبد الله (1995)**:**

2-3 المحاكاة (Simulation)

تعتبر المحاكاة من الوسائل المهمة في حل المشاكل التي يصعب حلها بالطرائق التحليلية او العددية ويعتمد اسلوب المحاكاة على توليد ارقام عشوائية.

ويمكن تعريف المحاكاة على انها تقليد سلوك نظام حقيقي من خلال بناء أنموذج بسيط يمثل هذا النظام ثم اختباره، يمثل أنموذج المحاكاة سلوك النظام المطلوب تقليده من خلال دراسة التفاعلات الحاصلة بين مكونات ذلك النظام وتمثل مخرجات الأنموذج من خلال مقياس ليبين اداء ذلك النظام.

ويمكننا القول بان المحاكاة هي عملية تقليد لظاهرة حقيقية عن طريق وجود مجموعة من المعادلات الحسابية الرياضية، نظرياً اي ظاهرة يمكن اختصارها على هيئة أو تحويلها الى بيانات حسابية أو معادلات يمكن محاكاتها، وكذلك يعبر عن المحاكاة على انها طريقة رياضية تعمل على ايجاد بيانات مشابهة لبيانات حقيقية من اجل حل المشاكل التي تقع ضمن إطار البحث والتي تبرز عند توليد احجام مختلفة بمعلمات مختلفة بنماذج متعددة من البيانات، اذ يتم تصميم عينة من المجتمع النظري لتمثيل الظاهرة بدلاً من المجتمع الحقيقي.

وعلى الرغم من ذلك فان عملية المحاكاة بحد ذاتها تعتبر صعبة للغاية وخصوصا فيما يتعلق بالظاهر الطبيعية التي يحيط بها عديد من العوامل، وغالبا تكون المشاكل صعبة ومعقدة ليس من السهولة تحليلها من خلال الاثبات الرياضي للنظرية الاحصائية، و بالإضافة الى عملية التقليد التي تقوم بها المحاكاة فأنها تستعمل ايضاً لاختبار مختلف النظريات الجديدة، وللمحاكاة مرونة وحرية عالية في اختيار حجوم العينات العشوائية التي تمثل مجتمع الدراسة افضل تمثيل وكذلك القدرة العالية على التنوع بتنوع الاخطاء العشوائية ونتيجة لما تقدم فننا نلجأ الى تحويل هذه النظريات الى بيانات افتراضية تحاكي الواقع الحقيقي ومن ثم اختيار عدد من العينات العشوائية منها ليتم التوصل الى افضل طريقة لحل هذه المشاكل بأقل كلفة واقصر وقت ممكن، لذلك فان المحاكاة تعمل على اختزال الجهد والوقت والتكلفة.

وقد تم كتابة برنامج المحاكاة باستعمال برمجة (MATLAB R2017B) ويعد هذا البرنامج من البرامج ذات القدرة الكبيرة في المجال البرمجي والرياضي.

الفصل الثالث

الجانب النظري (الطرق الإحصائية المستخدمة)

أولا: الاطار النظري للتحليل العنقودي

ثانيا: مفهوم المركبات الرئيسية

ثالثا: طريقة الإمكان الأعظم في التقدير

رابعا: طريقة بيز في التقدير

خامسا: الخوارزمية الجينية (مفاهيم عامة)

سادسا: الانحدار اللوجيستي

1-3الاطار النظري للتحليل العنقودي

ان العنقدة او التحليل العنقودي هي طريقة النموذجية لتجميع نقاط البيانات (العناصر) ضمن محيط التصنيف غير الموجه ، فيتم تقسيم مجموعة من البيانات الى عدد من المجاميع الجزئية او العناقيد بالاعتماد على تشابه العناصر، حيث تملك العناصر داخل العنقود الواحد درجة عالية من التشابه بينما تملك العناصر المنتمية الى عناقيد اخرى ومختلفة درجة عالية من عدم التشابه، وتتم عملية تصنيف العناصر الى عناقيد على اساس المقاييس الموضوعة على هذه العناصر، فمثلا اذا توفرت لدينا مجموعة من البيانات $[X= x_1,.....x_n]$ التي توضع في موجه خاص العناصر المنتمي الى فضاء العينة (R^n) فان الهدف هو تجزئة او تحديد المجاميع الجزئية او العناقيد لعناصر متشابهة على اساس مجموعة متجهات خواص العناصر $[X= x_1,.....X_n=x_n]$ تتضمن العنقدة عدة خطوات محددة منها تحديد المسافة الملائمة يهدف التحليل العنقودي الى تصنيف عينة المشاهدات الى فئتين متنافيتين او أكثر بالاعتماد على تشكيلات من فئات المتغيرات وذلك لغرض اكتشاف صفات مشتركة تنظم المشاهدات (الافراد) وتقسمها الى مجاميع تتمتع بنفس الخواص. وبفرض لدينا مصفوفة المشاهدات التالية (**Alvin c.Rencher,2002)**:

$$Y = \begin{pmatrix} y_1' \\ y_2' \\ . \\ . \\ . \\ y_n' \end{pmatrix} = (y_{(1)}, y_{(2)}, \ldots\ldots\ldots, y_{(p)}) \qquad (1\text{-}3)$$

بحيث ان :-

y_i' : يمثل متجه المشاهدات

$y_{(j)}$: يمثل متجه المتغيرات

فان الغرض من التحليل هو اكتشاف نمط معين ينظم المشاهدات المبينة في المصفوفة (1) أعلاه، والتي تتمتع عناصرها بخواص مشتركة، تمكننا من اجراء تنسيق هذه المشاهدات في مجموعات معينة.

اذ تتمثل مراحل اجراء التحليل العنقودي (Cluster analysis) بمجموعة من الخطوات تتضمن ايجاد ما يلي:

اولا: مصفوفة القرابة (The proximity between objects)

وهي مصفوفة متماثلة عدد صفوفها مساويا الى عدد اعمدتها. حيث يبدأ التحليل العنقودي عادة بتكوين تلك المصفوفة والتي تمثل عناصرها أحد مقاييس المسافة بين المشاهدات، والفكرة تتلخص بربط الوحدات التي تتشابه مع بعضها في مجموعات منفصلة. بحيث يمكن تمثيل الشكل العام لتلك المصفوفة كما يلي: - (Wolfgang Hardle(2003) (ch.11

$$D=\begin{bmatrix} d_{11} & d_{12} & . & . & . & d_{1n} \\ . & d_{22} & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ d_{n1} & d_{n2} & . & . & . & d_{nn} \end{bmatrix} \qquad (2\text{-}3)$$

بحيث ان:

D: مصفوفة القرابة ذات بعد $(n\times p)$

n:- تمثل المشاهدات

P:- تمثل المتغيرات

وان عناصر المصفوفة تمثل المسافة بين المشاهدات أو المتغيرات وتقاس تلك المسافة بأحد الصيغ التالية:-

1. باستخدام مربع المسافة الاقليدية (Euclidean distance square)

ان مربع المسافة الاقليدية بالنسبة للمتغيرين (x,y) يمكن كتابته وفق الصيغة التالية:

$$d^2(x,y)=(x-y)'(x-y)=\sum_{j=1}^{p}(x_j-y_j)^2 \qquad \text{----------(3-3)}$$

ويمكن اعادة كتابة الصيغة (3-3) اعلاه، وذلك كما يلي:-

$$d^2(x,y)=(v_x-v_y)^2+p(\bar{x}-\bar{y})^2+2v_xv_y(1-r_{xy}) \qquad \text{----------(3-4)}$$

بحيث ان: -

$$v_x^2 = \sum_{i-1}^{p}(x_i - \overline{x})^2$$

$$\overline{x} = \sum_{j=1}^{p} x_j / p$$

$$r_{xy} = \frac{\sum_{j=1}^{p}(x_j - \overline{x})(y_j - \overline{y})}{\sqrt{\sum_{j=1}^{p}(x_j - \overline{x})^2 \sum_{j=1}^{p}(y_j - \overline{y})^2}}$$

2. باستخدام المسافة الاقليدية (Euclidean distance)

ان المسافة الاقليدية بالنسبة للمتغيرين (x,y) يمكن كتابته وفق الصيغة التالية:

$$d(x,y) = \sqrt{\sum_{j=1}^{p}(x_j - y_j)^2} \quad \text{----------(3-5)}$$

3. باستخدام مسافة القطاع (City Block distance)

ان مسافة القطاع بالنسبة للمتغيرين (x,y) يمكن كتابته وفق الصيغة التالية:

$$d(x,y) = \sum_{j=`}^{p}\left|x_j - y_j\right| \quad \text{----------(3-6)}$$

ثانيا: توزيع المفردات في مجموعات (Agglomeration Schedule)

وهي المرحلة الثانية من التحليل العنقودي والتي تلي حساب المسافة بين المفردات حيث يتم توزيع المفردات في مجموعات وذلك باستخدام: -

أ. الطريقة الهرمية (Hierarchical cluster analysis)

تستخدم الطريقة الهرمية لإيجاد مجاميع كفؤة للمفردات وتمتاز بانها لا تتطلب معرفة مسبقة بعدد المجموعات حيث يتم وفق هذه الطريقة توزيع المفردات في مجموعات وفق اسلوبين: -

1 اسلوب الخلاف (The divisive technique)

يطبق هذا الاسلوب بافتراض وجود مجموعة واحدة للبيانات يتم تقسيمها الى مجموعات جزئية، وهذه المجموعات الجزئية يتم تقسيمها ايضا الى مجموعات جزئية اصغر وتستمر الى ان نحصل لكل مفردة مجموعة جزئية خاصة بها.

2 أسلوب التكتل (Agglomerative technique)

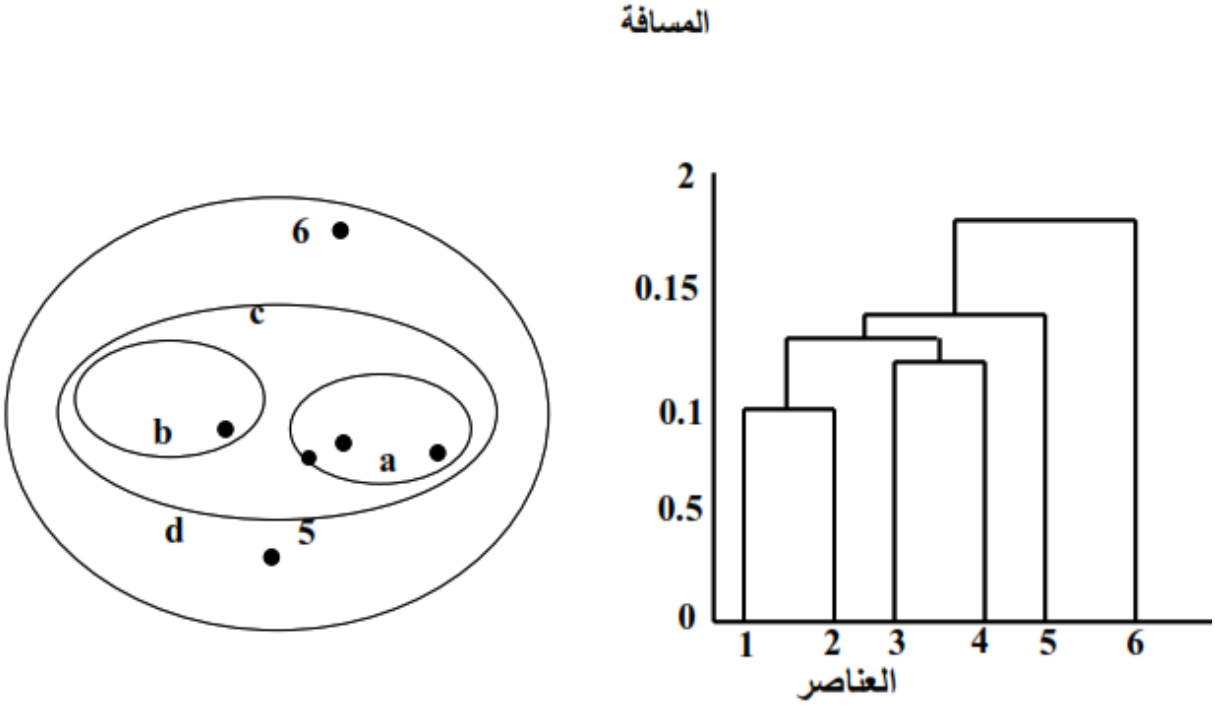
يطبق هذا الاسلوب بافتراض ان كل مفردة تصف مجموعة جزئية خاصة بها. ثم يتم تجميع المجموعات الجزئية المتشابهة في مجموعات جزئية أكثر شمولا وتتكرر هذه العملية عدة مرات حتى نصل الى مجموعة جزئية واحدة تكون شاملة لجميع البيانات. ولأجل حساب المسافة بين المجموعات الجزئية فهناك عدة طرق اهمها :- (Alvin c.Rencher(2002) ch.14)

- طريقة الجار الاقرب (الربط الاحادي) Single linkage (Nearest neighbor)

يتم باستخدام هذه الطريقة ايجاد أصغر مسافة لكل زوج من المجاميع ودمجهما معا، وذلك وفق الصيغة التالية: -

$$D(A,B)=\min\{d(y_i,y_j)\} \quad \text{----------(3-7)}$$

ويطلق على هذا الاسلوب بطريقة الربط المفرد، لانها تبدأ مع كل النقاط كعناقيد مفردة ومن ثم يتم اضافة الترابطات الاقوى بين النقاط لتجميع العناصر وتشكيل العناقيد، وتقود العناقيد الناتجة الى سلسلة طويلة من الترابطات كما في الشكل التالي:

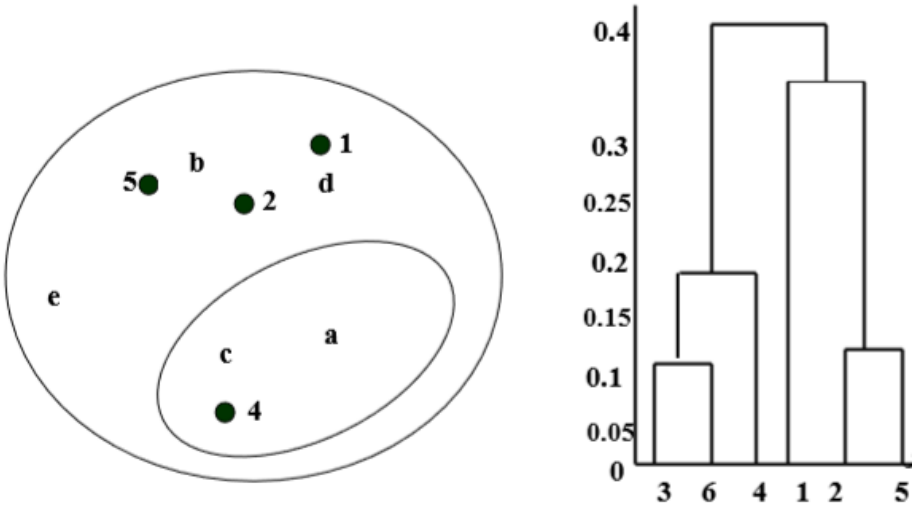


الشكل (3-1) طريقة الجار الأقرب (الربط الأحادي) (Wolfgang Hardle(2003) ch.11)

- طريقة الجار الابعد (الربط التام) Complete linkage (Farthest neighbor)

يتم باستخدام هذه الطريقة ايجاد اكبر مسافة لكل زوج من المجاميع ودمجهما معا، وذلك وفق الصيغة التالية:-

$$D(A,B)=\max\{d(y_i,y_j)\} \quad \text{----------(3-8)}$$

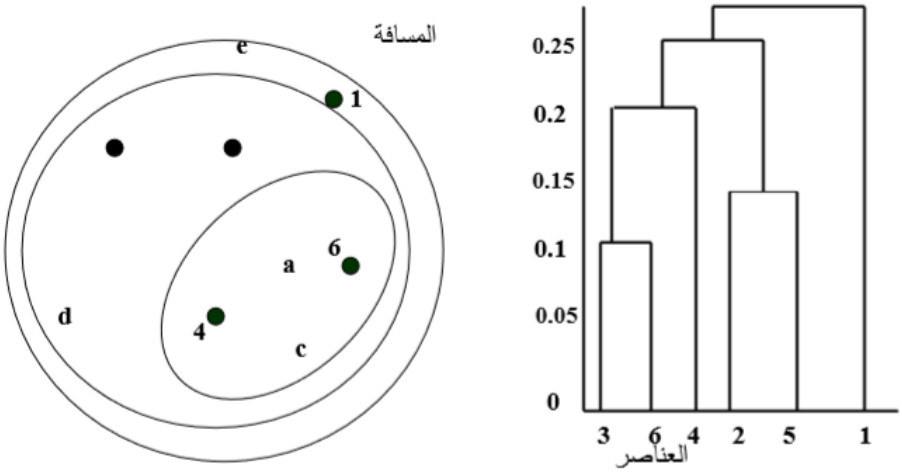


الشكل (3-2) طريقة الجار الأقرب (الربط التام)

- طريقة الربط بالاعتماد على المعدل (Average linkage) (Wolfgang Hardle(2003) ch.11)

يتم باستخدام هذه الطريقة دمج مجموعتين وذلك بالاعتماد على معدل المسافة بين نقطة من المجموعة الاولى ونقطة من المجموعة الثانية وذلك وفق الصيغة التالية: (Wolfgang Hardle(2007) p.205

$$D(A,B)=\frac{1}{n_A n_B}\sum_{i=1}^{n_A}\sum_{j=1}^{n_B} d(y_i,y_j) \quad \text{----------(3-9)}$$



الشكل (3-3) طريقة الربط بالاعتماد على المعدل (Wolfgang Hardle(2003) ch.11)

- طريقة المركز Centroid

وتتلخص هذه الطريقة بحساب المتوسط العام عن طريق جمع حاصل ضرب متوسط كل مجموعة بعدد مفرداتها وقسمتها على عدد المفردات الكلي، وكما يلي:

$$D(A,B) = d(\bar{y}_A, \bar{y}_B)$$ ----------(3-10)

بحيث ان :-

$$\bar{\mathbf{y}}_A = \sum_{i=1}^{n_A} \mathbf{y}_i / n_A$$

$$\bar{\mathbf{y}}_{AB} = \frac{n_A \bar{\mathbf{y}}_A + n_B \bar{\mathbf{y}}_B}{n_A + n_B}$$ ----------(3-11)

- طريقة وورلد (Ward's method)

تفضل هذه الطريقة على بقية الطرائق الهرمية التجميعية السابقة التي اقترحت من قبل Ward عام 1963 والتي يطلق عليها في بعض الاحيان بطريقة اصغر تباين minimum variance) method) لأنها تستخدم اسلوب تحليل التباين لحساب المسافات بين العناقيد والمعطاة حسب الصيغة التالية:

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

حيث ان :

d_{ij}^2: تمثل المسافة بين العنقود i والعنقود j و n,m يمثلان عدد العناصر في العنقودين I,j

وتحاول هذه الطريقة العمل على حساب مجموع مربعات الخطأ بين كل زوج من العناصر ومن ثم على ربط الزوجين الذي يعطي اصغر مجموع مربعات خطأ (SSE) والذي يتم حسابه وفق الصيغة التالية:

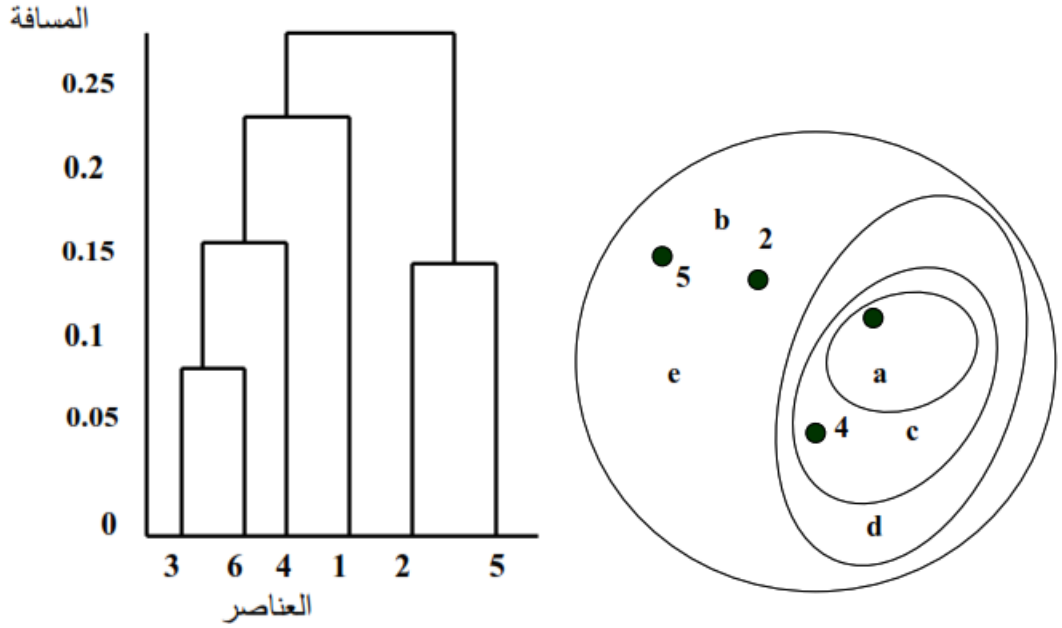
$$SSE = \sum_{j=1}^{n} \left[\sum_{i=1}^{n_j} x_{ij}^2 - \frac{1}{n_j} \left(\sum_{i=1}^{n_j} x_{ij} \right)^2 \right]$$

حيث ان:

n: تمثل عدد العناصر الكلية

n_j: عدد العناصر في العنقود j

ويعرف التقارب بين عنقودين بطريقة Ward على انه الزيادة في مربع الخطأ الناتجة عند ادماج عنقودين لذلك تستخدم هذه الطريقة نفس دالة الهدف المستخدمة في عنقدة C means وباجراء بعض الخطوات البسيطة نلاحظ ان هذه الطريقة تشابه طريقة معدل زوج المجموعة عندما يكون التقارب بين نقطتين ليمثل مربع المسافة بينهما كما موضح بالشكل التالي:



الشكل (3-4) طريقة وورلد لربط العنقدة ward لمجموعة من العناصر (1,2,3,4,5,6) ضمن العناقيد a,b,c,d,e (Wolfgang Hardle(2003) ch.11)

وبمعنى اخر ان هذه الطريقة تعتمد في عملية الدمج بين المجاميع على مربع المسافة بداخل المجاميع والتي يمكن ايجادها لربط مجموعتين مثل (A,B) كما يلي :-

$$SSE_A = \sum_{i=1}^{n_A} (\mathbf{y}_i - \overline{\mathbf{y}}_A)'(\mathbf{y}_i - \overline{\mathbf{y}}_A)$$

$$SSE_B = \sum_{i=1}^{n_B} (\mathbf{y}_i - \overline{\mathbf{y}}_B)'(\mathbf{y}_i - \overline{\mathbf{y}}_B)$$

$$SSE_{AB} = \sum_{i=1}^{n_{AB}} (\mathbf{y}_i - \overline{\mathbf{y}}_{AB})'(\mathbf{y}_i - \overline{\mathbf{y}}_{AB})$$

بحيث ان :-

n_a: يمثل عدد المفردات في المجموعة الأولى

n_b : يمثل عدد المفردات في المجموعة الثانية

$n_{ab}= n_a+ n_b$

$$\bar{\mathbf{y}}_{AB} = (n_A\bar{\mathbf{y}}_A + n_B\bar{\mathbf{y}}_B)/(n_A + n_B)$$

وان SSE_{AB}،SSE_B،SSE_A تمثل مجموع المربعات بداخل المجموعة (A,B,AB) على التوالي وان هذه الطريقة تعتمد في الاساس على عملية الدمج للمجموعتين (A,B) اللتان تمتلكان اصغر SSE والمعرف كما يلي :-

$$I_{AB} = \mathrm{SSE}_{AB} - (\mathrm{SSE}_A + \mathrm{SSE}_B).$$

ب. طريقة المتوسطات (K-means cluster analysis)

ان اول من استخدم مصطلح k-means كان جيمس ماكوين في عام 1967 بيد ان الفكرة خلف هذا المصطلح ترجع الى هوجو شتاين هاوس في عام 1957، والتطبيق الكلاسيكي للخوارزمية تم اقتراحه من قبل ستيوارت لويد في عام 1957 كتقنية تطبيقية للتضمين الرقمي، الا ان النشر الاول لم يتم حتى عام 1982 وفي عام 1965 نشر E.W.Forgy نفس الطريقة ولهذا يتم تسمية الخوارزمية احيانا على اسمه، تطوير اضافي على الخوارزمية تم نشرها من قبل هارتيجان و وونج في 1979 و 1975.

ان الخوارزمية التصنيفية (k- means clustering) هي طريقة لتكميم المتجهات في الاصل في علم معالجة الاشارة، والتي اشتهر في استخدامها في تطبيقات التصنيف خلال عملية التنقيب في البيانات، الهدف من هذه الخوارزمية هو تقسيم عدد من العناصر (بيانات n) الى عدد k من الاقسام والتي فيها ينضوي كل عنصر الى القسم ذي النقطة المركزية الاقرب (المتوسط) حيث تمثل النقطة المركزية الاساس الذي يتم عليه تقسيم البيانـات وتصـنيفها ولهذا اتت التسمية (k-means clustering) تتطلب هذه الطريقة معرفة مسبقة بعدد المجموعات لذلك فانه يمكن توزيع المفردات عليها بطريقة مستوية. وبهذه الطريقة يتم توزيع المفردات على المجموعات على أساس قاعدة معينة كخطوة أولى ويتم حساب متوسطات هذه المجموعات ثم يتم إعادة خلط هذه المفردات وإعادة توزيعها من جديد على أساس متوسطات المجموعات كخطوة ثانية وفي

نهاية هذه الخطوة يتم حساب متوسطات المجموعات ويتم تكرار هذه العملية لحين استقرار المفردات في مجموعات معينة بحيث لا تتغير المجموعة الخاصة باي مفردة.

1- وصف الخوارزمية التصنيفية:

نأخذ العناصر المعطاة ($x_1,\ldots\ldots x_n$) حيث كل عنصر يمثل متجها بعده d بعد تطبيق الخوارزمية على العناصر يتم تقسيمها حسب التشابه بينها الى عدد k من الاجزاء$s:s=\{s_1\ldots\ldots s_k\}$ بحيث ان k هي اصغر او مساوية الى n بحيث يؤخذ القيمة الدنيا لمجموع التربيع بين كل عنصر وبين النقاط المركزية والتي عددها k حيث ان العلاقة الرياضية تعطى بالصيغة التالية:

$$\mathrm{Arg_s}\ \min \sum\sum \|x-\mu_i\|^2$$

حيث:

μ_i : هي متوسط العناصر في الجزء s_i

وبشكل عام يمكن تثبيت خطوات الخوارزمية العادية من خلال ما يلي:

- تحديد k مركزا (centroids) للعناقيد بشكل عشوائي.
- حساب المسافة بين كل نقطة وبين جميع المراكز، ويتم ذلك باستخدام البعد الاقليدي، ويعطى البعد الاقليدي d_{ij} حسب العلاقة التالية:

$$d_{ij} = \sqrt{\sum_{k=}^{n} (x_{ik} - x_{jk})^2}$$

حيث ان:

N: تمثل عدد نقاط البيانات

X_{ik}: احداثيات الخاصية k للنقطة i

X_{jk}: احداثيات الخاصية k للنقطة j(يكون في العادة احداثيات المركز)

- تجميع الكائنات لأقرب مركز لها.
- حساب المتوسط الجديد لكل عنقود
- تكرار الخطوات من 2 الى 4 حتى حصول الاستقرار (عدم وجود كائنات تنتقل ضمن التجمعات) او حتى التكرار لعدد محدد من المرات.
- نهاية الخوارزمية.

يعتمد اداء هذه الخوارزمية على المواقع الاولية لمراكز التجمعات، ومن المفضل تنفيذ هذه الخوارزمية عدة مرات مع اختلاف المراكز في كل مرة، وتندرج هذه الخوارزمية تحت الهدف بدون اشراف لأنها تصنف النقطة تلقائيا اعتمادا على شرط معين (مثل اقرب مسافة الى مراكز التجمع).

2- تطبيقات خوارزمية (k-means clustering)

تعتبر خوارزمية (k-means clustering) من ابسط خوارزميات التجميع واكثرها فعالية، وهنالك الكثير من تطبيقاتها، اهمها:

- التعليم بدون اشراف (unsupervised learning) للشبكات العصبية.
- تمييز النماذج (Pattern recognitions)
- التصنيف (Classification)
- معالجة الصورة (Image processing)
- الرؤيا بواسطة الالة (machine vision)
- الذكاء الصنعي (Artificial intelligent)

وبشكل اساسي اذا كنت تملك عدة امثلة وكل مثال يملك عدة خصائص، وترغب بتصنيف الامثلة وفق خصائصها فبإمكانك تطبيق هذه الخوارزمية، وفيما يلي مبدا تطبيق خوارزمية k-means clustering) على تعلم الالة (machine learning) او التنقيب عن البيانات:

يتم اسناد كل مثال اعتمادا على خاصية التجميع تلقائيا الى احد التجمعات، وهذا يدعى التعليم بدون اشراف لان الخوارزمية تصنف المثال تلقائيا اعتمادا على شرط معطى (مثلا اقرب مسافة الى مركز التجمع) لا نحتاج الى الاشراف على الخوارزمية بان نقول لها ان عملية التصنيف صحيحة او خاطئة، تعتمد عملية التعلم على امثلة التدريب المقدمة للخوارزمية، ولدينا خياران في عملية التدريب هذه:

- تدريب غير منتهي: تعتبر جميع البيانات المقدمة الى الخوارزمية تلقائيا امثلة تدريب.
- التدريب المنتهي: بعد ان يتم اعتبار عملية التدريب منتهية (اعطاء معلومات صحيحة عن المتوسط) نبدأ عند ذلك بجعل الخوارزمية تصنف تجمع الامثلة الجديدة، وهذا يتم ببساطة من خلال اسناد المثال الجديد الى اقرب مركز تجمع دون اعادة حساب المراكز الجديدة وفي هذه الحالة تكون مراكز التجمعات الناتجة بعد انتهاء عملية التدريب مراكز ثابته.

3- مزايا خوارزمية (k-means clustering)

1. ذات فعالية عالية
2. سهلة التنفيذ
3. تتعامل مع القيم المستمرة والقيم المتقطعة (الاسمية)

4- مساوئ خوارزمية (k-means clustering)

- تحديد عدد التجمعات k يتم عشوائيا وقبل معالجة الامثلة
- حساسة للحالة الاولية، اذ يؤدي اختيار حالات اولية متعددة الى اعطاء نتائج مختلفة للتجمعات، ونتيجة لذلك قد تقع الخوارزمية في مشكلة النهاية المحلية.
- شكل التجمع دائري لأنه يعتمد على حساب المسافة.

2-3 :الاطار النظري تحليل المركبات الرئيسة

يسعى تحليل المركبات الرئيسية الى أيجاد عوامل (factors) أو توليفات خطية (linear combination) تسمى بالمكونات الرئيسية لتركيبة من المتغيرات تفسر معظم القياس الكلي للقيم الأصلية .

3-2-1: مفهوم تحليل المركبات الرئيسية

نفرض أن

$$Y \sim N_P(\mu, S) \quad \ldots (3\text{-}12)$$

$$Y = XB + e$$

حيث إن :- (X) يحوي على (m) من المتغيرات وبتطبيق طريقة المركبات الرئيسية سنحصل على مركبات رئيسية أقل من (μ)

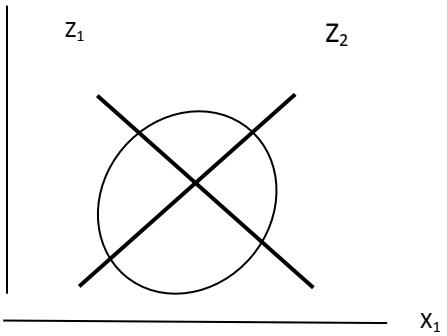
$P_{C1} = a_{11}x_1 + a_{21}x_2 + \ldots\ldots\ldots + a_{m1}x_m$

$P_{C2} = a_{12}x_1 + a_{22}x_2 + \ldots\ldots\ldots + a_{m2}x_m$

ويمكن كتابتها على شكل مصفوفة.

$P_c = XA$

ومن خصائص هذه المركبات أنها تكون متعامدة ولها خاصية تقليل مجموع مربعات المسافات العمودية للنقاط على المحاور وهي نفس خاصية خط الانحدار ،كما في الشكل (3-5):



شكل (3-5) محاور المركبات الرئيسية

أما الاساس الرياضي في تحليل المركبات الرئيسية فهو يتم بتدوير محاور جديدة للمتغيرات يتم تكوين مركبات رئيسية تحل محل المتغيرات الأصلية في التفسير وذلك بضرب كل متغير بمصفوفة متعامدة نرمز لها بالرمز A

$$Z=AX_i \qquad \ldots (3\text{-}13)$$

وحيث أن (A) مصفوفة متعامدة أي

$$A'A = I$$

$$Z_i'Z_i = (AX_i)'(AX_i)$$

$$= X_i'AA'X_i$$

$$= X_i'X_i$$

ويمكن كتابتها على شكل مصفوفة

$$Z = AX \quad \ldots\ldots\ldots (3\text{-}14)$$

أي إن المتغير (X_i) أصبح معبراً عنها بالتركيبة الخطية للمركبة الرئيسية .وللتركيبة الخطية لمصفوفة التباين والتباين المشترك ل(Z)

$$S_z = \begin{pmatrix} S^2_{Z_1} & 0 & . & . & 0 \\ 0 & S^2_{Z_2} & . & . & 0 \\ . & . & . & . & . \\ . & . & . & . & . \\ 0 & 0 & . & . & S^2_{Z_p} \end{pmatrix}$$

وإن مجموع التباين الكلي

$$\sum_{i=1}^{p} S^2 z_i = S^2 z_1 + S^2 z_2 + \ldots\ldots\ldots + S^2 z_p$$

والتركيبة الخطية الأولى

$Z_1 = a_{11}x_1 + a_{21}x_2 + \ldots\ldots\ldots + a_{p1}x_m$

لها اكبر تباين ممكن

$A'_i = (a_{11}, a_{21}, \ldots\ldots\ldots\ldots, a_{p1})$

ومن خلال المتجهات المميزة (eigenvectors) للجذور المميزة (eigenvalues) (X_i) للمصفوفة (S) حسب المعادلة التالية

$$(S - \lambda I) = 0$$

ولأجل أن يكون λ_i هو اكبر تباين ممكن يجب أن تكون المتجهات المميزة لها خاصية التعامد أي

$$a_1^2 + a_2^2 + \ldots\ldots\ldots\ldots\ldots + a_p^2 = 1 \quad a' a = 1$$

2-2-3 خصائص المركبات الرئيسية المحسوبة من المصفوفة (S) :-

إن جميع الجذور المميزة (λ_i) تكون اكبر من الصفر $\lambda_i > 0$ ومجموع الجذور يساوي مجموع عناصر قطر المصفوفة (S)

$$\sum_{i=1}^{p} \lambda_i = Tr(S) \qquad \ldots (3\text{-}15)$$

$$Tr(S) = \sum_{i=1}^{p} var\ (xi) \qquad \ldots (3\text{-}16)$$

$$And\ |S| = \prod_{j=1}^{p} \lambda_j \ldots (3\text{-}17)$$

أما الاهمية النسبية للمركبة (i,……,1)

$$\frac{Var\ (Z_i)}{\sum_{i=1}^{p} Var\ (Z_i)} = \frac{\lambda_i}{Tr(S)} = \frac{\lambda_i}{\sum \lambda_i} \qquad \ldots (3\text{-}18)$$

يمكن حساب الجذور المميزة من مصفوفات الارتباط R حسب الصيغة التالية .

$$\left|R-\hat{\lambda}^{*}I\right|=0 \quad \ldots (3\text{-}19)$$

حيث أن

$$R=\begin{pmatrix} 1 & \rho_{12} & . & . & \rho_{1p} \\ \rho_{21}. & 1 & . & . & \rho_{2p} \\ . & . & 1 & . & . \\ . & . & . & 1 & . \\ \rho_{p1}. & \cdot\rho_{p2} & . & . & 1 \end{pmatrix}$$

$$\therefore Tr(R)=\rho=\sum_{j=1}^{p}\hat{\lambda}^{*}_{j} \quad \ldots (3\text{-}20)$$

$$\therefore |R|=\prod_{j=1}^{p}\hat{\lambda}^{*}_{j} \quad \ldots (3\text{-}21)$$

$$\frac{\text{Var}(Z^{*}_{j})}{\sum_{j=1}^{p}\text{Var}(Z^{*}_{j})}=\frac{\hat{\lambda}^{*}_{j}}{\text{Tr(R)}}=\frac{\hat{\lambda}^{*}_{j}}{p} \quad \ldots (3\text{-}22)$$

3-2-3.اختبار المعنوية للمركبات الرئيسية :-

ان الاختبارات المتعلقة بالمركبات الرئيسية هي اختبارات متعلقة بالجذور والمتجهات المميزة . أي لاختبار الفرضية القائلة ان جميع الجذور المميزة للمصفوفة (S) أو (R) متساوية .

$$H_0=\hat{\lambda}_1=\hat{\lambda}_2=\ldots\ldots\ldots\ldots=\hat{\lambda}_p$$

$$H_1=\hat{\lambda}_1\neq\hat{\lambda}_2\neq\ldots\ldots\ldots\ldots\neq\hat{\lambda}_p$$

ولاختبار هذه الفرضية نستخدم اختبار (χ^2) حسب الصيغة التالية :

$$\chi^2=\ln|S|-\rho\ln(\frac{tr_{(S)}}{\rho}) \quad \ldots (3\text{-}23)$$

وتقارن قيمة (χ^2) الحسابية مع قيمة (χ^2) الجدولية الاتية .

$$v = \frac{1}{2}(k-1)(k+2)$$

نرفض فرضية العدم H_0 اذا كانت

$$\chi_0^2 \geq \chi_{a,v}^2$$

3-3: طريقة الإمكان الأعظم في التقدير

تعتبر من الطرق المهمة في التقدير وأن مقدراتها تتمتع بخصائص ممتازة، وهذه المقدرات تجعل لوغاريتم دالة الإمكان في نهايتها العظمى،

3-3-1.مفهوم طريقة الامكان الاعظم

اذا كانت لدينا عينة عشوائية $(x_1, x_2, \dots, x_n)$ من التوزيع الأحتمالي المعرف (2) فأن دالة الإمكان ستكون:

$$L(x_1, x_2, \dots, x_n; p, b) = \prod_{i=1}^{n} f(x_i; p, b)$$

$$= p^n b^n \prod_{i=1}^{n} x_i^{b-1} (1 + x_i^b)^{-(p+1)} \qquad (3\text{-}24)$$

وبإدخال اللوغاريتم الطبيعي على طرفي المعادلة (3-24) نحصل على:

$$ln\, L = n \ln p + n \ln b + (b-1) \sum_{i=1}^{n} \ln x_i - (p+1) \sum_{i=1}^{n} \ln(1 + x_i^b)$$

وباشتقاق طرفي المعادلة اعلاه نحصل على:

$$\frac{\partial \ln L}{\partial p} = \frac{n}{p} - (1) \sum_{i=1}^{n} \ln(1 + x_i^b) \qquad (3\text{-}25)$$

$$\frac{\partial \ln L}{\partial b} = \frac{n}{b} + \sum_{i=1}^{n} \ln x_i - (p+1) \sum_{i=1}^{n} \frac{x_i^b (1) \ln x_i}{(1 + x_i^b)} \qquad (3\text{-}26)$$

وعند مساواة المعادلة (3-25) للصفر نحصل على:

$$\Rightarrow \hat{p}_{MLE} = \frac{n}{\sum_{i=1}^{n} \ln(1 + x_i^{\hat{b}})} \qquad (3\text{-}27)$$

فإذا افترضنا أن $\hat{b} = 1$ ستكون صيغة:

$$\hat{p}_{MLE} = \frac{n}{\sum_{i=1}^{n} \ln(1 + x_i)} \qquad (3-28)$$

وبالإمكان اعطاء قيمة أولية أخرى للمعلمة b وتطبيق طريقة النقطة الصامدة لتطوير مقدر الأمكان الأعظم للمعلمة $\hat{p}_{MLE}$، ومن المعالدلة (3-26) نجد أيضاً أن:

$$\frac{n}{\hat{b}} + \sum_{i=1}^{n} \ln(x_i) = (\hat{p}+1) \sum_{i=1}^{n} \frac{x_i^{\hat{b}}(1) \ln x_i}{(1+x_i^{\hat{b}})} \qquad (3\text{-}29)$$

وهي أيضاً معادلة ضمنية تتضمن $\hat{b}, \hat{p}$ بالأمكان تعويض مقدر الأمكان الأعظم للمعلمة $(\hat{p})$ من المعادلة (3-28)، في المعادلة (3-29) واستخدام طرائق عددية لإيجاد مقدر الإمكان الأعظم للمعلمة $\hat{b}_{MLE}$.

وبعد ايجاد مقدري الإمكان الأعظم للمعلمتين $\hat{b}, \hat{p}$ يكون مقدر الأمكان الأعظم لدالة المعولية لتوزيع $Burr\ Type\ XII$ هو:

$$\hat{R}_{MLE}(x) = \left(1 + x^{\hat{b}_{MLE}}\right)^{-\hat{p}_{MLE}} (3\text{-}30)$$

4-3 :مفهوم طريقة بيز في التقدير

في هذا الجزء من البحث سيتم التركيز على أيجاد مقدر بيز لمعلمة الشكلp باعتبار أنها متغير عشوائي يتبع توزيع كاما بالمعلمات (α_1, β_1)، حيث أن توزيع المشاهدات x هو $Burr\ Type\ XII$ وكما معرف في المعادلة (2):

$$f(x; p, b) = pbx^{b-1}(1+x^b)^{-(p+1)} \qquad x, p, b > 0$$

وأن المعلمة p هي متغير عشوائي يتبع توزيع كاما،

$$p \sim Gamma\ (\alpha_1, \beta_1)$$

$$g_1(p) = \frac{\beta_1^{\alpha_1}}{\Gamma(\alpha_1)} p^{\alpha_1 - 1} e^{-\beta_1 p} \qquad p > 0 \qquad (3\text{-}31)$$

$$0 \qquad o/w$$

ولإيجاد المقدر البيزي وفق دالة خسارة معرفة، لابد أولاً من أيجاد التوزيع اللاحق $g_1(p|x)$:

$$g_1(p|x) = \frac{\prod_{i=1}^{n} f(x_i, p, b)\ g_1(p)}{\int_0^{\infty} \prod_{i=1}^{n} f(x_i, p, b)\ g_1(p)\, dp} (3-32)$$

وحيث أن:

$$\Pi f(x_i, p, b) = p^n\ b^n\ \Pi\ x_i^{b-1}\left(1 + x_i^b\right)^{-(p+1)}$$

$$f(x) = \int_0^{\infty} \Pi f(x_i, p, b)\ g_1(p) dp$$

$$= \frac{\beta_1^{\alpha_1}}{\Gamma(\alpha_1)} b^n \Pi\, x_i^{b-1} \int_0^\infty p^{n+\alpha_1-1} e^{-\beta_1 p}\, \Pi\left(1+x_i^b\right)^{-(p+1)} dp \qquad (3\text{-}33)$$

ويمكن كتابة المقدار $\Pi\left(1+x_i^b\right)^{-(p+1)}$ بالصيغة $e^{-(p+1)\sum \ln\left(1+x_i^b\right)}$

ثم يتم جمع هذا المقدار مع $e^{-\beta_1 p}$

$$\Rightarrow\ e^{-p\left[\beta_1+\sum \ln\left(1+x_i^b\right)\right]} \Rightarrow\ e^{-pT}$$

علماً بأن:

$$T = \beta_1 + \sum \ln\left(1+x_i^b\right)$$

$$\therefore f(x) = \frac{\beta_1^{\alpha_1} b^n \Pi\, x_i^{b-1}}{\Gamma(\alpha_1)} e^{-\sum \ln\left(1+x_i^b\right)} \int_0^\infty p^{n+\alpha_1-1} e^{-pT} dp$$

وحسب قاعدة كاما في التكامل:

$$\int_0^\infty x^{n-1} e^{-x} dx = \Gamma(n)$$

فأن:

$$\int_0^\infty p^{n+\alpha_1-1} e^{-pT} dp = \frac{1}{T^{n+\alpha_1}} \Gamma(n+\alpha_1)$$

وطبقاً لهذه التكاملات تكون:

$$g_1(p|x) = \frac{\frac{\beta_1^{\alpha_1} b^n \Pi\, x_i^{b-1}}{\Gamma(\alpha_1)} p^{n+\alpha_1-1} e^{-pT} e^{-\sum \ln\left(1+x_i^b\right)}}{\frac{\beta_1^{\alpha_1} b^n \Pi\, x_i^{b-1}}{\Gamma(\alpha_1)} e^{-\sum \ln\left(1+x_i^b\right)} \frac{\Gamma(n+\alpha_1)}{T^{n+\alpha_1}}}$$

$$g_1(p|x) = \frac{T^{n+\alpha_1}}{\Gamma(n+\alpha_1)} p^{n+\alpha_1-1} e^{-pT} \quad (3\text{-}34)$$

$$\sim \Gamma\left(n+\alpha_1, \beta_1 + \sum \ln\left(1+x_i^b\right)\right)$$

وتحت دالة خسارة تربيعية

$$loss = \ (p-d)^2 \qquad (3\text{-}35)$$

فأن دالة المخاطرة Risk function هي:

$$Risk = E(loss) = \ E(p-d)^2$$

$$= \int (p-d)^2 f(p|x) dp \qquad (3-36)$$

مقدر بيز للمعلمة p هو متوسط التوزيع اللاحق

$$\hat{p}_{Bayes} = E(p|x) \Rightarrow posterior\ mean\ for\ known\ \ b$$

$$\hat{p}_{Bayes(1)} = \frac{n+\alpha_1}{\beta_1+\sum \ln(1+x_i^b)} \qquad (3\text{-}37)$$

حيث أن (α_1, β_1, b) ثوابت معلومة.

أما المقدر البيزي الثاني فيعتمد على:

$$g_2(p) = \frac{k}{p^c} \qquad (3\text{-}38)$$

وهو التوزيع الثاني المقترح للمعلمة p وطبقاً لذلك لابد من أيجاد التوزيع اللاحق:

$$g_2(p|x) = \frac{\Pi f(x;p,b) g_2(p)}{f(x)}$$

$$f(x) = \int_0^\infty \Pi\ f(x;p,b) g_2(p) dp$$

$$= b^n \Pi\ x_i^{b-1} \int_0^\infty p^n \left(1+x_i^b\right)^{-(p+1)} \frac{k}{p^c}\, dp$$

$$= b^n \Pi\ x_i^{b-1} k \int_0^\infty p^{n-c} \left(1+x_i^b\right)^{-(p+1)}\, dp$$

$$\Pi\left(1+x_i^b\right)^{-(p+1)} = e^{-(p+1)\sum \ln\left(1+x_i^b\right)}$$

$$= b^n \Pi\ x_i^{b-1} k \int_0^\infty p^{n-c} e^{-(p+1)\sum \ln\left(1+x_i^b\right)}\, dp$$

$$= b^n \Pi\ x_i^{b-1} e^{-\sum \ln\left(1+x_i^b\right)}\ k \int_0^\infty p^{n-c} e^{-p\sum \ln\left(1+x_i^b\right)}\, dp$$

$$\int_0^\infty p^{n-c} e^{-pT}\, dp = \frac{1}{T^{n-c+1}} \int_0^\infty (Tp)^{n-c} e^{-pT}\, dp = \frac{1}{T^{n-c+1}}\ \Gamma(n-c+1)$$

$$\therefore f(x) = b^n \Pi\ x_i^{b-1} e^{-\sum \ln\left(1+x_i^b\right)}\ k\ \frac{\Gamma(n-c+1)}{T^{n-c+1}}$$

$$g_2(p|x) = \frac{b^n \Pi\ x_i^{b-1} k\ p^{n-c} e^{-(p+1)\sum \ln\left(1+x_i^b\right)}}{b^n \Pi\ x_i^{b-1} e^{-\sum \ln\left(1+x_i^b\right)}\ k\ \frac{\Gamma(n-c+1)}{T^{n-c+1}}}$$

$$g_2(p|x) = \frac{T^{n-c+1} p^{n-c} e^{-pT}}{\Gamma(n-c+1)} \qquad (3-39)$$

$$T = \sum \ln(1 + x_i^b)$$

وهو أيضاً توزيع كاما

$$g_2(p|x) \sim Gamma\ (n - c + 1, T)$$

وفي حالة دالة خسارة تربيعية يكون مقدر بيز للمعلمة p هو:

$$\hat{p}_{Bayes(2)} = \frac{n-c+1}{T} = \frac{n-c+1}{\sum \ln(1+x_i^b)} \qquad (3\text{-}40)$$

ولقيم c المعلومة وقيم b المقدرة بطريقة الأمكان الأعظم يمكن الحصول على مقدرات بيزية جديدة مقترحة للمعلمة p والتي تم تسميتها $\hat{p}_{Bayes(2)}$.

3-5: مفهوم الخوارزمية الجينية

مع التقدم العلمي الهائل، والتسارع الاتي (الدليلي) الذي تشهده التكنولوجيا في الوقت الراهن ،ظهرت أنماط جديدة من الأنظمة وسميت بالأنظمة الذكية ،وبالتالي فتح المجال أمام العالم نحو نوع جديد من التكنولوجيا أطلق عليه اسم تكنولوجيا الذكاء الصنعي(Artificial Intelligence (Technology التي سرعان ما طورت واستخدمت في العديد من التطبيقات اليوم. ومن هنا جاءت الخوارزميات الجينية ([GA [Genetic Algorithm) كآلية للتعامل مع كائن ما (Object) وجعله يتمتع بالذكاء الصنعي . إن اختراع تقنيات الذكاء الصنعي ([Al [Artificial Intelligence) فتح بوابات جديدة وشكل قفزة نوعية في علوم الحاسوب فقد تنوعت تعاريف الذكاء الصنعي وتتعدد، لكن مهما اختلفت تعاريف الذكاء الصنعي ،فإن مضمونه واحد، وجوهره هو الخبرة والمعرفة والذاكرة، فكل تطبيقاته تتمحور حول مفهوم الأمثلة (Optimization) و معرفة الحل الأنسب للمشاكل المطروحة ، بطريقة فعالة . الأمثلة هي عملية معالجة مسألة ما بطريقة مضبوطة بهدف التوصل إلى أفضل النتائج الممكنة، وانطلاقا من أهميتها في حل المشاكل، تعد الأمثلة من أهم فروع علم الحاسوب، والهندسة والعديد من الأنظمة الأخرى، والتي تعددت وسائل تطبيقها وتطورت مع الزمن، وصولا إلى تقنيات التطور (Evolutionary Strategies) والخوارزميات الجينية (Genetic Algorithm) اللتان أصبحتا الآن مع تشكلان ما يسمى الحوسبة التطورية (Evolutionary Computation).

في بدايات العقد السادس من القرن العشرين (1962 م) تم اختراع الخوارزميات الجينية من قبل جون هولاند (John Holland) في الولايات المتحدة الأمريكية والتي تم تطويرها من قبل تلامذته في جامعة ميتشيغان (Michigan) بين عامي 1960 و 1980. تعد الخوارزميات

الجينية تمثيل للمعتقد السائد بأن الذكاء البشري يخلق مع الإنسان ويتم اكتسابه عن طريق الوراثة بشكل كبير ،فهي محاكاة لعملية التزاوج بين الكائنات الحية من النوع نفسه، وقد استخدمت لها العديد من المصطلحات من علم الوراثة مثل : الجيل والوالدين والعبور والطفرة ...إلخ تحاول الخوارزميات الجينية الوصول إلى الحل الأنسب لمشكلة ما ،وذلك اعتمادا على مبدأ العالم داروين في الاصطفاء الطبيعي القائم على الاحتفاظ بالميزات والصفات الجيدة الموجودة في جيل الآباء ونقلها إلى جيل الأبناء بهدف الحصول على ذرية قوية تتمتع بأفضل صفات جيل السلف على أقل تقدير (البقاء للأقوى) . جميع الكائنات الحية تتألف من خلايا ، وكل خلية تحتوي العدد نفسه من سلاسل ال DNA التي تسمى الكروموسومات (Chromosomes) مما يضفي طابعة شخصية للكائن . كل كروموسوم يمكن تقسيمه إلى الجينات (Genes) المكونة له ،والتي تعطي الكائن سمة محددة ،مثل لون العين فعندما يختلف ترتيب الجينات تختلف السمة التي تعطيها، مما يؤثر في الكائن ككل . هناك ترابط وثيق بين المعلومات السابقة والخوارزميات الجينية التي تعمل بنفس مبدأ انقسام الخلية وتكاثر الخلايا الحية، حيث تشير الكروموسومات إلى الحل المحتمل للمسألة، والجينات هي عبارة عن بت (Bit) واحد أو مجموعة من البنات المتتالية التي ترمز إلى عنصر محدد من الحل المحتمل، بحيث تكون مجموعة احتمالات كل بت هي إما 0 أو 1، وبناء عليه يتم اتخاذ القرار فيما يتعلق بجينات الجيل التالي، وصولا إلى الحل الأنسب. وبعد أن تبلورت فكرة الخوارزميات الجينية وبدأت تطبيقاتها تتوسع في منتصف العقد الثامن من القرن العشرين حيث شملت نطاق واسع من المواضيع ،ففي عام 1992 استخدم العالم جون كوزا (John Koza) الخوارزميات الجينية لكي يطور برامج لتأدية مهام محددة، فانتقلت حينها الخوارزميات الجينية إلى حيز البرمجة ،وأطلق عليها اسم البرمجة الجينية (Genetic Programming)

3-5-1الخوارزميات الجينية (Genetic Algorithm and its mechanism)

أن الخوارزمية هي عبارة عن مجموعة من التعليمات المتسلسلة التي تهدف لحل مسألة أو إنجاز هدف معين . الخوارزمية الجينية هي طريقة من طرائق البحث العشوائي ستخدم في الحوسبة وذلك بهدف معرفة الحلول الصحيحة أو الأقرب للحل الأمثل من بين مجموعة الحلول الممكنة التي تشكل ما يسمى فضاء البحث (Search Space) وذلك عبر مجموعة من الخطوات التي تعتمد على المقارنة وإيجاد المسافة بين الحلول ومن ثم انتقاء الحلول المناسبة والاعتماد عليها في تشكيل حلول أخرى (الجيل الثاني) أكثر ملائمة وأقرب للحل الأمثل. يتم في أثناء تطبيق الخوارزمية الجينية توصيف المشكلة المطروحة بتمثيل الصبغيات

(Chromosomes) الممثلة للحلول بواحدة من طرائق الترميز. بعد ذلك طبق مجموعة من العمليات الرياضية المستنبطة من العمليات البيولوجية كالعبور (CrossOver) والانتخاب (Selection) والطفرة (Mutation) للحصول في نهاية المطاف على مجموعة من الصبغيات التي تمثل الجيل الأخير ،وباختيار الصبغي الأفضل يتم الوصول إلى الحل الأمثل الذي نريد البحث عنه في المسألة المطروحة ،فخوارزمية البحث هذه تبدأ من مجموعة من الحلول العشوائية وتنتهي باختيار واحد من مجموعة "أبناء" من الحلول.

2-5-3 عناصر الخوارزميات الجينية : (GA Elements)

على الرغم من اختلاف الخوارزميات الجينية حسب فروع الحوسبة التطورية ،إلا أنها تشترك على الأقل بالعناصر التالية :.بحبوح- ح. ريشة 2007(ص 106).

1- تعداد سكاني من الصبغيات (Populations of chromosomes)

التي تمثل مجموعة البحث أو فضاء البحث (Search Space) وهي عبارة عن مجموعة حلول المسألة

2- الانتقاء (Selection)

يتمثل باختيار الصبغيات المناسبة ك "والدين" من أجل القيام بعملية التزاوج فيما بينها، ولكن عملية

الاختيار هذه لا تكون عشوائية ، وإنما تعتمد على كفاءة الصبغي.

3- الكفاءة :وهي معامل يعطي لكل صبغي قيمة معينة تدل على مدى كفاءة الصبغي (اقترابه من الحل) وبناء عليه تتم عملية انتقاء الصبغيات .

4- العبور (Crossover): بعد اختيار الصبغيات الملائمة من الجيل الأول ويحصل العبور الذي يتم من خلاله تكوين صبغيات جديدة (نسل جديد) اعتمادا على الصبغيات الأم .

5- الطفرة العشوائية (Random Mutation)

بعد تشكيل النسل الجديد ، يتم إحداث طفرات (تغييرات في صيغتها الصبغية) ،وهذا الأمر يساعد في الوصول للحل بشكل أسرع.

3-5-3 : ترميز الصبغيات (Chromosomes encoding)

إذا أردنا أن نحل المسألة بشكل برمجي ولابد من ترميز كل صبغي بشكل يسهل التعامل معه من قبل الحاسوب وذلك حسب المسألة المطروحة. إذا كانت المتحولات متقطعة (Discrete Variables) فهي ذات عدد محدود ومن ثم يمكننا استخدام التمثيل الثنائي لترميز جميع الحالات ،أما إذا كانت المتحولات مستمرة (Continuous variables) فهي ذات عدد غير منته وهذا يتطلب القيام بحل ما . من الحلول المقترحة إيجاد درجة السماحية (Precision) ومن ثم نقوم بتقسيمها على عدد الخانات الثنائية من أجل الترميز، و يمكن أيضا تحديد عدد الخانات الثنائية ، ومن ثم تحديد حجم الخطوة بناء عليها .

4-5-3 : طرائق الترميز : (Encoding types)

يتم ترميز الحلول كصبغيات تمثل سلاسل من الخانات التي تدعى الجينات ((genesفي كل خانة يوجد رقم أو محرف يمثل قيم معينة أو معاملات لحل المسألة ،وعادة ما رمز الصبغيات كسلاسل من الأرقام (الثنائية - الحقيقية – الصحيحة - الثمانية - الست عشرية) ،ولكل نوع من الأرقام فوائده تبعا لنوع المسألة المطروحة .

5-5-3 : دالة الكفاءة

وهي عبارة عن تابع رياضي يستخدم في تحديد كفاءة الصبغي ،وعادة يكون الحل (الصبغي) الأفضل هو الذي تكون قيمة دالة الكفاءة عنده أكبرية أو أصغرية حسب نوع المسألة.
آ.بحبوح- ح. ريشة 2007(ص 106).

6-5-3 : آلية عمل الخوارزميات الجينية : (GA working mechanism)

عندما نقوم بتحديد المسألة المراد حلها ، تبدأ الخوارزمية عملها وفق المراحل التسلسلية التالية :

1- التعداد السكاني الابتدائي (Initial population)

تقوم الخوارزمية في هذه المرحلة بتشكيل مجموعة مكونة من n صبغي كل منها مكون من | جين (خانة) ،إن هذه الصبغيات تمثل مجموعة من الحلول الابتدائية للمسألة.

2- نقوم بتحديد دالة الكفاءة f (x) لكل صبغي x من التعداد، ومن ثم تقوم الخوارزمية بحساب كفاءة كل صبغي على حدة.

3- تكرر الخوارزمية مجموعة من العمليات حتى يتم الوصول إلى n سلالة جديدة، هذه العمليات هي كالآتي:

- تبدأ عملية الانتخاب ((Selectionحيث يتم اختيار زوج (أو أزواج) من صبغيات التعداد السكاني الحالي ،وذلك اعتمادا على قيمة الاحتمال الرياضي Probability) , of selection) الذي يتناسب مع دالة الكفاءة طردة في حال كانت القيمة الأمثلية للحل هي الأكبرية في دالة الكفاءة - أو عكسة – في حال العكس - كما يمكن أن يتم اختيار الصبغي ذاته أكثر من مرة . إن عملية تحديد قيمة ،p تتم باتباع إحدى خوارزميات الاختيار .
- يتم استدعاء إجرائية العبور (CrossOver) حيث يتم تحديد نقطة ما من الصبغيين الوالدين وفق قيمة الاحتمال (, Probability of crossover) وبعد ذلك يحصل العبور اعتماد على تلك النقطة)، إذا لم تتمكن الخوارزمية من تحديد نقطة عبور ،فإن الصبغيين الولدين يكونان مطابقان تماما للوالدين. من الممكن أن يتم اختيار أكثر من نقطة عبور واحدة (point crossover) في بعض الحالات. .. يحصل التطافر (Mutation) على كل جين من الصبغيين الجديدين باحتمال (probability of) mutation وذلك بتغيير قيمة الجين إلى قيمة جديدة مختلفة لزيادة فرصة الاقتراب من الحل الأمثل ،ومن ثم يتم وضع الصبغيين الجديدين في التعداد الجديد. بعد تنفيذ هذه العمليات عدد من التكرارات ريتشكل لدينا n صبغي جديد ،وهنا في حال كان nفردية يتم التخلص من صبغي عشوائية من التعداد.

. 4- بعد تنفيذ إجرائيات الخوارزمية الجينية يتم استبدال التعداد السابق بالتعداد الحالي .

5- وأخيرا يتم تكرار العمليات بدءا من العملية 2.

إن كل تكرار لهذه العملية يدعى الجيل (generation) كما يطلق على مجموعة كل الأجيال اسم التشغيل (run) حيث أن في نهاية كل تشغيل يوجد غالباً صبغي واحد أو أكثر يكون ذو كفاءة عالية مقارنة مع أبناء جيله (Melanie, M. " (page 8

3-5-7: معايير الاختيار والعبور والطفرة في الخوارزميات الجينية (GA's Selection, crossover, mutation standards)

كنا قد تطرقنا إلى آلية عمل الخوارزمية الجينية، وعلمنا أن هذه الخوارزمية هي خوارزمية للبحث وتبدأ بعشوائية (فوضوية) وتتسق تدريجيا وصولا إلى النظام (الحل الأمثل

الذي يرضي المستخدم . ومن أجل أن يتم الانتقال من العشوائية إلى الاتساق ،لابد من ضبط عمل الخوارزمية وفق عدة معايير تضمن للخوارزمية اتخاذ القرار السليم في أثناء عملها . هذه المعايير تتحكم في الاختيار والعبور والطفرة، من أجل الوصول إلى الهدف. أولا: معايير الاختيار : Selection) standards وهي عبارة عن خوارزميات معينة تسهم في جعل الخوارزمية تتخذ قرارها باختيار زوج الصبغيات الأفضل ، وفيما يأتي بعض منها :

1- الاختيار وفق مبدأ العجلة المتدحرجة : (Roulette Wheel Selection)

لنفترض أن لدينا عجلة مقسمة إلى قطاعات مختلفة المساحة (بعضها صغير وبعضها كبير نسبية)، إذا قمنا بإدارة العجلة عشوائية فإنها سوف تستقر بحيث يكون مؤشر دال على أحد القطاعات. في هذه الخوارزمية ويتم الاختيار بالاعتماد على القطاع المشار إليه من قبل المؤشر، بحيث أن مساحة كل قطاع تتناسب طردة مع قيمة ps للفرد الممثل بهذا القطاع. يتم حساب pc اعتمادا على العلاقة الرياضية التالية :

f(x)

ps (x) = Ef0)

حيث : (x) ps احتمالية انتخاب الفرد *. (x) f قيمة الكفاءة للفرد . n عدد أفراد الجيل. بعد تشكيل التعداد الابتدائي تقوم الخوارزمية الجينية بحساب قيمة كفاءة كل صبغي على حدة ،ومن ثم قيمة احتمالية اختيار كل فرد من العلاقة السابقة. الآن نحتاج إلى تدوير العجلة واختيار الفرد الذي أشار المؤشر على القطاع الخاص به، ونلاحظ أنه كلما زادت كفاءة (ملائمة الفرد ، زادت مساحة القطاع المعبر عنه ،وزادت إمكانية اختياره من أجل التزاوج.

في بعض الحالات الخاصة وقد ينتج عن الاختيار العشوائي للعجلة المتدحرجة فرد ذو كفاءة قليلة ،وبالتالي يتم استخدامه على حساب فرد آخر أكثر ملائمة . من هذا المبدأ تم استخدام طريقة حكم النخبة والتي تعتمد على نسخ الأفراد الذين يتمتعون بكفاءة عالية من الجيل السلف إلى الجيل الخلف ،وذلك لضمان الاستفادة منهم في الجيل الثاني ،أما بقية الأفراد فيتم اختيارهم بالطرق العادية

2- معايير العبور : (Crossover standards)

بعد الاختيار وتلجأ الخوارزمية الجينية لإحدى الطرق الآتية من أجل تحديد نقطة العبور بين كل زوج من الصبغيات الآباء :

- العبور البسيط : (Simple crossover)

بعد اختيار مجموعة الصبغيات ،يتم تطبيق هذا النوع من العبور وفق مرحلتين :

- يتم انتقاء صبغيات عشوائية من العينة المختارة .
- كل زوج من الصبغيات يختبر العبور ،حيث يتم توليد رقم عشوائي [4,1]

حيث ا يعبر عن طول الصبغي (عدد الجينات لكل صبغي) ، ومن ثم يتم خلق صبغيين جديدين عن طريق تبادل الجينات الواقعة بين الموقعين 1 + k و 1

Magalhães-Mendes, J. (2013 (page167).

- العبور ثنائي النقطة : (Two - point crossover)

هذا النوع من العبور مشابه للعبور البسيط ،ولكن مع اختلاف هو أنه يتم اختيار نقطتي عبور بدلا من واحدة ،فإذا تم تحديد نقطتي العبور

3- العبور المنسق (:Uniform crossovr)

في هذا النمط من العبور ،نحتاج أولا إلى قيمة (probability of crossover) . p يحددها المستخدم قبل تنفيذ الخوارزمية . إذا كان لدينا 1 خانة (جين) في كل من الصبغيين الوالدين وتقوم الخوارزمية بمقارنة قيمة كل جين في الصبغي الأول مع القيمة , pcفإذا كانت لدينا قيمة ترتيبها iii< 1 وقيمتها أكبر أو تساوي pc نقوم بتبديل الجين i من الصبغي الأول مع نظيره من الصبغي الثاني ،وإلا يبقى جين الصبغي الأول في موضعه في النسل الأول ،وكذلك الأمر بالنسبة لجين الصبغي الثانية.

6-3 نموذج الانحدار اللوجيستي (Logistic Regression Model)

يرى (Lea) بأنه وإن كانت هناك عدة أساليب إحصائية طورت لتحليل البيانات ذات المتغيرات التابعة التصنيفية، إلا أن تحليل الانحدار اللوجيستي يتمتع بعدة مميزات تجعله ملائما للاستخدام في مثل حالات كهذه ويوضح (Gebotys) أهمية تحليل الانحدار اللوجيستي بقوله: "إن الانحدار اللوجيستي هو أداة أكثر قوة، لأنه يقدم اختباراً لدلالة المعاملات، كما انه يعطي الباحث فكرة عن مقدار تأثير المتغير المستقل على متغير الاستجابة الثنائية، وبالإضافة إلى ذلك فإن الانحدار اللوجيستي يرتب تأثير المتغيرات، مما يسمح للباحث بالاستنتاج بأن متغيراً ما يعتبر أقوى من المتغير الآخر في فهم ظهور النتيجة المطلوبة، كما أن تحليل الانحدار اللوجيستي يمكنه أن يتضمن المتغيرات النوعية وحدوداً للتفاعلات" .

$$p(x)=\frac{\exp(\alpha+\beta \mathrm{X})}{1+\exp(\alpha+\beta \mathrm{X})} \quad ; 0<P<1$$

ويمكن كتابة هذه الصيغة كما يلي:

$$p(x)=\frac{1}{1+\exp[-(\alpha+\beta \mathrm{X})]}$$

حيث أن:

α و β : المعاملات المقدرة من البيانات.

X : هي المتغير المستقل.

exp : هي (e) الأساس اللوغاريتمي الطبيعي ويساوي تقريبا2.718.

$p(x)$: هي المتغير التابع y .

أما في حالة النموذج اللوجيستي المتعدد (Multiple Logistic Regression)،فإن النموذج يأخذ الصورة التالية:

$$p(x)=\frac{1}{1+\exp(-Z)}$$

حيث أن :

$$Z=\alpha+\beta_1 \mathrm{X}_1+\beta_2 \mathrm{X}_2+\ldots\ldots\ldots\ldots\ldots+\beta_p \mathrm{X}_p$$

وتعتبر نماذج الانحدار اللوجيستي حالة خاصة من حالات نماذج الانحدار العامة (Generalized linear models) ويطلق على هذه النماذج أحيانا نماذج اللوجت (Logit Model) وتستخدم عندما نرغب في التنبؤ بوجود صفة معينة أو ظاهرة أو خاصية معينة بالاعتماد على قيم متغير أو مجموعة من المتغيرات المستقلة الأخرى التي لها علاقة بالمتغير التابع تماما كما هو الحال في نماذج الانحدار العامة، وتستخدم معاملات النموذج اللوجيستي في تقدير نسبة الإمكان أو المفاضلة (Odds Ratio) لكل من المتغيرات المستقلة في النموذج.

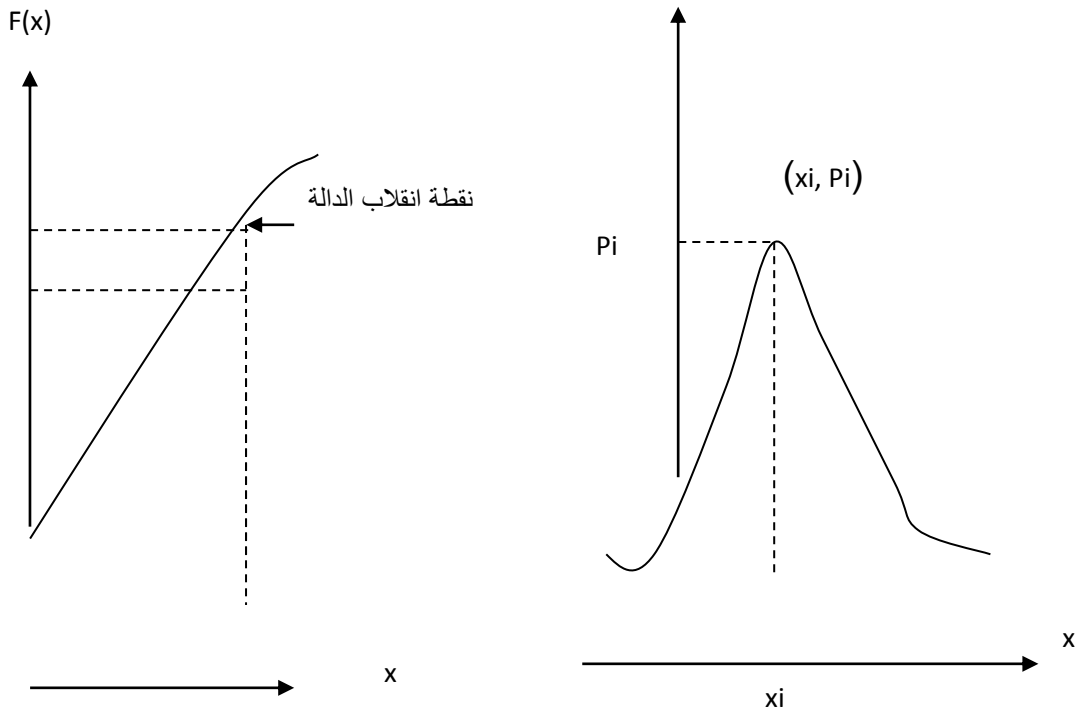
ومما سبق، فإن استخدام نموذج الانحدار اللوجيستي (Logistic Regression) والذي اعتمدت عليه كثير من الدراسات السابقة ،يعتبر من أكثر الأساليب الإحصائية شيوعا في التعامل مع البيانات الوصفية. إلا أن هذا النموذج يواجه مشكلة عدم القدرة على التعامل مع البيانات التي تعتمد على التغير في النقطة الزمنية التي تسبق حدوث الحدث، وتحتوي على بيانات اختفاء Censored Data وخاصة في الدراسات المقطعية (Cross Sectional) . لذلك كان هناك البديل الأكثر تعاملا مع هذه الحالة وهو انحدار كوكس

1-6-3 طبيعة النموذج اللوجستي (Logistic Model)

يستخدم النموذج اللوجستي بشكل واسع في الأبحاث الحياتية وذلك لتمثيل العلاقة بين متغير نسبة الاستجابات (P) حيث الاستجابة ثنائية (Binary Response) في مثل هذه الأبحاث – ومقياس الجرعة (X) الذي يمثل المتغيرات التوضيحية (Explanatory variables)إذ يعتقد بوجود تأثير كبير لهذه المتغيرات في احتمال الاستجابة P . وهو نموذج انحدار لا

خطي يلاءم البيانات الثنائية (Binary Data) التي يكون فيها المتغير المعتمد هو متغير وهمي يمثل الاستجابة التي أما أن تكون موجودة (إيجابية) وتأخذ القيمة (1) أو غير موجودة (سلبية) وتأخذ القيمة (5) ، ولكون الباحث الإحصائي يهتم بوجود (تحقق) الاستجابات عند كل مستوى من مستويات المتغير التوضيحي الذي يعتقد بعلاقته بها ، لذلك فان متغير نسبة الاستجابة المشاهدة هو نسبة الاستجابات المتحققة (الايجابية) عند كل مستوى من مستويات المتغير التوضيحي ، وهي نسبة متغيره من مستوى إلى آخر ويقترب توزيعها كثيرا من التوزيع الطبيعي عندما يكون حجم العينة كبيرا وذلك استناداً الى خاصية توزيع ذي الحدين الذي يؤول إلى التوزيع الطبيعي بزيادة حجم العينة ، وهذه الخاصية تعرف بأنها (الاتجاه نحو الاعتدال بزيادة حجم العينة) .

وكما هو موضح في الشكل (1) والشكل (2)



شكل (3-6) توزيع P

الشكل (3-7) المنحنى اللوجستي

فلو فرضنا ان

xi : المستوى i للتغير التوضيحي X

ni : حجم الاستجابات الكلية (الايجابية والسلبية) المشاهدة عند المستوى i

ri : عدد الاستجابات الموجودة (الايجابية) المشاهدة عند المستوى i

فأن :

$$pi = pr(y = 1|x) = \frac{ri}{ni}$$

هو نسبة الاستجابة المتحققة (الايجابية) المشاهدة

وان :

$$1 - pi = pr(y = 0|x) = \frac{ni - ri}{ni}$$

هو نسبة (احتمال) الاستجابة السلبية المشاهدة

وحيث ان pi , pi -1 هما دالة الى X فيمكن كتابتهما بدلالة X وفق النموذج الاتي المسمى (النموذج اللوجستي Logistic Model) :

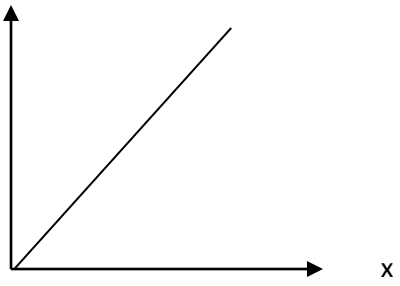
$$p = \frac{\exp(B_0 + B_1 X_1)}{1 + \exp(B_0 + B_1 X_1)} \quad .....(3-41)$$

$$\therefore 1 - p = \frac{1}{1 + \exp(B_0 + B_1 X_1)} .....(3-42)$$

2-6-3 التحويل الخطي للنموذج اللوجستي وأسبابه (The linear Transformation)

يصنف النموذج اللوجستي ضمن النماذج اللاخطية التي يمكن تحويلها إلى نماذج خطية وتسمى هذه النماذج بالنماذج الخطية ضمنيا (جوهريا). ويميل الاحصائيون عادة إلى التحويل الخطي لهذه النماذج لإزالة انحناءات معلماتها وذلك لتأثير هذه الانحناءات السلبي في حالة وجودها على خصائص مقدرات المربعات الصغرى لها (L.S estimators) ومن ثم قيم الاستجابة التي يتم التنبؤ بها ، حيث تكون هذه المقدرات في الغالب متحيزة ولا تتوزع توزيعا طبيعيا وتبايناتها لا تكون أصغر ما يمكن مما يجعل نتائج الاختبارات مضللة. وقد قام الباحث Berkson منذ العام 1944 بتحويل علاقة الانحدار اللاخطية بين المتغير التوضيحي

X ومتغير نسبة الاستجابة P في النموذج اللوجستي إلى علاقة إنحدار خطية ، برسم لوغاريتم النسبة $\frac{p}{1-p}$ مقابل قياسات المتغير التوضيحي X ، كما هو واضح في الشكل (8-3)



حيث :

$Ln\frac{p}{1-p} = z$ متغير الاستجابة الخطية

X : المتغير التوضيحي

شكل (8-3)

العلاقة الخطية بين المتغير التوضيحي X

ونسبة الاستجابة المحولة Z

ولبيان ذلك رياضيا نعود إلى الدالتين(3-41) و(3-42)فبقسمة الأولى على الثانية :

$$\frac{p}{1-p} = \frac{\frac{\exp(B_0 + B_1X_1)}{1+\exp(B_0 + B_1X_1)}}{\frac{1}{1+\exp(B_0 + B_1X_1)}} = \exp(B_0 + B_1X_1)$$

وبأخذ لوغاريتم الطرفين يكون :

$$Ln\frac{p}{1-p} = Z = B_0 + B_1X_1 \quad .....(3-43)$$

وهذه دالة انحدار خطي لـ Z على X .

3-6-3 متوسط وتباين متغير الاستجابة Z :

سبق وأن أشار الباحث إلى أن توزيع P في النموذج اللوجستي هو قريب من التوزيع الطبيعي –الشكل 2-2 وبعد تحويل P إلى Z بالتحويل اللوغاريتمي لإزالة تأثير انحناءات معلمات النموذج اللوجستي في تقدير قيم الاستجابة المتوقعة ، لابد من تحديد القيمة المتوقعة لـ Z وتباينه في التوزيع الطبيعي .بالرجوع إلى الدالتين (3-39) و (3-40) فأن :

$$p = \frac{\exp(B_0 + B_1 X_1)}{1 + \exp(B_0 + B_1 X_1)} \quad \& \quad 1 - p = \frac{1}{1 + \exp(B_0 + B_1 X_1)}$$

نلاحظ ان p =g (x) ، أي ان p هي دالة الى x

ولكن $Ln\frac{p}{1-p} = z \Leftarrow$ z = f(p) ، أي ان z هو دالة لـ P

$$\therefore z = f[g(x)]$$

(قاعدة السلسلة) chain Rule $\& \frac{\partial z}{\partial x} = \frac{\partial z}{\partial p} \cdot \frac{\partial p}{\partial x}$

$$\because \frac{\partial z}{\partial p} = \frac{1-p}{p} \quad \cdot \quad \frac{1-p+p}{(1-p)^2} = \frac{1}{p(1-p)} \qquad (3-44)$$

$$\& \frac{\partial p}{\partial x} = \frac{B_1 \exp(B_0 + B_1 X_1)}{[1 + \exp(B_0 + B_1 X_1)]^2} = B_1 \frac{1}{1 + \exp(B_0 + B_1 X_1)} * \frac{\exp(B_0 + B_1 X_1)}{1 + \exp(B_0 + B_1 X_1)}$$

-

$$\therefore \frac{\partial p}{\partial x} = B_1 p(1-p) \qquad (3-45)$$

وبتعويض (3-44) و (3-45) في الطرف الايمن من قاعدة السلسلة يكون :

$$\frac{\partial z}{\partial x} = \frac{B_1 p(1-p)}{p(1-p)} = B_1 \qquad (3-46) ، \frac{\partial^2 z}{\partial x^2} = 0$$

وهذا يعني أن انحدار Z على X_1 هو انحدار خطي بمعامل خطي ثابت مقداره (B_1) وبذلك تكون القيمة المتوقعة لـ Z وتباين Z هي كما يلي :

$$E(z) = B_0 + B_1 X_1 \quad \& \quad v(z) = \frac{1}{nipi(1-pi)} = \delta^2 i$$

الفصل الرابع

الجانب التطبيقي

ويشتمل هذا الفصل على:

أولا: الجانب التجريبي

1- المحاكاة وتقدير معادلات الانحدار، الخوارزمية الجينية
2- المقارنة بين الطرق الكلاسيكية والخوارزمية الجينية

ثانيا: الجانب العملي

1- تطبيق افضل الطرق التي تم التوصل اليها في الجانب التجريبي
2- تطبيق التحليل العنقودي في تصنيف المتغيرات للبيانات الحقيقية

4-1: الجانب التجريبي

سيتم في هذا الفصل عرض ومناقشة نتائج أسلوب المحاكاة للطرق المستخدمة في حل نموذج الانحدار اللوجستي باستعمال حجوم عينات مختلفة وبمعالم مختلفة وتم استخدام معيار المقارنة (mse) للمقارنة بين الطرق وكذلك تم استخدام الخوارزمية الجينية لجميع هذه الطرق وتمت المقارنة بينهما.

4-1-1المحاكاة وتقدير معادلات الانحدار الخوارزمية الجينية

- توليد الارقام العشوائية تتبع توزيع منتظم U(0,1)
- بالاعتماد على هذه الارقام العشوائية نولد المتغيرات التوضيحية حيث سنولد (5) متغيرات و(10) متغيرات
- نولد الدالة f(x) حيث ان:-

$$f(x) = \frac{e^{xB}}{1+e^{xB}} \quad x=5 , 10$$

- نولد الخطأ العشوائي ei والذي يتوزع توزيع برنولي باحتمال (pi)
- نوجد المتغير المعتمد yi وكما يلي:

yi = 1 if Zi >= 0.5

yi = 0 if Zi < 0.5

حيث ان

Zi= f(x) + ei

- نوجد معيار المقارنة متوسط مربعات الخطأ Mse(y):

وتم اجراء عدد من تجارب المحاكاة، حيث تم توليد10 متغيرات توضيحية وبأحجام مختلفة (n=100,500,2000,5000) وبواقع نموذجين اي سوف يكون هناك تنفيذ (16) تجربة بواقع تكرار (1000) لكل تجربة وكما في الجدول (1-4):

جدول (4-1) احجام العينات حسب النموذج

احجام العينات / المتغيرات التوضيحية				
النموذج الاول	100	500	2000	5000
النموذج الثاني	100	500	2000	5000

المصدر: (اعداد الباحث بواسطة البرنامج الاحصائي ماتلاب اصدار 2017)

اما القيم الافتراضية للمتغيرات التوضيحية الـ (10) ستكون للنموذج الاول قيم اولية للمعلمات المقدرة بطريقة المربعات الصغرى والنموذج الثاني قيم اولية للمعلمات المعيارية المقدرة بطريقة المربعات الصغرى وكانت كما في الجدول (4-2):

جدول (4-2) القيم الافتراضية لمعالم نموذج الانحدار

Bo	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10
0.007	0.231	0.024	0.006	0.085	0.065	0.017	0.038	0.097	0.011	0.055
0.007	0.012	0.023	0.005	0.018	0.004	0.006	0.057	0.009	0.049	0.087

المصدر: (اعداد الباحث بواسطة البرنامج الاحصائي ماتلاب اصدار 2017)

وقد تم الحصول على النتائج التالية:

جدول (3-4) قيم (mse) للطرق الكلاسيكية للنموذج الاول

حجم العينات	Mle (الإمكان الأعظم)	Wls (الموزونة)	PC (المركبات الرئيسية)	Bayse (طريقة بيز)	Best (الأفضل)
100	0.0146	0.0044	10.0363	0.1182	Wls
500	0.0008	0.00075	8.83659	0.03827	Wls
2000	0.00026	0.00015	4.94819	0.00527	Wls
5000	0.00011	0.00062	3.96701	0.00164	Wls

المصدر: (اعداد الباحث بواسطة البرنامج الاحصائي ماتلاب اصدار 2017)

جدول (4-4) قيم (mse) للطرق الكلاسيكية للنموذج الثاني

حجم العينات	Mle (الإمكان الأعظم)	Wls (الموزونة)	PC (المركبات الرئيسية)	Bayse (طريقة بيز)	Best (الأفضل)
100	0.02465	0.00452	12.79204	0.10203	Wls
500	0.00037	0.00014	0.68232	0.03271	Wls
2000	0.00045	0.00018	1.86864	0.00468	Wls
5000	0.00001	0.000065	1.47796	0.00141	Wls

المصدر: (اعداد الباحث بواسطة البرنامج الاحصائي ماتلاب اصدار 2017)

من خلال نتائج الجداول (3-4)، (4-4) نلاحظ بان طريقة Wls هي افضل طريقة لجميع حجوم العينات وللنموذجين كون قيم متوسط مربعات الخطأ (mse) هي الأقل عند استعمال عينات بحجوم مختلفة، وتبقى الأفضلية لطريقة المربعات الموزونة وبازدياد حجم العينة.

- نتائج الخوارزمية الجينية:

جدول (4-5) قيم (mse) للطرق الكلاسيكية في الخوارزمية الجينية للنموذج الاول

حجم العينات	MleGe (الإمكان الأعظم)	WlsGe (الموزونة)	PCGe (المركبات الرئيسية)	BayseGe (طريقة بيز)	Best (الأفضل)
100	0.0016	0.0040	0.0072	0.0042	MleGe
500	0.00038	0.00071	0.00046	0.00003	BayseGe
2000	0.00006	0.00013	0.00019	0.00014	MleGe
5000	0.000022	0.000061	0.000054	0.000039	MleGe

المصدر: (اعداد الباحث بواسطة البرنامج الاحصائي ماتلاب اصدار 2017)

جدول (4-6) قيم (mse) للطرق الكلاسيكية في الخوارزمية الجينية للنموذج الثاني

حجم العينات	MleGe (الإمكان الأعظم)	WlsGe (الموزونة)	PCGe (المركبات الرئيسية)	BayseGe (طريقة بيز)	Best (الأفضل)
100	0.00150	0.00402	0.00716	0.00431	MleGe
500	0.00035	0.00013	0.00053	0.00004	BayseGe
2000	0.00005	0.00014	0.00021	0.00015	MleGe
5000	0.000020	0.000061	0.000055	0.000042	MleGe

المصدر: (اعداد الباحث بواسطة البرنامج الاحصائي ماتلاب اصدار 2017)

من خلال نتائج الجداول (4-5)، (4-6) نلاحظ بان طريقة MleGe في الخوارزمية الجينية هي الافضل في حجوم العينات (100و 200و500) في كلا النموذجين في حين ظهرت طريقة BayseGe في الخوارزمية الجينية هي الافضل لحجم العينة (500) وفي كلا النموذجين ايضاً.

2-1-4 المقارنة بين نتائج الطرق الكلاسيكية والخوارزمية الجينية

سيتم التعرف في هذا المبحث على تحديد افضل الطرق المستخدمة في التقدير وذلك بالاعتماد على اقل معيار متوسط مربعات الخطأ (MSE) ولحجوم عينات مختلفة وبمعالم مختلفة

جدول (4-7) قيم (mse) للطرق الكلاسيكية و الخوارزمية الجينية ولحجوم عينات مختلفة وبمعالم مختلفة للنموذج الاول

		Classic (الطرق الكلاسيكية)	Genetic (الخوارزمية الجينية)	الافضل
n=100	Mle	0.0146	0.0016	Genetic
	Wls	0.0044	0.0040	Genetic
	PC	10.0363	0.0072	Genetic
	Bayes	0.1182	0.0042	Genetic
n=500	Mle	0.0008	0.00038	Genetic
	Wls	0.00075	0.00071	Genetic
	PC	8.83659	0.00046	Genetic
	Bayes	0.03827	0.00003	Genetic
n=2000	Mle	0.00026	0.00006	Genetic
	Wls	0.00015	0.00013	Genetic
	PC	4.94819	0.00019	Genetic
	Bayes	0.00527	0.00014	Genetic
n=5000	Mle	0.00011	0.000022	Genetic
	Wls	0.00062	0.000061	Genetic
	PC	3.96701	0.000054	Genetic
	Bayes	0.00164	0.000039	Genetic

المصدر: (اعداد الباحث بواسطة البرنامج الاحصائي ماتلاب اصدار 2017)

جدول (8-4) قيم (mse) للطرق الكلاسيكية و الخوارزمية الجينية ولحجوم عينات مختلفة وبمعالم مختلفة للنموذج الثاني

		Classic (الطرق الكلاسيكية)	Genetic (الخوارزمية الجينية)	Best
n=100	Mle	0.02465	0.00150	Genetic
	Wls	0.00452	0.00402	Genetic
	PC	12.79204	0.00716	Genetic
	Bayes	0.10203	0.00431	Genetic
n=500	Mle	0.00037	0.00035	Genetic
	Wls	0.00014	0.00013	Genetic
	PC	0.68232	0.00053	Genetic
	Bayes	0.03271	0.00004	Genetic
n=2000	Mle	0.00045	0.00005	Genetic
	Wls	0.00018	0.00014	Genetic
	PC	1.86864	0.00021	Genetic
	Bayes	0.00468	0.00015	Genetic
n=5000	Mle	0.00001	0.000020	Genetic
	Wls	0.000065	0.000061	Genetic
	PC	1.47796	0.000055	Genetic
	Bayes	0.00141	0.000042	Genetic

المصدر: (اعداد الباحث بواسطة البرنامج الاحصائي ماتلاب اصدار 2017)

من خلال الجداول (7-4)، (8-4) نلاحظ بالنسبة للنموذج الأول كانت افضل الطرق عند حجم العينة (100) بين الطرق الكلاسيكية هي طريقة المربعات الصغرى الموزونة اذ بلغت قيمة متوسط مربعات الخطأ (0.0044) وهي الاقل، وعند تطبيق الطرق في الخوارزمية الجينية تبين ان طريقة الإمكان الأعظم (Mle) (0.0016) هي الأفضل، وبزيادة حجم العينة (n=500) ظهرت طريقة المربعات الصغرى الموزونة هي الأفضل من بين الطرق الكلاسيكية (0.00075) بينما ظهرت طريقة بيز هي الأفضل (0.00003) في الخوارزمية الجينية، وبزيادة حجم العينة الى (2000) تبين ان طريقة المربعات الصغرى الموزونة (Wls) هي الأفضل أيضا بين الطرق الكلاسيكية بينما ظهرت طريقة الإمكان الأعظم (Mle) هي الأفضل (0.00006) في الطريقة الجينية، وعند زيادة حجم العينة الى 5000 تبين ان طريقة الإمكان الأعظم (Mle) هي الأفضل بين الطرق الكلاسيكية (0.00011) وظهرت الطريقة ذاتها هي الأفضل عند تطبيق الخوارزمية الجينية أيضا.

ومن ذلك تبين ان طريقة الخوارزمية الجينية هي الافضل في جميع حجوم العينات وللنموذجين.

من خلال النتائج التي تقدمت سوف نستخدم طريقة Wls في الطرق الكلاسيكية وطريقة الخوارزمية الجينية للـ Mle عند تطبيق البيانات الاصلية المتعلقة بجواز السفر السوداني.

4-2: الجانب العملي

4-2-1تطبيق أفضل الطرق التي تم التوصل اليها في الجانب التجريبي

سيتم في هذا الجانب وصف وعرض البيانات المستخدمة في التحليل، وتطبيق طريقة المربعات الصغرى الموزونة (WLS) وطريقة الخوارزمية الجينية بالنسبة للبيانات الحقيقية.

1- وصف البيانات:

تتكون البيانات من عدة متغيرات الخاصة بجواز السفر السوداني وكما يلي:

- Skin (لون البشرة)
- Eyes (لون العين)
- Marital status (الحالة الزوجية)
- Nationality (القومية)
- Marks (العلامات)
- Gender (نوع الجنس)
- Country (الإقليم)

وبلغت حجم العينة (20000) وتم كتابة برنامج الجانب التطبيق باستعمال برمجة (MATLAB R2017B) وكذلك باستعمال برنامج (SPSS V24) عند تطبيق التحليل العنقودي في التصنيف.

2- نتائج تطبيق طريقة Wls باستعمال الانحدار اللوجستي

عند تطبيق هذه الطريقة على بيانات التطبيق العملي ظهرت النتائج التالية:

جدول (9-4) نتائج الانحدار اللوجستي بطريقة (Wls)

المتغيرات المستقلة	المتغير المعتمد	قيمة المعالم	اختبار t	الدلالة المعنوية
Skin (لون البشرة)	Gender (الجنس)	0.003	1.451	غير دال (5%) ودال (10%)
Eyes (لون العين)		0.006	1.788	دال
marital_status (الحالة الزوجية)		-0.001	0.432	غير دال
Nationality (القومية)		-0.004	1.019	غير دال
Marks (العلامات)		-0.0009	0.673	غير دال
قيمة الثابت		1.052	0.021	غير دال

قيمة t الجدولية عند مستوى دلالة (0.05) ودرجة حرية (19998) = 1.65

قيمة t الجدولية عند مستوى دلالة (0.10) ودرجة حرية (19998) = 1.28

المصدر: (اعداد الباحث بواسطة البرنامج الاحصائي ماتلاب اصدار 2017)

من خلال الجدول (9-4) نلاحظ بان:

- المتغير (skin) : غير دال عند مستوى دلالة (0.05) اما عند مستوى دلالة (0.10) فقد ظهرت دلالة معنوية حيث بلغت قيمة t المحسوبة (1.45) وهي اكبر من قيمتها الجدولية والبالغة (1.28)، وقد بلغت قيمة معامل بيتا (0.003) وهي قيمة موجبة اي ان العلاقة بين المتغير (skin) والمتغير (gender) علاقة طردية.

- المتغير (eyes) : دال عند مستوى دلالة (0.10) و مستوى دلالة (0.05) حيث بلغت قيمة t المحسوبة (1.788) وهي اكبر من قيمتها الجدولية والبالغة (1.28) عند المستوى (0.10) والبالغة (1.65) عند مستوى دلالة (0.05) ، وقد بلغت قيمة معامل بيتا (0.006) وهي قيمة موجبة اي ان العلاقة بين المتغير (eyes) والمتغير (gender) علاقة طردية.
- المتغير (marital status) : غير دال عند مستوى دلالة (0.10) و مستوى دلالة (0.05) حيث بلغت قيمة t المحسوبة (0.432) وهي اصغر من قيمتها الجدولية والبالغة (1.28) عند المستوى (0.10) والبالغة (1.65) عند مستوى دلالة (0.05) ، وقد بلغت قيمة معامل بيتا (0.001−) وهي قيمة سالبة اي ان العلاقة بين المتغير (marital status) والمتغير (gender) علاقة عكسية.
- المتغير (Nationality) : غير دال عند مستوى دلالة (0.10) و مستوى دلالة (0.05) حيث بلغت قيمة t المحسوبة (1.019) وهي اصغر من قيمتها الجدولية والبالغة (1.28) عند المستوى (0.10) والبالغة (1.65) عند مستوى دلالة (0.05) ، وقد بلغت قيمة معامل بيتا (0.004−) وهي قيمة سالبة اي ان العلاقة بين المتغير (Nationality) والمتغير (gender) علاقة عكسية.
- المتغير (marks) : غير دال عند مستوى دلالة (0.10) و مستوى دلالة (0.05) حيث بلغت قيمة t المحسوبة (0.673) وهي اصغر من قيمتها الجدولية والبالغة (1.28) عند المستوى (0.10) والبالغة (1.65) عند مستوى دلالة (0.05) ، وقد بلغت قيمة معامل بيتا (0.0009−) وهي قيمة سالبة اي ان العلاقة بين المتغير (marks) والمتغير (gender) علاقة عكسية.

كذلك نلاحظ بان قيمة الثابت لم يظهر بها دلالة معنوية حيث بلغت قيمة t المحسوبة (0.673) وهي اصغر من قيمتها الجدولية والبالغة (1.28) عند المستوى (0.10) والبالغة (1.65) عند مستوى دلالة (0.05) وقد بلغت قيمة الثابت (1.052) وهي قيمة موجبة اما النموذج العام فهو:

$$Y=1.052 + 0.003X1 + 0.006X2 - 0.001X3 - 0.004X4 - 0.0009X5$$

حيث ان:

Y: يمثل gender

X1: يمثل skin

X2: يمثل eyes

X3: يمثل marital status

X4: يمثل Nationality

X5: يمثل marks

3- تطبيق الخوارزمية الجينية :

عند تطبيق هذه الطريقة على بيانات التطبيق العملي فظهرت النتائج كما يلي:

جدول (10-4) نتائج الانحدار اللوجستي بطريقة الخوارزمية الجينية

المتغيرات المستقلة	المتغير المعتمد	قيمة المعالم	اختبار t	الدلالة
Skin (لون البشرة)	Gender (الجنس)	-0.227	17.06	دال
Eyes (لون العين)		0.108	8.31	دال
marital_status (الحالة الزوجية)		-0.081	4.84	دال
Nationality (القومية)		-0.513	44.85	دال
Marks (العلامات)		0.432	29.14	دال
قيمة الثابت		-0.242	0.237	غير دال

قيمة t الجدولية عند مستوى دلالة (0.05) ودرجة حرية (19998) = 1.65

قيمة t الجدولية عند مستوى دلالة (0.10) ودرجة حرية (19998) = 1.28

المصدر: (اعداد الباحث بواسطة البرنامج الاحصائي ماتلاب اصدار 2017)

من خلال الجدول (10-4) نلاحظ بان:

- المتغير (skin) : دال عند مستوى دلالة (0.10) ومستوى دلالة (0.05) حيث بلغت قيمة t المحسوبة (17.06) وهي اكبر من قيمتها الجدولية والبالغة (1.28) عند المستوى (0.10) والبالغة (1.65) عند مستوى دلالة (0.05) ، وقد بلغت قيمة معامل بيتا (-0.227) وهي قيمة سالبة اي ان العلاقة بين المتغير (skin) والمتغير (gender) علاقة عكسية.

- المتغير (eyes) : دال عند مستوى دلالة (0.10) و مستوى دلالة (0.05) حيث بلغت قيمة t المحسوبة (8.31) وهي اكبر من قيمتها الجدولية والبالغة (1.28) عند المستوى (0.10) والبالغة (1.65) عند مستوى دلالة (0.05) ، وقد بلغت قيمة معامل بيتا (0.108) وهي قيمة موجبة اي ان العلاقة بين المتغير (eyes) والمتغير (gender) علاقة طردية.
- المتغير (marital status) : دال عند مستوى دلالة (0.10) ومستوى دلالة (0.05) حيث بلغت قيمة t المحسوبة (4.84) وهي اكبر من قيمتها الجدولية والبالغة (1.28) عند المستوى (0.10) والبالغة (1.65) عند مستوى دلالة (0.05) ،وقد بلغت قيمة معامل بيتا (0.081−) وهي قيمة سالبة اي ان العلاقة بين المتغير (marital status) والمتغير (gender) علاقة عكسية.
- المتغير (Nationality) : دال عند مستوى دلالة (0.10) ومستوى دلالة (0.05) حيث بلغت قيمةtالمحسوبة (44.85) وهي اكبر من قيمتها الجدولية والبالغة (1.28) عند المستوى (0.10) والبالغة (1.65) عند مستوى دلالة (0.05) ،وقد بلغت قيمة معامل بيتا (0.513−) وهي قيمة سالبة اي ان العلاقة بين المتغير (Nationality) والمتغير (gender) علاقة عكسية.
- المتغير (marks) : دال عند مستوى دلالة (0.10) ومستوى دلالة (0.05) حيث بلغت قيمةt المحسوبة (29.14) وهي اكبر من قيمتها الجدولية والبالغة (1.28) عند المستوى (0.10) والبالغة (1.65) عند مستوى دلالة (0.05) ،وقد بلغت قيمة معامل بيتا (0.432) وهي قيمة موجبة اي ان العلاقة بين المتغير (marks) والمتغير (gender) علاقة طردية.

كذلك نلاحظ بان قيمة الثابت لم يظهر بها دلالة معنوية حيث بلغت قيمة t المحسوبة (0.237) وهي اصغر من قيمتها الجدولية والبالغة (1.28) عند المستوى (0.10) والبالغة (1.65) عند مستوى دلالة (0.05) وقد بلغت قيمة الثابت (−0.242) وهي قيمة سالبة اما النموذج العام فهو:

$$Y=-0.242 - 0.227X1 + 0.108X2 - 0.081X3 - 0.513X4 + 0.432X$$

2-2-4 تطبيق التحليل العنقودي في تصنيف المتغيرات للبيانات الحقيقية

سيتم في هذا الجانب عرض نتائج التحليل العنقودي باستخدام الطرق الكلاسيكية في التصنيف والمتمثلة بجدول تحليل التباين، والمسافة بين مراكز المجموعات، ومتوسطات المجموعات في المجموعات المختلفة، فضلا عن مصفوفة القرابة وخطوات التجميع وعضوية المتغيرات وكما يلي:

1: جدول تحليل التباين

تم حساب جدول تحليل التباين باتجاه واحد بالنسبة لمتغيرات البحث وكما يلي:

جدول تحليل التباين باتجاه واحد بالنسبة لمتغيرات البحث

جدول (11-4) جدول تحليل التباين بالنسبة لمتغيرات البحث

	المجاميع		الخطأ			
	متوسط المربعات	درجة الحرية	متوسط المربعات	درجة الحرية	قيمة F	معنوية الاختبار
Zscore(skin) القيم المعيارية (البشرة)	1155.504	3	.827	19996	1397.579	.000
Zscore(eyes) القيم المعيارية (لون العين)	1956.445	3	.707	19996	2768.719	.000
Zscore(marital_status القيم المعيارية (الحالة الزوجية)	836.837	3	.875	19996	956.823	.000
Zscore(nationality_by) القيم المعيارية (القومية)	2293.610	3	.656	19996	3496.146	.000
Zscore(marks) القيم المعيارية (العلامات)	3965.711	3	.405	19996	9787.667	.000
Zscore(gender) القيم المعيارية (الجنس)	30.887	3	.996	19996	31.027	.000
هنا تم استخدام اختبارات F فقط للأغراض الوصفية لأنه تم اختيار المجموعات لوصف الاختلافات بين الحالات في مجموعات مختلفة.						

المصدر: (اعداد الباحث بواسطة البرنامج الاحصائي SPSS ver. 20)

يلاحظ من الجدول (4-11) ، بان متوسط متغير الجنس له اقل فروق بين المجموعات وبما يعادل (31.027) في حين ان متوسط (marks) له اكبر فروق بين المجموعات وبما يعادل (9787.667) علما ان جميع المتغيرات معنوية تحت مستوى (0.05) وهذا يعني ان هنالك فروق معنوية بين المجاميع وغير متجانسة لمتغيرات الدراسة.

2: المسافة بين مراكز المجموعات (Distances between Final Cluster Centers)

من اجل قياس المسافة بين مراكز المجموعات للمحافظات وذلك لمعرفة مدى تقارب او تباعد تلك المجموعات عن بعضها البعض تم إيجاد ما يلي:

الجدول (4-12) المسافة بين مراكز المجموعات بالنسبة لمتغيرات الدراسة

المجاميع	1	2	3	4
1		2.055	2.088	1.935
2	2.055		1.934	2.078
3	2.088	1.934		2.045
4	1.935	2.078	2.045	

المصدر: (اعداد الباحث بواسطة البرنامج الاحصائي SPSS ver. 20)

يلاحظ من الجدول (4-12) ، بان متوسط المجموعة الثالثة ابعد ما يكون (2.088) عن محافظات المجموعة الأولى. في حين ان متوسط محافظات المجموعة الثانية أقرب ما يكون (1.934) عن محافظات المجموعة الثالثة.

3: متوسطات المتغيرات في المجموعات المختلفة

تم توزيع المتغيرات في أربع مجاميع وكما يلي:

- متوسطات المتغيرات في المجموعات المختلفة بالنسبة لمتغيرات الدراسة

يبين الجدول (4-13) متوسطات المتغيرات في المجموعات المختلفة بالنسبة لمتغيرات الدراسة

جدول (13-4) متوسطات المتغيرات في المجموعات المختلفة بالنسبة لمتغيرات الدراسة

	Cluster			
	1	2	3	4
Zscore(skin) القيم المعيارية (البشرة)	-.28970-	.49033	-.52025-	.31733
Zscore(eyes) القيم المعيارية (لون العين)	-.60920-	.43710	-.45634-	.63374
Zscore(marital_status القيم المعيارية (الحالة الزوجية)	-.25170-	-.43857-	.32796	.37355
Zscore(nationality_by) القيم المعيارية (القومية)	-.56521-	-.59600-	.55393	.62825
Zscore(marks) القيم المعيارية (العلامات)	.82931	-.73828-	-.81064-	.69892
Zscore(gender) القيم المعيارية (الجنس)	-.06014-	.10269	.02061	-.06235-

المصدر: (اعداد الباحث بواسطة البرنامج الاحصائي SPSS ver. 20)

يلاحظ من الجدول (13-4) بأن أعلى متوسط بالنسبة لمتغير (skin) كان ضمن محافظات المجموعة الثانية، حيث بلغ (0.49033) من الانحراف المعياري، أما بالنسبة لمتغير eyes كان ضمن محافظات المجموعة الثالثة وهكذا لبقية المتغيرات.

4: مصفوفة القرابة لمتغيرات البحث

جدول (14-4)

مصفوفة القرابة Proximity Matrix باستخدام الطريقة الهرمية للمفردات (المحافظات)

Proximity Matrix

Case	Skin (البشرة)	Eyes (لون العين)	marital_status (الحالة الزوجية)	nationality_ (القومية)	Marks (العلامات)	Gender (الجنس)
Skin (البشرة)	1.000	.006	.015	.000	.002	.010
Eyes (لون العين)	.006	1.000	.016	.000	.003	.013
marital_status (الحالة الزوجية)	.015	.016	1.000	.002	.002	.003
nationality (القومية)	.000	.000	.002	1.000	.002	.007
Marks (العلامات)	.002	.003	.002	.002	1.000	.005
Gender (الجنس)	.010	.013	.003	.007	.005	1.000

المصدر: (اعداد الباحث بواسطة البرنامج الاحصائي SPSS ver. 20)

من الجدول (14-4) بان أقرب مسافة بين المتغيرات كانت بين المتغيرين marks skin اذ بلغت 0.002 مقاسة بمربع المسافة الاقليدية

5: خطوات التجميع بالنسبة لمتغيرات الدراسة

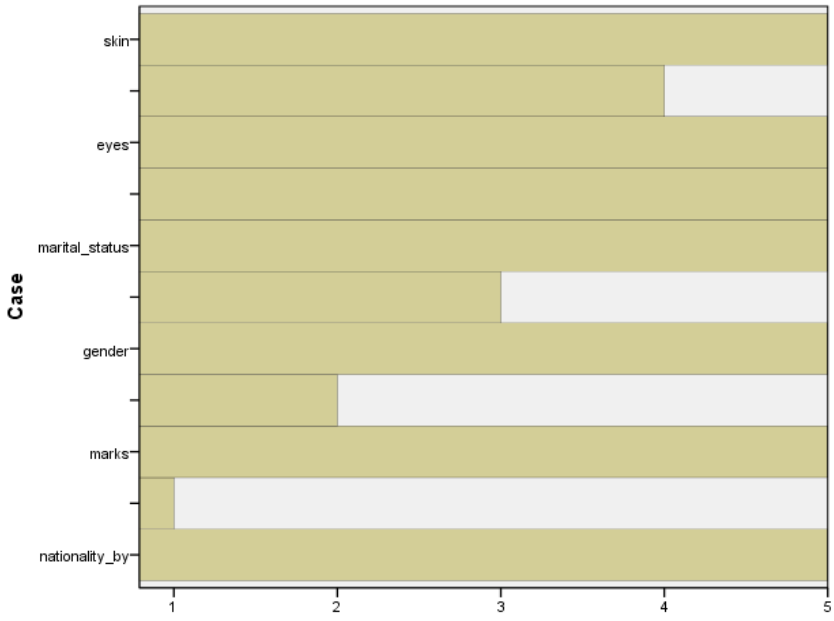
الجدول (15-4) خطوات التجميع:

جدول (4-15) خطوات التجميع (Agglomeration Schedule) بالطريقة الهرمية

الخطوة التالية	ظهور المجموعة للمرة الاولى		المعاملات	دمج المجاميع		الخطوات
	2 المجموعة	1 المجموعة		المجموعة 2	المجموعة 1	
2	0	0	.016	3	2	1
3	1	0	.010	2	1	2
4	0	2	.009	6	1	3
5	0	3	.003	5	1	4
0	0	4	.002	4	1	5

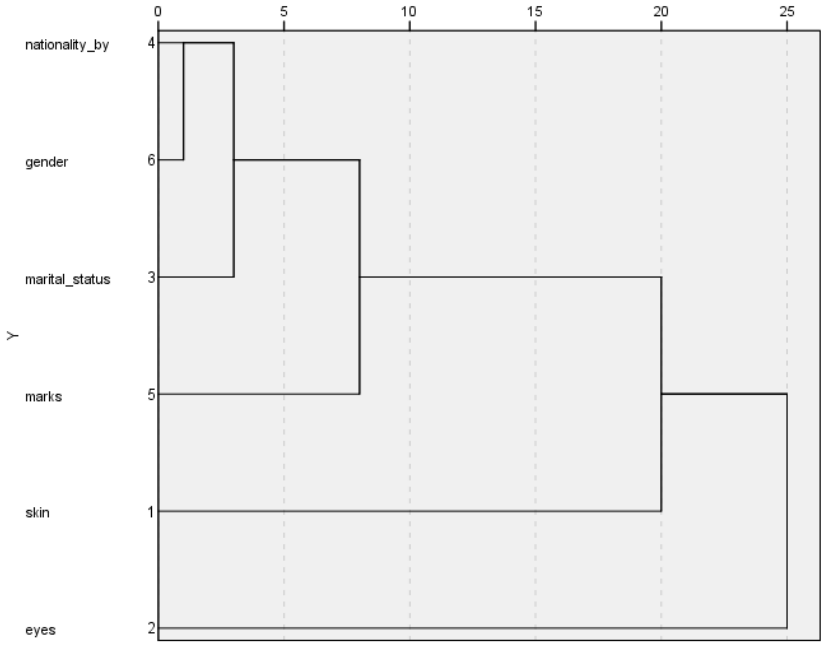
المصدر: (اعداد الباحث بواسطة البرنامج الاحصائي SPSS ver. 20)

من الجدول (4-15) انه في الخطوة الأولى تم دمج المفردة الثانية والمتمثلة بالمتغير eyes بالمفردة الثالثة والمتمثلة بالمفردة (marital status) كون المسافة بينهما والمقاسة بمربع المسافة الاقليدية اقل ما يمكن والبالغة (0.016) وبعد ذلك يتم الانتقال الى الخطوة الثانية والمتمثلة بدمج المفردة الأولى والمتمثلة بالمتغير (skin) مع المجموعة الثانية والمتشكلة ضمن الخطوة الأولى وبعد ذلك يتم الانتقال الى الخطوة الثالثة وهكذا الى بقية الخطوات وكما موضح بالشكلين التاليين.



شكل (4-1) الالواح الجليدية باستخدام الطريقة الهرمية

المصدر: (اعداد الباحث بواسطة البرنامج الاحصائي SPSS ver. 20)



شكل (3-10) المخطط الهرمي لمتغيرات الدراسة

المصدر: (اعداد الباحث بواسطة البرنامج الاحصائي SPSS ver. 20)

6:عضوية المتغيرات

يبين الجدول التالي عضوية كل متغير بالمجموعة التي تنتمي اليه

جدول (16-4) عضوية كل متغير بالمجموعة التي تنتمي اليه

عضوية المجاميع	
Case	المجموعة 4
Skin(لون البشرة)	1
Eyes(لون العين)	1
marital_status (الحالة الزوجية)	1
nationality_by (القومية)	2
Marks (العلامات)	3
Gender (الجنس)	4

المصدر: (اعداد الباحث بواسطة البرنامج الاحصائي SPSS ver. 20)

فعلى سبيل المثال فان المتغير (skin) ولون العين والحالة ينتمي الى المجموعة الأولى وبالنسبة القومية والعلامات ونوع الجنس تعنقدت كل في مجموعة منفردة، حسب درجة التقارب.

الفصل الخامس

أولا: النتائج

ثانيا: التوصيات

أولا: النتائج

بعد اجراء الدراسة النظرية والتطبيقية، توصل الباحث للنتائج التالية:

5) من نتائج الجانب التجريبي ظهرت افضل طريقة هي طريقة المربعات الصغرى الموزونة من الطرق الكلاسيكية في تقدير الانحدار اللوجيستي.
6) ظهرت طريقة الإمكان الأعظم المعتمدة على الخوارزمية الجينية كأفضل طريقة ولجميع حجوم العينات في تقدير الانحدار اللوجيستي.
7) تفوقت جميع الطرق المعتمدة على الخوارزمية الجينية من الطرق الكلاسيكية.
8) في الجانب التطبيقي ظهرت المتغيرات لون العين، لون البشرة ذات تأثير معنوي على نوع الجنس بالطرق الكلاسيكية.
9) في الجانب التطبيقي عند تطبيق الخوارزمية الجينية اظهرت المتغيرات المستقلة (skin,eyes,martial-status,Nationality,marks) معنوية إحصائية تحت مستوى (10%،5%) باستثناء قيمة الحد الثابت غير دال احصائيا عند نفس مستوى المعنوية.
10) متوسط متغير الجنس له اقل فروق بين المجموعات وبما يعادل (31.027) في حين ان متوسط (marks) له اكبر فروق بين المجموعات وبما يعادل (9787.667) علما ان جميع المتغيرات معنوية تحت مستوى (0.05) وهذا يعني ان هنالك فروق معنوية بين المجاميع وغير متجانسة لمتغيرات الدراسة.
11) تم تصنيف الأقاليم الى اربع مجاميع (عناقيد) حسب درجة القرابة والمقاسة بمربع المسافة الاقليدية
12) ظهرت بعض المتغيرات في المجموعة الأولى والمتمثلة بكل من (skin,eyes,martial status) وهذا يدل على تجانسها ، في تعنقد واحد، وباقي المتغيرات كل في تعنقدات مختلفة.

ثانيا: التوصيات:

من النتائج السابقة، قد توصل الباحث الى التوصيات التالية:

1) نوصي باستخدام الطرق المعتمدة على الخوارزمية الجينية في تقدير معلمات الانحدار اللوجيستي.
2) يمكن اعتماد نتائج خوارزميات العنقدة في التنقيب عن البيانات خاصة عندما تكون البيانات كبيرة جدا، حيث تتمكن النتائج من الوصول الى المفردة المطلوبة بعد المرور بالطرق الأقصر لهذه النتائج.
3) ان تقنيات التحليل العنقودي يمكن ان تتعامل مع البيانات الحكومية الالكترونية والذكية وهنا يتطلب اعتماد نتائج العنقدة في الوصول الى قواعد البيانات الضخمة.
4) استعمال خوارزميات الذكاء الاصطناعي (Artificial intelligence) ومنها الخوارزمية الجينية مع خوارزميات التحليل العنقودي في تنقيب البيانات، وهذا ما يساهم في تقليل الجهد والوقت في الوصول الى المعلومة.
5) يمكن استعمال خوارزميات جينية مهجنة في دراسات مشابهة لأغراض تحليل البيانات ذات الكم الهائل للبحث عن المعلومة المطلوبة.
6) نوصي بإضافة حقول أخرى لجواز السفر الالكتروني لضمان زيادة رصانة وامنية الجواز ومنها تخزين معلومات السفر في صور رقمية مثل التأشيرات الالكترونية واختام الدخول/ الخروج لدعم إجراءات اكثر فاعلية ورصانة لجواز السفر سيما تقنية التعرف على الوجه باستخدام البيانات البيوميترية.

المصادر والمراجع

قائمة المصادر والمراجع

- **القران الكريم**

أولا: المراجع باللغة العربية

1/ الكتب

1- احسن، طيار وشلابي، عمار ، (2007) " التنقيب في البيانات واتخاذ القرارات" كلية العلوم الاقتصادية، جامعة 20 اوت.

2- العلاق، بشير عباس ، (2005)" الإدارة الرقمية والتطبيقات" مركز الامارات للدراسات والبحوث الذاتية، العدد 83

3- العلي، عبد القادر وقنديل، عامر إبراهيم والطمري، غسان .2006 " المدخل الى إدارة المعرفة وإدارة السيرة" جامعة حلب

4- مصطفى، فؤاد عبيد (1995) " تقنيات التنقيب في قواعد البيانات واستكشاف في المعلومات المخبأة فيها" جامعة بغداد.

2/ البحوث المنشورة

5- الاسدي، نضال حسين. خروفه، شهلة حازم احمد، (2007)" استخدام الخوارزمية الجينية المهجنة لتصنيف صور الأقمار الصناعية" المجلة العراقية للعلوم الإحصائية العدد 11،.

6- اسماء، ط.د. بن عبد الله، (2017) استثمار تكنلوجيا المعلومات والاتصالات في مجال الخدمة العمومية (ضعف الأداء المؤسسي كدافع للاستثمار) دراسة استكشافية حول مشروع جواز السفر البيوميتري في الجزائر.

7- أيوب، نور كاظم ، (2017)" دالة صلاحية جديدة لاستخدام الخوارزمية الجينية في كسر شفرات النصوص العربية والإنكليزية المشفرة بطريقة نابساك – ميركل هيلمن " مجلة جامعة بابل ، العلوم الصرفة والتطبيقية العدد 6.

8- بحبوح، أسامة اسعد، (2007)" تأثير متغيرات الخوارزميات الجينية في مسائل إيجاد الحل الأمثل" مجلة جامعة دمشق للعلوم الهندسية، المجلد الثالث والعشرون، العدد الثاني،.

9- البدراني، مها عبد الاله محمد ، (2007) " استخدام الخوارزمية الجينية في تطابق أنماط الحرف الإنكليزي" قسم البرمجيات كلية علوم الحاسبات والرياضيات جامعة الموصل.

10- الجبور، دكتور محمد عودة ، (1987) مكافحة تزوير الجوازات ووثائق السفر، دار النشر بالمركز العربي للدراسات الأمنية والتدريب بالرياض.

11- الجمل، محمد محمود محمد، 2009 " الثروة الحيوانية والمراعي في بعض قرى شمال محافظة نابلس" جامعة النجاح الوطنية- كلية الدراسات العليا
12- حجوز، محمد مصطفى، 2014" تحسين خوارزميات K-MEANS" مجلة جامعة البعث – المجلد 36 – العدد 6-
13- خروفه، شهلة حازم احمد 2009" استخدام الخوارزمية الجينية لتحديد خصائص صور الأبنية" قسم العلوم الأساسية، كلية طب الاسنان ، جامعة الموصل ، العراق
14- الدوري، زكريا مطلب ، (2004) التنقيب في البيانات" الجامعة التكنلوجية
15- رحيمه، وليد عبد الله، (1995) " استخدام التحليل العنقودي وتحليل الانحدار في تشخيص امراض القلب" الجامعة المستنصرية.
16- الشمرتي، حامد سعد نور. التميمي، سهاد علي " استخدام التحليل العنقودي لتقييم مؤشرات البيئة والتنمية المستدامة في محافظات العراق للسنوات 2007-2011" مجلة الإدارة والاقتصاد، السنة الخامسة والثلاثون عدد 93 لسنة2012.
17- طه، حذيفة حازم. حسين، محمد زيد، 2012" استخدام التحليل العنقودي لتصنيف نوعية المياه الجوفية في ابار منطقة بعشيقة في محافظة نينوى" المجلة العراقية للعلوم الإحصائية.
18- عبد الستار العلي، عامر إبراهيم قنديلجي، غسان العمري،2006 المدخل إلى إدارة المعرفة، دار المسيرة للنشر و التوزيع و الطباعة، الطبعة الأولى، عمان، ، ص. 157.
19- عودة، حياة كاظم،2009 " دراسة تحليلية للمشكلات الإنتاجية والمالية والإدارية والتسويقية لمشاريع تربية فروج اللحم في محافظة الديوانية" مجلة الفرات للعلوم الزراعية.
20- قارطي، حورية، مداوي، ايمان ،2017، دراسة اثر استخدام الإدارة الالكترونية من طرف مصالح الخدمة العمومية في تحسين جودة الخدمة العمومية في الجزائر: دراسة استطلاعية حول خدمة استخراج جواز السفر البيوميتري، مجلة مجاميع المعرفة، الجمهورية الجزائرية العدد 4
21- الكريم، عوض محمد عوض،2012 " تصميم وتطبيق نظام كشف لاختراقات الشبكة مبني على العميل الذكي" جامعة السودان للعلوم والتكنولوجيا، كلية علوم الحاسوب وتقانة المعلومات، ص.ب. 407، الخرطوم، السودان.
22- الكلاك، اسراء نذير. شعبان، رقية زيدان،2008 " الخوارزمية الجينية في جدولة العمليات مع عدم إمكانية القطع" مجلة تنمية الرافدين العدد 89،
23- متراس، بان احمد حسن ، ثابت، همسة معن محمد2007 " استخدام الخوارزمية الجينية في حل مسألة النقل" المجلة العراقية للعلوم الإحصائية العدد 1 ،.

24- مجلة جامعة دمشق للعلوم الهندسية - المجلد الثالث والعشرون (1995)- العدد الثاني آ.بحبوح- ح. ريشة (ص 106).
25- مجلة وحدة البحث في تنمية وإدارة الموارد البشرية، المجلد 8، العدد 2
26- محمد، محمد بكري عبيد2015 " تحديد العوامل المؤثرة في مرض السكري باستخدام طرائق متعدد المتغيرات" دراسة حالة ولاية شمال كردفان" جامعة سودان للعلوم والتكنلوجيا، كلية الدراسات العليا.
27- هيثم يعقوب يوسف واخرون ، 2009" استخدام التحليل العنقودي لتقييم ابعاد دليل التنمية البشرية في العراق لعام " مجلة ديالى للبحوث الإنسانية 2011 مجلد .49

3/ شبكة المعلومات الدولية (الانترنت)

28- Ramachandran M, Pushpa , (2001).”Mining for gold Wipro technologies”, December . www.wipro.com

ثانيا: المراجع باللغة الانجليزية

29- A Self-Help Test 2002"DM Review Magazine", March. P 2

30- Alvin c.Rencher2002"Methods of multivariate analysis" second edition , John wiley and sons.

31- Canadas department of external affairs

32- Daft, Richard L. “Organization Theory and Design” Seventh Edition, South-Western College Publishing, U.S.A, 2001.p258

33- Daniel C Turack " The passport in international law" London ,Lexington books,1972, p.15

34- EC Media Group, August 2000.p 1

35- Going Abroad "Areport on passport" P 5

36- M.Halkidi & Y.Batistakis & M. Vazoirgiannis , (2001). "clustering Algorithm and usability measures”

37- Magalhães-Mendes, J. "A comparative study of crossover operators for genetic algorithms to solve the job shop scheduling problem."(page 168).

38- Magalhães-Mendes, J.. (2013)A comparative study of crossover operators for genetic algorithms to solve the job shop scheduling problem. WSEAS transactions on computers, 12(4), 164-173. (page167).

39- Melanie, M. "An introduction to genetic algorithms." (page 8)

40- Noonan, Jack. (2000)"Data mining strategies (DM Review magazine), EC media Group, July

41- Nopiah, Z. M., Khairir, M. I., Abdullah, S., Baharin, M. N., & Arifin, A. (2010, February). Time complexity analysis of the genetic algorithm clustering method. In Proceedings of the 9th WSEAS international conference on Signal processing, robotics and automation (pp. 171-176). World Scientific and Engineering Academy and Society (WSEAS). (page 171).

42- Romny,Marshall B, (2000)"Accounting information system"9th USA

43- Seiner, Robert S. : The IRM Test: How Did "IT" Get This Way?

44- Turack ;OP.cit,.p 17

45- Wolfgang Hardle, 2007"Multivariate statistics" Barlin and Praue

46- Wu, Jonathan: What is Data Mining?, "DM Review Magazine",

الملاحق

يضم البرامج المصممة لتنفيذ الجانب التطبيقي

أولا: برنامج المحاكاة للجانب التجريبي باستخدام برنامج Matlab R2017b

ثانيا: البرنامج التطبيقي للبيانات الاصلية Matlab R2017b

البرنامج الأول (برنامج المحاكاة) البرنامج التجريبي

```
clc, clear all
n=input('n=');
r=1000;
p=10;
rand('state',0);
randn('state',0);
%Bo=0.007;B1=0.231;B2=0.024;B3=0.006;B4=0.085;B5=0.065;B6=0.017;B7=0.
038;B8=0.097;B9=0.001;B10=0.055;
Bo=0.007;B1=0.012;B2=0.023;B3=0.005;B4=0.018;B5=0.004;B6=0.006;B7=0.0
57;B8=0.009;B9=0.049;B10=0.087;
B=[B1;B2;B3;B4;B5;B6;B7;B8;B9;B10];
ii=ones(n,1);
for j=1:r
x=random('uniform',0,1,n,p);
pi=exp(Bo+B1.*x(:,1)+B2.*x(:,2)+B3.*x(:,3)+B4.*x(:,4)+B5.*x(:,5)+B6.*
x(:,6)+B7.*x(:,7)+B8.*x(:,8)+B9.*x(:,9)+B10.*x(:,10))./(1+exp(Bo+B1.*
x(:,1)+B2.*x(:,2)+B3.*x(:,3)+B4.*x(:,4)+B5.*x(:,5)+B6.*x(:,6)+B7.*x(:
,7)+B8.*x(:,8)+B9.*x(:,9)+B10.*x(:,10)));
for i=1:n
ei(i)=binornd(1,pi(i));
end
ZZ=exp(x*B)./(1+exp(x*B))+ei';
for i=1:n
    if ZZ(i)>=0.5
        y(i)=1;
    else if ZZ(i)<0.5
        y(i)=0;
        end
    end
end
y=floor(ZZ);

%---------------------------------
%disp(' MLE Methods  ')
%disp('----------------------')
Bm=inv(x'*x)*x'*y;
pim=exp(x*Bm)./(1+exp(x*Bm));
vm=diag(pim.*(1-pim));
pm=pim;
Bhadm=Bm+inv(x'*vm*x)*x'*(y-pm);
yhad=Bhadm(1).*x(:,1)+Bhadm(2).*x(:,2)+Bhadm(3).*x(:,3)+Bhadm(4).*x(:
,4)+Bhadm(5).*x(:,5)+Bhadm(6).*x(:,6)+Bhadm(7).*x(:,7)+Bhadm(8).*x(:,
8)+Bhadm(9).*x(:,9)+Bhadm(10).*x(:,10);
mm=0;
for i=1:n

yhad(i)=Bhadm(1).*x(i,1)+Bhadm(2).*x(i,2)+Bhadm(3).*x(i,3)+Bhadm(4).*
x(i,4)+Bhadm(5).*x(i,5)+Bhadm(6).*x(i,6)+Bhadm(7).*x(i,7)+Bhadm(8).*x
(i,8)+Bhadm(9).*x(i,9)+Bhadm(10).*x(i,10);
    mseMLE=mm+(yhad(i)-y(i))^2;
end
mseMLE=(mseMLE/(n-p));

% start gentic algorithm
%----------------------------
```

```
 bb=rand;
 Pm=length(y);
 popsize=20;
parant2=rand(19,10);
%%%%%%%%%%%%%%%%%%%%%%
FF1=Bhadm(1);
FF2=Bhadm(2);
FF3=Bhadm(3);
FF4=Bhadm(4);
FF5=Bhadm(5);
FF6=Bhadm(6);
FF7=Bhadm(7);
FF8=Bhadm(8);
FF9=Bhadm(9);
FF10=Bhadm(10);
parant1=[FF1  FF2 FF3 FF4 FF5 FF6 FF7 FF8 FF9 FF10];

%crossover using Arthmitac
%----------------------------
childd11=bb*parant1+(1-bb)*parant2(1,:);%الكروموسوم الاول
childd12=(1-bb)*parant1+bb*parant2(1,:);%الكروموسوم الثاني
%%%%%%%%%%%%%%%%%%
childd21=bb*parant1+(1-bb)*parant2(2,:);%الكروموسوم الثالث
childd22=(1-bb)*parant1+bb*parant2(2,:);%الكروموسوم الرابع
%%%%%%%%%%%%%%%%
childd31=bb*parant1+(1-bb)*parant2(3,:);%الكروموسوم الخامس
childd32=(1-bb)*parant1+bb*parant2(3,:);%الكروموسوم السادس
%%%%%%%%%%%%%%%
childd41=bb*parant1+(1-bb)*parant2(4,:);%الكروموسوم السابع
childd42=(1-bb)*parant1+bb*parant2(4,:);%الكروموسوم الثامن
%%%%%%%%%%%%%%%
childd51=bb*parant1+(1-bb)*parant2(5,:);%الكروموسوم التاسع
childd52=(1-bb)*parant1+bb*parant2(5,:);%الكروموسوم العاشر
%%%%%%%%%%%%%%%
childd61=bb*parant1+(1-bb)*parant2(6,:);%الكروموسوم الحادي عشر
childd62=(1-bb)*parant1+bb*parant2(6,:);%الكروموسوم الثاني عشر
%%%%%%%%%%%%%%%
childd71=bb*parant1+(1-bb)*parant2(7,:);%الكروموسوم الثالث عشر
childd72=(1-bb)*parant1+bb*parant2(7,:);
%%%%%%%%%%%%%%%
childd81=bb*parant1+(1-bb)*parant2(8,:);
childd82=(1-bb)*parant1+bb*parant2(8,:);
%%%%%%%%%%%%%%%
childd91=bb*parant1+(1-bb)*parant2(9,:);
childd92=(1-bb)*parant1+bb*parant2(9,:);
%%%%%%%%%%%%%%%
childd101=bb*parant1+(1-bb)*parant2(10,:);
childd102=(1-bb)*parant1+bb*parant2(10,:);
%%%%%%%%%%%%%%%
childd111=bb*parant1+(1-bb)*parant2(11,:);
childd112=(1-bb)*parant1+bb*parant2(11,:);
%%%%%%%%%%%%%%%
childd121=bb*parant1+(1-bb)*parant2(12,:);
childd122=(1-bb)*parant1+bb*parant2(12,:);
%%%%%%%%%%%%%%%
childd131=bb*parant1+(1-bb)*parant2(13,:);
childd132=(1-bb)*parant1+bb*parant2(13,:);
%%%%%%%%%%%%%%%
childd141=bb*parant1+(1-bb)*parant2(14,:);
childd142=(1-bb)*parant1+bb*parant2(14,:);
```

```
%%%%%%%%%%%%%%%
childd151=bb*parant1+(1-bb)*parant2(15,:);
childd152=(1-bb)*parant1+bb*parant2(15,:);
%%%%%%%%%%%%%%%
childd161=bb*parant1+(1-bb)*parant2(16,:);
childd162=(1-bb)*parant1+bb*parant2(16,:);
%%%%%%%%%%%%%%%
childd171=bb*parant1+(1-bb)*parant2(17,:);
childd172=(1-bb)*parant1+bb*parant2(17,:);
%%%%%%%%%%%%%%%
childd181=bb*parant1+(1-bb)*parant2(18,:);
childd182=(1-bb)*parant1+bb*parant2(18,:);
%%%%%%%%%%%%%%%
childd191=bb*parant1+(1-bb)*parant2(19,:);
childd192=(1-bb)*parant1+bb*parant2(19,:);
%%%%%%%%%%%%%%%
A1=[childd11 ;childd12];
A2=[childd21 ;childd22];
A3=[childd31 ;childd32];
A4=[childd41 ;childd42];
A5=[childd51 ;childd52];
A6=[childd61 ;childd62];
A7=[childd71 ;childd72];
A8=[childd81 ;childd82];
A9=[childd91 ;childd92];
A10=[childd101; childd102];
A11=[childd111 ;childd112];
A12=[childd121 ;childd122];
A13=[childd131 ;childd132];
A14=[childd141 ;childd142];
A15=[childd151 ;childd152];
A16=[childd161 ;childd162];
A17=[childd171 ;childd172];
A18=[childd181 ;childd182];
A19=[childd191 ;childd192];
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%


CA11=[A1;A2;A3;A4;A5;A6;A7;A8;A9;A10;A11;A12;A13;A14;A15;A16;A17;A18;
A19] ; % مصفوفة الابناء بعد اجراء الطفرة
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%


% comput fitness for each chrom.
%--------------------------------
for i=1:38
    if CA11(i,:)==[CA11(i,1) CA11(i,2) CA11(i,3) CA11(i,4) CA11(i,5)
CA11(i,6) CA11(i,7) CA11(i,8) CA11(i,9) CA11(i,10)]
    J(i)=0;
    for h=1:n
       J(i)=J(i)+(y(h)-
((CA11(i,1)*x(h,1))+(CA11(i,2)*x(h,2))+(CA11(i,3)*x(h,3))+(CA11(i,4)*
x(h,4))+(CA11(i,5)*x(h,5))+(CA11(i,6)*x(h,6))+(CA11(i,7)*x(h,7))+(CA1
1(i,8)*x(h,8))+(CA11(i,9)*x(h,9))+(CA11(i,10)*x(h,10))));

    end
    end
end
CA11(:,11)=J';
```

```
CA11;
FF=0;
kF1=min(CA11(:,11));
% comput MSE for GE Algrothem
%------------------------------
for i=1:38
    for j=1:11
        if kF1==CA11(i,j)
            FF=FF+1;
        end
    end
    if FF==1
       for j=1:11
           if kF1==CA11(i,j)
              ffg1mle=([CA11(i,j-10) CA11(i,j-9) CA11(i,j-8)
CA11(i,j-7) CA11(i,j-6) CA11(i,j-5) CA11(i,j-4) CA11(i,j-3) CA11(i,j-
2) CA11(i,j-1)]);
              M1=0;
              for i=1:n

yhad2(i)=ffg1mle(1).*x(i,1)+ffg1mle(2).*x(i,2)+ffg1mle(3).*x(i,3)+ffg
1mle(4).*x(i,4)+ffg1mle(5).*x(i,5)+ffg1mle(6).*x(i,6)+ffg1mle(7).*x(i
,7)+ffg1mle(8).*x(i,8)+ffg1mle(9).*x(i,9)+ffg1mle(10).*x(i,10);
                  M1=M1+(y(i)-yhad2(i))^2;
              end
              MSEMleG=M1/(n*Pm*popsize-p);

           end
       end

    end
end
 %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

%-------------------------------------
%disp(' WLS Methods  ')
%disp('----------------------')
wi=diag(pim.*(1-pim));
zi=log(pim./(1-pim));
Bhadw=inv(x'*wi*x)*x'*wi*zi;
yhad=Bhadw(1).*x(:,1)+Bhadw(2).*x(:,2)+Bhadw(3).*x(:,3)+Bhadw(4).*x(:
,4)+Bhadw(5).*x(:,5)+Bhadw(6).*x(:,6)+Bhadw(7).*x(:,7)+Bhadw(8).*x(:,
8)+Bhadw(9).*x(:,9)+Bhadw(10).*x(:,10);
mm=0;
for i=1:n

yhad(i)=Bhadw(1).*x(i,1)+Bhadw(2).*x(i,2)+Bhadw(3).*x(i,3)+Bhadw(4).*
x(i,4)+Bhadw(5).*x(i,5)+Bhadw(6).*x(i,6)+Bhadw(7).*x(i,7)+Bhadw(8).*x
(i,8)+Bhadw(9).*x(i,9)+Bhadw(10).*x(i,10);
mseWLS=mm+(yhad(i)-y(i))^2;
end
mseWLS=(mseWLS/(n-p));

% start gentic algorithm
%------------------------------
 bb=rand(20,2);
parant2=rand(19,10);
%%%%%%%%%%%%%%%%%%%%%%%
FF1=Bhadw(1);
FF2=Bhadw(2);
```

```
FF3=Bhadw(3);
FF4=Bhadw(4);
FF5=Bhadw(5);
FF6=Bhadw(6);
FF7=Bhadw(7);
FF8=Bhadw(8);
FF9=Bhadw(9);
FF10=Bhadw(10);
parant1=[FF1  FF2 FF3 FF4 FF5 FF6 FF7 FF8 FF9 FF10];

%crossover using Arthmitac
%---------------------------
childd11=bb(1,1)*parant1+(1-bb(1,2))*parant2(1,:);%الكروموسوم الاول
childd12=(1-bb(1,1))*parant1+bb(1,2)*parant2(1,:);%الكروموسوم الثاني
%%%%%%%%%%%%%%%%%%
childd21=bb(2,1)*parant1+(1-bb(2,2))*parant2(2,:);%الكروموسوم الثالث
childd22=(1-bb(2,1))*parant1+bb(2,2)*parant2(2,:);%الكروموسوم الرابع
%%%%%%%%%%%%%%%%
childd31=bb(3,1)*parant1+(1-bb(3,2))*parant2(3,:);%الكروموسوم الخامس
childd32=(1-bb(3,1))*parant1+bb(3,2)*parant2(3,:);%الكروموسوم السادس
%%%%%%%%%%%%%%%
childd41=bb(4,1)*parant1+(1-bb(4,2))*parant2(4,:);%الكروموسوم السابع
childd42=(1-bb(4,1))*parant1+bb(4,2)*parant2(4,:);%الكروموسوم الثامن
%%%%%%%%%%%%%%%
childd51=bb(5,1)*parant1+(1-bb(5,2))*parant2(5,:);%الكروموسوم التاسع
childd52=(1-bb(5,1))*parant1+bb(5,2)*parant2(5,:);%الكروموسوم العاشر
%%%%%%%%%%%%%%%
childd61=bb(6,1)*parant1+(1-bb(6,2))*parant2(6,:);%الكروموسوم الحادي
عشر
childd62=(1-bb(6,1))*parant1+bb(6,2)*parant2(6,:);%الكروموسوم الثاني
عشر
%%%%%%%%%%%%%%%
childd71=bb(7,1)*parant1+(1-bb(7,2))*parant2(7,:);%الكروموسوم الثالث
عشر
childd72=(1-bb(7,1))*parant1+bb(7,2)*parant2(7,:);
%%%%%%%%%%%%%%%
childd81=bb(8,1)*parant1+(1-bb(8,2))*parant2(8,:);
childd82=(1-bb(8,1))*parant1+bb(8,2)*parant2(8,:);
%%%%%%%%%%%%%%%
childd91=bb(9,1)*parant1+(1-bb(9,2))*parant2(9,:);
childd92=(1-bb(9,1))*parant1+bb(9,2)*parant2(9,:);
%%%%%%%%%%%%%%%
childd101=bb(10,1)*parant1+(1-bb(10,2))*parant2(10,:);
childd102=(1-bb(10,1))*parant1+bb(10,2)*parant2(10,:);
%%%%%%%%%%%%%%%
childd111=bb(11,1)*parant1+(1-bb(11,2))*parant2(11,:);
childd112=(1-bb(11,1))*parant1+bb(11,2)*parant2(11,:);
%%%%%%%%%%%%%%%
childd121=bb(12,1)*parant1+(1-bb(12,2))*parant2(12,:);
childd122=(1-bb(12,1))*parant1+bb(12,2)*parant2(12,:);
%%%%%%%%%%%%%%%
childd131=bb(13,1)*parant1+(1-bb(13,2))*parant2(13,:);
childd132=(1-bb(13,1))*parant1+bb(13,2)*parant2(13,:);
%%%%%%%%%%%%%%%
childd141=bb(14,1)*parant1+(1-bb(14,2))*parant2(14,:);
childd142=(1-bb(14,1))*parant1+bb(14,2)*parant2(14,:);
%%%%%%%%%%%%%%%
childd151=bb(15,1)*parant1+(1-bb(15,2))*parant2(15,:);
childd152=(1-bb(15,1))*parant1+bb(15,2)*parant2(15,:);
%%%%%%%%%%%%%%%
childd161=bb(16,1)*parant1+(1-bb(16,2))*parant2(16,:);
```

```
childd162=(1-bb(16,1))*parant1+bb(16,2)*parant2(16,:);
%%%%%%%%%%%%%%%
childd171=bb(17,1)*parant1+(1-bb(17,2))*parant2(17,:);
childd172=(1-bb(17,1))*parant1+bb(17,2)*parant2(17,:);
%%%%%%%%%%%%%%%
childd181=bb(18,1)*parant1+(1-bb(18,2))*parant2(18,:);
childd182=(1-bb(18,1))*parant1+bb(18,2)*parant2(18,:);
%%%%%%%%%%%%%%%
childd191=bb(19,1)*parant1+(1-bb(19,2))*parant2(19,:);
childd192=(1-bb(19,1))*parant1+bb(19,2)*parant2(19,:);
%%%%%%%%%%%%%%%
A1=[childd11 ;childd12];
A2=[childd21 ;childd22];
A3=[childd31 ;childd32];
A4=[childd41 ;childd42];
A5=[childd51 ;childd52];
A6=[childd61 ;childd62];
A7=[childd71 ;childd72];
A8=[childd81 ;childd82];
A9=[childd91 ;childd92];
A10=[childd101; childd102];
A11=[childd111 ;childd112];
A12=[childd121 ;childd122];
A13=[childd131 ;childd132];
A14=[childd141 ;childd142];
A15=[childd151 ;childd152];
A16=[childd161 ;childd162];
A17=[childd171 ;childd172];
A18=[childd181 ;childd182];
A19=[childd191 ;childd192];
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%


CA11=[A1;A2;A3;A4;A5;A6;A7;A8;A9;A10;A11;A12;A13;A14;A15;A16;A17;A18;
A19] ; % مصفوفة الابناء بعد اجراء الطفرة
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%


% comput fitness for each chrom.
%--------------------------------
for i=1:38
    if CA11(i,:)==[CA11(i,1) CA11(i,2) CA11(i,3) CA11(i,4) CA11(i,5)
CA11(i,6) CA11(i,7) CA11(i,8) CA11(i,9) CA11(i,10)]
    J(i)=0;
    for h=1:n
       J(i)=J(i)+(y(h)-
((CA11(i,1)*x(h,1))+(CA11(i,2)*x(h,2))+(CA11(i,3)*x(h,3))+(CA11(i,4)*
x(h,4))+(CA11(i,5)*x(h,5))+(CA11(i,6)*x(h,6))+(CA11(i,7)*x(h,7))+(CA1
1(i,8)*x(h,8))+(CA11(i,9)*x(h,9))+(CA11(i,10)*x(h,10))));

    end
    end
end
CA11(:,11)=J';

CA11;
FF=0;
kF1=min(CA11(:,11));
% comput MSE for GE Algrothem
%--------------------------------
```

```
for i=1:38
    for j=1:11
        if kF1==CA11(i,j)
            FF=FF+1;
        end
    end
    if FF==1
       for j=1:11
           if kF1==CA11(i,j)
              ffg1w=([CA11(i,j-10) CA11(i,j-9) CA11(i,j-8) CA11(i,j-
7) CA11(i,j-6) CA11(i,j-5) CA11(i,j-4) CA11(i,j-3) CA11(i,j-2)
CA11(i,j-1)]);
              M1=0;
              for i=1:n

yhad2(i)=ffg1w(1).*x(i,1)+ffg1w(2).*x(i,2)+ffg1w(3).*x(i,3)+ffg1w(4).
*x(i,4)+ffg1w(5).*x(i,5)+ffg1w(6).*x(i,6)+ffg1w(7).*x(i,7)+ffg1w(8).*
x(i,8)+ffg1w(9).*x(i,9)+ffg1w(10).*x(i,10);
                  M1=M1+(y(i)-yhad2(i))^2;
              end
              MSEWLSG=M1/(n*Pm*popsize-p);

          end
       end

    end
end
 %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

 %--------------Principal Components methods------------
S=(1/n-1)*x'*x;
[eg ev]=eig(S);
V=eg;
D=ev;
Z=x*V;
%----principal components of data x----------
COEFF = pca(x);
evect=COEFF;
pc=x*evect;
varpc=std(Z);
eval=diag(varpc);
%----------matrix v for beta = V *g----------
a=zeros(p+1,p+1);
a(1,1)=1;
b=zeros(1,p);
%v=[a(:,1) [b;evect]];
v=evect;
%------------PCLR ALL pc----------------
op_mod_all=glmfit(pc,y,'binomial','link','logit');
g_all=op_mod_all(1:p);
Bhat_all=v*g_all;
pihat_allpc=exp(x*Bhat_all)./(1+exp(x*Bhat_all));
MSEallpc=nansum(pihat_allpc-y).^2/(n-p);



% start gentic algorithm
%-----------------------------
 bb=rand(20,2);
parant2=rand(19,10);
%%%%%%%%%%%%%%%%%%%%%%%%
```

```
FF1=Bhat_all(1);
FF2=Bhat_all(2);
FF3=Bhat_all(3);
FF4=Bhat_all(4);
FF5=Bhat_all(5);
FF6=Bhat_all(6);
FF7=Bhat_all(7);
FF8=Bhat_all(8);
FF9=Bhat_all(9);
FF10=Bhat_all(10);
parant1=[FF1  FF2 FF3 FF4 FF5 FF6 FF7 FF8 FF9 FF10];

%crossover using Arthmitac
%----------------------------
childd11=bb(1,1)*parant1+(1-bb(1,2))*parant2(1,:);%الكروموسوم الاول
childd12=(1-bb(1,1))*parant1+bb(1,2)*parant2(1,:);%الكروموسوم الثاني
%%%%%%%%%%%%%%%%%%
childd21=bb(2,1)*parant1+(1-bb(2,2))*parant2(2,:);%الكروموسوم الثالث
childd22=(1-bb(2,1))*parant1+bb(2,2)*parant2(2,:);%الكروموسوم الرابع
%%%%%%%%%%%%%%%%
childd31=bb(3,1)*parant1+(1-bb(3,2))*parant2(3,:);%الكروموسوم الخامس
childd32=(1-bb(3,1))*parant1+bb(3,2)*parant2(3,:);%الكروموسوم السادس
%%%%%%%%%%%%%%%%
childd41=bb(4,1)*parant1+(1-bb(4,2))*parant2(4,:);%الكروموسوم السابع
childd42=(1-bb(4,1))*parant1+bb(4,2)*parant2(4,:);%الكروموسوم الثامن
%%%%%%%%%%%%%%%%
childd51=bb(5,1)*parant1+(1-bb(5,2))*parant2(5,:);%الكروموسوم التاسع
childd52=(1-bb(5,1))*parant1+bb(5,2)*parant2(5,:);%الكروموسوم العاشر
%%%%%%%%%%%%%%%%
childd61=bb(6,1)*parant1+(1-bb(6,2))*parant2(6,:);%الكروموسوم الحادي
عشر
childd62=(1-bb(6,1))*parant1+bb(6,2)*parant2(6,:);%الكروموسوم الثاني
عشر
%%%%%%%%%%%%%%%%
childd71=bb(7,1)*parant1+(1-bb(7,2))*parant2(7,:);%الكروموسوم الثالث
عشر
childd72=(1-bb(7,1))*parant1+bb(7,2)*parant2(7,:);
%%%%%%%%%%%%%%%%
childd81=bb(8,1)*parant1+(1-bb(8,2))*parant2(8,:);
childd82=(1-bb(8,1))*parant1+bb(8,2)*parant2(8,:);
%%%%%%%%%%%%%%%%
childd91=bb(9,1)*parant1+(1-bb(9,2))*parant2(9,:);
childd92=(1-bb(9,1))*parant1+bb(9,2)*parant2(9,:);
%%%%%%%%%%%%%%%%
childd101=bb(10,1)*parant1+(1-bb(10,2))*parant2(10,:);
childd102=(1-bb(10,1))*parant1+bb(10,2)*parant2(10,:);
%%%%%%%%%%%%%%%%
childd111=bb(11,1)*parant1+(1-bb(11,2))*parant2(11,:);
childd112=(1-bb(11,1))*parant1+bb(11,2)*parant2(11,:);
%%%%%%%%%%%%%%%%
childd121=bb(12,1)*parant1+(1-bb(12,2))*parant2(12,:);
childd122=(1-bb(12,1))*parant1+bb(12,2)*parant2(12,:);
%%%%%%%%%%%%%%%%
childd131=bb(13,1)*parant1+(1-bb(13,2))*parant2(13,:);
childd132=(1-bb(13,1))*parant1+bb(13,2)*parant2(13,:);
%%%%%%%%%%%%%%%%
childd141=bb(14,1)*parant1+(1-bb(14,2))*parant2(14,:);
childd142=(1-bb(14,1))*parant1+bb(14,2)*parant2(14,:);
%%%%%%%%%%%%%%%%
childd151=bb(15,1)*parant1+(1-bb(15,2))*parant2(15,:);
childd152=(1-bb(15,1))*parant1+bb(15,2)*parant2(15,:);
```

```
%%%%%%%%%%%%%%%
childd161=bb(16,1)*parant1+(1-bb(16,2))*parant2(16,:);
childd162=(1-bb(16,1))*parant1+bb(16,2)*parant2(16,:);
%%%%%%%%%%%%%%%
childd171=bb(17,1)*parant1+(1-bb(17,2))*parant2(17,:);
childd172=(1-bb(17,1))*parant1+bb(17,2)*parant2(17,:);
%%%%%%%%%%%%%%%
childd181=bb(18,1)*parant1+(1-bb(18,2))*parant2(18,:);
childd182=(1-bb(18,1))*parant1+bb(18,2)*parant2(18,:);
%%%%%%%%%%%%%%%
childd191=bb(19,1)*parant1+(1-bb(19,2))*parant2(19,:);
childd192=(1-bb(19,1))*parant1+bb(19,2)*parant2(19,:);
%%%%%%%%%%%%%%%
A1=[childd11 ;childd12];
A2=[childd21 ;childd22];
A3=[childd31 ;childd32];
A4=[childd41 ;childd42];
A5=[childd51 ;childd52];
A6=[childd61 ;childd62];
A7=[childd71 ;childd72];
A8=[childd81 ;childd82];
A9=[childd91 ;childd92];
A10=[childd101; childd102];
A11=[childd111 ;childd112];
A12=[childd121 ;childd122];
A13=[childd131 ;childd132];
A14=[childd141 ;childd142];
A15=[childd151 ;childd152];
A16=[childd161 ;childd162];
A17=[childd171 ;childd172];
A18=[childd181 ;childd182];
A19=[childd191 ;childd192];
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%


CA11=[A1;A2;A3;A4;A5;A6;A7;A8;A9;A10;A11;A12;A13;A14;A15;A16;A17;A18;
A19] ; % مصفوفة الابناء بعد اجراء الطفرة
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%


% comput fitness for each chrom.
%--------------------------------
for i=1:38
    if CA11(i,:)==[CA11(i,1) CA11(i,2) CA11(i,3) CA11(i,4) CA11(i,5)
CA11(i,6) CA11(i,7) CA11(i,8) CA11(i,9) CA11(i,10)]
    J(i)=0;
    for h=1:n
       J(i)=J(i)+(y(h)-
((CA11(i,1)*x(h,1))+(CA11(i,2)*x(h,2))+(CA11(i,3)*x(h,3))+(CA11(i,4)*
x(h,4))+(CA11(i,5)*x(h,5))+(CA11(i,6)*x(h,6))+(CA11(i,7)*x(h,7))+(CA1
1(i,8)*x(h,8))+(CA11(i,9)*x(h,9))+(CA11(i,10)*x(h,10))));

    end
    end
end
CA11(:,11)=J';

CA11;
FF=0;
kF1=min(CA11(:,11));
```

```
% comput MSE for GE Algrothem
%------------------------------
for i=1:38
    for j=1:11
        if kF1==CA11(i,j)
            FF=FF+1;
        end
    end
    if FF==1
       for j=1:11
           if kF1==CA11(i,j)
              ffg1w=([CA11(i,j-10) CA11(i,j-9) CA11(i,j-8) CA11(i,j-
7) CA11(i,j-6) CA11(i,j-5) CA11(i,j-4) CA11(i,j-3) CA11(i,j-2)
CA11(i,j-1)]);
              M1=0;
              for i=1:n

yhad2(i)=ffg1w(1).*x(i,1)+ffg1w(2).*x(i,2)+ffg1w(3).*x(i,3)+ffg1w(4).
*x(i,4)+ffg1w(5).*x(i,5)+ffg1w(6).*x(i,6)+ffg1w(7).*x(i,7)+ffg1w(8).*
x(i,8)+ffg1w(9).*x(i,9)+ffg1w(10).*x(i,10);
                  M1=M1+(y(i)-yhad2(i))^2;
              end
              MSEPC=M1/(n*Pm*popsize-p);

          end
       end

    end
end
 %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

 %-------------------------------------
%disp(' Bayse Methods  ')
%disp('----------------------')

z1=exp(x*Bm)/1+exp(x*Bm);

aj=max((1./x).*log(abs(z1./(1-z1)))-sum(x*Bm));
bj=min((1./x).*log(abs(z1./(1-z1)))-sum(x*Bm));
BhadBays1=aj./n;
BhadBays2=bj./n;
yhad=BhadBays1(1).*x(:,1)+BhadBays1(2).*x(:,2)+BhadBays1(3).*x(:,3)+B
hadBays1(4).*x(:,4)+BhadBays1(5).*x(:,5)+BhadBays1(6).*x(:,6)+BhadBay
s1(7).*x(:,7)+BhadBays1(8).*x(:,8)+BhadBays1(9).*x(:,9)+BhadBays1(10)
.*x(:,10);
mm=0;
for i=1:n

yhad(i)=BhadBays1(1).*x(i,1)+BhadBays1(2).*x(i,2)+BhadBays1(3).*x(i,3
)+BhadBays1(4).*x(i,4)+BhadBays1(5).*x(i,5)+BhadBays1(6).*x(i,6)+Bhad
Bays1(7).*x(i,7)+BhadBays1(8).*x(i,8)+BhadBays1(9).*x(i,9)+BhadBays1(
10).*x(i,10);
mseBays1=mm+(yhad(i)-y(i))^2;
end
mseBays1=mseBays1/n;

yhad=BhadBays2(1).*x(:,1)+BhadBays2(2).*x(:,2)+BhadBays2(3).*x(:,3)+B
hadBays2(4).*x(:,4)+BhadBays2(5).*x(:,5)+BhadBays2(6).*x(:,6)+BhadBay
s2(7).*x(:,7)+BhadBays2(8).*x(:,8)+BhadBays2(9).*x(:,9)+BhadBays2(10)
.*x(:,10);
```

```
mm=0;
for i=1:n

yhad(i)=BhadBays2(1).*x(i,1)+BhadBays2(2).*x(i,2)+BhadBays2(3).*x(i,3
)+BhadBays2(4).*x(i,4)+BhadBays2(5).*x(i,5)+BhadBays2(6).*x(i,6)+Bhad
Bays2(7).*x(i,7)+BhadBays2(8).*x(i,8)+BhadBays2(9).*x(i,9)+BhadBays2(
10).*x(i,10);
mseBays2=mm+(yhad(i)-y(i))^2;
end
mseBays2=(mseBays2/(n-p));

% start gentic algorithm
%-----------------------------
 bb=rand;
parant2=rand(19,10);
%%%%%%%%%%%%%%%%%%%%%%%%
FF1=BhadBays2(1);
FF2=BhadBays2(2);
FF3=BhadBays2(3);
FF4=BhadBays2(4);
FF5=BhadBays2(5);
FF6=BhadBays2(6);
FF7=BhadBays2(7);
FF8=BhadBays2(8);
FF9=BhadBays2(9);
FF10=BhadBays2(10);
parant1=[FF1  FF2 FF3 FF4 FF5 FF6 FF7 FF8 FF9 FF10];

%crossover using Arthmitac
%-----------------------------
childd11=bb*parant1+(1-bb)*parant2(1,:);%الكروموسوم الاول
childd12=(1-bb)*parant1+bb*parant2(1,:);%الكروموسوم الثاني
%%%%%%%%%%%%%%%%%%%
childd21=bb*parant1+(1-bb)*parant2(2,:);%الكروموسوم الثالث
childd22=(1-bb)*parant1+bb*parant2(2,:);%الكروموسوم الرابع
%%%%%%%%%%%%%%%%
childd31=bb*parant1+(1-bb)*parant2(3,:);%الكروموسوم الخامس
childd32=(1-bb)*parant1+bb*parant2(3,:);%الكروموسوم السادس
%%%%%%%%%%%%%%%%
childd41=bb*parant1+(1-bb)*parant2(4,:);%الكروموسوم السابع
childd42=(1-bb)*parant1+bb*parant2(4,:);%الكروموسوم الثامن
%%%%%%%%%%%%%%%%
childd51=bb*parant1+(1-bb)*parant2(5,:);%الكروموسوم التاسع
childd52=(1-bb)*parant1+bb*parant2(5,:);%الكروموسوم العاشر
%%%%%%%%%%%%%%%%
childd61=bb*parant1+(1-bb)*parant2(6,:);%الكروموسومالحادي عشر
childd62=(1-bb)*parant1+bb*parant2(6,:);%الكروموسوم الثاني عشر
%%%%%%%%%%%%%%%%
childd71=bb*parant1+(1-bb)*parant2(7,:);%الكروموسوم الثالث عشر
childd72=(1-bb)*parant1+bb*parant2(7,:);
%%%%%%%%%%%%%%%%
childd81=bb*parant1+(1-bb)*parant2(8,:);
childd82=(1-bb)*parant1+bb*parant2(8,:);
%%%%%%%%%%%%%%%%
childd91=bb*parant1+(1-bb)*parant2(9,:);
childd92=(1-bb)*parant1+bb*parant2(9,:);
%%%%%%%%%%%%%%%%
childd101=bb*parant1+(1-bb)*parant2(10,:);
childd102=(1-bb)*parant1+bb*parant2(10,:);
%%%%%%%%%%%%%%%%
childd111=bb*parant1+(1-bb)*parant2(11,:);
```

```
childd112=(1-bb)*parant1+bb*parant2(11,:);
%%%%%%%%%%%%%%%
childd121=bb*parant1+(1-bb)*parant2(12,:);
childd122=(1-bb)*parant1+bb*parant2(12,:);
%%%%%%%%%%%%%%%
childd131=bb*parant1+(1-bb)*parant2(13,:);
childd132=(1-bb)*parant1+bb*parant2(13,:);
%%%%%%%%%%%%%%%
childd141=bb*parant1+(1-bb)*parant2(14,:);
childd142=(1-bb)*parant1+bb*parant2(14,:);
%%%%%%%%%%%%%%%
childd151=bb*parant1+(1-bb)*parant2(15,:);
childd152=(1-bb)*parant1+bb*parant2(15,:);
%%%%%%%%%%%%%%%
childd161=bb*parant1+(1-bb)*parant2(16,:);
childd162=(1-bb)*parant1+bb*parant2(16,:);
%%%%%%%%%%%%%%%
childd171=bb*parant1+(1-bb)*parant2(17,:);
childd172=(1-bb)*parant1+bb*parant2(17,:);
%%%%%%%%%%%%%%%
childd181=bb*parant1+(1-bb)*parant2(18,:);
childd182=(1-bb)*parant1+bb*parant2(18,:);
%%%%%%%%%%%%%%%
childd191=bb*parant1+(1-bb)*parant2(19,:);
childd192=(1-bb)*parant1+bb*parant2(19,:);
%%%%%%%%%%%%%%%
A1=[childd11 ;childd12];
A2=[childd21 ;childd22];
A3=[childd31 ;childd32];
A4=[childd41 ;childd42];
A5=[childd51 ;childd52];
A6=[childd61 ;childd62];
A7=[childd71 ;childd72];
A8=[childd81 ;childd82];
A9=[childd91 ;childd92];
A10=[childd101; childd102];
A11=[childd111 ;childd112];
A12=[childd121 ;childd122];
A13=[childd131 ;childd132];
A14=[childd141 ;childd142];
A15=[childd151 ;childd152];
A16=[childd161 ;childd162];
A17=[childd171 ;childd172];
A18=[childd181 ;childd182];
A19=[childd191 ;childd192];
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%


CA11=[A1;A2;A3;A4;A5;A6;A7;A8;A9;A10;A11;A12;A13;A14;A15;A16;A17;A18;
A19];
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%


% comput fitness for each chrom.
%--------------------------------
for i=1:38
    if CA11(i,:)==[CA11(i,1) CA11(i,2) CA11(i,3) CA11(i,4) CA11(i,5)
CA11(i,6) CA11(i,7) CA11(i,8) CA11(i,9) CA11(i,10)]
    J(i)=0;
    for h=1:n
```

```
       J(i)=J(i)+(y(h)-
((CA11(i,1)*x(h,1))+(CA11(i,2)*x(h,2))+(CA11(i,3)*x(h,3))+(CA11(i,4)*
x(h,4))+(CA11(i,5)*x(h,5))+(CA11(i,6)*x(h,6))+(CA11(i,7)*x(h,7))+(CA1
1(i,8)*x(h,8))+(CA11(i,9)*x(h,9))+(CA11(i,10)*x(h,10))));

    end
    end
end
CA11(:,11)=J';

CA11;
FF=0;
kF1=min(CA11(:,11));
% comput MSE for GE Algrothem
%-------------------------------
for i=1:38
    for j=1:11
        if kF1==CA11(i,j)
            FF=FF+1;
        end
    end
    if FF==1
       for j=1:11
           if kF1==CA11(i,j)
              ffg1Bays2=([CA11(i,j-10) CA11(i,j-9) CA11(i,j-8)
CA11(i,j-7) CA11(i,j-6) CA11(i,j-5) CA11(i,j-4) CA11(i,j-3) CA11(i,j-
2) CA11(i,j-1)]);
              M1=0;
              for i=1:n

yhad2(i)=ffg1Bays2(1).*x(i,1)+ffg1Bays2(2).*x(i,2)+ffg1Bays2(3).*x(i,
3)+ffg1Bays2(4).*x(i,4)+ffg1Bays2(5).*x(i,5)+ffg1Bays2(6).*x(i,6)+ffg
1Bays2(7).*x(i,7)+ffg1Bays2(8).*x(i,8)+ffg1Bays2(9).*x(i,9)+ffg1Bays2
(10).*x(i,10);
                  M1=M1+(y(i)-yhad2(i))^2;
              end
              MSEBays2G=M1/(n*Pm*popsize-p);

          end
       end

    end
end
 %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

end
format long
disp(' Clasic Methods  ')
disp('----------------------')
[mseMLE mseWLS MSEallpc mseBays2]'

disp(' Genitac Methods  ')
disp('----------------------')
[MSEMleG MSEWLSG MSEPC MSEBays2G]'
```

البرنامج الثاني (البرنامج التطبيقي) للبيانات الاصلية

```
clc, clear all
filename = 'data';
x=xlsread(filename);
y=x(:,6);x1=x(:,1);x2=x(:,2);x3=x(:,3);x4=x(:,4);x5=x(:,5);
n=length(y);p=5;
i=ones(n,1);
x=[i x1 x2 x3 x4 x5];
%disp(' WLS Methods  ')
%disp('----------------------')
Bm=inv(x'*x)*x'*y;
pim=exp(x*Bm)./(1+exp(x*Bm));
wi=diag(pim.*(1-pim));
zi=log(pim./(1-pim));
Bhadw=inv(x'*wi*x)*x'*wi*zi;
yhad=Bhadw(1)+Bhadw(2).*x(:,2)+Bhadw(3).*x(:,3)+Bhadw(4).*x(:,4)+Bhad
w(5).*x(:,5)+Bhadw(6).*x(:,6);
mm=0;
for i=1:n

yhad(i)=Bhadw(1)+Bhadw(2).*x(i,2)+Bhadw(3).*x(i,3)+Bhadw(4).*x(i,4)+B
hadw(5).*x(i,5)+Bhadw(6).*x(i,6);
mseWLS=mm+(yhad(i)-y(i))^2;
end
mseWLS=(mseWLS/(n-p));



ei=(y-yhad).^2;
for i=1:p+1
    SE(i)=sum(ei)/((n-2)*sum((x(:,i)-mean(x(:,i))).^2));
end
SE(1)=sum(ei);
SE=(sqrt(SE))';
tbhad=Bhadw./SE;



%---------------------------------
%disp(' MLE Methods  ')
%disp('----------------------')
vm=diag(pim.*(1-pim));
pm=pim;
Bhadm=Bm+inv(x'*vm*x)*x'*(y-pm);
yhad=Bhadm(1)+Bhadm(2).*x(:,2)+Bhadm(3).*x(:,3)+Bhadm(4).*x(:,4)+Bhad
m(5).*x(:,5)+Bhadm(6).*x(:,6);
mm=0;
for i=1:n

yhad(i)=Bhadm(1)+Bhadm(2).*x(i,2)+Bhadm(3).*x(i,3)+Bhadm(4).*x(i,4)+B
hadm(5).*x(i,5)+Bhadm(6).*x(i,6);
    mseMLE=mm+(yhad(i)-y(i))^2;
end
mseMLE=(mseMLE/(n-p));

% start gentic algorithm
%-----------------------------
 bb=rand;
 Pm=length(y);
 popsize=20;
```

```
parant2=rand(popsize,6);
%%%%%%%%%%%%%%%%%%%%%%%%
FF1=Bhadm(1);
FF2=Bhadm(2);
FF3=Bhadm(3);
FF4=Bhadm(4);
FF5=Bhadm(5);
FF6=Bhadm(6);

parant1=[FF1  FF2 FF3 FF4 FF5 FF6];

%crossover using Arthmitac
%----------------------------
childd11=bb*parant1+(1-bb)*parant2(1,:);%الكروموسوم الاول
childd12=(1-bb)*parant1+bb*parant2(1,:);%الكروموسوم الثاني
%%%%%%%%%%%%%%%%%%
childd21=bb*parant1+(1-bb)*parant2(2,:);%الكروموسوم الثالث
childd22=(1-bb)*parant1+bb*parant2(2,:);%الكروموسوم الرابع
%%%%%%%%%%%%%%%%
childd31=bb*parant1+(1-bb)*parant2(3,:);%الكروموسوم الخامس
childd32=(1-bb)*parant1+bb*parant2(3,:);%الكروموسوم السادس
%%%%%%%%%%%%%%%%
childd41=bb*parant1+(1-bb)*parant2(4,:);%الكروموسوم السابع
childd42=(1-bb)*parant1+bb*parant2(4,:);%الكروموسوم الثامن
%%%%%%%%%%%%%%%%
childd51=bb*parant1+(1-bb)*parant2(5,:);%الكروموسوم التاسع
childd52=(1-bb)*parant1+bb*parant2(5,:);%الكروموسوم العاشر
%%%%%%%%%%%%%%%%
childd61=bb*parant1+(1-bb)*parant2(6,:);%الكروموسوم الحادي عشر
childd62=(1-bb)*parant1+bb*parant2(6,:);%الكروموسوم الثاني عشر
%%%%%%%%%%%%%%%%
A1=[childd11 ;childd12];
A2=[childd21 ;childd22];
A3=[childd31 ;childd32];
A4=[childd41 ;childd42];
A5=[childd51 ;childd52];
A6=[childd61 ;childd62];
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

CA11=[A1;A2;A3;A4;A5;A6] ; % مصفوفة الابناء بعد اجراء الطفرة
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%


xx=[CA11(:,2) CA11(:,3) CA11(:,4) CA11(:,5) CA11(:,6)] ;
[nn m]=size(xx);
ii=ones(nn,1);
xx=[ii xx];

% comput fitness for each chrom.
%--------------------------------
for i=1:12
    if CA11(i,:)==[CA11(i,1) CA11(i,2) CA11(i,3) CA11(i,4) CA11(i,5)
CA11(i,6)]
    J(i)=0;
    for h=1:n
       J(i)=J(i)+(y(h)-
((CA11(i,1)*x(h,1))+(CA11(i,2)*x(h,2))+(CA11(i,3)*x(h,3))+(CA11(i,4)*
x(h,4))+(CA11(i,5)*x(h,5))+(CA11(i,6)*x(h,6))));
```

```
    end
    end
end
CA11(:,7)=J';
yy=J';
%xx=abs(xx);

Bm=inv(xx'*xx)*xx'*yy;
%zz=abs(xx*Bm);
zz=Bm(1)+Bm(2)*xx(:,2)+Bm(3)*xx(:,3)+Bm(4)*xx(:,4)+Bm(5)*xx(:,5)+Bm(6
)*xx(:,6);
%pim=exp(zz)./(1+exp(zz));
pim=zz;
vm=diag(pim.*(1-pim));
pm=pim;
BhadmG=(Bm+inv(xx'*vm*xx)*xx'*(yy-pm))/(n*popsize);
yhad=BhadmG(1)+BhadmG(2).*xx(:,2)+BhadmG(3).*xx(:,3)+BhadmG(4).*xx(:,
4)+BhadmG(5).*xx(:,5)+BhadmG(6).*xx(:,6);
nn=length(yy);
mm=0;
for i=1:nn

yhadG(i)=BhadmG(1)+BhadmG(2).*xx(i,2)+BhadmG(3).*xx(i,3)+BhadmG(4).*x
x(i,4)+BhadmG(5).*xx(i,5)+BhadmG(6).*xx(i,6);
    mseMLEG=mm+(yhadG(i)-yy(i))^2;
end
mseMLEG=(mseMLEG/(n*popsize));

eii=(yy-yhadG').^2;
for i=1:p+1
    SE(i)=sum(eii)/((n-2)*sum((xx(:,i)-mean(xx(:,i))).^2));
end
SE(1)=sum(eii);
SE=(sqrt(SE));
tbhadG=(BhadmG./SE)*(n*popsize);



format long
disp(' Clasic Methods  ')
disp('----------------------')
Bhadw
tbhad

disp(' Genitac Methods  ')
disp('----------------------')
BhadmG
tbhadG
```