جامعة البصرة
كلية الادارة والاقتصاد
قسم الإحصاء

إستخدام أنموذجي آلة المتجه الداعم "SVM" والإنحدار اللوجستي "LRM" في تصنيف البيانات مع تطبيق عملي على مرضى داء السكري في مستشفى الموانئ العام في البصرة

رسالة مقدمة

الى مجلس كلية الإدارة والإقتصاد / جامعة البصرة

وهي جزء من متطلبات نيل درجة الماجستير علوم في الإحصاء

للطالبة

أحمد عبد الصمد حبيب ثامر الجبوري

بإشراف

أ.د فوزية غالب عمر السعدون

1439هـ 2018 م

بسم الله الرحمن الرحيم

إنا فَتَحنا لَكَ فَتحاً مُبِيناً

صدق الله العلي العظيم

سورة الفتح

الآية (1)

إقرار المشرف

أشهد أن أعداد هذه الرسالة الموسومة بـ "مقارنة بين أنموذجي آلة المتجه الداعم "SVM" والإنحدار اللوجستي "LRM" في تصنيف البيانات مع تطبيق عملي على مرضى داء السكري في مستشفى الموانئ العام في البصرة " للطالب "أحمد عبد الصمد حبيب ثامر" قد جرى تحت اشرافي في قسم الإحصاء/ كلية الإدارة والاقتصاد /جامعة البصرة وهي جزء من متطلبات نيل شهادة ماجستير في علوم الاحصاء ولأجله وقعت.

أ.د. فوزية غالب عمر

المشرف

توصية رئيس قسم الاحصاء

بناءً على التوصية المقدمة من الأستاذ المشرف. أحيل هذه الرسالة الى لجنة المناقشة لدراستها وبيان الرأي فيها.

م.د. وليد اميه رودين

رئيس قسم الاحصاء

إقرار الخبير اللغوي

أشهد أن أعداد هذه الرسالة الموسومة بـ " **مقارنة بين أُنموذجي آلة المتجه الداعم "SVM" والإنحدار اللوجستي "LRM" في تصنيف البيانات مع تطبيق عملي على مرضى داء السكري في مستشفى الموانئ العام في البصرة** " قد أنيطت بي مهمة تقويمها لغوياً وهي الان مستوفية شروط السلامة اللغوية، ولأجله وقعت.

الخبير اللغوي

أ.د. صباح عبدالكريم مهدي

كلية الإدارة والاقتصاد/ جامعة البصرة

بسم الله الرحمن الرحيم

إقرار لجنة مناقشة

نشهد أننا أعضاء لجنة المناقشة، قد أطلعنا على رسالة طالب الماجستير(احمد عبدالصمد حبيب ثامر)الموسومة ((أستخدام أنموذجي آلة المتجه الداعم SVMوالانحدار اللوجستي LRMفي تصنيف البيانات مع تطبيق عملي على مرضى داء السكري في مستشفى الموانئ العام في البصرة) وقد ناقشناه في محتوياتها وفيما له علاقة بها، وهي جديرة بالقبول لنيل درجة الماجستير في الإحصاء وبتقدير(جيد جداً).

الأستاذ الدكتور
فوزية غالب عمر
عضواً ومشرفاً

المدرس الدكتور
وداد أدور وادي
عضواً

الأستاذ المساعد الدكتور
معاني احمد محمد
عضواً

الأستاذ الدكتور
طاهر ريسان دخيل
رئيساً

مصادقة مجلس الكلية:
صادق مجلس كلية الادارة والاقتصاد/جامعة البصرة على اقرار لجنة المناقشة

الأستاذ المساعد الدكتور
يوسف علي عبد الاسدي
عميد كلية الإدارة والاقتصاد

الإهداء

الى سبب وجودي في هذه الحياة

والدتي ،

والدي الذي شجعني على إكمال الدراسة ووقف الـى جانبي في جميع المراحل

إلى زوجتي وأبنائي عادل ، سجاد ، فاطمة ، محمد ، علي

الى أرواح الشهداء علي ، عبد العزيز ، عادل

وكل أرواح شهداء العراق الحبيب

اهدي هذا الجهد المتواضع

احمد

شكر وتقدير "

الحمد لله على ما كان ونستعين به على ما يكون ، وصلى الله تعالى على سيدنا محمد وعلى آله الطيبين الطاهرين وصحبه المنتجبين . وبعد ... نحمد الله ونشكره على نعمه وعطائه إن هداني و وفقني بكتابة وإنجاز هذه الدراسة .

قال رسول الله (صلى الله عليه وآله وسلم) : ((لم يشكر الله من لم يشكر الناس))

في البداية أتقدم بالشكر الى الأُستاذة فوزية غالب عمر لقبولها الإشراف على رسالتي ، وعلى ما ابدته من ملاحظات أسهمت في انجاز هذه الدراسة ، كما اتقدم بالشكر الى الدكتور وليد ميه رئيس قسم الإحصاء على دعمه العلمي المتواصل

كما أشكر رئيس وأعضاء لجنة المناقشة من الأساتذة الأفاضل على ما سيبدونه من آراء وتوجيهات سديدة ستغني البحث العلمي

كما أشكر جميع أساتذة قسم الإحصاء المحترمين الذين كان لهم الفضل لما وصلت اليه اليوم من أيام الدراسة الأولية ولغاية أكمال رسالة الماجستير .

والشكر موصول الى إدارة مركز الغدد الصم في مستشفى الموانئ العام لتعاونهم معي في جمع البيانات والمساعدة في تقديم أي معلومة وبالإخص الدكتور أحمد عبيد شرهان والسادة العاملين في قسم الملفات وهم الأُستاذ حبيب عبد الرضا عبيد ، والأُستاذ صادق فاخر غالب ، والأُستاذ علي عبد الواحد .

وأخيراً شكري وتقديري الى والدي الحبيب الذي كان يشجعني ويعطيني الحافز لإتمام دراستي .

الباحث

المحتــوى

قائمة الجداول

الاشكال والرسومات

جدول المختصرات

ت	إسم الطريقة باللغة العربية	إسم الطريقة باللغة الأنكليزية	إسم الطريقة مختصر
1	آلة المتجه الداعم	Support vector machine	SVM
2	أُنموذج الإنحدار اللوجستي	Logistic regression model	LRM

مستخلص البحث

تم في هذا البحث دراسة عملية التمييز أو التصنيف للبيانات الإحصائية بإستعمال آلة المتجه الداعم الثنائي الإستجابة ، وأُنموذج الإنحدار اللوجستي الثنائي الإستجابة ، وبالإعتماد على نسبة التصنيف الصحيح للمشاهدات لكلا الطريقتين ، وقد تم إستعمال المحاكاة اولاً في تطبيق الطريقتين على حجوم عينات مختلفة ولتباينات مختلفة ووسط حسابي مختلف ومن ثم تمت المقارنة بين الطريقتين ، وبعد ذلك تم تطبيق الطريقتين في الجانب العملي وعلى بيانات حقيقية لمرضى داء السكري تم الحصول عليها من مركز الغدد الصم في مستشفى الموانئ العام في البصرة ، ومن ثم المقارنة بين الطريقتين المستعملتين في الدراسة وقد توصلت الدراسة الى أن طريقة أو أُسلوب آلة المتجه الداعم كانت الأفضل دقة في التصنيف سواء بإستعمال البيانات الحقيقية في الجانب التطبيقي العملي أو بإستعمال المحاكاة في الجانب التجريبي ولمختلف حجوم العينات الصغيرة والمتوسطة والكبيرة وخاصة عند تداخل البيانات.

الفصل الاول

"المقدمة والاستعراض المرجعي"

(Introduction and Reference review)

1-1المقدمة :-

ان تطور التكنولوجيا وازدياد التعامل معها الكترونيا ادى الى إمكانية الاستفادة من هذا التطور الهائل في عالم الالكترونيات وتكنلوجيا المعلومات ، وتسخيرها في مجال الدراسات والابحاث العلمية ، ومن ضمنها التصنيف الاحصائي للبيانات المطلوبة في الدراسة ، لذا سيتم إستخدام طريقة آلة المتجه الداعم Support Vector Machine (SVM) وذلك لأن هذه الطريقة حازت في السنوات الاخيرة على اهتمام كبير من قبل الباحثين في العالم والتفكير في ايجاد تقنيات جديدة للتقدير والتنبؤ والتصنيف ، كذلك تم إستخدام طريقة أُخرى في هذه الدراسة وهي أُنموذج الإنحدار اللوجستي Logistic Regression Model (LRM) إذ بالإمكان إستخدام الطريقتين عندما يكون متغير الاستجابة ذا صفتين وممكن ايضاً إستخدامهما للتصنيف متعدد الحالات ، لذا كان من الضروري ايجاد اساليب علمية وعملية في تصنيف البيانات احصائيا بدلا من الاساليب العادية مثل التحليل التمييزي(Discriminant Analysis) وأُسلوب العنقدة (Clustring) .

وقد قسم البحث الى خمسة فصول هي :

الفصل الأول : المقدمة ومنهجية البحث وأهمية البحث والاستعراض المرجعي الذي يشتمل على المؤلفات ونتاج البحوث التي ينبغي دراستها تاريخياً لأنها تمثل البدايات الأولية .

الفصل الثاني : الجانب النظري الذي يبدأ فيه التمهيد في عرض طريقة (SVM) وأُنموذج الإنحدار اللوجستي (LRM) ، اعتماداً على مبدأ التصنيف الذي يرى فيه أهم مايميز خصائص آلة المتجه الداعم وأُنموذج الإنحدار اللوجستي (LRM) وما يمكن ان ينتج عنهما من نتائج منطقية .

الفصل الثالث : الجانب التجريبي الذي تم فيه إستخدام المحاكاة لحجوم عينات مختلفة صغيرة ومتوسطة وكبيرة ، وذلك من اجل التعرف والوقوف على صحة النموذج بشكل علمي ، وعند مستوى عالٍ من الثقة .

الفصل الرابع : الجانب التطبيقي وهي المعطيات التي منها نبدأ الاستدلال ومبادئ التصنيف التي نستدل على وفقها ، ثم نتائج الإستدلال من التطبيق التجريبي التي تشكل جزءاً من هيكله ، ومن ثم المقارنة مع الأُسلوب الآخر للبرهنة على دقة معطيات آلة المتجه الداعم (SVM) .

الفصل الخامس : الاستنتاجات والتوصيات

1-2 مشكلة البحث :

تكمن مشكلة البحث في وجود صعوبة في تصنيف البيانات عندما يكون التداخل بين الاصناف والمجموعات المختلفة شديداً الى درجة لايمكن بسهولة تحديد انتماء اي مفردة او مشاهدة الى اي من المجموعات.

1-3 هدف البحث :

يهدف البحث الى المقارنة بين طريقة آلة المتجه الداعم (SVM) وأُنموذج الإنحدار اللوجستي (LRM) وبيان كفاءة أي من الطريقتين من حيث دقة التصنيف وعلى أساس التصنيف الصحيح لمشاهدات المتغير المعتمد.

1-4 أهمية البحث:

تكمن أهمية البحث في إيجاد طريقة ملائمة للبيانات في مجال التصنيف الثنائي (Binary Classification) تواكب التطور الحاصل في مجال تصنيف البيانات .

1-5 الاستعراض المرجعي :

نستعرض هنا ما وقع في أيدينا من نتاجات الباحثين السابقين في موضوع آلة المتجه الداعم ، والإنحدار اللوجستي ، فالمعرفة تبنى دائماً عن طريق دمج مادة مشاهدة من أعمال الآخرين ، أي نتاج لبحوث وأعمال كثيرة ، ونعيد تجديدها وبناءها ودمجها مع الحاضر استعداداً للمستقبل لتقديم مقدمة واضحة عن آلة المتجه الداعم Support Vector Machine(SVM) وبعض التطبيقات وبناء نماذج لمشكلات خاصة إذ أن آلة المتجه الداعم (SVM) قد نجحت في العديد من التطبيقات ذات العوائد المتميزة في التصنيف (وبدون الحاجة الى معلومات أو معرفة أولية) ، كونها مؤكد الوصول من خلالها الى نتائج مقنعة مع درجة عالية من الدقة .

ومن الامثلة على تلك الجهود ما يأتي :-

في العام (2002) نشر الباحثون [19] بحثاً تم فيه تطبيق طريقة SVM للتنبؤ بخصائص أو خاصية GalNAc-transferase إذ تم فحص الاتساق الذاتي واختبار جاكنايف jackknife test لمجموعة بيانات تدريب طريقة SVM ، وكان معدل التصحيح للاتساق الذاتي واختبار جاكنايف jackknife test يصل الى 100% و 84.9% على التوالي .

وفي العام نفسه قام الباحث [31] بإجراء دراسة تهدف الى المقارنة بين التحليل التمييزي والإنحدار اللوجستي ثنائي الإستجابة للتنبؤ بنجاح الطلاب في برنامج حول الإبتكار في تكساس وقد قام بتصنيف الطلاب الى فئيتن فئة الناجحين وغير الناجحين وقد اعطت دالة الإنحدار اللوجستي الثنائي دقة حوالي 94% ودالة التحليل التمييزية 79% .

وفي عام (2003) قدم الباحثون [18] بحثاً إستخدموا فيه SVM كطريقة لتصنيف البروتينات الى طبقات على اساس التمييز الوظيفي للبروتينات ، واظهرت نتائج الدراسة التي توصل اليها الباحثون من خلال اجراء اختبارات على سبعة اصناف من وظائف البروتين ، أن دقة آلة المتجه الداعم SVM في تصنيف هذه الفئات من البروتين كانت في حدود المدى ما بين 86-96% .

وفي العام (2004) استخدم [48] طريقة SVM لتطوير نماذج ارتباط QSAR في مجال الانشطة البيولوجية وهي تعني علاقة النشاط الكمي الهيكلي Quantitative structure activity relationship التي تربط التركيبات الجزيئية بدرجة نشاطاتها الحيوية ، إذ تمت مقارنة SVM مع اساليب اخرى مثل الانحدار الخطي المتعدد ، ودالة الأساس الإشعاعي للشبكة العصبية الإصطناعية radial basis function ، فكانت القدرة التنبؤية لنموذج SVM أعلى من تلك التي حصل عليها بواسطة القاعدة الاشعاعية للشبكة العصبية RBFNN ونموذج الانحدار الخطي المتعدد.

وفي عام (2005) قامت الباحثة [2] بإجراء بحث تناولت فيه دراسة التأثيرات المباشرة وغير المباشرة لمجموعة من العوامل المؤثرة في الإصابة بمرض فقر الدم لمجموعتين من المجاميع السكانية ، الأكثر عرضة للإصابة بهذا المرض هم الأشخاص الذين تقل أعمارهم عن (18) سنة والنساء الحوامل وذلك بإستخدام الإنحدار اللوجستي (LRM) .

كما قام [43] في العام نفسه بإستعمال آلة المتجه الداعم SVM للتصنيف وذلك لكشف واستخراج الشروحات للمصطلحات العلمية والتقنية في مجال البيولوجيا الجزيئية مثل تلك الموجودة في الطب والتخصصات ذات الصلة او حتى المقالات البحثية وبالاعتماد على علم تمييز الأنماط ، واعطت نتائج الاختبارات دقة تعرف عالية بلغت حوالي (74%) .

وفي العام (2006) كتب [20] بحثاً إستعمل SVM في مجال تمييز الأنماط وتمييز الكلام واللغات وكان أُسلوب الباحث هو إستعمال دالة النواة (Kernel Function) للتعامل مع المدخلات عالية الأبعاد

وتوصل الباحث الى عدة استنتاجات ، أهمها أن ميزة استعمال اسلوب الأنوية للتعامل مع فضاء الخصائص الواسع لـ SVM يجعل من الممكن جمع جميع متجهات الداعمة في نموذج واحد وبتعقيدات حسابية منخفضة و أن بناء SVM على اساس مصنف ذي متوسط مربعات خطأ بسيط ينتج نظام أكثر دقة وأخيرا فإن هذا الأُسلوب يعد تنافسياً للأساليب الأخرى مثل نماذج Gaussian mixture models (GMMs) في مجال تمييز الكلام واللغة .

وفي العام نفسه نشرت [7] دراسة تناولت فيها التحليل المميز والتطرق الى بعض الأنواع التمييزية المتقدمة في هذا المجال وتطبيقها على نوعين من أمراض القلب لبناء نموذج إحتمالي للتمييز بينها، وقد أظهر إنموذج الإنحدار اللوجستي تفوقاً واضحاً على بقية النماذج المستخدمة في البحث من حيث قلة نسبة التصنيف الخاطئ مقارنة ببقية النماذج .

وفي عام (2007) قدمت الباحثة [21] أُطروحة دكتوراه إقترحت فيها ثلاث خوارزميات جديدة لتطبيق التعلم النشط لـSVM بطريقة شبه مراقب او شبه اشراف للاستفادة من ميزة البيانات غير المسماة ، عن طريق التقليل من عدد التجارب اللازمة وتحقيق وفورات في كلفة مسائل التصنيف والوقت المستغرق في تجميع البيانات وهذه الميزة من شأنها أن تسرع من عملية التعلم.

وفي العام (2009) قام الباحث [13] بنشر دراسة أجرى فيها مقارنة بين ثلاث خوارزميات للتعلم الآلي هي (التحليل الخطي المميز ، شجرة القرار . الة المتجه الداعم) للتمييز بين اصوات الضفادع والطيور وكانت النتائج 94.96% لآلة المتجه الداعم و 89.20% لشجرة القرار و 71.45% للتحليل المميز الخطي .

وفي عام (2010) قام [33] بإستعمال الإنحدار اللوجستي في مجال التجارة كطريقة للمقارنة بهدف بناء نموذج أفضل للتنبؤ بعائد الأسهم للشركات وبشكل أكثر فاعلية وكفاءة وإستعمل عدة طرق للتقييم بين الطرائق شائعة الإستخـدام في مجـال تنقيب البيانات وهي كل من مصنف آلة المتجه الداعم والإنحدار اللوجستي وتحليل الدوال التمييزية .

وفي العام ذاته قام [40] باستخدام آلة المتجه الداعم لإجراء الكشف المبكر التلقائي عن الأمراض التي تصيب اوراق بنجر السكرعن طريق تمييز الاوراق المصابة من غير المصابة وتوصل الى دقة

تصنيف وصلت الى 97% وايضاً تم الكشف عن الأمراض النباتية اعتمادا على نوع ومرحلة المرض ، وكانت دقة التصنيف بين 65% و 90% .

وفي عام (2011) أجرى الباحث [10] دراسة لتحليل بعض العوامل المؤثرة في الإصابة بمرض اللثة مثل الترسبات الكلسية وسوء التغذية وغيرها من العوامل ، بإستعمال أُنموذج الإنحدار اللوجستي وخلصت الدراسة الى مجموعة من الإستنتاجات كان أهمها أن عامل الترسبات الكلسية يعد السبب الرئيس في الإصابة بمرض اللثة .

وفي العام ذاته نشر الباحث [41] بحثاً اقترح فيه طريقة اسماها SVM الموازية للتنبؤ بمرض السكري للتعامل مع الحجوم الكبيرة جداً من العينات وذلك من خلال الاعتماد على مسح لبيانات متعلقة بمعالم الجسم المختلفة من أشخاص مصابين وغير مصابين والهدف من هذا البحث هو التنبؤ بشكل صحيح بإمكانية اصابة اي شخص بالمرض مستقبلاً .

كما قدم الباحث [49] في العام نفسه أُطروحة دكتوراه ناقش في هذه الأُطروحة طريقة لتحسين هيكلية Support Vector Machines (SVM) للتعامل مع مسألة الهامش الكبير (Large margin) في مسائل SVM التمييزية وناقش الباحث SVM للتصنيف الثنائي (Binary classification) وكيفية تعميمها لتكون متعددة الاصناف Multi classification) .

وفي العام نفسه قام الباحثان [12] بإجراء دراسة استعملا فيها تقنية آلة المتجه الداعم support vector machine (SVM) مع تقنية مميز فيشر الخطي (Fisher Linear Discriminator) في مجال الإخفاء والكشف عن الصورة التي تحتوي الرسائل السرية وعند المقارنة بين الطريقتين كانت الافضلية لتقنية SVM على تقنية FLD من ناحية نسبة الكشف ومقدار الخطأ بالرغم من ان تقنية FLD كانت الاسرع في زمن التنفيذ.

وفي عام (2012) نشرت الباحثة [34] رسالة ماجستير إذ إستعملت الباحثة الانحدار اللوجستي في دراسة مرض داء السكري من النوع الثاني فقط وتحليل المخاطر من جراء الإصابة بهذا المرض ليشكل نموذجاً لإنتشار مرض إعتمام عدسة العين (cataract) ، ودراسة تأثير إستعمال دواء الستاتينس (statins) الذي يوصف عادة لمرضى النوع الثاني الذين يعانون من مرض إعتمام عدسة العين (cataract) .

وفي العام ذاته قدم الباحث [39] دراسة في مجال آلة التعلم لإكتشاف طرائق جديدة في مجال التعلم شبه اشراف (او شبه مراقب) وفي هذا البحث تم اقتراح طريقة جديدة غير مألوفة سابقاً ممكن استخدامها في مجال آلة التعلم والتنقيب عن البيانات وذلك بتوأمة طريقة لابلاسيان وآلة متجه الداعم أُطلق على هذه الطريقة (lap-Tsvm) لمسألة التصنيف.

كما قدمت الباحثة [42] في العام نفسه رسالة ماجستير إستعملت فيها الانحدار اللوجستي المتعدد لمعرفة ما إذا كانت عوامل مثل العمر والجنس والوضع المهني والحالة الإجتماعية والتدخين واستهلاك الكحول وارتفاع ضغط الدم تسهم في التشخيص السريري لمرض السكري.

وفي العام نفسه أيضاً قام الباحثان [46] بإقتراح إستعمال طريقة التحليل المميز الخطي مع آلة المتجه الداعم وذلك لغرض تحسين نظام تمييز الوجه ومن خلال مقارنة هذه الطريقة المقترحة ، مع طريقة تحليل المميز الخطي والشبكة العصبية الاصطناعية تبين ان الطريقة المقترحة (طريقة تحليل المميز الخطي وآلة المتجه الداعم) تعطي نتائج افضل من حيث معدل التعرف الصحيح .

وفي العام (2013) قدم الباحث [45] رسالة ماجستير استعمل فيها الإنحدار اللوجستي لدراسة أثر العرق في الإصابة بمرض السكري من خلال مقارنة لعوامل الاصابة بالسكري بين السكان الكنديين الاصليين والمهاجرين و الافراد المولودين في كندا وكانت النتائج هي ارتفاع معدلات الاصابة بين جميع الافراد المذكورين .

وفي العام ذاته نشر الباحثان [32] بحثاً حاولا فيه تصميم مصنف للكشف عن مرض السكري بأقل تكلفة وأفضل إداء ، ويلبي طموحات العصر وقد استند الباحثان إلى قاعدة بيانات بيما الهندية عن مرض السكري واظهرت نتائج البحث الى ان طريقة SVM يمكن استخدامها في تشخيص الامراض .

وفي العام (2014) قام الباحث [17] بدراسة مرضى الصرع من خلال تصنيف بيانات في حالة وجود نوبات الصرع وفي حالة عدم وجودها باستخدام خوارزمية آلة المتجه الداعم(SVM). مع ذلك اظهرت الدراسة ان التصنيف بواسطة SVM تكون بدقة جيدة أفضل من خوارزميات تعلم الالة الاخرى مثل الشبكات العصبية الاصطناعية.

وفي العام نفسه قدمت الباحثة [44] رسالة ماجستير طبقت فيها الإنحدار اللوجستي الثنائي الوصفي ومتعدد المتغيرات لتحديد انتشار ومؤشرات المخاطر المرتبطة بمرض داء السكري الحملي (Gestational Diabetes Mellitus (GDM الذي من مضاعفاته إمكانية الإصابة بمرض داء السكري من النوع الثاني .

كما قام الباحثان [1] بتطبيق دالة الإنحدار اللوجستي في ميدان الرياضة من خلال بناء نموذج يفسر العلاقة الموجودة بين بعض العوامل الإجتماعية ومتغير ممارسة الأنشطة البدنية والرياضية في الأوساط الجامعية.

وفي عام (2015) إستطاع [38] مزج ثلاث تقنيات هي النمط الثنائي المحلي (LBP) ، وتحليل المكون الرئيسي (PCA) ، والة المتجه الداعم (SVM) لتقديم نظام مميز للوجه قادر على التعرف على الوجه اعتمادا على عينة واحدة للشخص وتمكن الباحث وعن طريق مزج آالة المتجه الداعم مع التقنيتين تقنية (LBP) وتقنية (PCA) من تحقيق هذه النتيجة وهذا يعطي دعماً وتأييداً لطريقة آلة المتجه الداعم وإستخداماتها في المجال التطبيقي .

وفي العام ذاته قدم الباحث [37] رسالة ماجستير إذ قام بإستعمال نموذج الإنحدار اللوجستي المتعدد ليتناسب مع عوامل مثل العمر والجنس والوضع المهني والتدخين واستهلاك الكحول ومستوى الكوليسترول وارتفاع ضغط الدم والتاريخ العائلي لمرض السكري كعوامل خطر الإصابة بمرض السكري وبإستخدام حزمة البرنامج الجاهز (spss)

وفي العام نفسه إستعمل الباحث [36] نموذج الإنحدار اللوجستي لتحديد انتشار البكتريا الدقيقة بين مرضى السكري النوع الاول والنوع الثاني ودراسة العلاقة بين المعلمات المسيطرة على مرض السكري مثل الهيموغلوبين A1C (HbA1C) ، وضغط الدم .

وفي العام (2016) قدم الباحث [16] رسالة ماجستير ركز الباحث في رسالته على مرض داء السكري من النوع الثاني وحدد في دراسته ثمانية عوامل ممكن ان يكون لها تأثير في الاصابة بالمرض وتم استخدام نموذج الانحدار اللوجستي الثنائي واوضحت النتائج عن نتائج تمييزية مقبولة بإستخدام نموذج الإنحدار اللوجستي للتحقق من وجود مرض داء السكري من النوع الثاني .

كما نشر الباحث [15] بحثاً بعنوان (Using One-Class SVM with Spam Classification) استعملت الباحثة فيه طريقة مقترحة في مجال مصنفات البريد الالكتروني تجمع بين

نسبة الربح وتقليل الكلفة وهي طريقة اختيار الخصائص مع تدريب SVM ذات الصنف الواحد لزيادة عملية الكشف ، و أظهرت النتائج دقة عالية تصل الى 100% وبنسبة خطأ أقل مع عدد خصائص يصل الى 5 خصائص .

وفي عام (2017) قام الباحثون [4] بإجراء دراسة تمت فيها المقارنة بين أُنموذج الإنحدار اللوجستي وانموذج الإنحدار الخطي المميز في القدرة التنبؤية، إتضح من خلال المقارنة أن نموذج الإنحدار اللوجستي أفضل من نموذج الدالة المميزة الخطية بإستعمال البيانات الأصلية ، أما بإستعمال المركبات الرئيسة بعد أن تم تقليص المتغيرات الى 5 عوامل رئيسة فكانت النتائج متساوية بين الدالتين .

الفصل الثاني

الجانب النظري

(Theoretical Part)

2-1 مفهوم التصنيف الثنائي للبيانات: (Binary classification)

يقصد بالبيانات الثنائية هي الحالة التي يكون فيها متغير الإستجابة (Y_i) (المتغير التابع) ثنائي الإستجابة فيعطى واحد مثلاً لوقوع الحدث وصفر لعدم وقوع الحدث ، ومن الأمثلة على ذلك دراسة عينة من مجتمع ما (x_i) من المصابين بداء السكري لغرض تمييز الافراد المصابين بداء السكري من النوع الأول والأفراد المصابين بداء السكري من النوع الثاني ، لذا فإن (Y_i) سوف يأخذ (1) إذا كان الفرد مصاب بداء السكري من النوع الأول و(0) إذا كان الفرد مصاب بداء السكري من النوع الثاني ، أو أخذ رأي شريحة معينة من السكان حول رأي ما فإذا كان جواب الشخص مع ذلك الرأي فإن المتغير (Y_i) سوف يكون (1) واذا كان جوابه ضد أو ليس مع ذلك الرأي سوف يأخذ (0) وغيرها من الدراسات التي تهتم بتحليل الظواهر المختلفة .

وهناك العديد من الأساليب والطرائق الإحصائية التي تعالج هذه الظواهر غير إننا سنهتم بالتركيز على أُسلوب آلة المتجه الداعم (SVM) ، و أُنموذج الإنحدار اللوجستي (LRM) كونهما من الأساليب الحديثة والمتطورة وخاصة أُسلوب آلة المتجه الداعم .

2-1-1 : آلة المتجه الداعم (SVM)

ويقسم أُسلوب أو تقنية آلة المتجه الداعم الى قسمين الأول آلة المتجه الداعم للتصنيف (support vector machine for classification) وهو موضوع هذا البحث ، والقسم الثاني هو آلة المتجه الداعم للإنحدار (Support Vector Machine For Regression)

2-1-2 : آلة المتجه الداعم للتصنيف (support vector machine for classification)

وهي إحدى أساليب تعلم الآلة (machine learning) قدمها في العام (1992) الباحث (vapnik) وهي عبارة عن خوارزمية تعلم عن طريق مشرف أو موجه (supervised) ، وتستند في عملها الى نظرية التعلم الإحصائية (Statistical Learning Theory)[26] .

لقد كان إكتشاف تقنية آلة المتجه الداعم في الأصل لحل مسائل تمييز الأنماط (Pattern Recognition) عن طريق تحديد المستوى الفاصل للبيانات ، إذ أن الهدف الأساسي من هذه التقنية هو إيجاد أفضل مستوى فاصل للبيانات المراد فصلها وتصنيفها الى صنفين[30] .

ويمكن إستخدامها في مسائل التصنيف الخطية وغير الخطية إذ يمكنها التصنيف بالإعتماد على مصنف خطي (Linear Classifier) ومصنف غير خطي (Nonlinear Classifier)[14] ، علماً أن المصنف غير الخطي أتى من بعض مسائل التصنيف التي لايكون لديها مستوى فاصل بسيط لكي يستخدم كمعيار فاصل للفصل إذ يتم إيجاده عبر إستخدام مفهوم الأنوية (Kernels) .

2-1-3 آلة المتجه الداعم الخطية (Linear Support Vector Machine) [25][21][35]

إذا فرضنا أنه كان لدينا L من النقاط بحيث أن كل قيمة x_i مدخلة لها D (dimensionality) إذ ان قيم y_i تساوي اما 1+ او 1− فتكون البيانات المدربة (training data) بالشكل التالي:

$$\{x_i, y_i\} \text{ where } i=1,\ldots\ldots L \ , \ y_i \in \{-1,+1\}, x \in R^D$$

وعلى إفتراض ان البيانات مفصولة خطيا فبالإمكان رسم خط لـ x_1ضد x_2 يقوم بفصل البيانات الى صنفين او مجموعتين اذا كانت D=2 ويكون المستوى الفاصل على الرسم لـ $x_1,\ldots\ldots x_D$ في حالة D >2 ، ويمكن التعبير عن المستوى الفاصل (Hyperplane) حسب الصيغة التالية :

$$w' x_i + b = 0$$

حيث أن :

$x_1,\ldots\ldots x_D$: متغيرات توضيحية تمثل خصائص المشاهدات

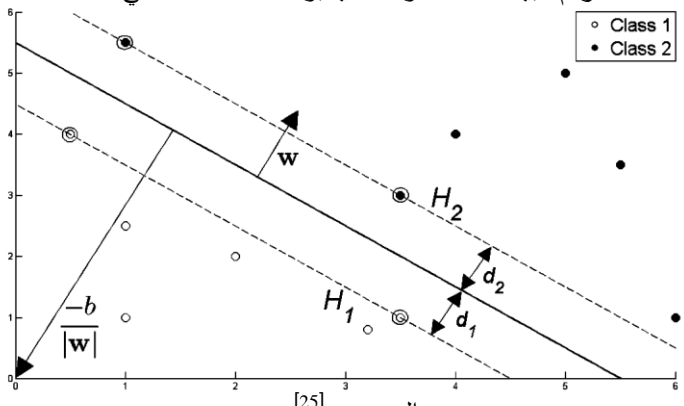
W : متجه الاوزان ويتم تطبيعها (normalization) للمستوى الفاصل (Hyperplane)

b: يمثل حد القطع أو التحيز (bias)

وتكون النقاط الاقرب الى المستوى الفاصل (Hyperplane) عبارة عن المتجهات الداعمة (support vectors) وتكون قيمتها أما 1+ للمجموعة الاولى و 1− للمجموعة الثانية ويمكن ملاحظتها من خلال الرسم التالي :

شكل (2-1)

رسم البيانات المفصولة خطياً بواسطة مصنف خطي



المصدر: [25]

H_1 : يمثل المستوي الثانوي الأول والذي تقع عليه المتجهات الداعمة للنوع الاول عندما $w'x_i + b = 1$

H_2 : يمثل المستوي الثانوي الثاني والذي تقع عليه المتجهات الداعمة للنوع الثاني عندما $w'x_i + b = -1$

d_1 : اقصر مسافة الى أقرب نقطة من النوع الاول

d_2 : أقصر مسافة الى أقرب نقطة من النوع الثاني

$\frac{-b}{\|w\|}$: تمثل المسافة العمودية من المستوى الفاصل (Hyperplane) الى نقطة الاصل .

والمستوى الذي بين المستويين H_1 و H_2 هو المستوى الفاصل الرئيسي (Hyperplane) حيث ان الهدف من الة المتجهات الداعمة support vector machines (SVM) هي جعل المستوى الفاصل (Hyperplane) ابعد مايمكن عن نقاط الصنفين .

وإن عملية تصنيف المشاهدات تتم على وفق الصيغتين الآتيتين :

$$w'x_i + b \geq +1 \qquad \text{for } y_i = +1 \ldots\ldots\ldots\ldots\ldots\ldots(1.2)$$

$$w'x_i + b \leq -1 \qquad \text{for } y_i = -1 \ldots\ldots\ldots\ldots\ldots\ldots(2.2)$$

فإذا كانت قيمة y اكبر من 1+ يعني أن المشاهدة تنتمي للمجموعة الاولى و اذا كانت قيمة y أقل من 1− فإن المشاهدة تنتمي للمجموعة الثانية وإذا كانت 1+ أو 1− فتكون عبارة عن الـ (support vectors machine) المتجهات الداعمة و هي النقاط التي تكون أقرب نقاط المجموعتين الى المستوى الفاصل (hyperplane) .

علماً أن الصيغتين المذكورتين آنفاً تم الحصول عليهما من الصيغة الآتية :

$$y_i(w'x_i + b) \geq 1$$

$$y_i(w'x_i + b) - 1 \geq 0 \ldots\ldots\ldots(3.2)$$

تعرف المسافة بين المتجهات الداعمة والمستوى الفاصل (Hyperplane) بهامش المتجهات الداعمة (SVM Margin) ولجعل المستوى الفاصل (Hyperplane) بعيد قدر الامكان عن المتجهات الداعمة (support vectors) نحتاج الى تعظيم الهامش

والمسافة بين المستويين H_1 و H_2 تكون مساوية لـ $\frac{2}{\|w\|} = \frac{2}{\sqrt{w'w}}$ ويتم تعظيمها كالتالي :

$$\text{Max } \frac{2}{\|w\|} = \text{Min } \frac{\|w\|}{2} \equiv \text{Min } \frac{w'w}{2}$$

وهي تكافئ دالة الهدف في حالة التقليل (التصغير) :

$$\text{Min}\|w\| \text{ s.t } y_i(w'x_i. +b)-1\geq 0 \quad \forall i$$

$$\text{Min}\frac{1}{2}\|w\|^2 \text{ يكافئ Minmizing} \|w\| \quad \text{و}$$

ولذلك نحتاج الى ايجاد الدالة:

$$\text{Min}\frac{1}{2}\|w\|^2 \text{ s.t } y_i(w'x_i +b)-1\geq 0 \quad \forall i \ldots\ldots (4.2)$$

ولكي يتم تحقيق قيود هذه الدالة نحتاج الى استعمال مضاعف لاكرانج :

$$L_p(w,b,\alpha) = \frac{1}{2}\|w\|^2 - \alpha_i[\, y_i(w'.x_i + b) - 1 \ \forall i \ ]$$

$$L_p(w,b,\alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{L}\alpha_i y_i(w'.x_i + b) + \sum_{i=1}^{L}\alpha_i$$

$$L_p(w,b,\alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{L}\alpha_i\, y_i w'.x_i + b\sum_{i=1}^{L}\alpha_i y_i + \sum_{i=1}^{L}\alpha_i \ldots\ldots (5.2)$$

ولما كانت الغاية هي ايجاد قيم w,b في حالة تقليل الدالة في الصيغة الاولية وقيم مضاعفات لاكرانج α_i في حالة تعظيم الدالة في الصيغة الثنائية لذلك سوف يتم أولاً إشتقاق الصيغة (5.2) بالنسبة لـ w و بالنسبة لـ b

$$\frac{\partial L_p}{\partial w} = 0 \ \rightarrow \ \frac{1}{2}2\|w\| - \sum_{i=1}^{L}\alpha_i y_i x_i = 0$$

$$w = \sum_{i=1}^{L}\alpha_i y_i x_i \ldots\ldots\ldots\ldots (6.2)$$

$$\frac{\partial L_p}{\partial b} = 0 \ \rightarrow \sum_{i=1}^{L}\alpha_i y_i = 0 \ldots\ldots\ldots\ldots (7.2)$$

وبتعويض المعادلتين (6.2) و (7.2) في (5.2) والتي تعتمد على α_i نحتاج الى تعظيم الدالة لذا نستخدم الصيغة الثنائية المقابلة للصيغة الاولية

$$L_p(w,b,\alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{L}\alpha_i y_i x_i \,.\sum_{j=1}^{L}\alpha_j y_j.x_j + b\sum_{i=1}^{L}\alpha_i y_i + \sum_{i=1}^{L}\alpha_i$$

$$L_p(w,b,\alpha) = \frac{1}{2}\sum_{i=1}^{L}\sum_{j=1}^{L}\alpha_i y_i x_i \,.\alpha_j y_j x_j - \sum_{i=1}^{L}\sum_{j=1}^{L}\alpha_i y_i x_i \,.\alpha_j y_j x_j + \sum_{i=1}^{L}\alpha_i$$

$$L_d(w,b,\alpha) = \sum_{i=1}^{L} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{L} \alpha_i \alpha_j y_i y_j x_i \ . x_j \ \dots\dots (8.2)$$

$$s.t \quad \alpha_i \geq 0 \quad \forall i \quad , \sum_{i=1}^{L} \alpha_i y_i = 0$$

علماً بان x_i هو مدور لـ x_j و y_i مدور y_j و α_i مدور α_j

$$L_d(w,b,\alpha) = \sum_{i=1}^{L} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{L} \alpha_i \ H_{ij} \ \alpha_j \ \text{ where } H_{ij} = y_i y_j x_i \ . x_j$$

$$L_d(w,b,\alpha) = \sum_{i=1}^{L} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{L} \alpha' H \, \alpha \ \dots\dots\dots\dots\dots\dots (9.2)$$

$$s.t \quad \alpha_i \geq 0 \quad \forall i \quad , \sum_{i=1}^{L} \alpha_i y_i = 0$$

وتعد هذه الصيغة (9.2) L_d هي الصيغة الثنائية للصيغة الاولية L_P .

والآن يتم الانتقال من حالة تقليل دالة الهدف في الصيغة الاولية الى حالة تعظيم دالة الهدف في الصيغة الثنائية

$$\underset{\alpha}{Max}\left[\sum_{i=1}^{L} \alpha_i - \frac{1}{2}\sum_{i=1}^{L} \alpha' \, H \, \alpha\right] s.t \ \alpha_i \geq 0 \ \forall i \ , \sum_{i=1}^{L} \alpha_i y_i = 0 \ \dots\dots (10.2)$$

وبالوصول الى الصيغة الثنائية وبالإعتماد على طريقة مضاعف للاكرانج من أجل تحقيقها لشروط كرش كان توكر* (Karush–Kuhn–Tucker) تعد المسألة قد وصلت الى حالة التحدب (convex quadratic optimization problem) او الحالة التي يمكن حلها للإقتراب نحو الحل الأمثل .

---

* شروط كرش كان توكر

1- يجب ان يكون القيد من النوع اكبر اويساوي صفر او اصغر او يساوي صفر

2- مجموع حاصل ضرب مضاعفات لاكرانج في القيد = صفر

3- مشتقة دالة الهدف – مشتقة القيد = صفر

4- مضاعفات لاكرانج اكبر او تساوي صفر

وبإستخدام البرمجة التربيعية* (Quaratic programming) سوف نحصل على قيم α_i ، ونعوضها في الصيغة (6.2) نحصل على قيم w_i .

أي نقطة تحقق المعادلة (7.2) تكون متجهاً داعماً (support vector) وسوف يرمز لها x_s سوف تكون بالشكل التالي :

$$y_s\ (x_s.w+b) =1$$

وبالتعويض عن قيمة w من الصيغة (6.2)

$$y_s(\sum_{m\in s} \alpha_m y_m x_m \,.\, x_s + b) = 1$$

وبضرب طرفي المعادلة في y_s وتبسيط المعادلة للحصول على قيمة b

$$y_s^2(\sum_{m\in s} \alpha_m y_m x_m \,.\, x_s + b) = \ y_s$$

$$b= y_s - \sum_{m\in s} \alpha_m y_m x_m \,.\, x_s$$

حيث ان : $y_s^2=1$ ، كذلك فإن x_m هي مدور x_s

وبقسمتها على العدد الكلي للمتجهات الداعمة نحصل على قيمة b

$$b=\frac{1}{Ns}\sum_{m\in s}(\ y_s - \sum_{m\in s} \alpha_m y_m x_m \,.\, x_s).....(11.2)$$

ومن خلال الصيغة التالية يمكننا ان نصنف اي نقطة جديدة يراد إختبارها لأي مجموعة تصنيف

$$f(x) = y = (\sum_{i=1}^{L} \alpha_i y_i x_i + b)$$

ونعوض الصيغة (6.2) في الصيغة المذكورة آنفاً فنحصل على الصيغة النهائية

$$y =(w'x+b).....(12.2)$$

2-1-4 : سوء تصنيف البيانات (misclassification) [25][28]

ويقصد بـ سوء التصنيف هو وجود بعض النقاط من كل صنف في جهة الصنف الاخر وهذه الحالة تسمى (misclassifications) أي سوء تصنيف إذ تكون البيانـات ليست مصنفة خطياً بصورة تامة ويتم

البرمجة التربيعية : هي نوع من انواع البرمجة غير الخطية إذ تكون مسألة الأمثلية فيها مقيدة خطياً بدالة هدف تربيعية أو تكون دالة الهدف خطية أو أحد القيود غير خطي .

هنا في هذه الحالة اضافة متغير وهمي (ξ_i) variable slack الى القيدين (1.2) و(2.2) وهذا مايطلق عليه تخفيف القيود (Soft margin SVM) إذ يكون المتغير الراكد ξ_i حيث i=1……L

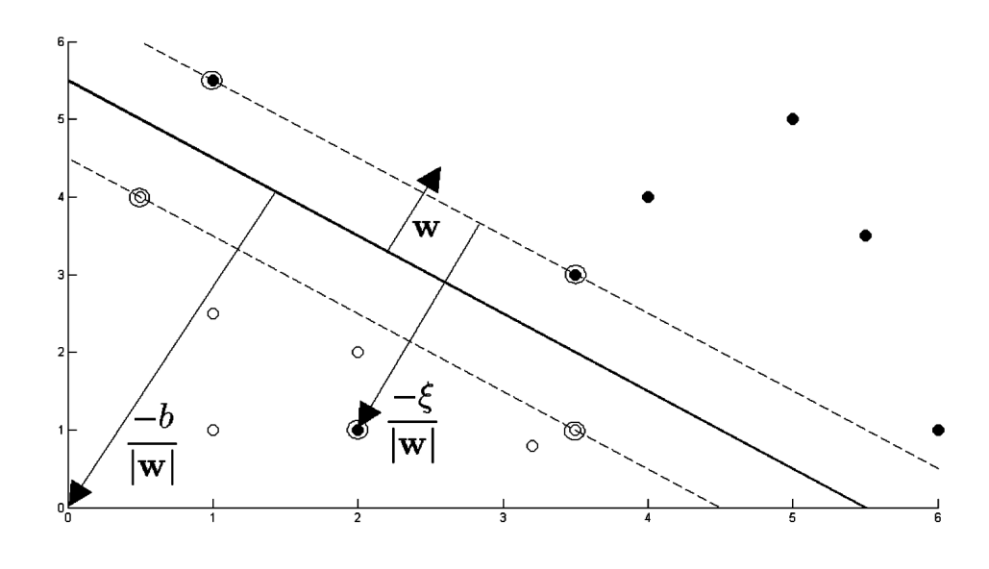
$$w'x_i + b \geq +1 - \xi_i \quad \text{for } y_i=+1$$
$$w'x_i + b \leq -1 + \xi_i \quad \text{for } y_i=-1$$
$$\xi_i \geq 0 \ \forall i$$

وبمكن جمعها في صيغة عامة بالشكل التالي :

$$w'x_i + b - 1 + \xi_i \geq 0 \quad \text{where } \xi_i \geq 0 \ \forall i$$

الشكل (2-2)

حالة سوء التصنيف الخطي



المصدر : [25]

$\frac{-\xi}{\|w\|}$: تمثل المسافة العمودية من المستوى الثانوي لأحد الصنفين لأي نقطة تم تصنيفها خطأ الى الصنف الآخر .

وتكون دالة الهدف في هذه الحالة :

$$\text{Min}\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{L}\xi_i \ \text{ s.t } y_i\,(w'.x_i + b) - 1 + \xi_i \geq 0 \quad \forall i \ldots\ldots\ldots (13.2)$$

C: معيار الموازنة بين مقدار الجزاء للمتغير الراكد وحجم الهامش

وباستخدام مضاعفات لاكرانج تكون دالة الهدف :

$$L_p = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{L}\xi_i - \sum_{i=1}^{L}\alpha_i[y_i\,(w'.x_i + b) - 1 + \xi_i] - \sum_{i=1}^{L}\mu_i\xi_i \ldots\ldots(14.2)$$

حيث $\alpha_i \geq 0\ ,\ \mu_i \geq 0\ \forall i$

μ_i : مضاعفات لاكرانج لقيد المتغير الوهمي $\xi_i \geq 0$

α_i : مضاعفات لاكرانج للقيد $w'x_i + b - 1 + \xi_i \geq 0$

ونقوم بإشتقاق الصيغة المذكورة آنفاً بالنسبة لـ w و b و ξ_i ونجعل المشتقة الجزئية مساوية للصفر:

$$L_p = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{L}\xi_i - \sum_{i=1}^{L}\alpha_i y_i\, w'.x_i + b\sum_{i=1}^{L}\alpha_i y_i + \sum_{i=1}^{L}\alpha_i - \sum_{i=1}^{L}\alpha_i\xi_i - \sum_{i=1}^{L}\mu_i\xi_i$$

$$\frac{\partial L_p}{\partial b} = 0 \ \rightarrow \frac{1}{2}2\|w\| + \sum_{i=1}^{L}\alpha_i y_i x_i = 0$$

$$w = \sum_{i=1}^{L}\alpha_i y_i x_i \ldots\ldots\ldots\ldots(15.2)$$

$$\frac{\partial L_p}{\partial b} = 0 \ \rightarrow \sum_{i=1}^{L}\alpha_i y_i = 0 \ldots\ldots\ldots\ldots(16.2)$$

$$\frac{\partial L_p}{\partial \xi} = 0 \ \rightarrow\ C = \alpha_i + \mu_i \ \ldots\ldots\ldots\ldots(17.2)$$

وبتعويض المعادلات الثلاث الاخيرة في المعادلة (14.2) وبالطريقة السابقة في الحصول على الصيغة الثنائية L_D تنتج معادلة شبيهة بالمعادلة (10.2) وكالاتي :

$$L_p = \frac{1}{2}\|w\|^2 + (\alpha_i + \mu_i)\sum_{i=1}^{L}\xi_i - \|w\|^2 + \sum_{i=1}^{L}\alpha_i - \sum_{i=1}^{L}\alpha_i\xi_i - \sum_{i=1}^{L}\mu_i\xi_i$$

$$L_p = \frac{1}{2}\|w\|^2 + \sum_{i=1}^{L}\alpha_i\xi_i \ + \sum_{i=1}^{L}\mu_i\xi_i - \|w\|^2 + \sum_{i=1}^{L}\alpha_i - \sum_{i=1}^{L}\alpha_i\xi_i - \sum_{i=1}^{L}\mu_i\xi_i$$

حيث أن :

$$\|w\|^2 = \sum_{i,j=1}^{L} \alpha_i \alpha_j y_i y_j x_i x_j$$

$$\max_{\alpha} \left[\sum_{i=1}^{L} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{L} \alpha' H \alpha \right]$$

$$s.t \;\; 0 \leq \alpha_i \leq C \;\; \forall i \text{ and } \sum_{i=1}^{L} \alpha_i y_i = 0$$

حيث : $H_{ij} = y_i y_j x_i \; . x_j$

وبالطريقة السابقة نفسها نجد قيم α_i بطريقة البرمجة التربيعية ومن ثم نحسب الاوزان w وبعدها نجد b بالاعتماد على مجموعة المتجهات الداعمة (set of support vectors) بايجاد المؤشرات i حيث $0 \leq \alpha_i \leq C$

$$b = \frac{1}{Ns} \sum_{m \in s} (y_s - \sum_{m \in s} \alpha_m y_m x_m \; . x_s) \ldots\ldots (18.2)$$

ومن خلال الصيغة التالية يمكننا ان نصنف اي نقطة جديدة يراد إختبارها لأي مجموعة تصنيف

$$f(x) = y = (\sum_{i=1}^{L} \alpha_i y_i x_i + b)$$

ونعوض الصيغة (15.2) في الصيغة المذكورة آنفاً فنحصل على الصيغة النهائية

$$y = (w' x + b) \ldots\ldots\ldots (19.2)$$

2-2الإنحدار اللوجستي: (Logestic Regression)

هو أُسلوب احصائي مرن يستخدم لتفسير العلاقة بين متغير الإستجابة (او المتغير التابع أو المتغير المعتمد) الذي يكون ثنائي الاستجابة (Binary or Dichotomous) مع واحد او أكثر من المتغيرات التوضيحية (او المتغيرات التفسيرية أو المستقلة) ذات الطبيعة الفئوية أو الطبيعة الكمية ، وهو يشبه نموذج الانحدار الخطي من ناحية تفسير العلاقة بين متغير الإستجابة والمتغيرات التوضيحية إلا ان الباحثين قد يجدون صعوبة في إستخدام الإنحدار الخطي عندما يكون المتغير التابع ثنائي الاستجابة ، وهذا الإختلاف ينعكس بدوره على افتراضات الإنحدار اللوجستي [5] .

لان الانحدار البسيط والمتعدد يكون مقيداً نوعاً ما بإشتراط أن يكون متغير الإستجابة متغيراً كمياً متصلاً بدلاً من أن يكون وصفياً منفصلاً [6] .

2-2-1 تعريف الإنحدار اللوجستي

للإنحدار اللوجستي (Logistic Regression) تعاريف و مفاهيم عدة منها :

أنه أُسلوب إحصائي لفحص العلاقة بين متغير الإستجابة النوعي و متغير واحد أو اكثر من المتغيرات التوضيحية ، أي أن الاسلوب الاحصائي المستخدم لفحص وتوفيق العلاقة بين المتغير التابع النوعي ثنائي القيمة ومتغير واحد او اكثر من المتغيرات التوضيحية (المستقلة) أياً كان نوعها ، ويسمى هنا بتحليل الانحدار اللوجستي الثنائي (Binary Logistic Regression).

ويعرف كذلك بأنه ذلك النوع من الإنحدار المستخدم في التنبؤ بقيم المتغيرات التابعة النوعية أو الفئوية بالإعتماد على مجموعة متغيرات مستقلة مختلطة ، كأن يكون قسم منها متغيرات مستمرة ، القسم الآخر تكون على شكل متغيرات متقطعة نوعية أو فئوية[6] .

أو يمكن تعريفه على أنه نموذج يستخدم للتنبؤ بإحتمالية وقوع حدث ما من عدم وقوعه وذلك بتمثيل البيانات على منحنى لوجستي ، ويستخدم النموذج متغيراً توضيحياً واحداً أو عدة متغيرات توضيحية متوقعة والتي يمكن ان تكون رقمية أو فئوية ، على سبيل المثال حدوث نوبة قلبية عند شخص ما خلال فترة زمنية معينة يمكن التنبؤ بها من خلال معلومات عن عمر المريض وجنسه ومنسب كتلة الجسم لديه[51] .

ويعرف كذلك على إنه أداة أكثر قوة لأنه يقدم إختباراً لدلالة المعاملات ، كما أنه يعطي الباحث الفكرة عن مقدار تأثير المتغير المستقل في متغير الإستجابة الثنائية[27] ، ويعرفه آخرون بقولهم الإنحدار اللوجستي هو نموذج من أكثر النماذج شيوعاً في تحليل البيانات الوصفية ، وهو أُسلوب إحصائي لفحص العلاقة بين متغير الإستجابة (المتغير التابع) ذي المستوى الوصفي ومتغير واحد أو أكثر من المتغيرات التوضيحية (المستقلة) والتي تسمى أحياناً متغيرات مصاحبة أو متغيرات مفسرة (Explantory Variable) بحيث تكون تلك المتغيرات التوضيحية أو المستقلة من أي نوع من مستويات القياس[22] .

كما يطلق على الإنحار اللوجستي أسماء أُخرى في التطبيقات المختلفة له مثل :النموذج اللوجستي ، نموذج اللوجيت ، والمصنف العام للإنتروبية ، ويستخدم الإنحدار اللوجستي بشكل واسع في مجالات الطب

والعلوم الإجتماعية وكما يستخدم في التسويق لحساب توقعات ميل المستهلك الى شراء منتج ما أو إمتناعه عن الشراء [51] ، وحتى في مجالات الهندسة والتأمين الصحي[23] .

2-2-2 : من أهم خصائص الإنحدار اللوجستي [5]

أ- لايفترض وجود علاقة خطية بين المتغير التابع والمتغيرات التوضيحية أو التفسيرية .

ب-المتغير التابع يجب أن يكون ثنائي التفرع (Binary) .

ج- لايفترض تساوي التباين ضمن كل فئة ، وهذا يجعل أُنموذج الإنحدار اللوجستي أكثر مرونة من بقية نماذج التنبؤ والتصنيف .

د- يجب أن تكون الفئات محددة وشاملة بحيث كل مفردة تنتمي الى فئة واحدة فقط .

هـ-يتم تقدير معاملات إنموذج الإنحدار اللوجستي بإستخدام دالة الإمكان الأعظم (Maximum Likelihod Method) وهي طريقة تحتاج الى عينة كبيرة الحجم نسبياً.

2-2-3 : تقدير معلمات إنموذج الإنحدار اللوجستي

يبنى إنموذج الإنحدار اللوجستي على فرض أساسي هو أن متغير الإستجابة أو المتغير التابع (y) هو متغير ثنائي الإستجابة ويأخذ الرتبة أما (1) بإحتمال (p) أو (0) بإحتمال (1-p) ، أي الى حدوث الإستجابة وعدم حدوثها .

وتأخذ دالة الإستجابة (الدالة اللوجستية) الشكل الآتي

$$f(z)=E(Y/z) = \frac{e^z}{1+e^z} \quad \ldots\ldots\ldots(20.2)$$

وهي مثل نظرية الإحتمالات تأخذ القيم أما الصفر أو الواحد ، وتكون عملية التصنيف خاضعة

لمخرجات هذه الدالة فإذا كانت قيمتها تساوي (واحداً) يعني ان المشاهدة تنتمي للمجموعة الأولى وإذا كانت قيمتها تساوي (صفراً) يعني أن المشاهدة تنتمي للمجموعة الثانية ، وتأخذ دالة الإستجابة مدخلات من سالب مالانهاية الى موجب ما لانهاية ، لكن المخرجات دائماً بين الواحد والصفر ، و يمثل المتغير z المتغيرات التوضيحية إذ f(z) تمثل الإحتمال لمخرج معين لمجموعة من المتغيرات التوضيحية ، ويقيس المتغير z مجموع مساهمات جميع المتغيرات التوضيحية المستخدم في هذا الإنموذج والتي تعرف باللوجت .

ويعرف المتغير z كالآتي :

$$z=\beta_0 +\beta_1 x_1 \ldots..(21.2)$$

أما في حالة الإنحدار اللوجستي المتعدد فيكتب بالشكل الآتي :

$$z= \beta_0+ \beta_1 x_1+ \beta_2 x_2+ \beta_3 x_3+\ldots\ldots+ \beta_k x_k \ldots\ldots(22.2)$$

حيث أن $\beta_0,\beta_1,\ldots\ldots\ldots,\beta_k$ تمثل معلمات النموذج [51]

ولتحويل الصيغة (20.2) الى الشكل الخطي يتم إستعمال ما يعرف بتحويلة لوجت (Logit Transformation) التي يعبر عنها بالصيغة الآتية :

$$g(z)=Ln\left[\frac{f(z)}{1-f(z)}\right]= \beta_0 + \beta_1 x_1+ \beta_2 x_2+ \beta_3 x_3+\ldots\ldots+ \beta_k x_k\ldots\ldots\ldots(23.2)$$

ويمكن إثبات ذلك من خلال الخطوات التالية :

$$g(z)=Ln\left[\frac{\frac{e^z}{1+e^z}}{1-(\frac{e^z}{1+e^z})}\right] \longrightarrow \quad g(z)=Ln\left[\frac{\frac{e^z}{1+e^z}}{(\frac{1+e^z-e^z}{1+e^z})}\right] \longrightarrow \quad g(z)=Ln\left[\frac{\frac{e^z}{1+e^z}}{(\frac{1}{1+e^z})}\right]$$

$$g(z)=Ln\left[\frac{e^z}{1+e^z} * \frac{1+e^z}{1}\right] \longrightarrow \quad g(z)=Ln[e^z] \longrightarrow \quad g(z)=z$$

$\because z= \beta_0 + \beta_1 x_1+ \beta_2 x_2+ \beta_3 x_3+\ldots\ldots+ \beta_k x_k$

$\therefore g(z)= Ln\left[\frac{f(z)}{1-f(z)}\right]= \beta_0 + \beta_1 x_1+ \beta_2 x_2+ \beta_3 x_3+\ldots\ldots+ \beta_k x_k$

إذ أن $p(Y=1/x)=f(x)$ عندما $Y=1$ و تمثل $p(Y=1/x)$ الإحتمال الشرطي (conditional probability) بأن تكون $Y=1$ عند قيمة معينة لـ (x) .

$p(Y=0/x)=1-f(x)$ عندما $Y-0$ وتمثل $p(Y=0/x)$ الإحتمال الشرطي(conditional probability) بأن تكون $Y=0$ عند قيمة معينة لـ (x) .

ولتقدير معلمات المعادلة (23.2) وهي $\beta_0, \beta_1,\ldots., \beta_k$ يتم إستخدام طريقة الإمكان الأعظم Maximum Likelihood Estimation (MLE) [5] . كما في المعادلة (27.2)

حيث يبدأ الإجراء في الإنحدار اللوجستي (Logistic Regression) بصيغة لأرجحية مشاهدة نمط حدوث السمة المطلوبة (Y=1) أو عدم حدوثها (Y=0) في عينة ما ، وهذا مايعرف اصطلاحاً بدالة الترجيح (Likelihood Function) ، وتفترض دالة الترجيح أن جميع الحالات مستقلة فإن إحتمال الحصول على البيانات المشاهدة للحالة i عندما تكون (Y=1) تعطى بالحد $p(x_i)$ في حين أن إحتمال الحصول على البيانات المشاهدة للحالة i والتي تكون فيها (Y=0) تعطى بالحد $1-P(x_i)$ إذ أن :

$$p(Y=1/x) = f(x) \quad \text{عندما } Y=1$$

$$p(Y=0/x)=1-f(x) \quad \text{عندما } Y=0 \quad \text{و}$$

وكما سبق أن تم الذكر بأن $f(x_i)$ هي معادلة النموذج اللوجستي للمتغير x [3]

$$\because \quad f(x) = \text{Logistic model}$$

$$\therefore f(x) = \frac{e^{(\beta_0+\sum \beta_i x_i)}}{1+e^{(\beta_0+\sum \beta_i x_i)}} \quad \ldots\ldots\ldots\ldots(24)$$

$$\therefore L= \frac{\prod_{i=1}^{n} \exp(\beta_0+\sum_{i=0}^{k} \beta_i x_i)}{\prod_{i=1}^{n}[1+\exp(\beta_0+\sum_{i=0}^{k} \beta_i x_i)}\ldots\ldots\ldots(25)$$

ويلاحظ من الصيغة المذكورة آنفاً أن الحالة التي يكون فيها Y_i تساوي الواحد ، فإن المعادلة تتقلص لتصبح P_i مرفوعة للقوة واحد والتي تساوي P_i و $(1-P_i)$ مرفوعة للأُس صفر تساوي الواحد الصحيح ، لذا عندما تكون $Y_i=1$ فإن قيمة الحالة تساوي إحتماليتها المتوقعة وبناءً على ذلك إذا كانت الحالة لديها قيمة إحتمال متوقعة مرتفعة للحدوث عندما تكون $Y_i=1$ فإن مساهمتها في الأرجحية تكون أعلى بالمقارنة بما لو كان احتمال ضعيف الحدوث .

أما بالنسبة للحالة عندما تكون $Y_i=0$ فإن المعادلة ستصبح $(1-P_i)$ لأن P_i مرفوعة للأُس صفر في حين أن $1-P_i$ مرفوعة للأُس واحد وبذلك فهي تساوي $1-P_i$ ولذا عندما تكون $Y_i=0$ ، فإن قيمة الحالة تساوي $1- P_i$ ، لذا إذا كانت الحالة لها إحتمال متوقع منخفض القيمة للحدوث أي عندما $Y_i=0$ ، فإن مساهمتها أكبر للأرجحية مقارنة بما لو كانت لها قيمة إحتمال مرتفعة ومثال ذلك إذا كانت $P_i=0.1$ فإن $1- P_i=0.9$ يعني أنها تؤثر أكثر مما لوكانت $P_i=0.9$ و $1- P_i=0.1$ ، لذا عندما تكون لدينا عينة من المشاهدات التوضيحية (التفسيرية) بحجم (n) فإن لكل زوج (x_i,y_i) حيث $i=1,2,\ldots\ldots,n$ ، وإن (y_i) تمثل رتبة متغير الإستجابة الثنائي للمفردة i و (x_i) تمثل قيمة المتغير المستقل للمفردة (i) فأن :

$$P(y_{i=1}/x)=f(x_i)^{y_i} \quad \text{عندما تكون } y_i=1$$

$$P(y_i=0/x)=[1-f(x_i)]^{1-y_i} \quad \text{عندما تكون } y_i=0$$

و يعبر عن دالة الإمكان بالصيغة الرياضية الآتية :

$$L(\beta) = \prod_{i=1}^{n} f(x_i)^{yi}[1 - f(x_i)]^{1-yi} \dots\dots(26.2)$$

وبأخذ اللوغاريتم الطبيعي للطرفين نحصل على المعادلة الآتية :

$$L(\beta) = Ln[L(\beta)] = \sum_{i=1}^{n} \{y_i \ln[f(x_i)] + (1 - y_i)\ln[1 - f(x_i)]\} \dots\dots(27.2)$$

ثم يتم إشتقاق المعادلة المذكورة آنفاً بالنسبة للمعلمات المراد تقديرها (β_i) وجعلها مساوية للصفر لينتج عدد من المعادلات التي لايمكن حلها إلا من خلال خوارزمية تكرارية ، تسمى خوارزمية المربعات الصغرى الموزونة التكرارية[5] (iteratively Weighted Least Squares Algorithm).

2-2-4 تقييم القوة التفسيرية لأُنموذج الإنحدار اللوجستي :

يتم إستعمال إحصاءة $R^2_{\text{cox \& snel}}$ أو إحصاءة $R^2_{\text{Nagelkerke}}$ لغرض إختبار القوة التفسيرية لنموذج الإنحدار اللوجستي أي إختبار جودة معادلة الإنحدار التقديرية في تفسير العلاقة بين متغير الإستجابة (التابع) والمتغيرلت التوضيحية (المستقلة) وهاتان الإحصائتان لهما هدف معامل التحديد R^2 نفسه المستخدم في الإنحدار الخطي المتعدد ،إلا أن $R^2_{\text{cox \& snel}}$ لا يمكن أن تصل قيمتها الى الواحد الصحيح في حين أن $R^2_{\text{Nagel kerke}}$ بإمكانها ذلك فحدود $R^2_{\text{Nagelkerke}}$ تمتد من (الصفر الى الواحد الصحيح) مما يجعلها أكثر موثوقية من $R^2_{\text{cox \& snel}}$ ومن الطبيعي أن تكون قيمتها أعلى $R^2_{\text{cox \& snel}}$ وتحسب إحصاءة $R^2_{\text{cox \& snel}}$ من الصيغة الرياضية الآتية[10] :

$$R^2_{\text{cox \& snel}} = 1 - \left[\frac{L0}{L1}\right]^{(2/n)} \dots\dots(28.2)$$

حيث أن :

L_0 : دالة الإمكان الأعظم في حالة الإنموذج المتضمن الحد الثابت فقط .

L_1 : دالة الإمكان الأعظم في حالة الإنموذج المتضمن جميع المتغيرات التوضيحية .

n : حجم العينة .

أما إحصاءة $R^2_{\text{Nagel kerke}}$ فتحسب من الصيغة الرياضية الآتية[10] :

$$R^2_{\text{Nagelkerke}} = \frac{R^2_{\text{cox \& snel}}}{1-[L0]\ (2/n)} \dots\dots(29.2)$$

5-2-2: الإختبارات الإحصائية الخاصة بأُنموذج الإنحدار اللوجستي

أولاً :- إختبار wald : تختبر إحصاءة wald التي تتبع توزيع مربع كاي (χ^2) وبدرجة حرية df=1 الدلالة الإحصائية لكل معلمة من معلمات إنموذج الإنحدار اللوجستي ، وذلك لإختبار الفرضية الصفرية القائلة إن تأثير معامل لوجت ما يساوي صفراً .

وتحسب إحصاءة wald وفق الصيغة الرياضية الآتية[9] :

$$Wald = \left[\frac{\beta_j}{S.E(\beta_j)}\right]^2 \quad \ldots\ldots\ldots\ldots\ldots\ldots(30.2)$$

حيث أن β_j : المعلمة المقدرة ذات الرتبة (j) .

$S.E(\beta_j)$: الخطأ المعياري للمعلمة المقدرة ذات الرتبة (j) .

أما فرضية الإختبار فهي :

$H_0 : \beta_j = 0$ vs $H_1 : \beta_j \neq 0$ j=1,2,……,k

(أي إختبار معنوية كل معلمة من معلمات إنموذج الإنحدار اللوجستي) وهو إختبار من طرفين[9] ، فإذا كانت القيمة الإحتمالية (significance Value) لإحصاءة Wald أقل من (0.05) عندئذ نرفض فرضية العدم[7] ، أي أن المعلمة (β_j) معنوية ولاتساوي صفراً في المجتمع الذي سحبت منه العينة[9] .

أما إذا تم إحتساب w بدلاً من w^2 فإن اختبار والد Wald.Test سوف يتبع توزيع Z [29] .

ثانياً :- إختبار Hosmer& Lemshow (H&L)

يعتمد هذا الإختبار على تجميع حالات العينة بناءً على قيم الإحتمالات المتوقعة على وفق إحدى استراتيجيتين للتجميع هما :

1- تجميع الحالات بناءً على المئينيات للإحتمالات المتوقعة .

2- أو تجميع الحالات بناءً على قيم ثابتة للإحتمالات المتوقعة .

وأياً كانت طريقة تجميع الحالات ، فإنه يتم تجميع المشاهدة والمتوقعة للحالات على وفق قيمتي y (y=1 و y=0) وذلك في كل فئة من مجموعات التصنيف ، ثم يتم حساب إحصاءة هوزمر –ليمشو لجودة المطابقة والتي يرمز لها بالرمز $\hat{C}$ بحيث يتم حسابها على وفق حساب إحصاءة مربع كاي

لبيرسون من الجدول g*2 للتكرارات المشاهدة والمتوقعة ، وتكتب معادلة احصاءة هوزمر – ليمشو بالشكل التالي :

$$\hat{C} = \sum_{k=1}^{g} \frac{(O_k - n'_k \bar{P}_k)^2}{n'_k \bar{P}_k (1-\bar{P}_k)} \quad \text{.....(31.2)}$$

حيث أن n'_k هي العدد الكلي للحالات في المجموعة k

$O_k = \sum_{i=1}^{n'_k} y_i$ أي أن O_k هي عدد الإستجابات y=1

$\bar{P}_k = \sum_{i=1}^{n'_k} \frac{P_i}{n'_k}$ ،وهي متوسط الإحتمالات المتوقعة للمجموعة k [3] .

وحيث أن الإحصاءة $\hat{C}$ تتبع توزيع مربع كاي بدرجات حرية تساوي g−2 [29] ، حيث g تمثل عدد المجموعات فإنها تستعمل لمعرفة فيما إذا كان النموذج يمثل البيانات بشكل جيد أم لا (Well−fitting model or not) ، من خلال تقييم الفرق بين القيم المشاهدة (observed value) والقيم المتوقعة (expected values) ، وإختبار الفرضية الآتية :

H_0 : عدم وجود فرق معنوي بين القيم المشاهدة والقيم المتوقعة .

H_1 : وجود فرق معنوي بين القيم المشاهدة والقيم المتوقعة .

فإذا كانت إحصاءة H&L أكبر من (0.05) عندئذ يكون الإنموذج ممثلاً بشكل جيد . نقبل فرضية العدم القائلة بعدم وجود إختلافات بين القيم المشاهدة والقيم المتوقعة [47] .

2−3 جدول التصنيف Classification Table :

هو جدول يوضح عدد الحالات المشاهدة التي تمتلك صفة ما وعدد الحالات المشاهدة التي لاتمتلك تلك الصفة في مقابل عدد الحالات المتوقعة التي تمتلك الصفة وعدد الحالات المتوقعة التي لاتمتلك تلك الصفة بحيث يوضح الجدول عدد الحالات التي تم تصنيفها بطريقة صحيحة وعدد الحالات التي تم تصنيفها بصورة خاطئة ، وتعتمد فكرة إستخدام هذا التحليل على أن النموذج إذا قام بتوقع تصنيف الحالات بشكل صحيح إعتماداً على معيار ما فإن ذلك يعطي برهاناً بأن النموذج يطابق البيانات المشاهدة ، أما الشكل العام لجدول التصنيف فهو كما في أدناه :

جدول (2-1)

الشكل العام لجداول التصنيف

التصنيف		المتوقع		المجموع
		Pالموجب	Nالسالب	
المشاهد	Pالموجب	TP الموجب الصحيح	FN السالب الخاطئ	P
	N السالب	FP الموجب الخاطئ	TN السالب الصحيح	P'
المجموع		Q	Q'	

المصدر: [24]

ويفسر الجدول المذكور آنفاً :

TP : هي المشاهدات الموجبة التي تم تصنيفها بشكل صحيح مشاهدات موجبة .

FP : هي المشاهدات الموجبة التي تم تصنيفها خطأً على أنها مشاهدات سالبة .

TN : هي المشاهدات السالبة التي تم تصنيفها بشكل صحيح مشاهدات سالبة .

FN : هي المشاهدات التي تم تصنيفها خطأً على أنها مشاهدات موجبة .

P : هي عبارة عن حاصل جمع TP+FN

P' : هي عبارة عن حاصل جمع FP+TN

Q : هي عبارة عن حاصل جمع TP+FP

Q' : هي عبارة عن حاصل جمع FN+TN

أما المشاهد فيقصد به المشاهدات الحقيقية والمتوقع يقصد به ما تم توقعه في عملية التصنيف مثلاً إذا كان العدد الحقيقي للمشاهدات الموجبة هو (100) ومن خلال عملية التصنيف تم توقع (90) منها بشكل صحيح و(10) بشكل خاطئ تم تصنيفها على انها مشاهدات سالبة ، وايضاً إذا كان العدد الحقيقي للمشاهدات السالبة (50) وتم تصنيف (45) بشكل صحيح و(5) منها تم تصنيفه على أنها مشاهدات موجبة . فيكون بذلك العدد الكلي المتوقع للمشاهدات الموجبة هو(95) بينما هو في الحقيقة كان (100) والعدد الكلي للمشاهدات السالبة المتوقعة هو (55) بينما هو في الحقيقة (50) .

2-3-1 قانون نسبة التصنيف الصحيح :

ويتم حسابها من خلال المعادلة التالية :

$$PC = \frac{TP + TN}{N} \%100 \dots\dots\dots (32.2)$$

حيث أن :

TP : عدد المشاهدات التي تم تصنيفها بصورة صحيحة في المجموعة الأولى .

TN : عدد المشاهدات التي تم تصنيفها بصورة صحيحة في المجموعة الثانية .

N : مجموع المشاهدات الكلي للمجموعتين .

الفصل الثالث

الجانب التجريبي

(Experimental Part)

الجانب التجريبي

نظراً للتقدم العلمي الهائل في مجال الحاسبات الإلكترونية وعلى المستويين (Hardware) و(Software) أصبح من السهل على أي باحث أن يستفيد من تلك التقنيات في مجال البحث العلمي لما توفره من إمكانات مما توفر عليه الكثير من الوقت والجهد ، ومن هنا جاء إستخدام المحاكاة بوصفها إحدى الأدوات التي تحاول إعادة عملية ما في ظروف إصطناعية مشابهة الى حد ما للظروف الطبيعية لتلك العملية أي هي ببساطة عبارة عن عملية تقليد للأداة الحقيقية[51] .

تستخدم المحاكاة لتوليد بيانات غير حقيقية تحاكي البيانات الحقيقية إذ تقوم بتوظيف نماذج تكون فيها العديد من الحالات الإفتراضية لتكون نتائج التحليل أكثر شمولاً وتعميماً [8] .

وغالباً ما يلجأ الى هذه الطريقة عندما تكون هناك صعوبة في الواقع الحقيقي في إستخدام التحليل المنطقي لتفسير بعض النظريات والمشكلات الإحصائية وبإستخدام البراهين الرياضية [11] ، وكذلك عند الصعوبة في الحصول على البيانات المطلوبة وصعوبة توافر الكم الهائل من هذه البيانات التي يتم توليدها بسهولة عن طريق استخدام المحاكاة وبإستخدام الحاسوب وفي الوقت نفسه تمتاز بالدقة .

3-1 مرحلة بناء تجربة المحاكاة : Stage of the Simulation Experiment

أولاً : توليد المتغيرات التوضيحية (التفسيرية) Explanatory Variables Generation

لتوليد متغيرات توضيحية تتبع التوزيع الطبيعي ، فإن طريقة Box-Muller تعد من أشهر الطرائق وأكثرها شيوعاً ، والتي تعتمد على توليد عددين عشوائيين يتبعان التوزيع المنتظم القياسي $U(0,1)$ ، ثم يتم تحويل هذين العددين الى متغيرين عشوائيين مستقلين z_1 و z_2 يتبعان التوزيع الطبيعي القياسي على وفق الصيغة التالية :

$$\left.\begin{aligned} Z_1 &= (-2\,Ln\,U_1)^{\frac{1}{2}} \,.\, \cos(2\,\pi\,U_2) \\ Z_2 &= (-2\,Ln\,U_1)^{\frac{1}{2}} \,.\, \sin(2\,\pi\,U_2) \end{aligned}\right\} \quad \dots\dots\dots(1-3)$$

ولتحويل المتغيرات من التوزيع الطبيعي القياسي الى التوزيع الطبيعي بمتوسط μ وتباين σ^2 يتم إستعمال التحويل التالي :

$$X_i = \mu + \sigma\, z \qquad , \quad i = 1\,,2\,,\dots,n \quad \dots\dots\dots(2-3)$$

إذ أن σ : الإنحراف المعياري للخطأ (e) .

$$X_i \sim i.i.d\ N\ (\mu\ ,\ \sigma^2)\quad i= 1,2,3,4,5$$

إذ يولد كل متغير توضيحي بصورة مستقلة عن الآخر .

ثانياً: اختيار القيم الإفتراضية فقد أُختيرت ثلاثة حجوم مختلفة هي (n=50) و (n=100) و (n=216) حجم العينة الكبيرة . وتم عرض ثلاثة مستويات للتباين هي $\sigma^2 = 1$ و $\sigma^2 = 1.25$ و $\sigma^2 = 1.5$ ، وتم إختيار قيم للمتوسط (μ=0,0.1,0.2,0.3,0.4,0.5,0.6,0.7) وتم تكرار هذه التجارب بمقدار (RP=1000) مرة .

ثالثاً : حساب متغير الإستجابة Respone Variable

يتم حساب المتغير لمعتمد Y_i مباشرة من خلال توليد مجموعتين متباينتين في الوسط الحسابي والتباين .

3-2 النماذج المستعملة :

وهي النماذج التي تم تناولها في الفصل النظري من قبل الباحث وهي كل من أُنموذج آلة المتجه الداعم (SVM) وأُنموذج الإنحدار اللوجستي (LRM) والتي تمت الإشارة اليهما بالصيغتين (23.2) و (32.2) على التوالي .

3-3 تنفيذ تجارب المحاكاة

تم الإعتماد على الإيعازات التي يوفرها لنا برنامج (R-Language) في توليد بيانات تتبع لكل أُنموذج من النماذج المستعملة الواردة في الفقرة 3-2 وتم إجراء التجارب الآتية :-

1- عندما يكون التباين ($\sigma^2 = 1$) والوسط الحسابي (μ=0,0.1,0.2,0.3,0.4,0.5,0.6,0.7) .

أ- التجربة الأولى عند حجم العينة (n=50)

ب- التجربة الثانية عند حجم العينة (n=100)

ت- التجربة الثالثة عند حجم العينة (n=216)

2- عندما يكون التباين ($\sigma^2 = 1.25$) والوسط الحسابي (μ=0,0.1,0.2,0.3,0.4,0.5,0.6,0.7) .

أ- التجربة الأولى عند حجم العينة (n=50)

ب- التجربة الثانية عند حجم العينة (n=100)

ت- التجربة الثالثة عند حجم العينة (n=216)

3- عندما يكون التباين ($\sigma^2 = 1.5$) والوسط الحسابي (μ=0,0.1,0.2,0.3,0.4,0.5,0.6,0.7) .

أ- التجربة الأولى عند حجم العينة (n=50)

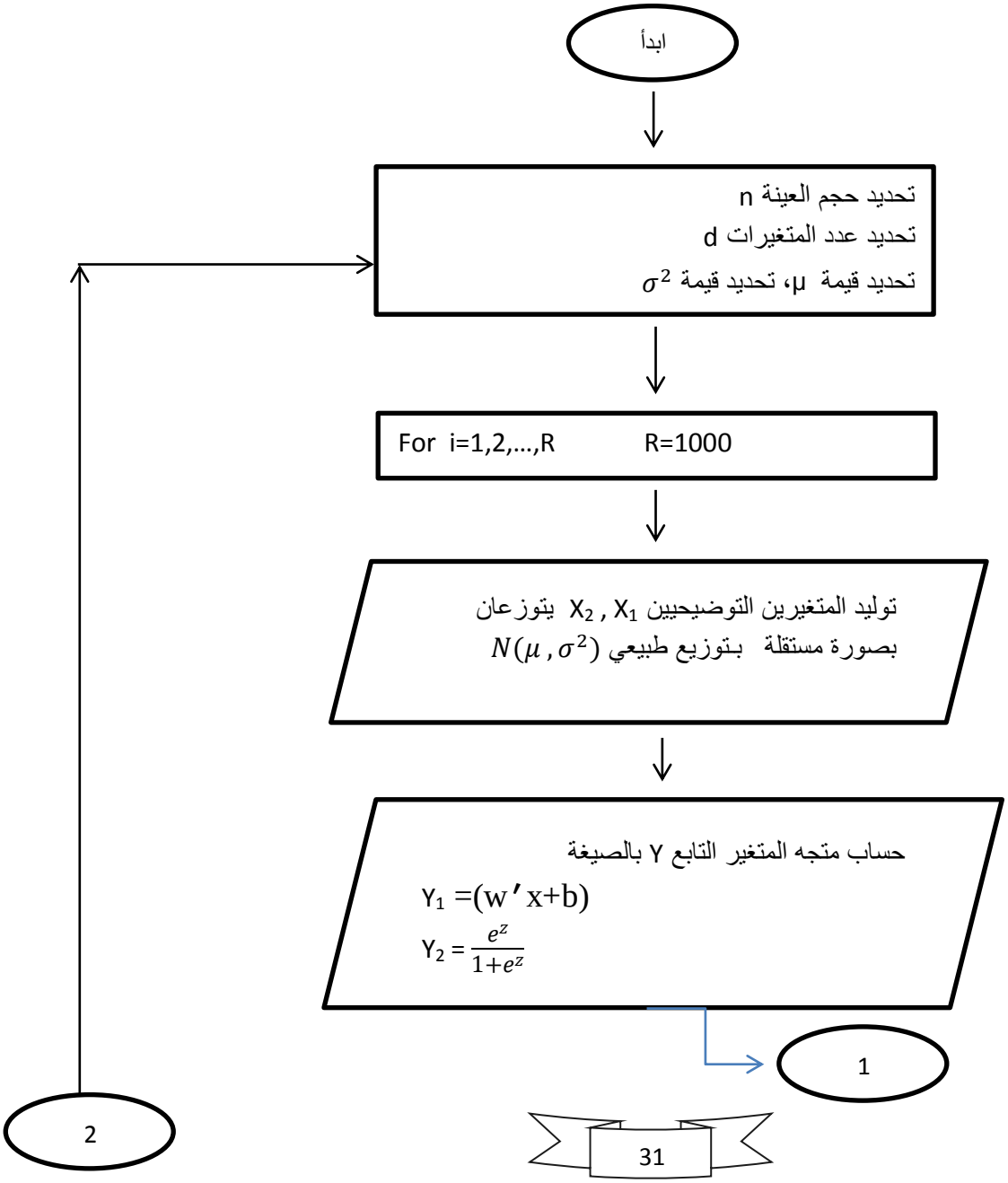
ب- التجربة الثانية عند حجم العينة (n=100)

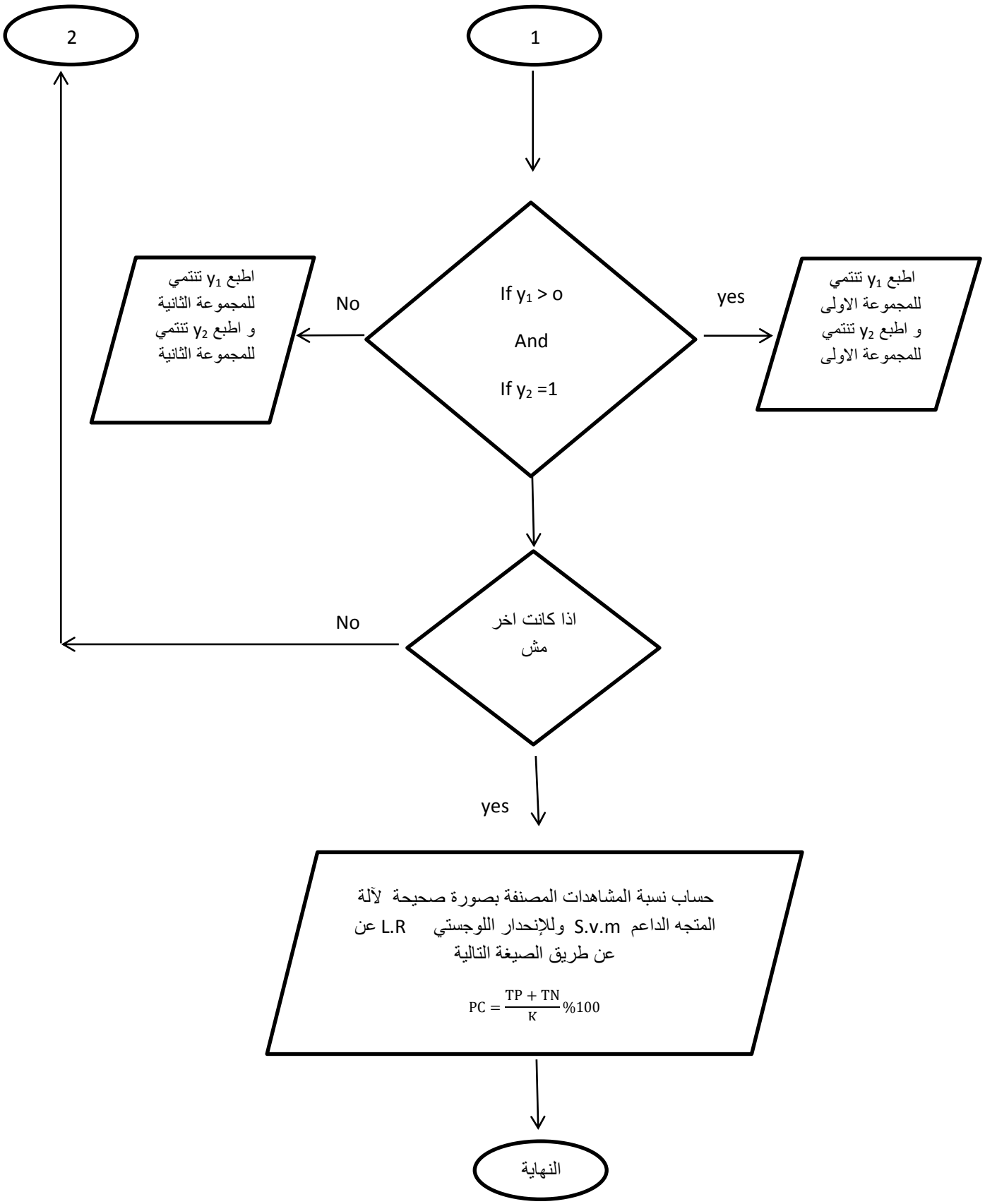
ت- التجربة الثالثة عند حجم العينة (n=216)

ولتوضيح الخطوات السابقة يتم إتباع المخطط الإنسيابي الآتي الذي يوضح آلية المحاكاة للمقارنة بين طرائق التصنيف ولجميع النماذج المستعملة .

الشكل (3-1)

المخطط الإنسيابي لآلية المحاكاة لنموذجي (SVM) و (LRM)




2
1
If $y_1 > o$
And
If $y_2 = 1$
No
اطبع y_1 تنتمي للمجموعة الثانية و اطبع y_2 تنتمي للمجموعة الثانية
yes
اطبع y_1 تنتمي للمجموعة الاولى و اطبع y_2 تنتمي للمجموعة الاولى
اذا كانت اخر مش
No
yes
حساب نسبة المشاهدات المصنفة بصورة صحيحة لآلة المتجه الداعم S.v.m وللإنحدار اللوجستي L.R عن
عن طريق الصيغة التالية
$$PC = \frac{TP + TN}{K} \%100$$
النهاية

4-3 تحليل نتائج المحاكاة : Analysis of the Simulation

أولاً:- أظهرت نتائج المحاكاة عند قيمة التباين ($\sigma^2 = 1$) ولعينة بحجم (n=50) أن طريقة SVM هي الأفضل وفي جميع حالات μ لغاية μ =0.7 حيث كانت نسبة التصنيف متساوية بين الطريقتين وكما هو مبين في الجدول (3-1) .

جدول (3-1)
نتائج التصنيف عند مستوى تباين ($\sigma^2 = 1$) وحجم عينة (n=50)

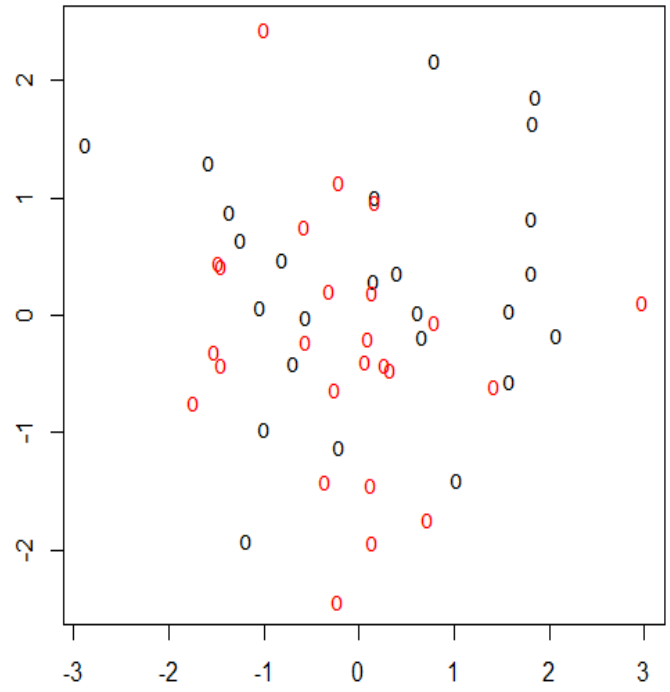
الوسط الحسابي μ	نسبة التصنيف بواسطة SVM	نسبة التصنيف بواسطة LRM
0.0	81 %	63 %
0.1	81 %	65 %
0.2	83 %	71 %
0.3	87 %	78 %
0.4	90 %	84 %
0.5	93 %	90 %
0.6	95 %	94 %
0.7	97 %	97 %

إعداد الباحث بالإعتماد على نتائج برنامج (R- language)

يظهر الجدول المذكور آنفاً نسب التصنيف الصحيح لكلا الطريقتين عند حجم العينة الصغيرة (n=50) والتباين ($\sigma^2 = 1$) إذ تراوحت نسبة التعرف الصحيح لـ SVM ما بين 81% عندما كانت قيمة μ=0 ولغاية 97% عندما أصبحت قيمة μ=0.7 أما نسبة التعرف الصحيح لـ LRM فبلغت 63% عند قيمة μ=0 حتى وصلت 97% عند قيمة μ=0.7 مما يعني تفوق SVM في جميع حالات (μ) ما عدا الحالة الأخيرة إذ تعادلت الآليتان في نسبة التعرف الصحيح بسبب تباعد بيانات المجموعتين عن بعضهما البعض، كذلك يمكن ملاحظة الفارق الواضح بين دقة SVM وLRM في نسبة التعرف الصحيح في الحالات التي يكون فيها التداخل شديداً بين مشاهدات المجموعتين عند قيم (μ=0,0.1,0.2,0.3) وعلى الرغم من إنخفاض نسبة التداخل عند قيم (μ=0.4,0.5,0.6) إلا أن التفوق كان أيضاً لصالح SVM ، كما يمكن توضيح عملية التداخل بين المشاهدات في هذه الحالة عندما يكون التباين $\sigma^2 = 1$ وحجم العينة n=50 من خلال الرسومات والأشكال التالية :

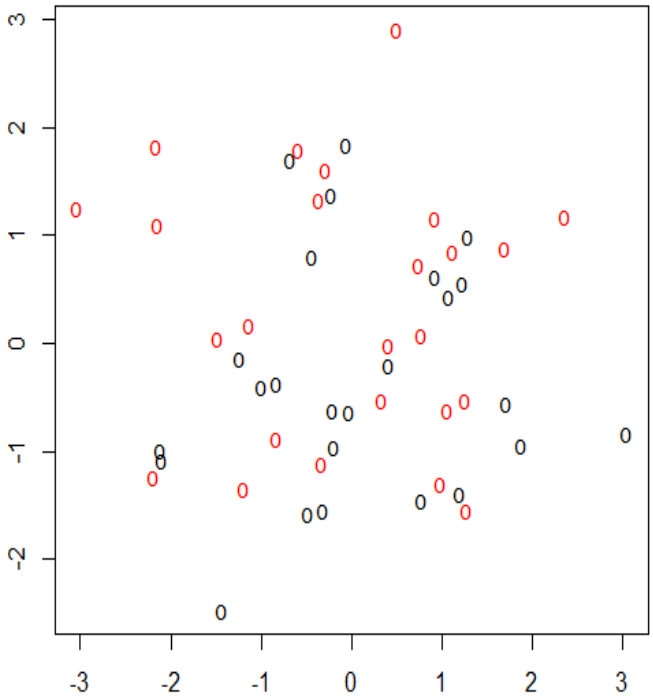
شكل (2-3)

عند مستوى $\sigma^2 = 1$ وحجم عينة n=50 و μ =.0



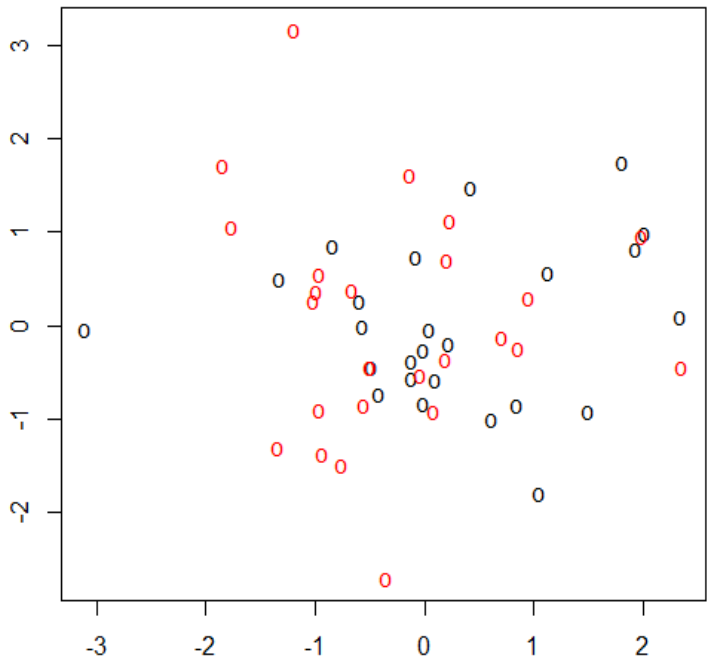
شكل (3-3)

عند مستوى $\sigma^2 = 1$ وحجم عينة n=50 و μ =0.1



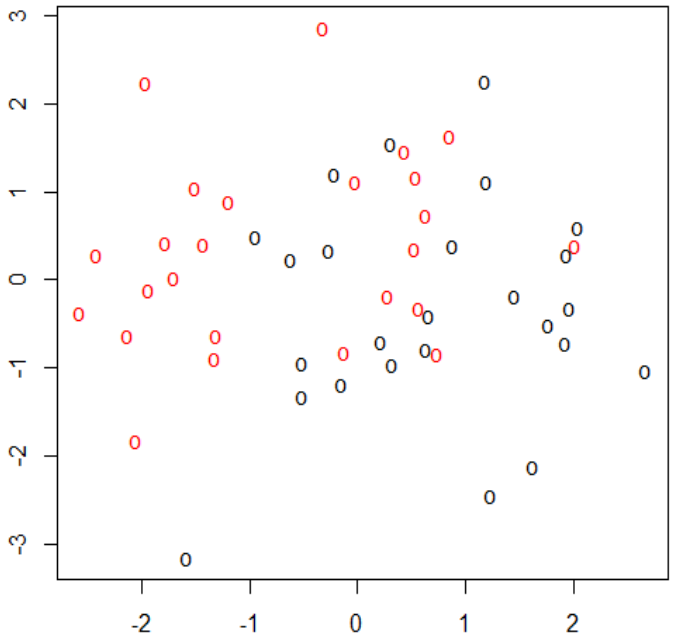
شكل (4-3)

عند مستوى $\sigma^2 = 1$ وحجم عينة n=50 و μ =0.2



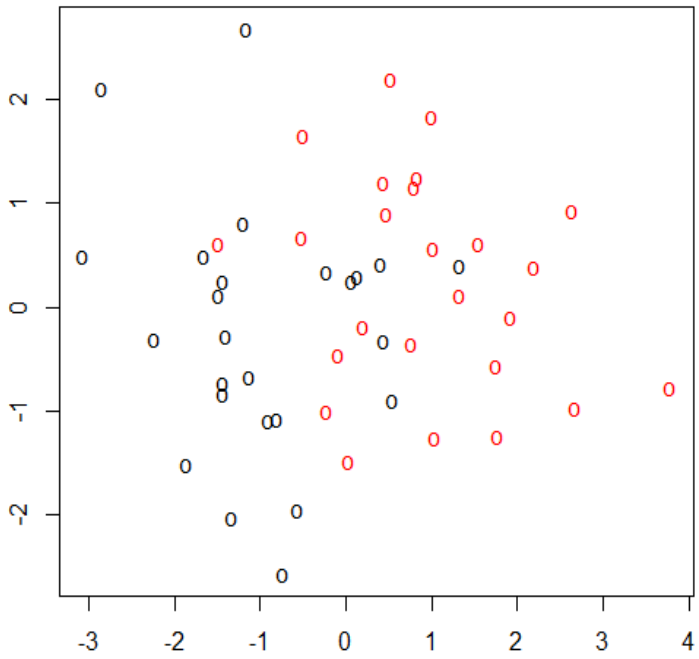
شكل (5-3)

عند مستوى $\sigma^2 = 1$ وحجم عينة n=50 و μ =0.3



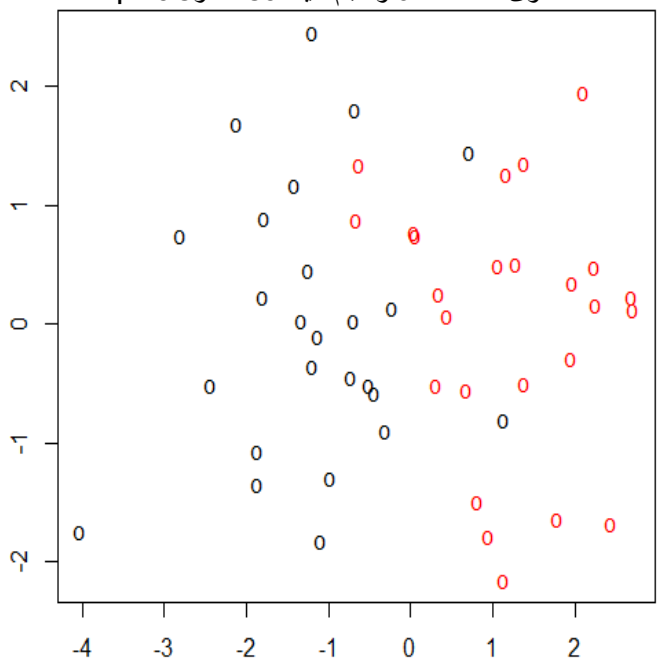
شكل (3-6)

عند مستوى $\sigma^2 = 1$ وحجم عينة n=50 و μ =0.4



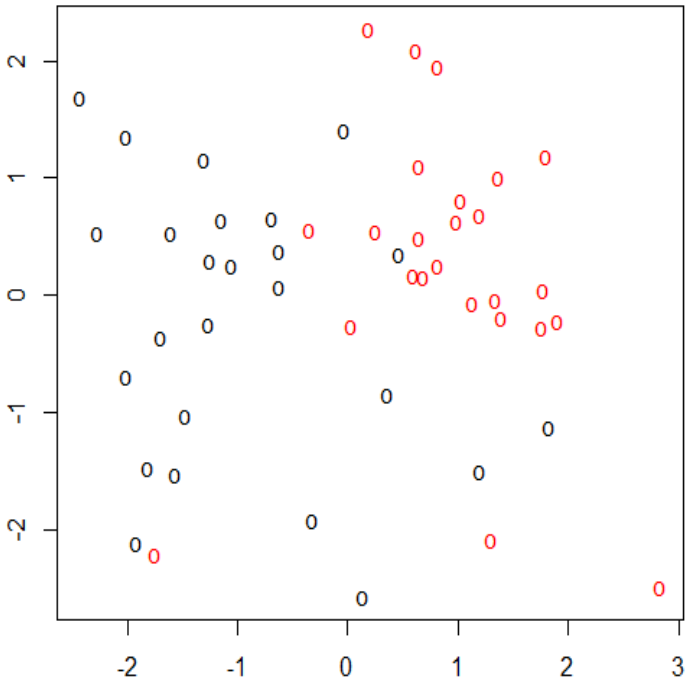
شكل (3-7)

عند مستوى $\sigma^2 = 1$ وحجم عينة n=50 و μ =0.5



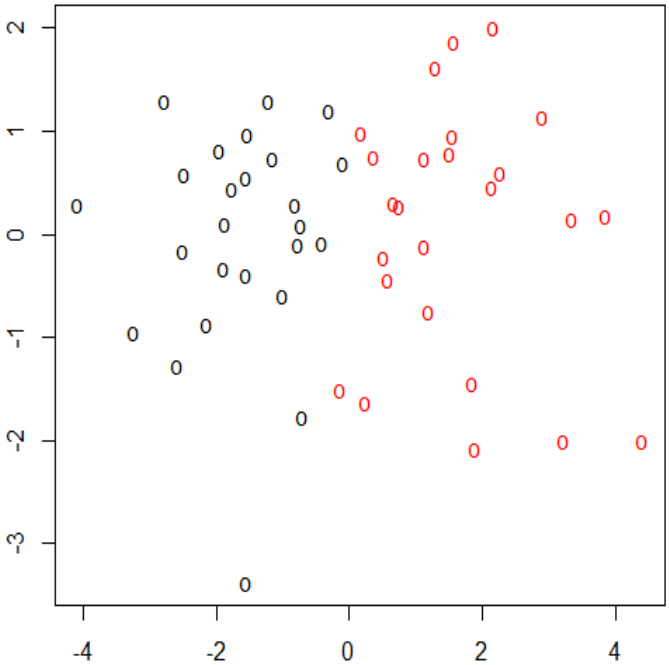
شكل (3-8)

عند مستوى $\sigma^2 = 1$ وحجم عينة n=50 و μ =0.6



شكل (3-9)

عند مستوى $\sigma^2 = 1$ وحجم عينة n=50 و μ =0.7



يلاحظ من الأشكال (3-2) و(3-3) و(3-4) وعندما يكون حجم العينة الصغيرة (n=50) إن التداخل بين البيانات يكون قوياً عندما تكون قيمة μ تساوي (0) أو (0.1) أو (0.2) أو (0.3) غير أن البيانات بدأت تنفصل بشكل أكبر عندما أصبحت قيمة μ تساوي (0.3) وكما موضح في الشكل (3-5) ، ونلاحظ من الأشكال (3-6) و(3-7) و(3-8) ان بيانات المجموعتين بدأت بالإنفصال والتباعد عن بعضها أكثر بشكل يمكن معه تمييز كل مجموعة عن الأُخرى ، أما الشكل (3-9) فيظهر الإنفصال التام لمشاهدات المجموعتين بحيث نلاحظ عدم وجود حتى ولو مشاهدة واحدة متداخلة مع مشاهدات المجموعة الأخرى . ونستنتج مما سبق أن المشاهدات بدأت من حالة التداخل الشديد بينها عندما كانت قيمة μ تساوي (0) وانتهى التداخل بين المجموعتين عندما أصبحت قيمة μ تساوي (0.7) .

كما أظهرت نتائج المحاكاة وعندما تكون قيمة التباين ($\sigma^2 = 1$) ولعينة بحجم (n=100) أن طريقة SVM هي الأفضل وفي جميع حالات μ وكما هو مبين في الجدول (3-2) .

جدول (3-2)
نتائج التصنيف عند مستوى تباين ($\sigma^2 = 1$) و حجم عينة (n=100)

الوسط الحسابي μ	نسبة التصنيف بواسطة SVM	نسبة التصنيف بواسطة LRM
0.0	75 %	59 %
0.1	76 %	62 %
0.2	79 %	69 %
0.3	83 %	77 %
0.4	88 %	83 %
0.5	91 %	88 %
0.6	94 %	92 %
0.7	96 %	95 %

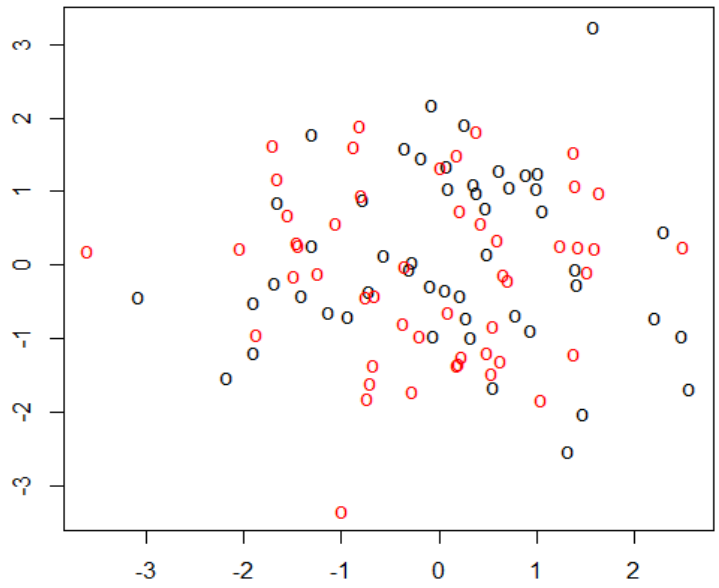
إعداد الباحث بالإعتماد على نتائج برنامج (R- language)

يظهر الجدول المذكور آنفاً نسب التعرف الصحيح لكلا الطريقتين عند حجم العينة المتوسطة (n=100) والتباين ($\sigma^2 = 1$) إذ تراوحت نسبة التعرف الصحيح لـ SVM ما بين 75% عندما كانت قيمة μ=0 ولغاية 96% عندما أصبحت قيمة μ=0.7 أما نسبة التعرف الصحيح LRM فقد بلغت 59% عند قيمة μ=0 وحتى وصلت 95% قيمة μ=0.7 مما يعني تفوق SVM في جميع حالات (μ) على الآلية الأُخرى، كذلك يمكن ملاحظة الفارق الواضح بين دقة SVM و LRM في نسبة التعرف الصحيح

في الحالات التي يكون فيها التداخل شديداً بين مشاهدات المجموعتين عند قيم (μ=0,0.1,0.2,0.3) وعلى الرغم من إنخفاض نسبة التداخل عند قيم (μ=0.4,0.5,0.6,0.7) إلا أن التفوق كان لـ SVM في جميع الحالات ، كما يمكننا توضيح الرسومات المرافقة لإجراء عملية التصنيف عند كل قيمة من قيم μ وعند التباين ($\sigma^2 = 1$) ولحجم العينة n=100 وملاحظة الإختلافات بينها ، وكما يلي :

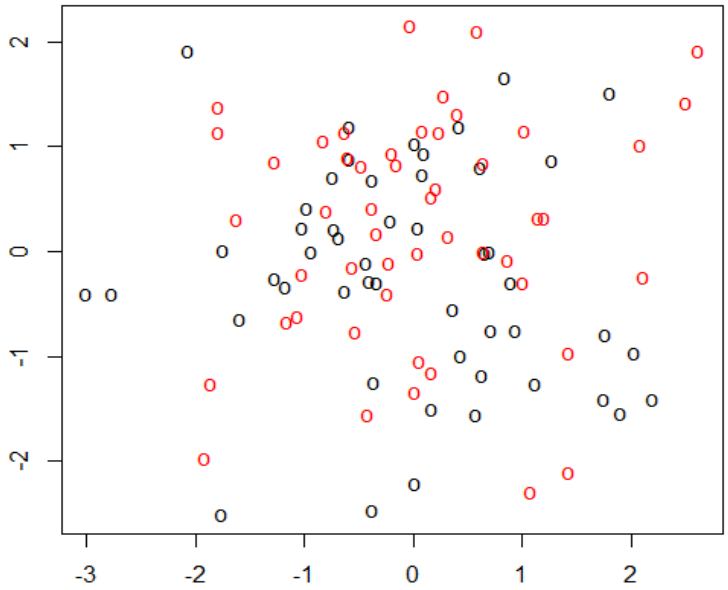
شكل (3-10)

عند مستوى $\sigma^2 = 1$ وحجم عينة n=100 و μ =0



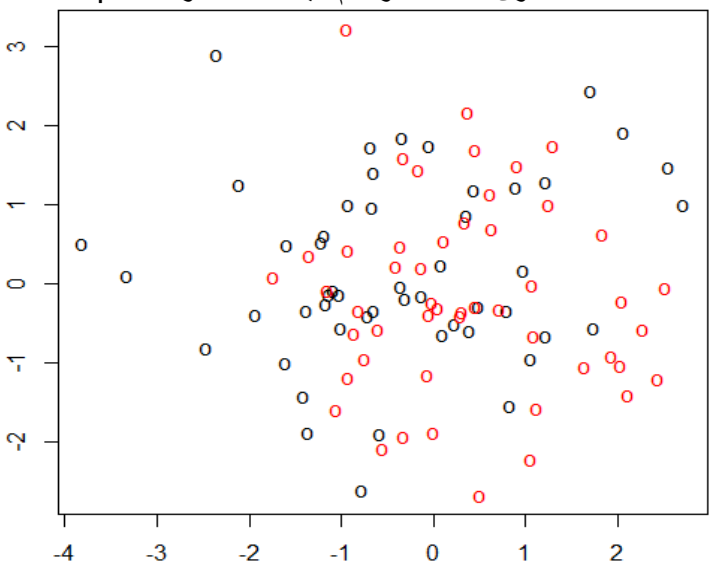
شكل (3-11)

عند مستوى $\sigma^2 = 1$ وحجم عينة n=100 و μ =0.1



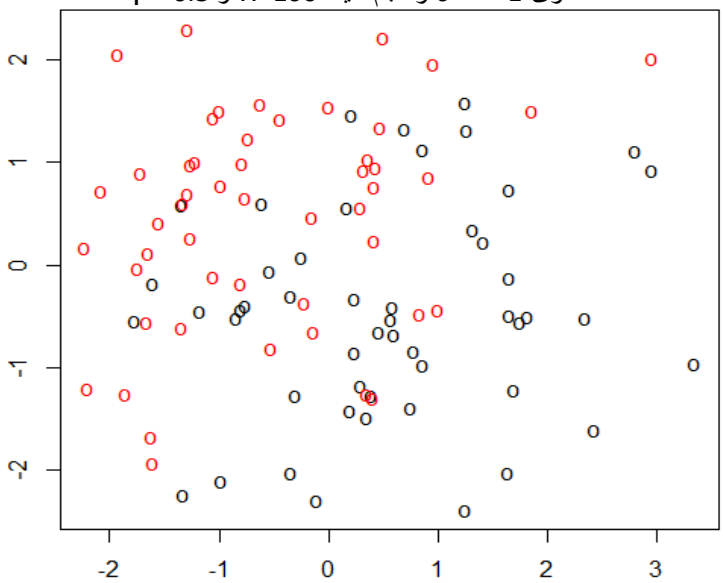
شكل (3-12)

عند مستوى $\sigma^2 = 1$ وحجم عينة n=100 و μ =0.2



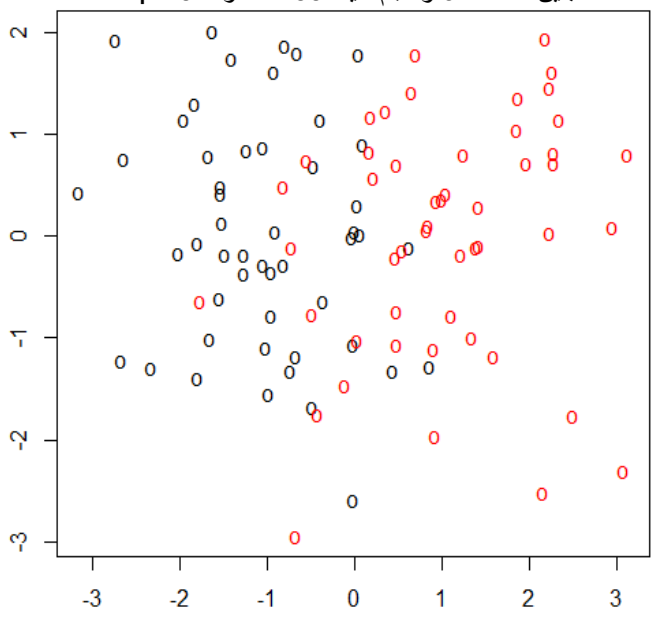
شكل (3-13)

عند مستوى $\sigma^2 = 1$ وحجم عينة n=100 و μ =0.3



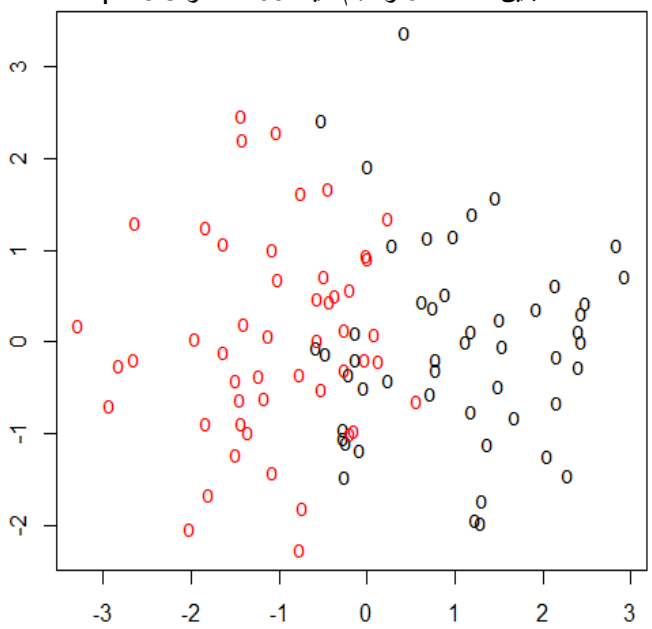
شكل (3-14)

عند تباين $\sigma^2 = 1$ وحجم عينة n=100 و μ =0.4



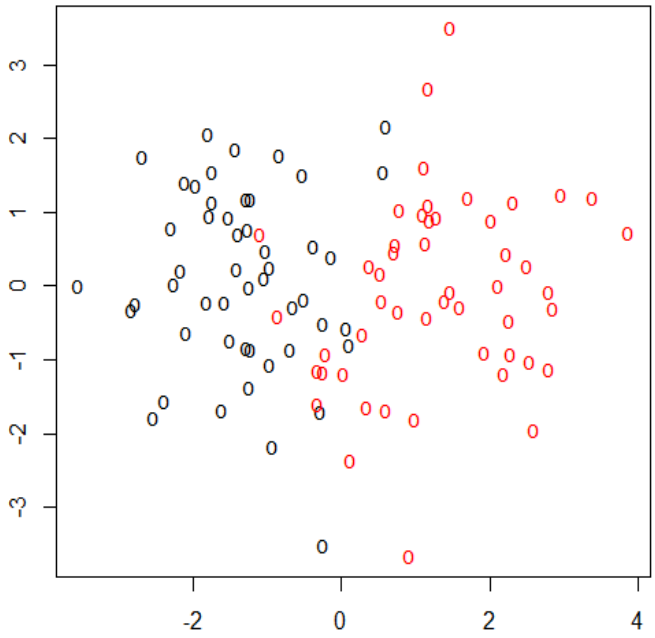
شكل (3-15)

عند تباين $\sigma^2 = 1$ وحجم عينة n=100 و μ =0.5



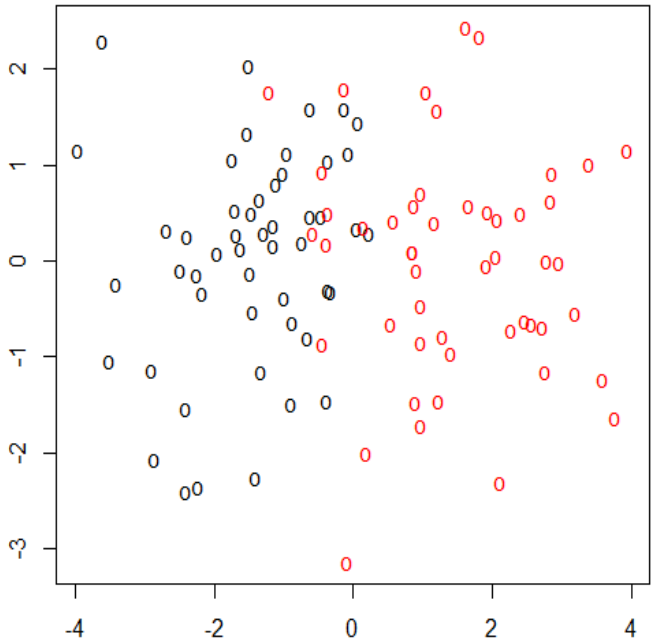
شكل (3-16)

عند تباين $\sigma^2 = 1$ وحجم عينة n=100 و μ =0.6



شكل (3-17)

عند تباين $\sigma^2 = 1$ وحجم عينة n=100 و μ =0.7



يلاحظ من الأشكال (3-10) و(3-11) و(3-12) أن التداخل بين المجموعتين قوي جداً إلا الشكل (3-13) فإنه أقلها شدة في التداخل عندما أصبحت قيمة μ =0.3 ، كما نلاحظ من الأشكال (3-14)

و(3−15) و(3−16) و(3−17) أن التباعد بين مشاهدات المجموعتين أصبح ملحوظاً وخصوصاً الشكل (3−17) الذي يمكننا من خلاله ملاحظة التباعد شبه التام بين المجوعتين عندما μ =0.7.

كما أظهرت نتائج المحاكاة عندما تكون قيمة التباين $(\sigma^2 = 1)$ ولعينة بحجم (n=216) أن طريقة SVM هي الأفضل وفي جميع حالات μ وكما هو مبين في الجدول (3−3) .

جدول (3−3)

نتائج التصنيف عند مستوى تباين $(\sigma^2 = 1)$ وحجم عينة n=216

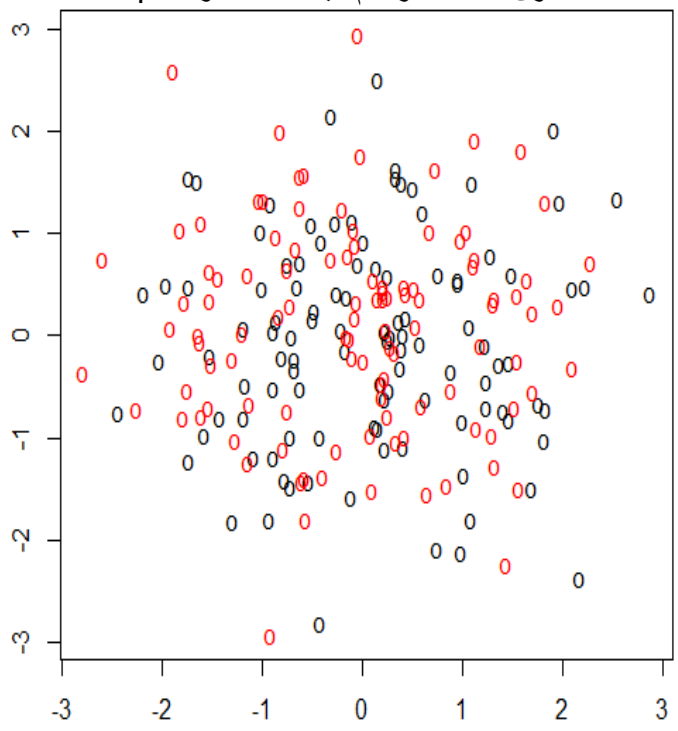
الوسط الحسابيμ	نسبة التصنيف بواسطةSVM	نسبة التصنيف بواسطةLRM
0.0	70 %	56 %
0.1	72 %	61 %
0.2	75 %	68 %
0.3	81 %	76 %
0.4	86 %	82 %
0.5	90 %	87 %
0.6	93 %	92 %
0.7	96 %	95 %

إعداد الباحث بالإعتمادعلى نتائج برنامج (R− language)

يوضح الجدول السابق نسب التعرف الصحيح لكلا الطريقتين عند حجم العينة الكبيرة (n=216) والتباين $(\sigma^2 = 1)$ إذ تراوحت نسبة التعرف الصحيح لـ SVM ما بين 70% عندما كانت قيمة μ=0 ولغاية 96% عندما أصبحت قيمة μ=0.7 أما نسبة التعرف الصحيح لـ LRM فبلغت 56% عند قيمة μ=0 حتى وصلت 95% عند قيمة μ=0.7 مما يعني تفوق SVM في جميع حالات μ. كذلك يمكن ملاحظة الفارق الواضح بين دقة SVM وLRM في نسبة التعرف الصحيح في الحالات التي يكون فيها التداخل شديداً بين مشاهدات المجموعتين عند قيم (μ=0,0.1,0.2,0.3, 0.4) وعلى الرغم من إنخفاض نسبة التداخل عند قيم (μ=,0.5,0.6,0.7) إلا أن التفوق كان لـ SVM في جميع الحالات ، كما يمكننا توضيح الرسومات المرافقة لإجراء عملية التصنيف عند كل قيمة من قيم μ وعند التباين $(\sigma^2 = 1)$ ولحجم عينة n=216 وملاحظة الإختلافات بينها كما يلي:

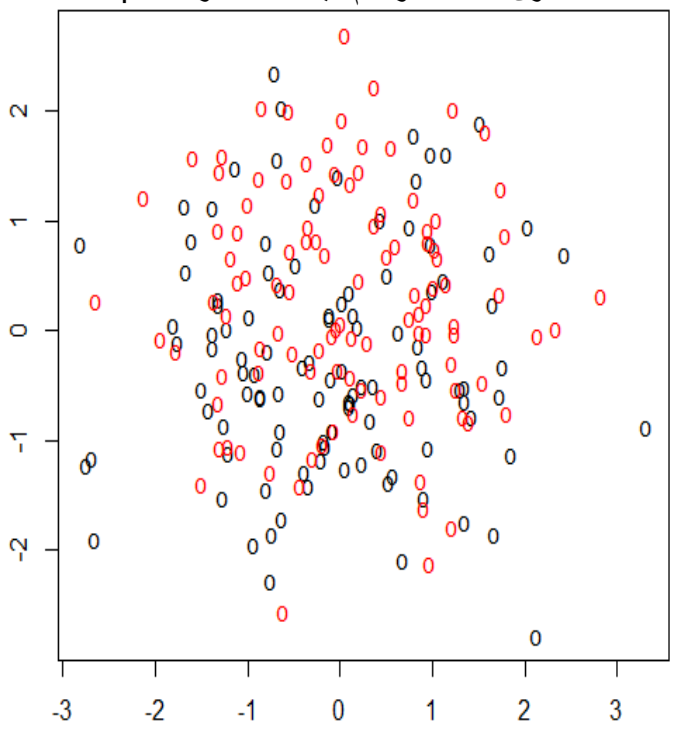
شكل (3-18)

عند مستوى $\sigma^2 = 1$ وحجم عينة n=216 و μ =0



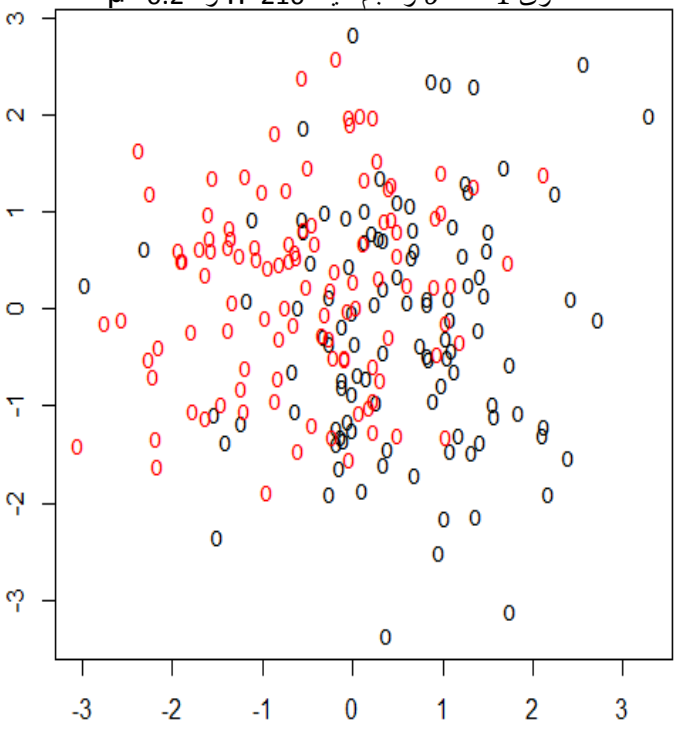
شكل (3-19)

عند مستوى $\sigma^2 = 1$ وحجم عينة n=216 و μ =0.1



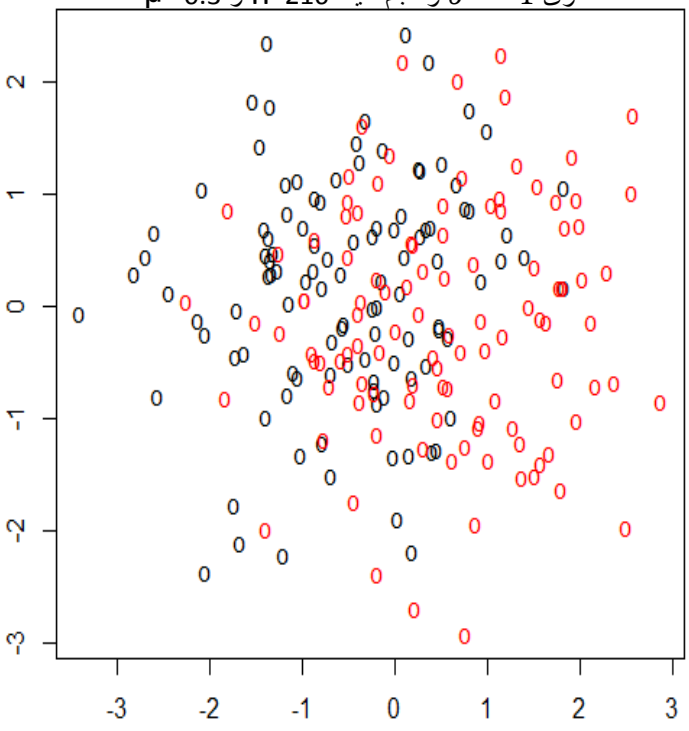
شكل (3-20)

عند مستوى $\sigma^2 = 1$ وحجم عينة n=216 و μ =0.2



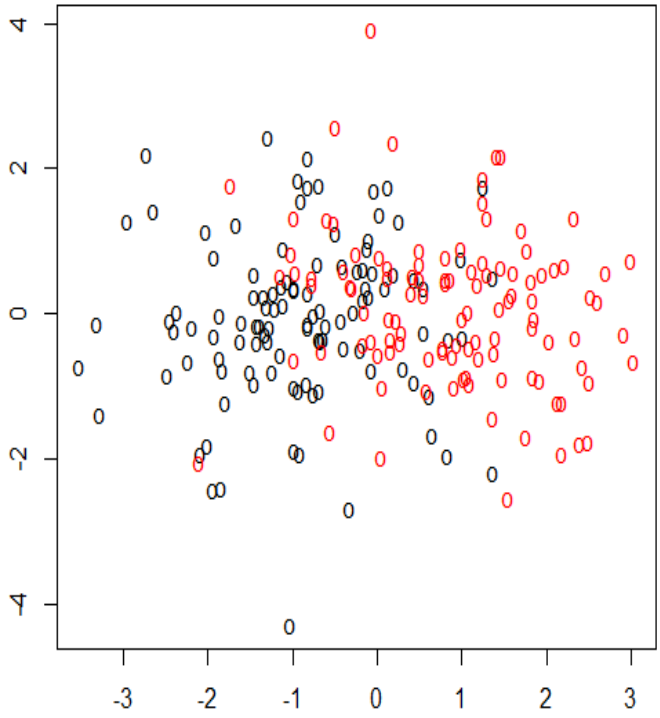
شكل (3-21)

عند مستوى $\sigma^2 = 1$ وحجم عينة n=216 و μ =0.3



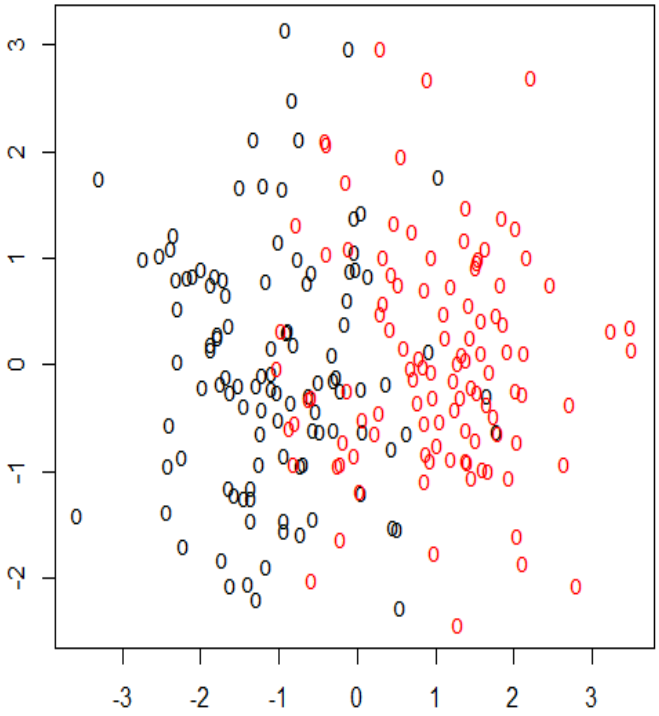
شكل (3-22)

عند مستوى تباين $\sigma^2 = 1$ وحجم عينة n=216 و μ =0.4



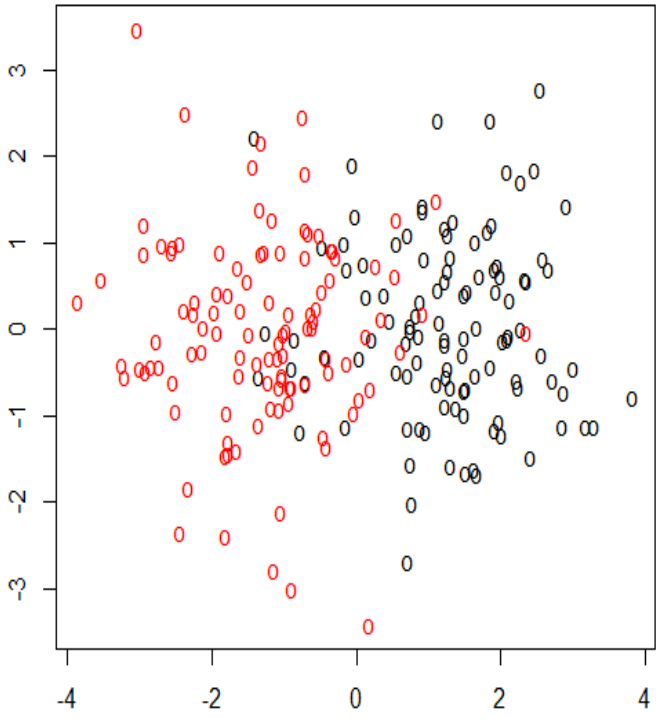
شكل (3-23)

عند مستوى تباين $\sigma^2 = 1$ وحجم عينة n=216 و μ =0.5



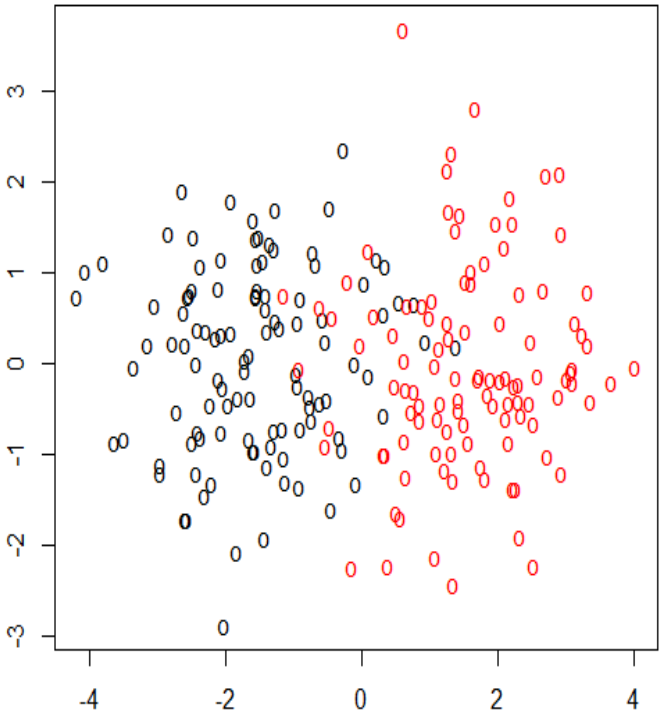
شكل (3-24)

عند مستوى تباين $\sigma^2 = 1$ وحجم عينة n=216 و μ =0.6



شكل (3-25)

عند مستوى تباين $\sigma^2 = 1$ وحجم عينة n=216 و μ =0.7



يلاحظ من الأشكال إبتداءً من (3−18) ولغاية الشكل (3−24) أن التداخل بدأ قوياً عندما كانت قيمة μ =0 و بدأ تدريجياً بالضعف عنما إزادت قيمة μ الى أن أصبح بإلامكان تمييز المجموعتين من خلال الشكل (3−22) و(3−23) و(3−24) ، كذلك نلاحظ من خلال الشكل (3−25) التباعد أو الإنفصال شبه التام بين المجموعتين وذلك عندما أصبحت قيمة μ =0.7 .

ثانياً: أظهرت نتائج المحاكاة عندما تكون قيمة التباين ($\sigma^2 = 1.25$)ولعينة بحجم (n=50) أن طريقة SVM كانت أفضل من طريقة LRM وفي جميع حالات μ وكما هو مبين في الجدول (3-4) .

جدول (3−4)

نتائج التصنيف عند مستوى تباين ($\sigma^2 = 1.25$) و حجم عينة n=50

الوسط الحسابي μ	نسبة التصنيف بواسطة SVM	نسبة التصنيف بواسطة LRM
0.0	81 %	63 %
0.1	81 %	64 %
0.2	82 %	69 %
0.3	85 %	74 %
0.4	87 %	80 %
0.5	90 %	84 %
0.6	92 %	89 %
0.7	94 %	93 %

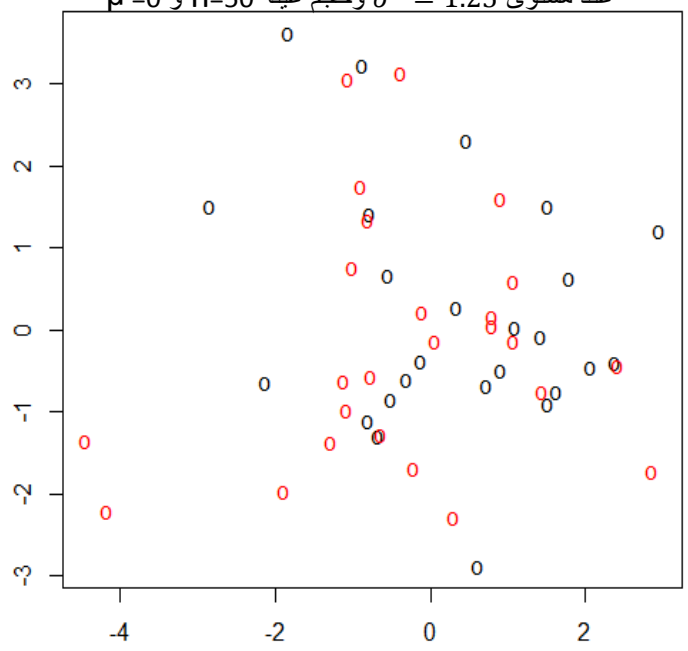
إعداد الباحث بالإعتماد نتائج برنامج (R− language)

يظهر الجدول المذكور آنفاً نسب التعرف الصحيح لكلا الطريقتين عند حجم العينة الصغيرة (n=50) والتباين ($\sigma^2 = 1.25$) إذ تراوحت نسبة التعرف الصحيح لـ SVM ما بين 81% عندما كانت قيمة μ=0 ولغاية 94% عندما أصبحت قيمة μ=0.7 أما نسبة التعرف الصحيح لـ LRM فتراوحت مابين 63% عند قيمة μ=0 ، و 93% عند قيمة μ=0.7 مما يعني تفوق SVM في جميع حالات μ. كذلك يمكن ملاحظة الفارق الواضح بين دقة SVM و LRM في نسبة التعرف الصحيح في الحالات التي يكون فيها التداخل شديداً بين مشاهدات المجموعتين عند قيم (μ=0,0.1,0.2,0.3, 0.4,0.5) وعلى الرغم من إنخفـــاض نسبة التداخـــل عـند قيم (μ=0.6,0.7) إلا أن التفوق كان لـ SVM في جميع الحالات .

كما يمكن توضيح عملية التداخل بين المشاهدات في الرسومات والأشكال المرافقة لكل عملية تصنيف وفي الحالة التي يكون التباين $\sigma^2 = 1.25$ وحجم العينة n=50 وكالتالي :

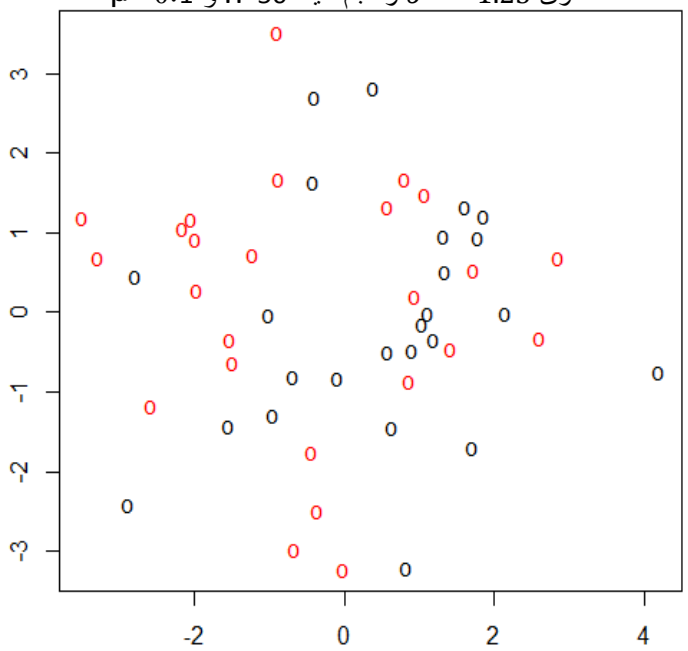
شكل (3-26)

عند مستوى $\sigma^2 = 1.25$ وحجم عينة n=50 و μ =0



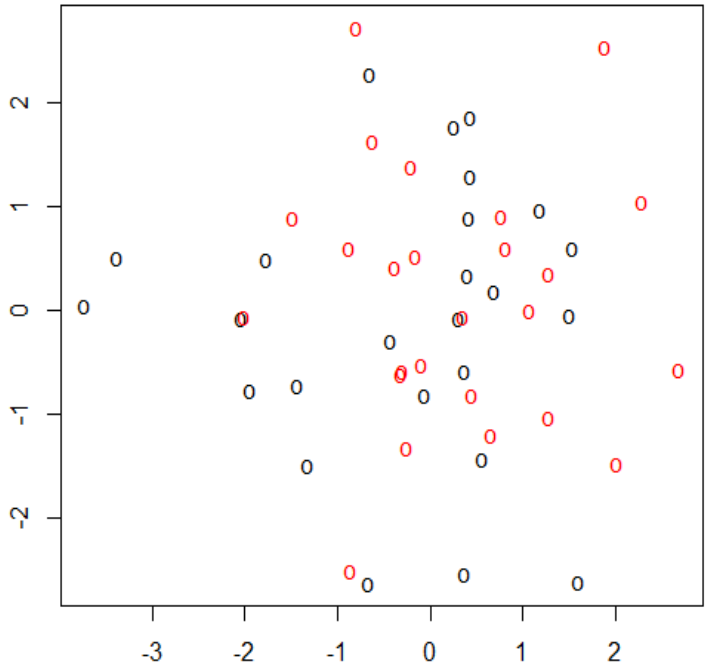
شكل (3-27)

عند مستوى $\sigma^2 = 1.25$ وحجم عينة n=50 و μ =0.1



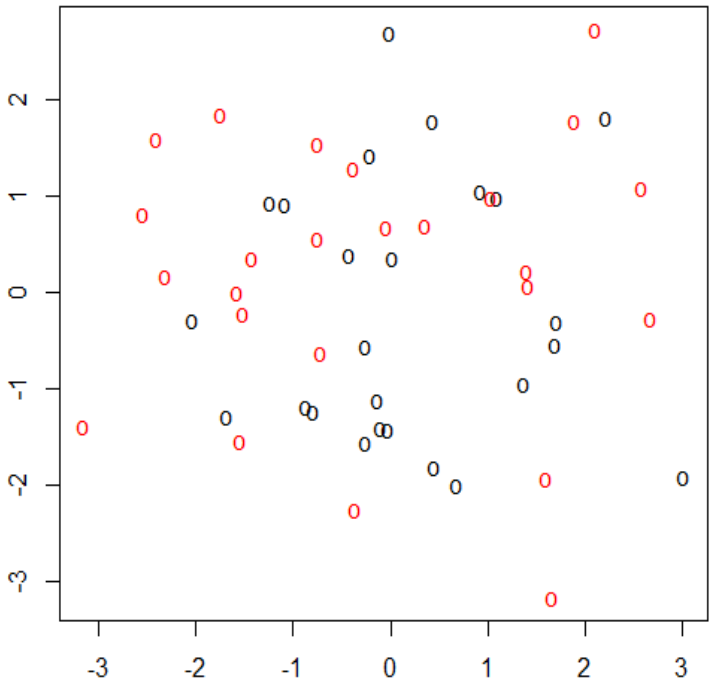
شكل (3-28)

عند مستوى $\sigma^2 = 1.25$ وحجم عينة n=50 و μ =0.2



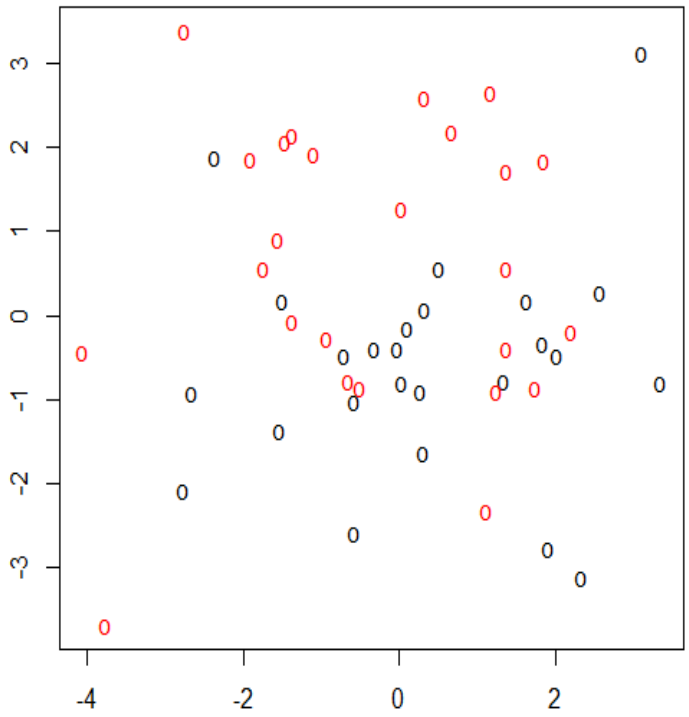
شكل (3-29)

عند مستوى $\sigma^2 = 1.25$ وحجم عينة n=50 و μ =0.3



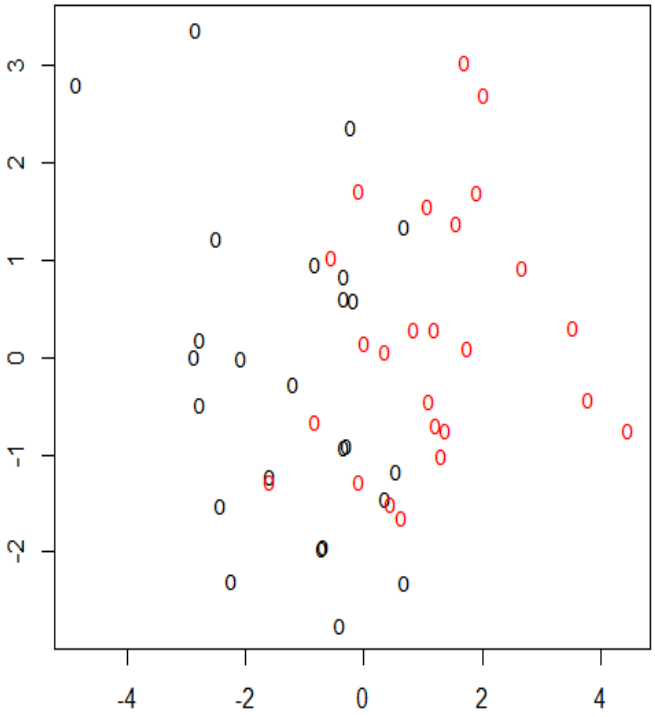
شكل (3-30)

عند مستوى $\sigma^2 = 1.25$ وحجم عينة n=50 و μ =0.4



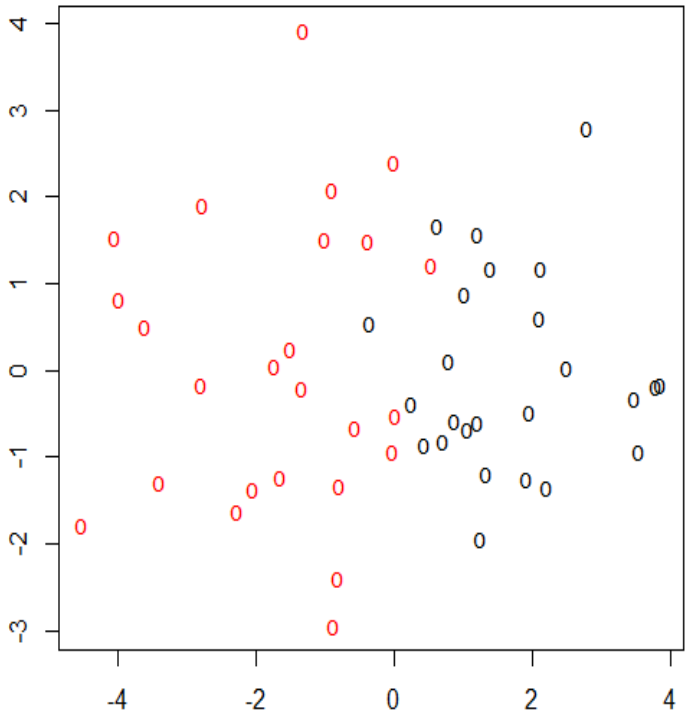
شكل (3-31)

عند مستوى $\sigma^2 = 1.25$ وحجم عينة n=50 و μ =0.5



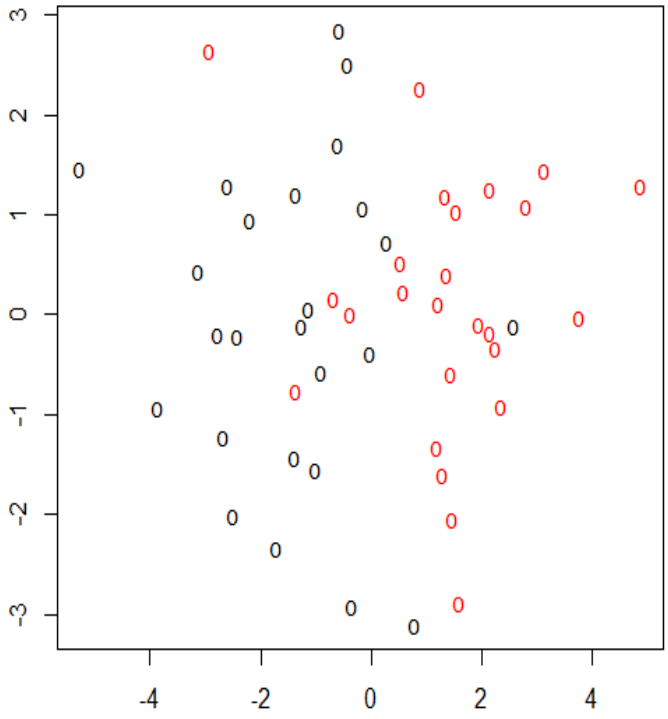
شكل (3-32)

عند مستوى $\sigma^2 = 1.25$ وحجم عينة n=50 و μ =0.6



شكل (3-33)

عند مستوى $\sigma^2 = 1.25$ وحجم عينة n=50 و μ =0.7



يلاحظ من الاشكال من (3-26) و(3-27) و(3-28) و(3-29) و(3-30) و(3-31) و(3-32) أن التداخل يكون قوياً بين مشاهدات المجموعتين ثم يبدأ تدريجياً بالإنخفاض كلما إزدادت قيمة (μ) حتى تصل قيمة (μ) الى 0.7 عندها يحدث الإنفصال شبه التام بين مشاهدات المجموعتين كما في الشكل(3-33) .

كما أظهرت نتائج المحاكاة عندما تكون قيمة التباين ($\sigma^2 = 1.25$) ولعينة بحجم n=100 أن طريقة SVM كانت أفضل من طريقة LRM وفي جميع حالات (μ) وكما مبين في الجدول (3-5):-

جدول (3-5)

نتائج التصنيف عند مستوى تباين ($\sigma^2 = 1.25$) وحجم عينة n=100

الوسط الحسابي μ	نسبة التصنيف بواسطة SVM	نسبة التصنيف بواسطة LRM
0.0	75 %	59 %
0.1	76 %	61 %
0.2	78 %	67 %
0.3	81 %	72 %
0.4	84 %	78 %
0.5	87 %	83 %
0.6	90 %	87 %
0.7	93 %	91 %

إعداد الباحث بالإعتماد على نتائج برنامج (R- language)

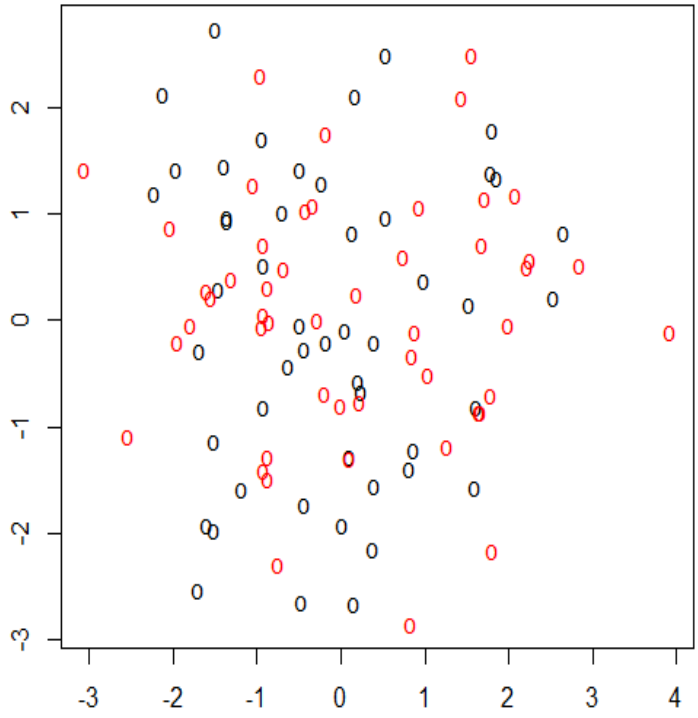
يظهر الجدول السابق نسب التعرف الصحيح لكلا الطريقتين عند حجم العينة المتوسطة (n=100) والتباين ($\sigma^2 = 1.25$) إذ تراوحت نسبة التعرف الصحيح لـ SVM ما بين 75% عندما كانت قيمة μ=0 ولغاية 93% عندما أصبحت قيمة μ=0.7 أما نسبة التعرف الصحيح لـ LRM فتراوحت مابين 59% عند قيمة μ=0 ، و 91% عند قيمة μ=0.7 مما يعني تفوق SVM في جميع حالات μ. كذلك يمكن ملاحظة الفارق الواضح بين دقة SVM وLRM في نسبة التعرف الصحيح في الحالات التي يكون فيها التداخل شديداً بين مشاهدات المجموعتين عند قيم (μ=0,0.1,0.2,0.3, 0.4) وعلى الرغم من إنخفاض نسبة التداخل عند قيم (μ=0.5,0.6,0.7) إلا أن التفوق كان لـ SVM في جميع الحالات أيضاً .

كما يلاحظ أن الفارق في دقة التصنيف الصحيح بين طريقة SVM و LRM إزدادت قليلاً عند قيمة μ=0.7 عند زيادة حجم العينة الى (n=100) عما كان عليه في العينة الصغيرة (n=50) وبمقدار نقطة واحدة مما يدل على تأثير تغير التباين من (1) الى (1.25) في درجة الـدقة عند μ=0.7 إذ كانت دقة التصنيف لـ SVM (0.93%) و دقة التصنيف لـ LRM (0.91%) إذ الفارق بينهما أصبح نقطتين بينما كانت دقتهما في حجم العينة (n=50) (0.94%) لـ SVM و (0.93%) لـ LRM أي أن الفارق كان نقطة في العينة الصغيرة وأصبح نقطتين في العينة المتوسطة (n=100) .

كما يمكن توضيح عملية التداخل بين مشاهدات المجموعتين في الرسومات والأشكال المرافقة لكل عملية تصنيف وفي الحالة التي يكون التباين فيها $\sigma^2 = 1.25$ وحجم العينة n=100 وكالتالي

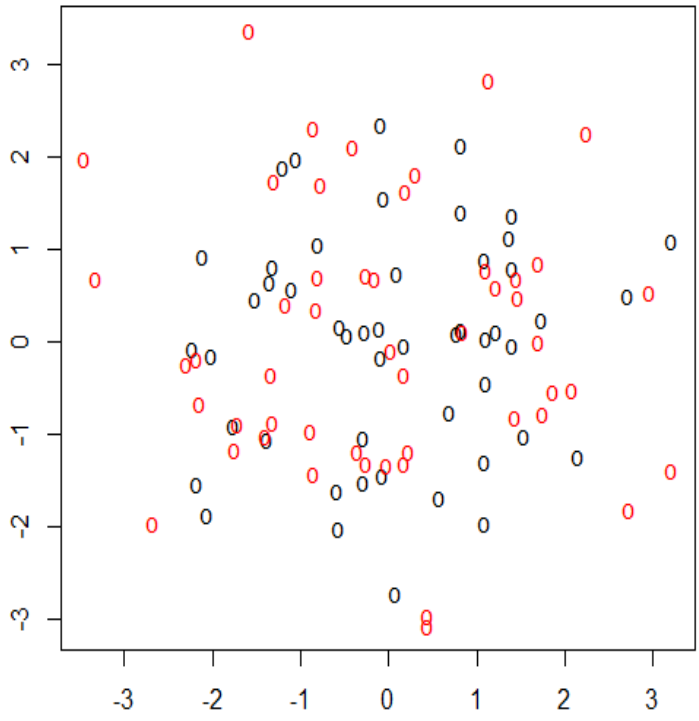
شكل (3-34)

عند مستوى $\sigma^2 = 1.25$ وحجم عينة n=100 و μ =0



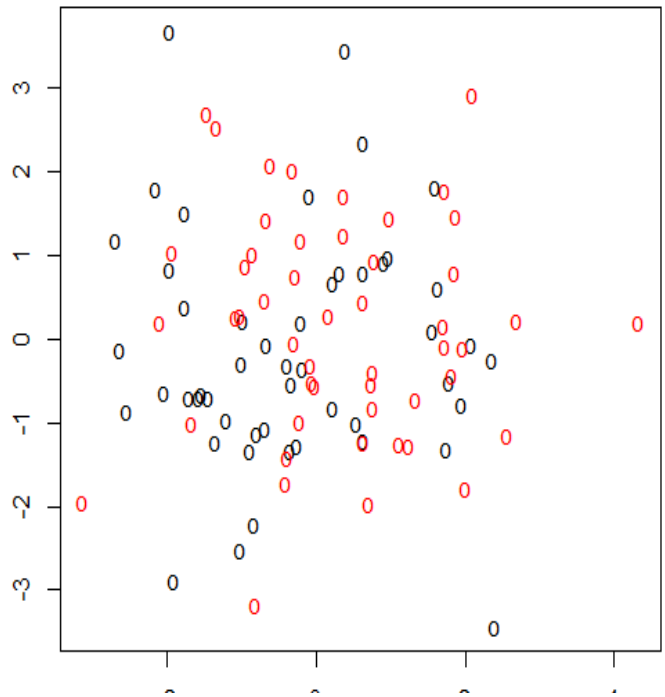
شكل (3-35)

عند مستوى $\sigma^2 = 1.25$ وحجم عينة n=100 و μ =0.1



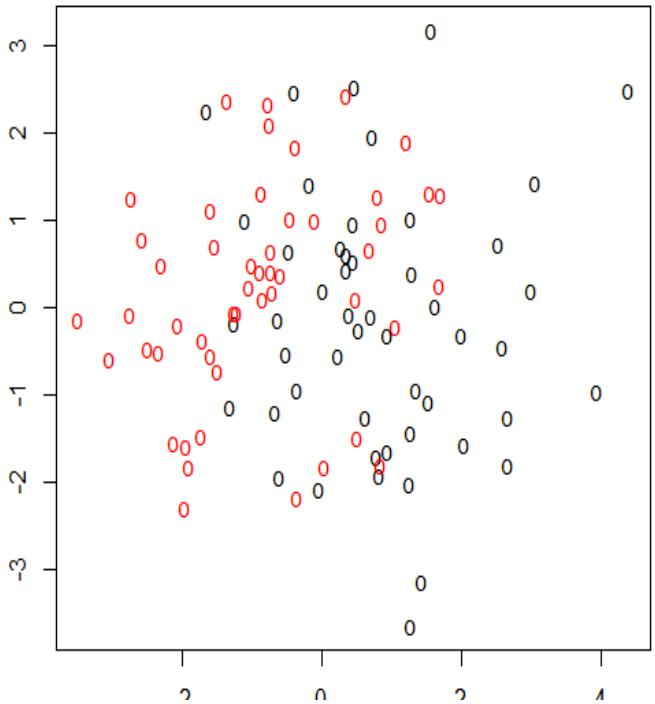
شكل (3-36)

عند مستوى $\sigma^2 = 1.25$ وحجم عينة n=100 و μ =0.2



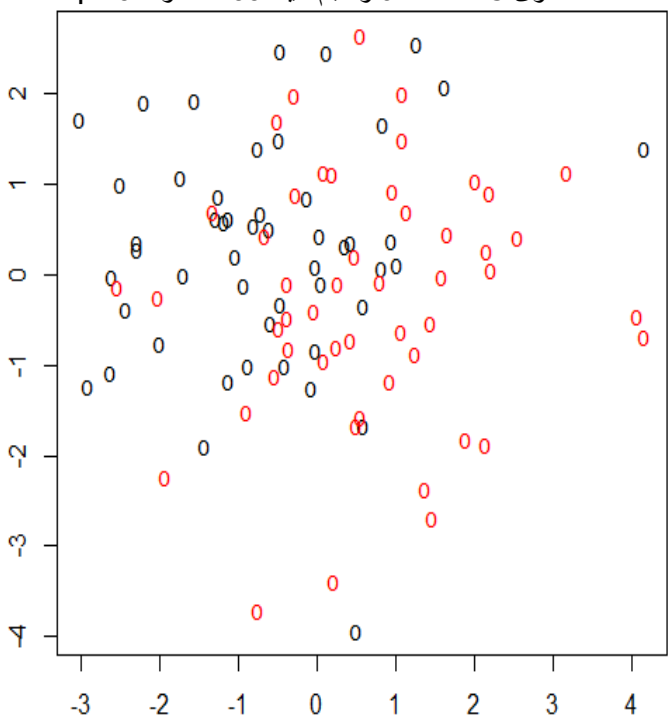
شكل (3-37)

عند مستوى $\sigma^2 = 1.25$ وحجم عينة n=100 و μ =0.3



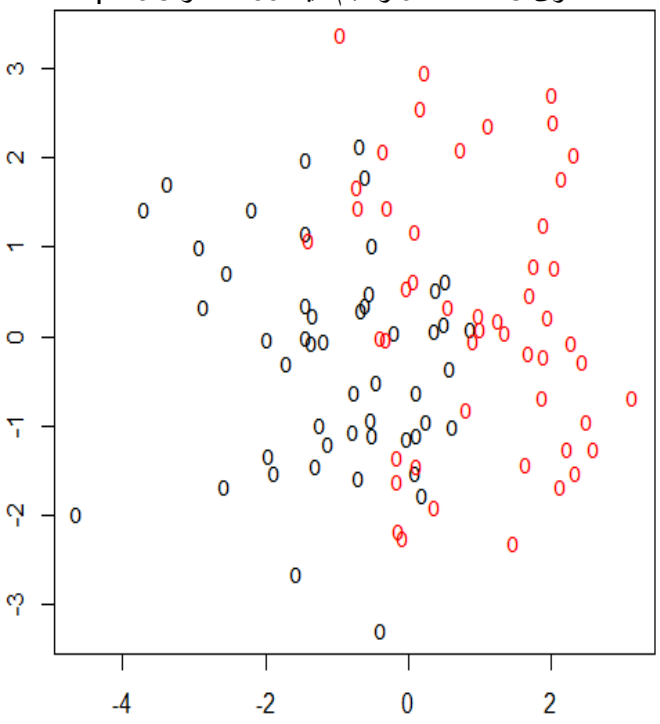
شكل (3-38)

عند مستوى $\sigma^2 = 1.25$ وحجم عينة n=100 و μ =0.4



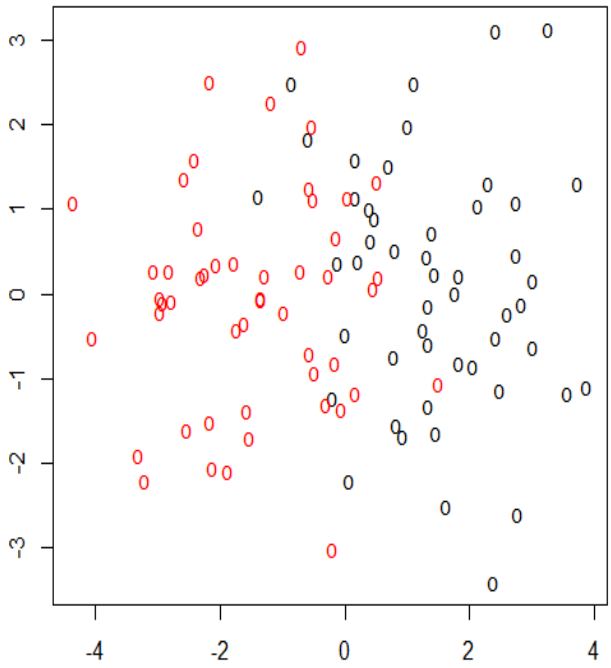
شكل (3-39)

عند مستوى $\sigma^2 = 1.25$ وحجم عينة n=100 و μ =0.5



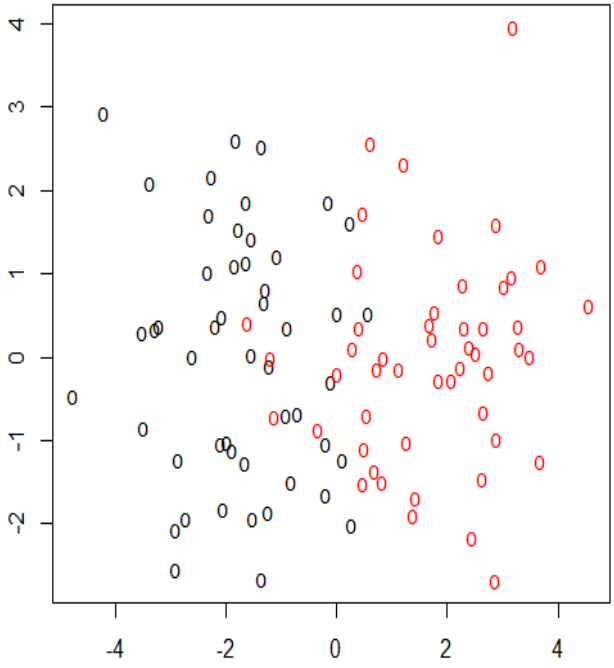
شكل (3-40)

عند مستوى $\sigma^2 = 1.25$ وحجم عينة n=100 و μ =0.6



شكل (3-41)

عند مستوى $\sigma^2 = 1.25$ وحجم عينة n=100 و μ =0.7



يلاحظ من الاشكال (3-34) و(3-35) و(3-36) و(3-37) و(3-38) و(3-39) و(3-40) أن التداخل يكون قوياً بين مشاهدات المجموعتين ثم يبدأ تدريجياً بالإنخفاض كلما إزدادت قيمة μ حتى تصل قيمة μ الى 0.7 عندها يحدث الإنفصال شبه التام وليس الإنفصال التام بين مشاهدات المجموعتين كما في الشكل (3-41) .

كما أظهرت نتائج المحاكاة عندما تكون قيمة التباين ($\sigma^2 = 1.25$) ولعينة بحجم n=216 أن طريقة SVM كانت أفضل من طريقة LRM وفي جميع حالات μ وكما مبين في الجدول (3-6):-

جدول (3-6)

نتائج التصنيف عند مستوى تباين ($\sigma^2 = 1.25$) و حجم عينة n=216

الوسط الحسابي μ	نسبة التصنيف بواسطة SVM	نسبة التصنيف بواسطة LRM
0.0	70 %	56 %
0.1	71 %	59 %
0.2	74 %	65 %
0.3	78 %	71 %

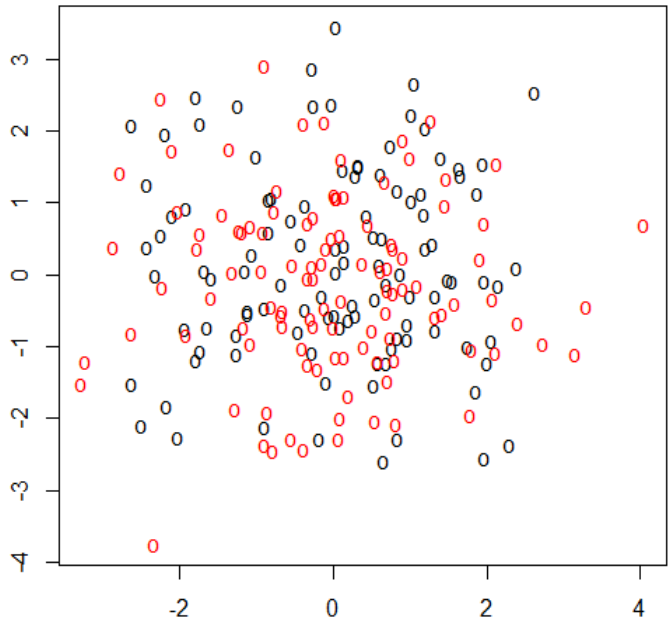
0.4	82 %	77 %
0.5	86 %	82 %
0.6	89 %	86 %
0.7	92 %	90 %

إعداد الباحث بالإعتماد على نتائج برنامج (R– language)

يوضح الجدول السابق نسب التعرف الصحيح لكلا الطريقتين عند حجم العينة الكبيرة (n=216) والتباين ($\sigma^2 = 1.25$) إذ تراوحت نسبة التعرف الصحيح لـ SVM ما بين 70% عندما كانت قيمة μ=0 ولغاية 92% عندما أصبحت قيمة μ=0.7 أما نسبة التعرف الصحيح لـ LRM فتراوحت مابين 56% عند قيمة μ=0 ، و 90% عند قيمة μ=0.7 مما يعني تفوق SVM في جميع حالات μ. كذلك يمكن ملاحظة الفارق الواضح بين دقة SVM و LRM في نسبة التعرف الصحيح في الحالات التي يكون فيها التداخل شديداً بين مشاهدات المجموعتين عند قيم (μ=0,0.1,0.2,0.3, 0.4) وعلى الرغم من إنخفاض نسبة التداخل عند قيم (μ=0.5,0.6,0.7) إلا أن التفوق كان لـ SVM في جميع الحالات ، كما يمكن توضيح عملية التداخل بين مشاهدات المجموعتين في الرسومات والأشكال المرافقة لكل عملية تصنيف وفي الحالة التي يكون التباين فيها =1.25 وحجم العينة n=216 وكالتالي :

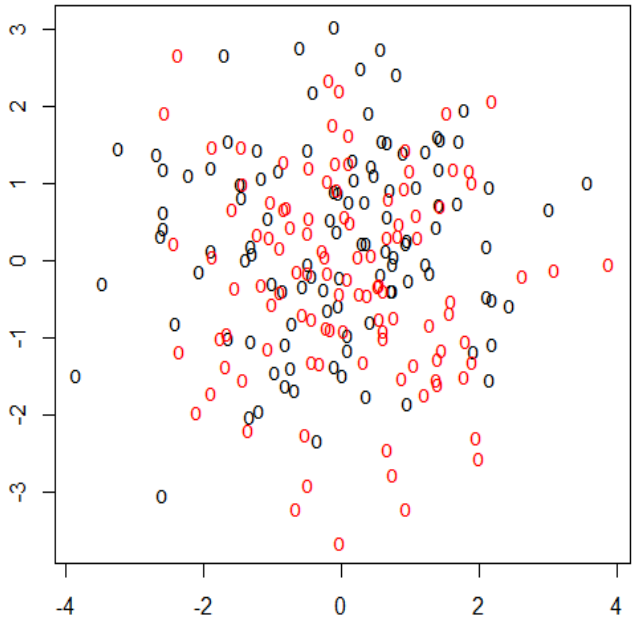
شكل (3-42)

عند مستوى $\sigma^2 = 1.25$ وحجم عينة n=216 و μ =0



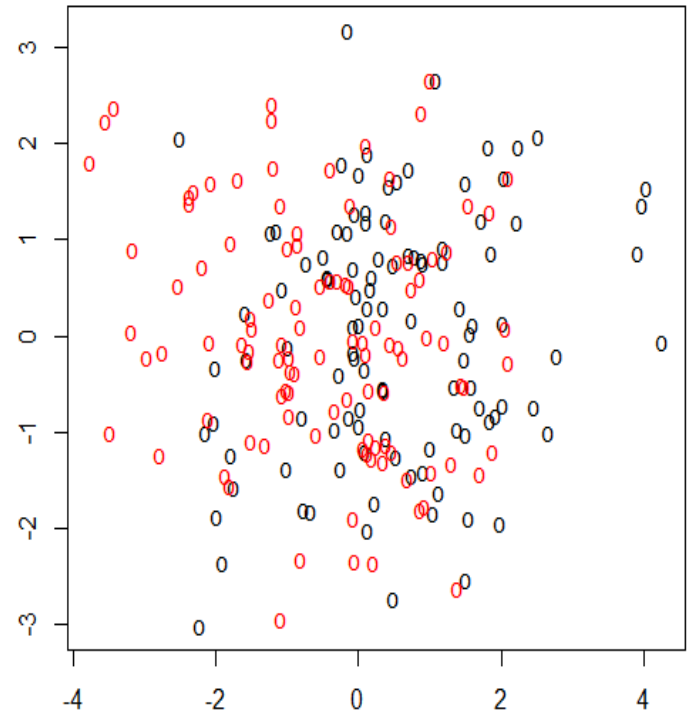
شكل (3-43)

عند مستوى $\sigma^2 = 1.25$ وحجم عينة n=216 و μ =0.1



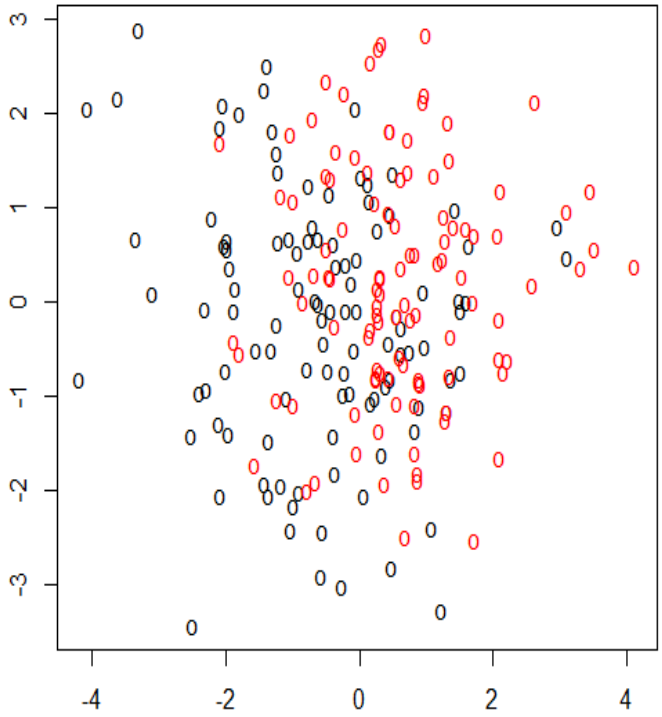
شكل (3-44)

عند مستوى $\sigma^2 = 1.25$ وحجم عينة n=216 و μ =0.2



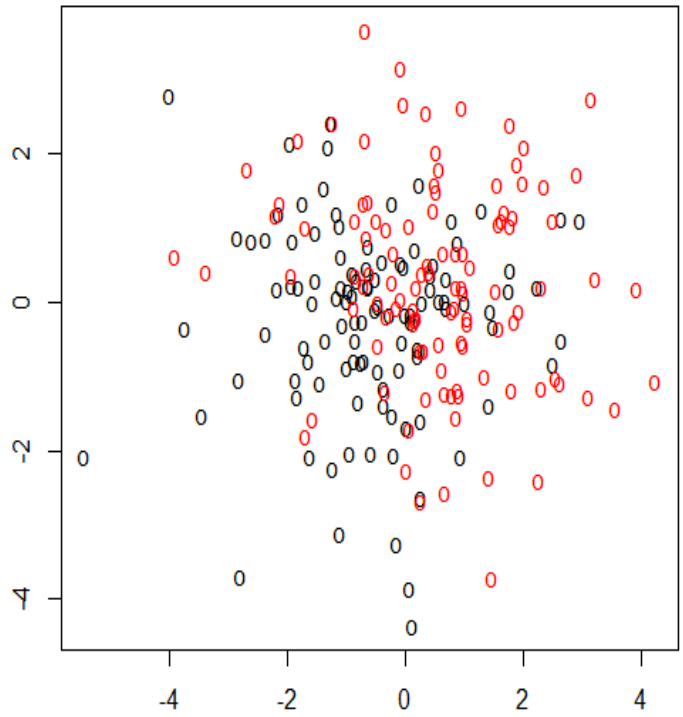
شكل (3-45)

عند مستوى $\sigma^2 = 1.25$ وحجم عينة n=216 و μ =0.3



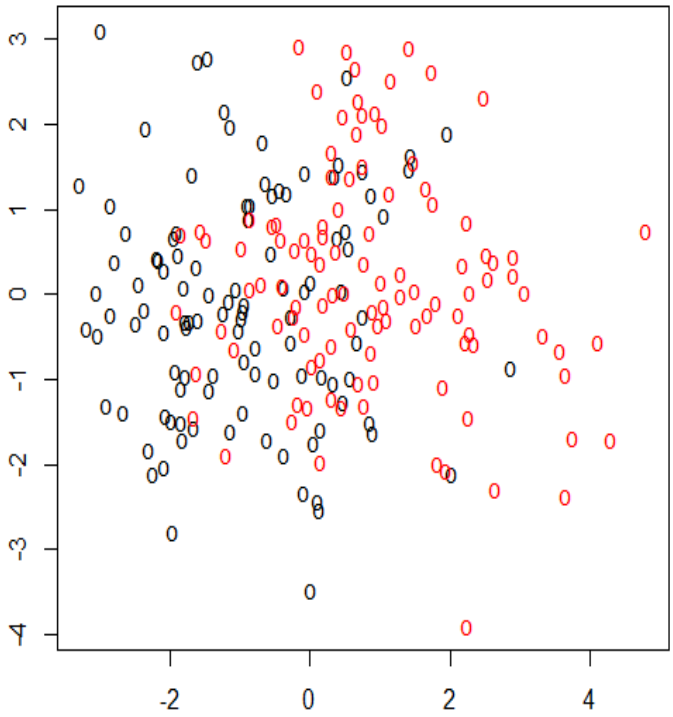
شكل (3-46)

عند مستوى $\sigma^2 = 1.25$ وحجم عينة n=216 و μ =0.4



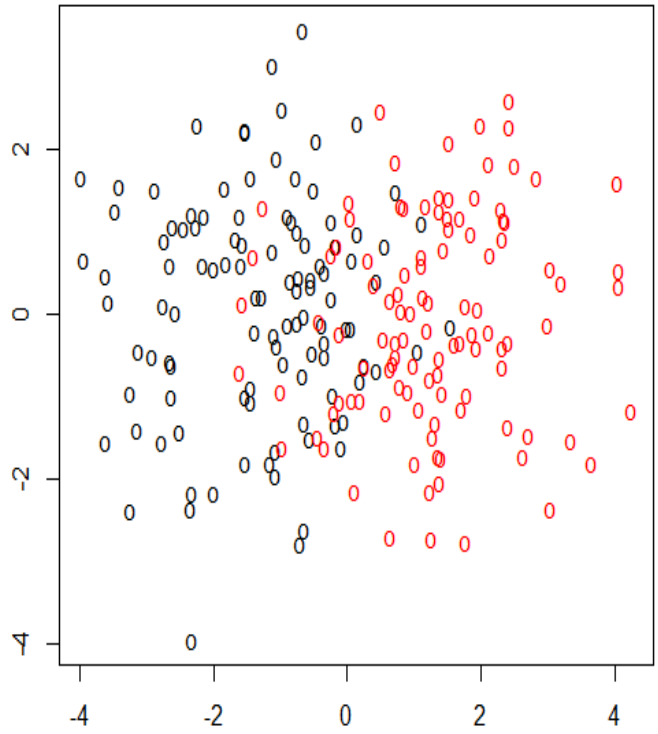
شكل (3-47)

عند مستوى $\sigma^2 = 1.25$ وحجم عينة n=216 و μ =0.5



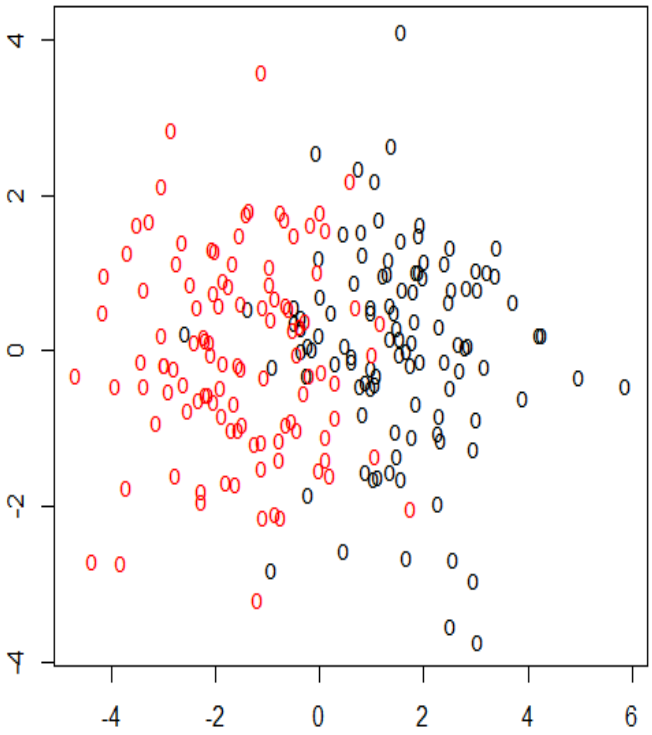
شكل (3-48)

عند مستوى $\sigma^2 = 1.25$ وحجم عينة n=216 و μ =0.6



شكل (3-49)

عند مستوى $\sigma^2 = 1.25$ وحجم عينة n=216 و μ =0.7



يلاحظ من الاشكال (3-42)و(3-43) و(3-44) و(3-45) و(3-46) و(3-47) و(3-48) أن التداخل يكون قوياً بين مشاهدات المجموعتين ثم يبدأ تدريجياً بالإنخفاض كلما إزدادت قيمة (μ)حتى تصل قيمة (μ) الى 0.7 عندها يحدث الإنفصال شبه التام وليس الإنفصال التام بين مشاهدات المجموعتين ، ونستنتج من ذلك أنه إذا إزدادت قيمة التباين فإن ذلك يؤدي الى زيادة مقدار التداخل بين مشاهدات المجموعتين ، وهذا عامل آخر يضاف الى حجم العينة يؤثر في التداخل بين مشاهدات المجموعتين ، كما في الشكل (3-49) .

ثالثاً: كما أظهرت نتائج المحاكاة وعندما تكون قيمة التباين($\sigma^2 = 1.5$) والعينة بحجم (n=50) أن طريقة SVM كانت أفضل من طريقة LRM وفي جميع حالات μ وكما هو مبين في الجدول (3-7) .

جدول (3-7)

نتائج التصنيف عند مستوى تباين ($\sigma^2 = 1.5$) وحجم عينة n=50

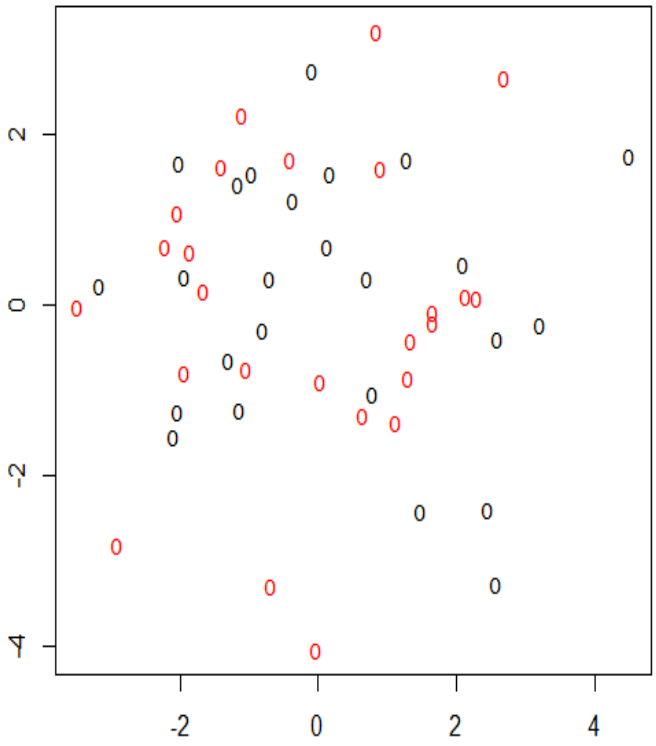
الوسط الحسابي μ	نسبة التصنيف بواسطة SVM	نسبة التصنيف بواسطة LRM
0.0	81 %	63 %
0.1	81 %	64 %
0.2	82 %	67 %
0.3	83 %	71 %
0.4	85 %	76 %
0.5	87 %	80 %
0.6	90 %	84 %
0.7	92 %	88 %

إعداد الباحث بالإعتماد على نتائج برنامج (R- language)

يظهر الجدول السابق نسب التعرف الصحيح لكلا الطريقتين عند حجم العينة الصغيرة (n=50) والتباين ($\sigma^2 = 1.5$) إذ تراوحت نسبة التعرف الصحيح لـ SVM ما بين 81% عندما كانت قيمة μ=0 ولغاية 92% عندما أصبحت قيمة μ=0.7 أما نسبة التعرف الصحيح لـ LRM فتراوحت النسبة مابين 63% عند قيمة μ=0 ، و 88% عند قيمة μ=0.7 مما يعني تفوق SVM في جميع حالات (μ) ، كذلك يمكن ملاحظة الفارق الواضح بين دقة SVM وLRM في نسبة التعرف الصحيح في الحالات عند جميع قيم (μ=0,0.1,0.2,0.3, 0.4,0.5,0.6,0.7) . كما يمكن توضيح عملية التداخل بين المشاهدات في الرسومات والأشكال المرافقة لكل عملية تصنيف وفي الحالة التي يكون التباين فيها ($\sigma^2 = 1.5$) وحجم العينة n=50 وكالتالي :

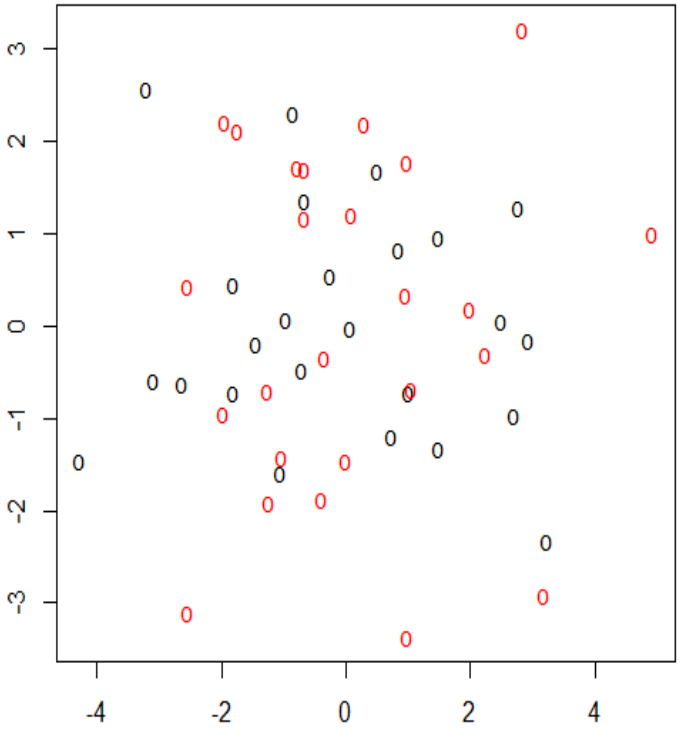
شكل (3-50)

عند مستوى $\sigma^2 = 1.5$ وحجم عينة n=50 و μ =0



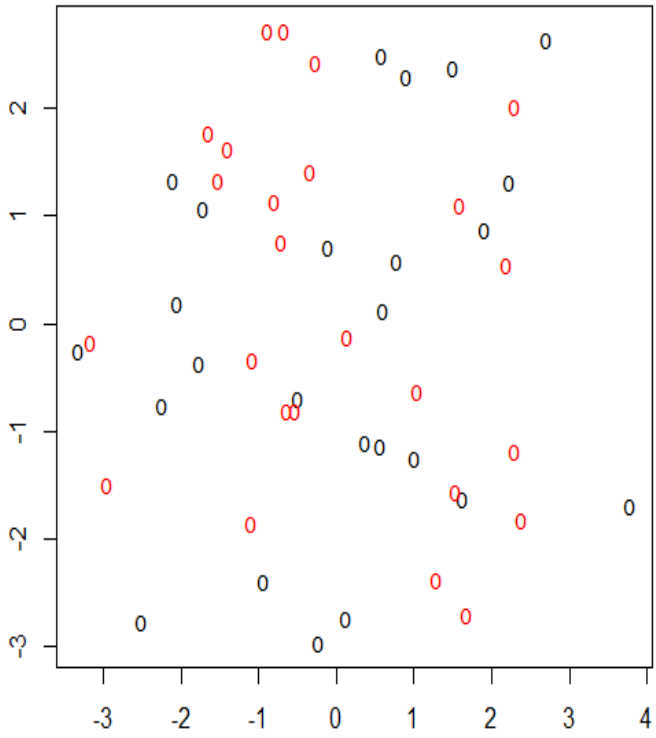
شكل (3-51)

عند مستوى $\sigma^2 = 1.5$ وحجم عينة n=50 و μ =0.1



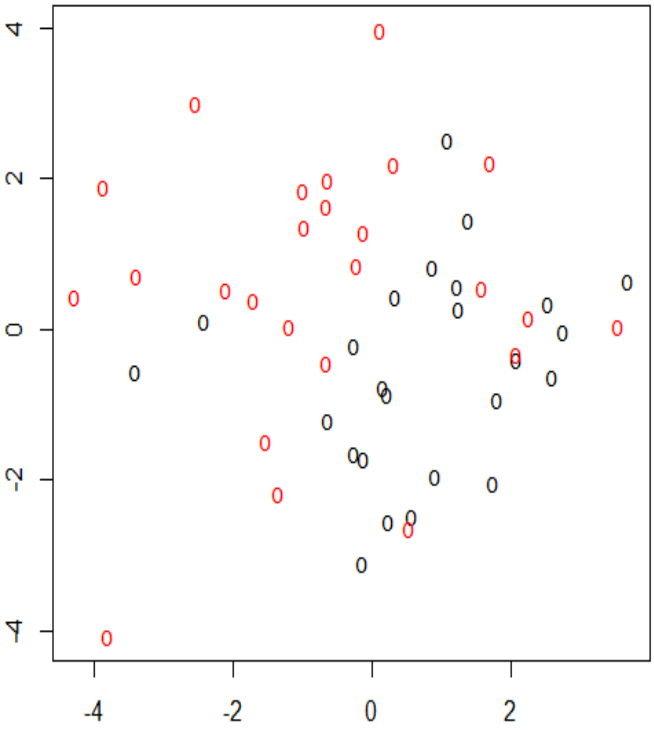
شكل (3-52)

عند مستوى $\sigma^2 = 1.5$ وحجم عينة n=50 و μ =0.2



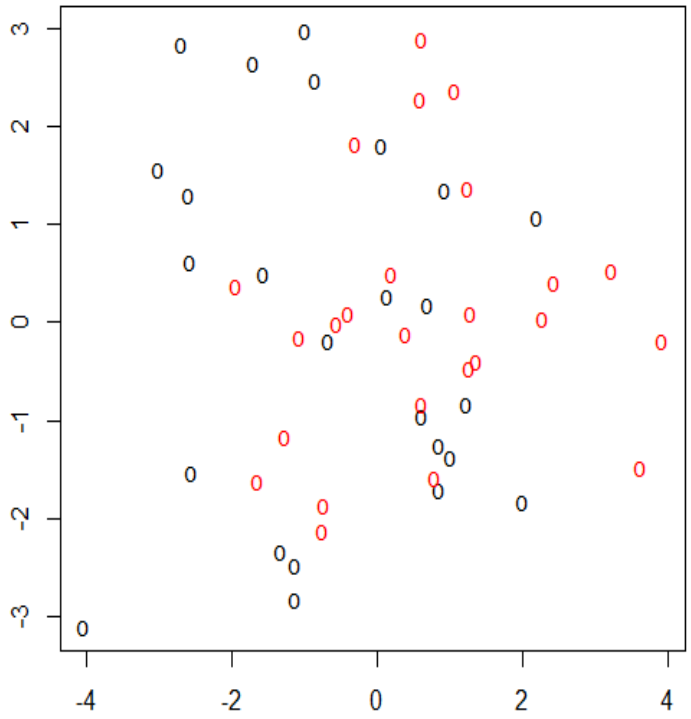
شكل (3-53)

عند مستوى $\sigma^2 = 1.5$ وحجم عينة n=50 و μ =0.3



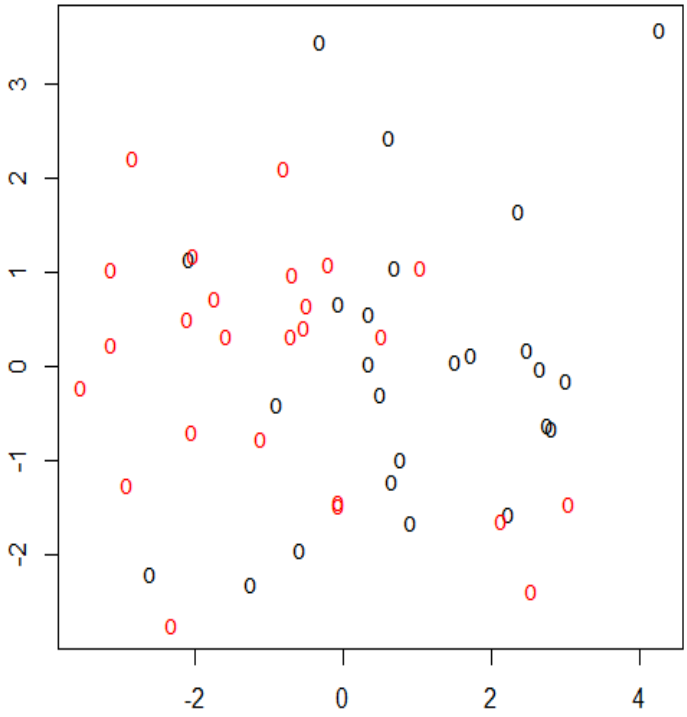
شكل (3-54)

عند مستوى $\sigma^2 = 1.5$ وحجم عينة n=50 و μ =0.4



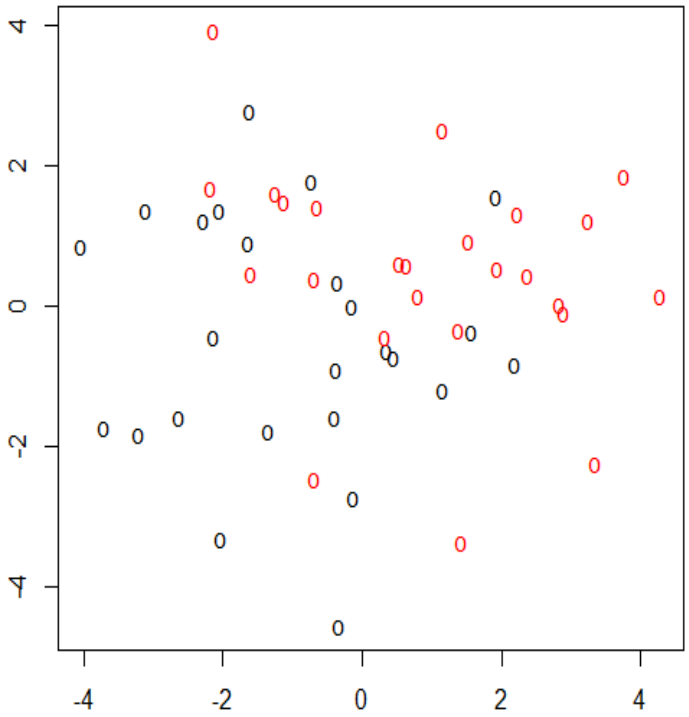
شكل (3-55)

عند مستوى $\sigma^2 = 1.5$ وحجم عينة n=50 و μ =0.5



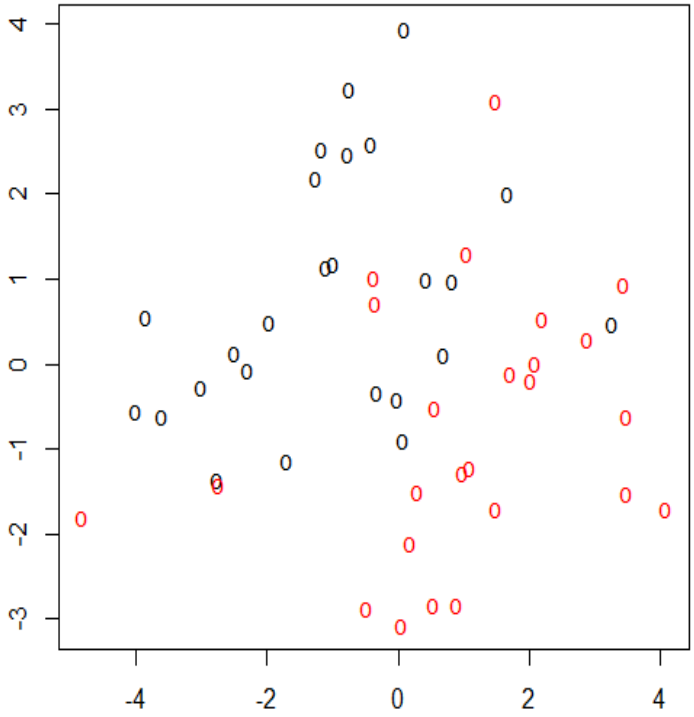
شكل (3-56)

عند مستوى $\sigma^2 = 1.5$ وحجم عينة n=50 و μ =0.6



شكل (3-57)

عند مستوى $\sigma^2 = 1.5$ وحجم عينة n=50 و μ =0.7



يلاحظ من الاشكال (3-50) و(3-51) و(3-52) و(3-53) و(3-54) و(3-55) و(3-56) أن التداخل يكون قوياً بين مشاهدات المجموعتين ثم يبدأ تدريجياً بالإنخفاض كلما إزدادت قيمة μ حتى تصل قيمة μ الى 0.7 عندها يحدث الإنفصال شبه التام بين مشاهدات المجموعتين كما في الشكل (3-57) .

كما أظهرت نتائج المحاكاة وعندما تكون قيمة التباين ($\sigma^2 = 1.5$) والعينة بحجم n=100 أن طريقة SVM كانت أفضل من طريقة LRM وفي جميع حالات μ وكما مبين في الجدول (3-8) الآتي:-

جدول (3-8)

نتائج التصنيف عند مستوى تباين ($\sigma^2 = 1.5$) و حجم عينة n=100

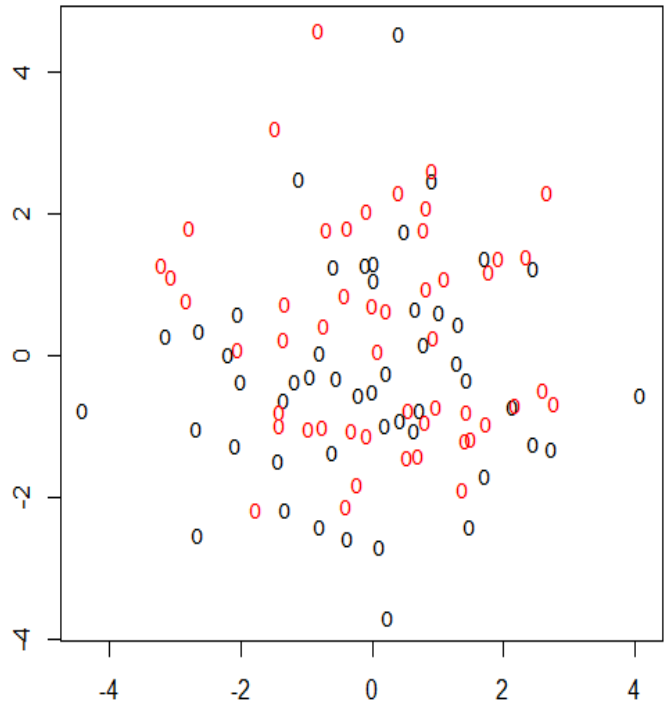
الوسط الحسابي μ	نسبة التصنيف بواسطة SVM	نسبة التصنيف بواسطة LRM
0.0	75 %	59 %
0.1	76 %	60 %
0.2	77 %	64 %
0.3	79 %	69 %
0.4	82 %	74 %
0.5	85 %	79 %
0.6	87 %	83 %
0.7	90 %	87 %

إعداد الباحث بالإعتماد على نتائج برنامج (R- language)

يظهر الجدول المذكور آنفاً نسب التعرف الصحيح لكلا الطريقتين عند حجم العينة المتوسطة (n=100) والتباين ($\sigma^2 = 1.5$) إذ تراوحت نسبة التعرف الصحيح لـ SVM ما بين 75% عندما كانت قيمة μ=0 ولغاية 90% عندما أصبحت قيمة μ=0.7 أما نسبة التعرف الصحيح لـ LRM فتراوحت النسبة مابين 59% عند قيمة μ=0 ، و 86% عند قيمة μ=0.7 مما يعني تفوق SVM في جميع حالات μ كذلك يمكن ملاحظة الفارق الواضح بين دقة SVM و LRM في نسبة التعرف الصحيح في الحالات عند جميع قيم (μ=0,0.1,0.2,0.3, 0.4,0.5,0.6,0.7) ، كما يمكن توضيح عملية التداخل بين المشاهدات في الرسومات والأشكال المرافقة لكل عملية تصنيف وفي الحالة التي يكون التباين فيها =1.5 وحجم العينة n=100 وكالتالي :

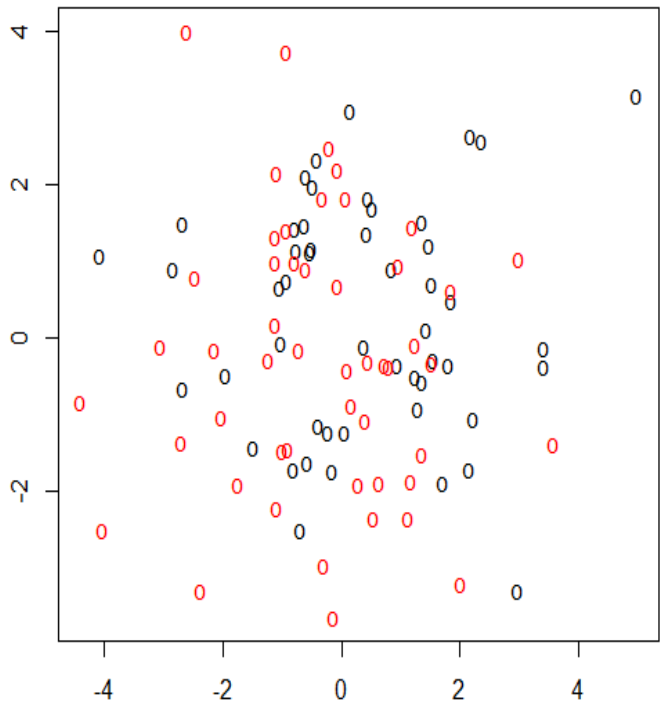
شكل (3-58)

عند مستوى $\sigma^2 = 1.5$ وحجم عينة n=100 و μ =0



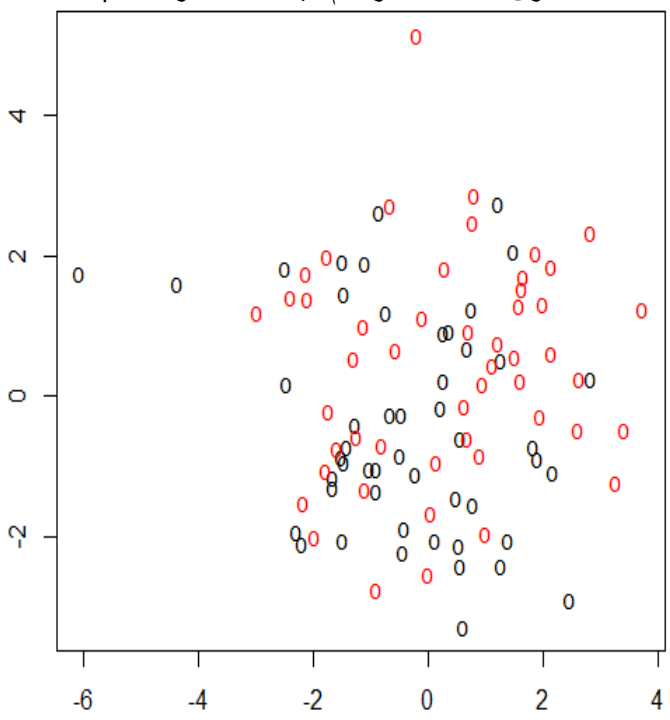
شكل (3-59)

عند مستوى $\sigma^2 = 1.5$ وحجم عينة n=100 و μ =0.1



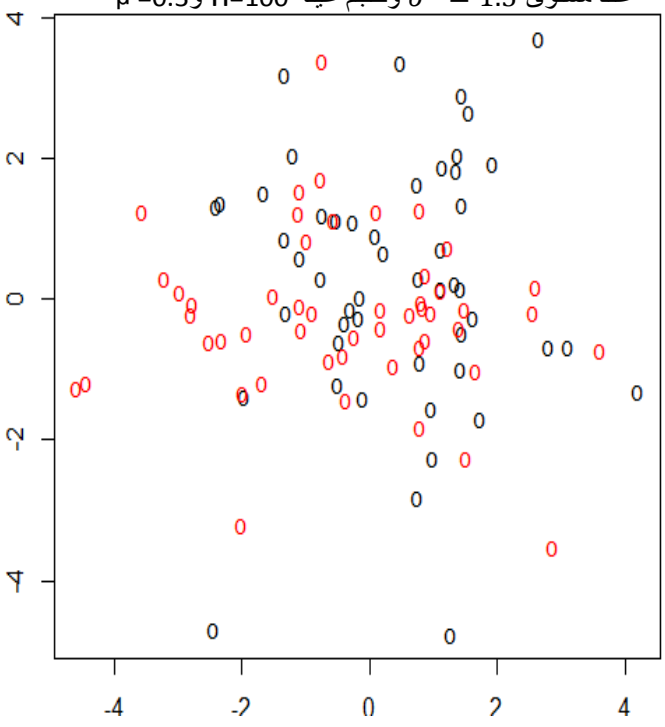
شكل (3-60)

عند مستوى $\sigma^2 = 1.5$ وحجم عينة n=100 و μ =0.2



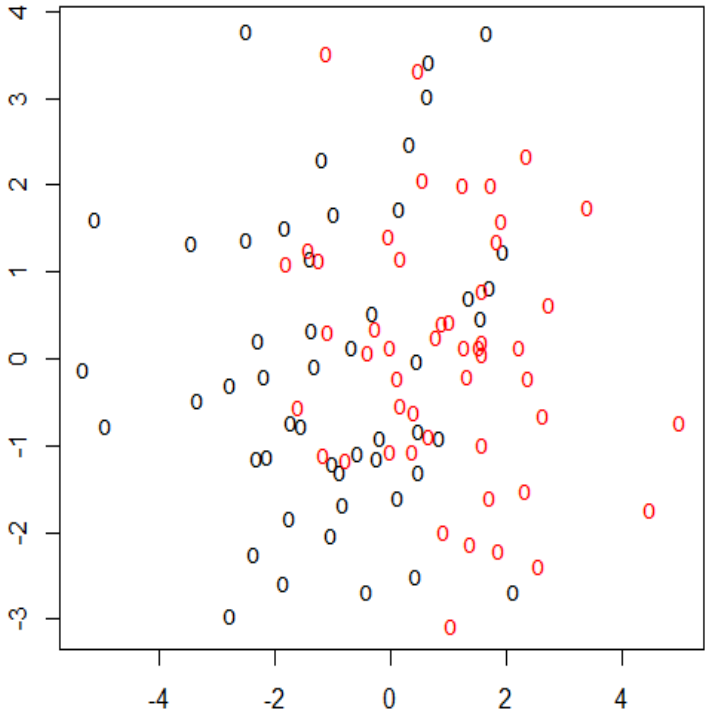
شكل (3-61)

عند مستوى $\sigma^2 = 1.5$ وحجم عينة n=100 و μ =0.3



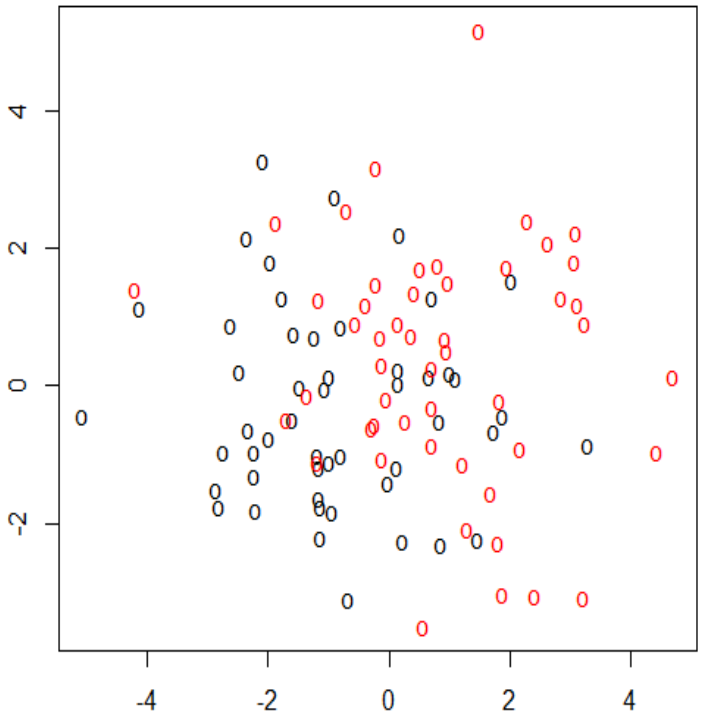
شكل (3-62)

عند مستوى تباين $\sigma^2 = 1.5$ وحجم عينة n=100 و μ =0.4



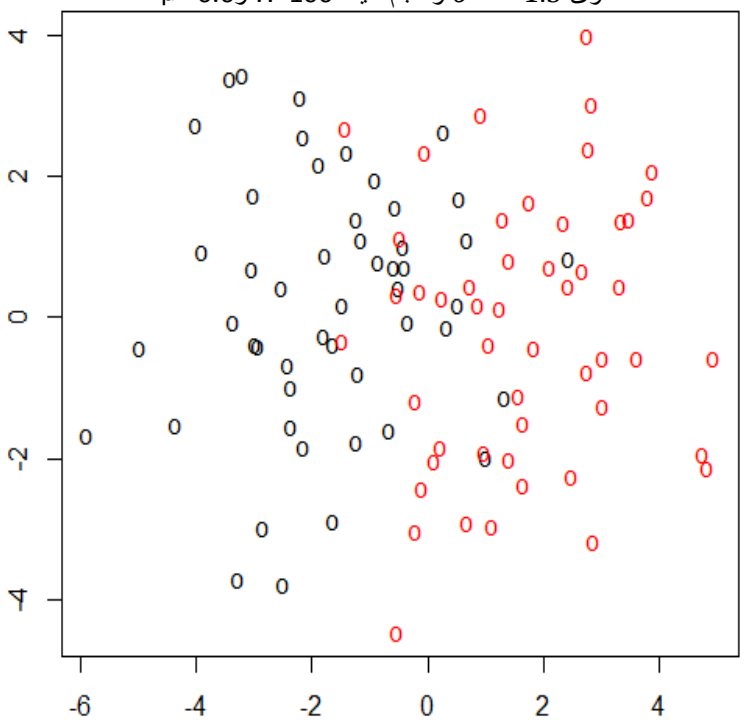
شكل (3-63)

عند مستوى تباين $\sigma^2 = 1.5$ وحجم عينة n=100 و μ =0.5



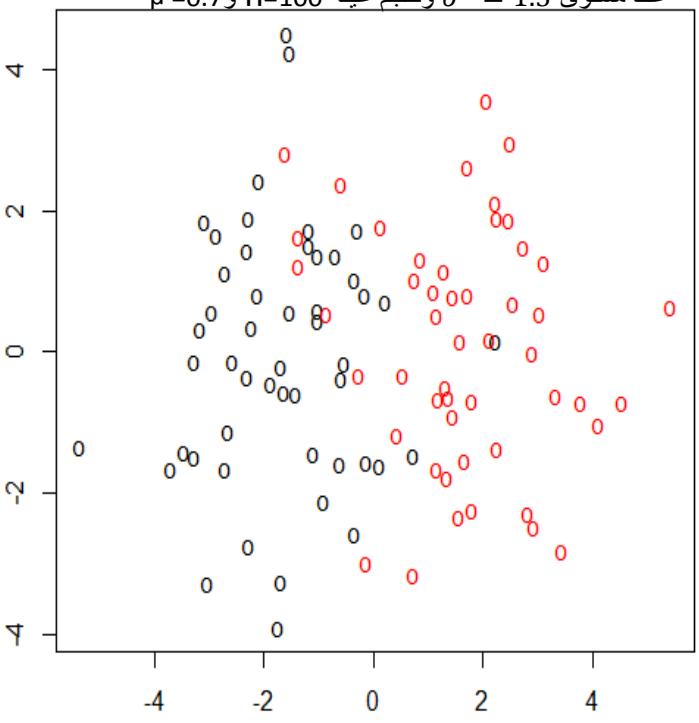
شكل (3-64)

عند مستوى تباين $\sigma^2 = 1.5$ وحجم عينة n=100 و μ =0.6



شكل (3-65)

عند مستوى تباين $\sigma^2 = 1.5$ وحجم عينة n=100 و μ =0.7



يلاحظ من الاشكال (3-58) و(3-59) و (3-60) و(3-61) و (3-62) و (3-63) و (3-64) أن التداخل يكون قوياً بين مشاهدات المجموعتين ثم يبدأ تدريجياً بالإنخفاض كلما إزدادت قيمة μ حتى تصل قيمة (μ) الى 0.7 عندها يحدث الإنفصال شبه التام بين مشاهدات المجموعتين كما في الشكل (3-65) .

كما أظهرت نتائج المحاكاة وعندما تكون قيمة التباين ($\sigma^2 = 1.5$) ولعينة بحجم n=216 أن طريقة SVM كانت أفضل من طريقة LRM وفي جميع حالات μ وكما مبين في الجدول (3-9):-

جدول (3-9)

نتائج التصنيف عند مستوى تباين ($\sigma^2 = 1.5$) وحجم عينة n=216

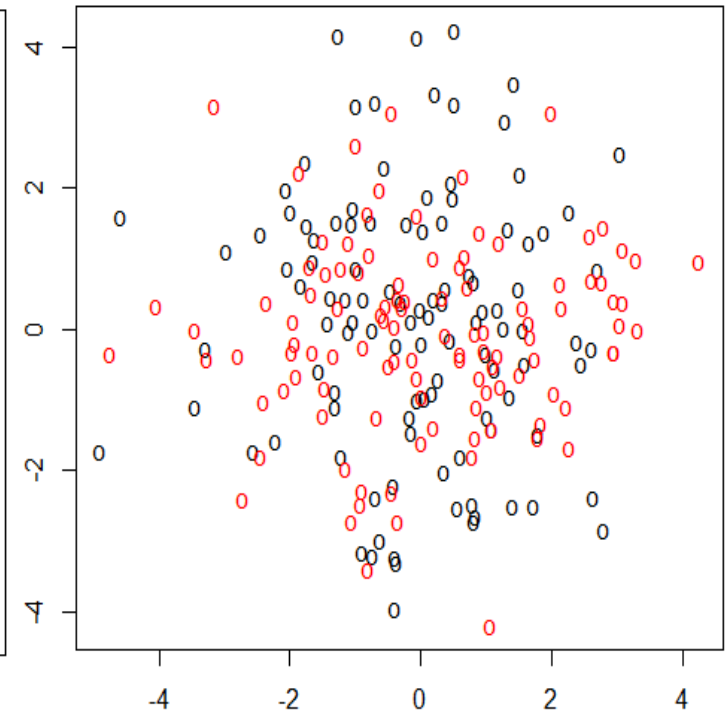
الوسط الحسابيμ	نسبة التصنيف بواسطةSVM	نسبة التصنيف بواسطةLRM
0.0	70 %	56 %
0.1	71 %	58 %
0.2	72 %	63 %
0.3	75 %	68 %
0.4	79 %	73 %
0.5	82 %	78 %
0.6	86 %	82 %
0.7	89 %	86 %

إعداد الباحث بالإعتماد على نتائج برنامج (R- language)

يظهر الجدول المذكور آنفاً نسب التعرف الصحيح لكلا الطريقتين عند حجم العينة الكبيرة (n=216) والتباين ($\sigma^2 = 1.5$) إذ تراوحت نسبة التعرف الصحيح لـ SVM ما بين 70% عندما كانت قيمة μ=0 ولغاية 89% عندما أصبحت قيمة μ=0.7 أما نسبة التعرف الصحيح لـ LR فتراوحت النسبة مابين 56% عند قيمة μ=0 ، و 86% عند قيمة μ=0.7 مما يعني تفوق SVM في جميع حالات μ كذلك يمكن ملاحظة الفارق الواضح بين دقة SVM وLRM في نسبة التعرف الصحيح في الحالات عند جميع قيم (μ=0,0.1,0.2,0.3, 0.4,0.5,0.6,0.7) ، كما يمكن توضيح عملية التداخل بين المشاهدات في الرسومات والأشكال المرافقة لكل عملية تصنيف وفي الحالة التي يكون التباين فيها ($\sigma^2 = 1.5$) وحجم العينة n=216 وكالتالي :

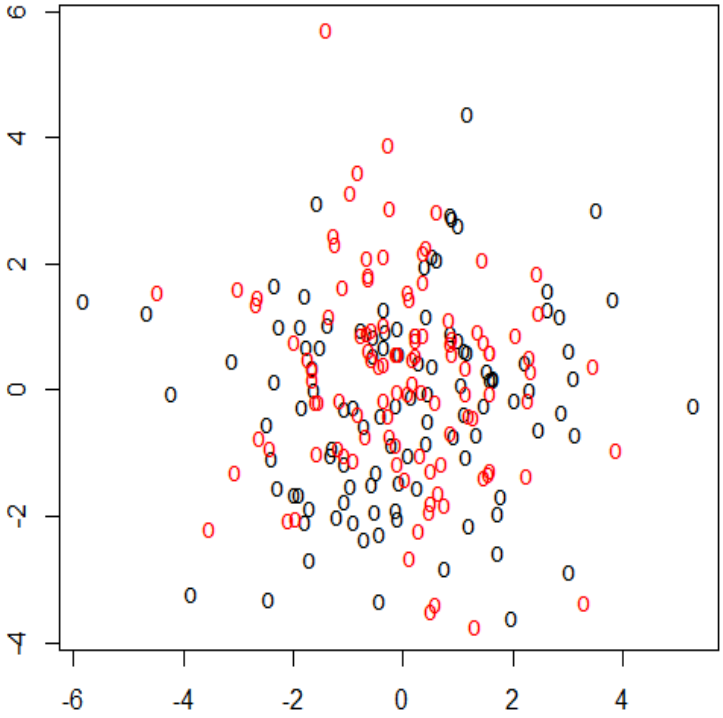
شكل (3-66)

عند مستوى $\sigma^2 = 1.5$ وحجم عينة n=216 و μ =0



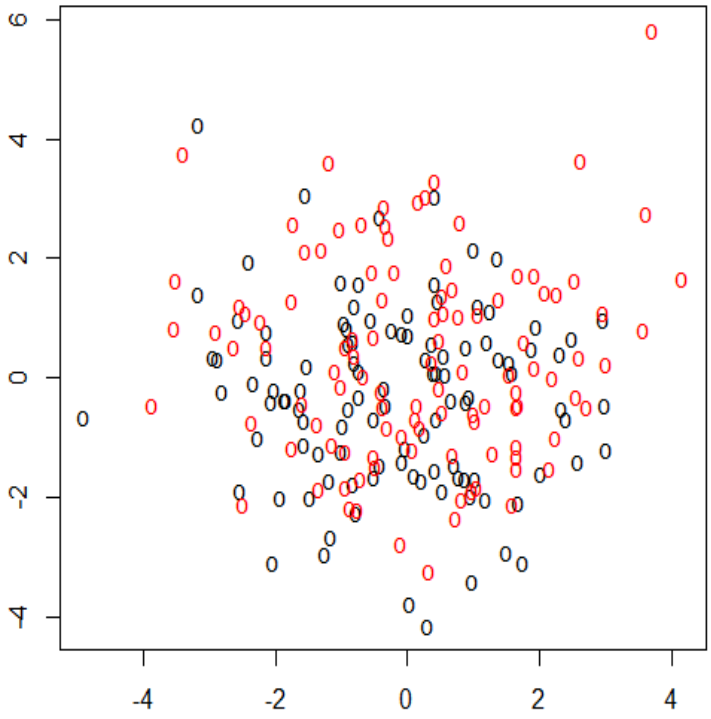
شكل (3-67)

عند مستوى $\sigma^2 = 1.5$ وحجم عينة n=216 و μ =0.1



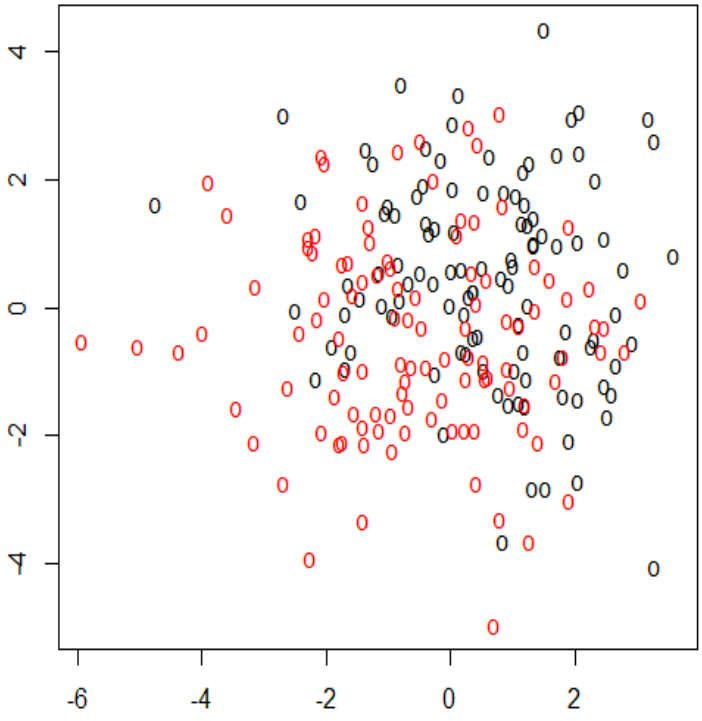
شكل (3-68)

عند مستوى $\sigma^2 = 1.5$ وحجم عينة n=216 و μ =0.2



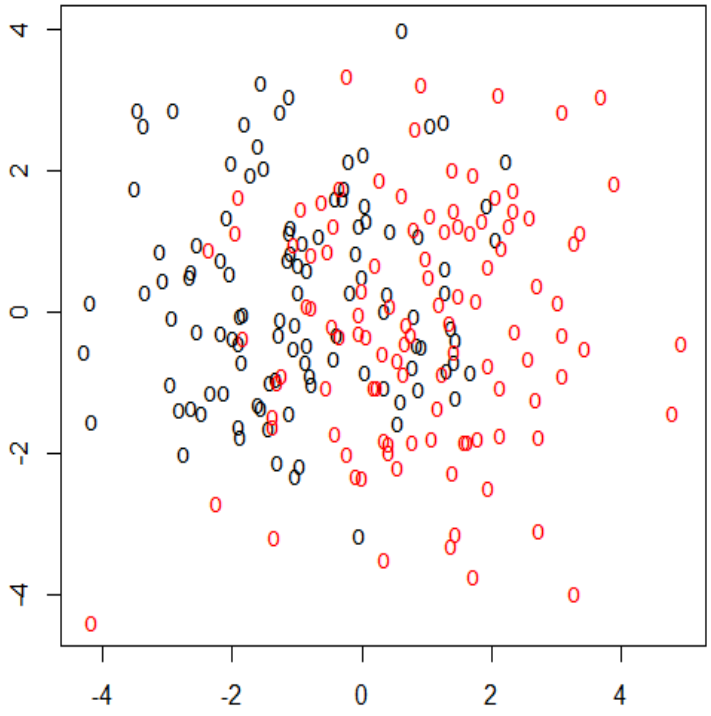
شكل (3-69)

عند مستوى $\sigma^2 = 1.5$ وحجم عينة n=216 و μ =0.3



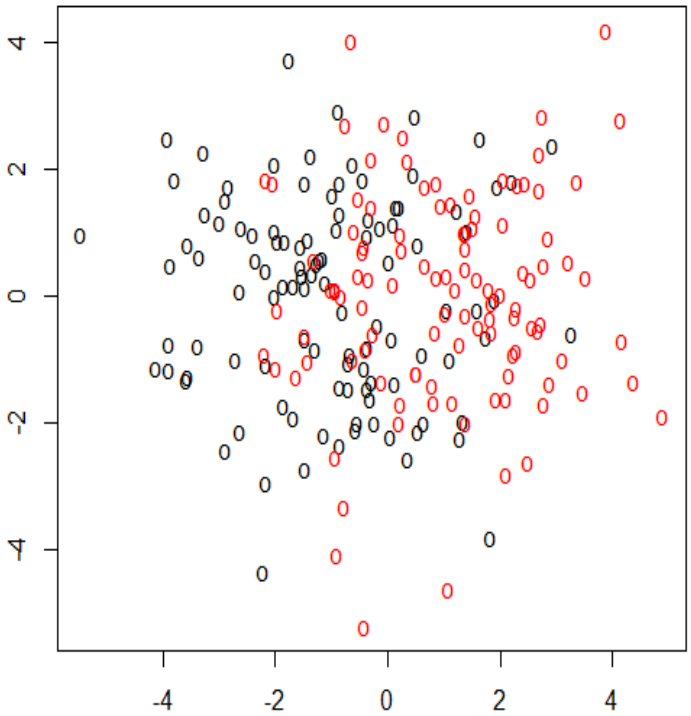
شكل (3-70)

عند مستوى $\sigma^2 = 1.5$ وحجم عينة n=216 و μ =0.4



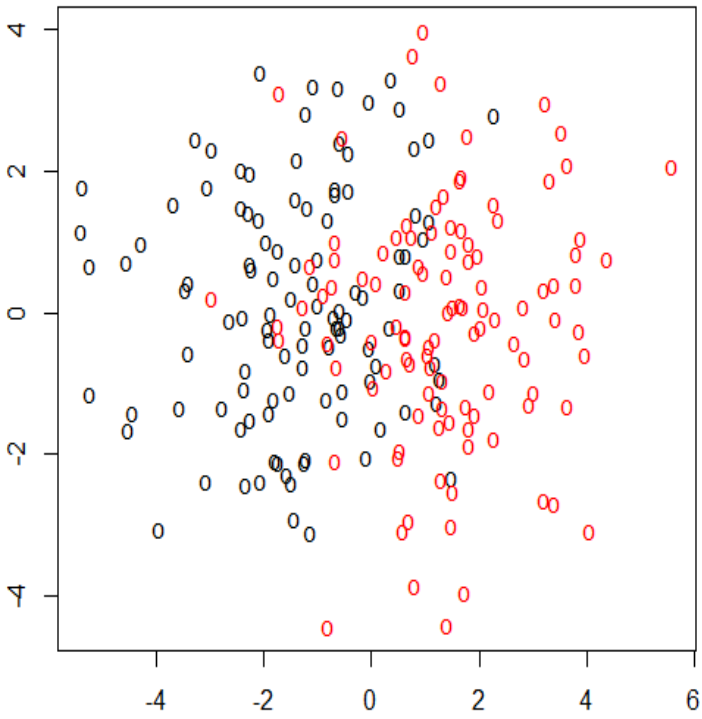
شكل (3-71)

عند مستوى $\sigma^2 = 1.5$ وحجم عينة n=216 و μ =0.5



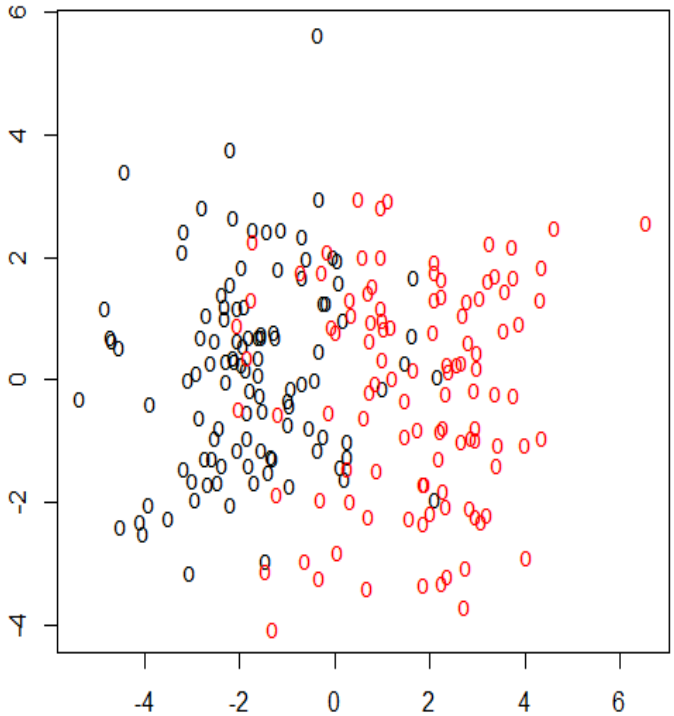
شكل (3-72)

عند مستوى $\sigma^2 = 1.5$ وحجم عينة n=216 و μ =0.6



شكل (3-73)

عند مستوى $\sigma^2 = 1.5$ وحجم عينة n=216 و μ =0.7



يلاحظ من الاشكال (3-66) و (3-67) و(3-68) و (3-69) و(70) و(3-71) و (3-72) أن التداخل يكون قوياً بين مشاهدات المجموعتين ثم يبدأ تدريجياً بالإنخفاض كلما إزدادت قيمة μ حتى تصل قيمة μ الى 0.7 عندها يحدث الإنفصال شبه التام بين مشاهدات المجموعتين كما في الشكل (3-73) .

الفصل الرابع

الجانب التطبيقي

(Practical Part)

4-1 : مفاهيم عامة عن مرض السكري

يعد مرض السكري من الامراض الخطيرة على صحة الإنسان ، وفي حالة تفاقم المرض وعدم اهتمام المريض المصاب بداء السكري واهماله لاخذ الدواء او تنظيم الطعام وعدم اتباعه لإرشادات الطبيب بدقة ممكن ان يؤدي ذلك الى حدوث مضاعفات خطيرة مثل إلحاق الضرر بالقلب والأوعية الدموية والعينين والكلى والأعصاب . وربما يؤدي الى بتر الأعضاء[54] .

وتشير اخر الاحصاءات لوزارة الصحة العراقية في عام2013 الى إصابة حوالي 10.2% من الشعب العراقي بداء السكري أي ما يعادل حوالي 3.5 مليون شخص . ويحتل العراق المرتبة التاسعة عربياً والثلاثين عالمياً [50] .

4-1-1 :- توصيف مرض السكري

مرض السكري هو مرض مزمن يحدث عندما يعجز البنكرياس عن إنتاج الأنسولين بكمية كافية أو عندما يعجز الجسم عن الإستخدام الفعال للأنسولين الذي ينتجه . والأنسولين هو هرمون ينظم مستوى السكر في الدم .

4-1-2 :- أسباب مرض السكري

لمرض السكري اسباب وعوامل عدة منها :

1- نتيجة لتراكم السكر في الدم بسبب عجز الجسم عن هضم السكر وتحويله الى الطاقة نتيجة لقلة إفراز مادة الأنسولين من البنكرياس والذي يرجع سبب حدوث المرض الى خلل في عضو البنكرياس في الجسم المسؤول عن إفراز مادة الأنسولين في الدم اللازمة ضبط مستوى السكر في الدم .

2- عامل الوراثة فعندما يكون أحد أفراد الأسرة مصاباً بالسكري سيؤدي إنتقاله بالجينات الوراثية إلى أحد أبنائه والذي ينتقل عن طريق إصابة الأُم أو إصابة الأب وهذه أحد أسباب مرض السكري ، كما لايمكن تجنب مرض السكري لأنه لايظهر أعراض إلا بعد عمر ثلاث السنوات وعمر الخمسة عشرة وبعد هذا السن عند البلوغ ومرحلة الشباب وسن الأربعين عند بعض الناس.

3- ومن الأسباب الأخرى السمنة التي تتعلق أيضاً بالوراثة فمرض السكري والسمنة تربطهما علاقة مغلقة فكل يؤدي إلى الآخر فالسمنة هي الإفراط في تناول الطعام مما يؤدي الى تغير حجم الجسم وزيادة الوزن مما يؤدي الى تراكم نسبة السكر في الدم ونتيجة لقلة الحركة والنشاط البدني لايمكن هضم السكر في الدم ،

بسبب عدم مقدرة خلايا بيتا على فرز الأنسولين الكافي لهضم كميات السكر المخزونة في الدم ، مما يرفع نسبة السكر مقارنة بالأنسولين في الدم .ويعاني حوالي 55% من المرضى المصابين بالنوع الثاني من السمنة[51] .

4- التقدم في العمر .

5- عوامل أخرى .

4-1-3 :- أنواع مرض السكري

أ- مرض السكري من النوع الأول

يتسم مرض داء السكري من النوع الأول الذي كان يعرف سابقاً بإسم داء السكري المعتمد على الأنسولين أو داء السكري الذي يبدأ في مرحلة الطفولة ، أو الشباب ما قبل سن الأربعين ، بنقص إنتاج الأنسولين ، أو لايستطيع الجسم من إنتاج الأنسولين إطلاقاً بسبب خسارة خلايا بيتا المنتجة للأنسولين في خلايا لانكرهانس بالبنكرياس مما يؤدي الى نقص الأنسولين[54] ، والسبب الرئيس لهذه الخسارة هو مناعة ذاتية تتميز بهجوم الخلايا المناعية على خلايا بيتا ويمثل حوالي 10% فقط من جميع الحالات ويقتضي تعاطي الأنسولين يومياً[51].

ب - مرض السكري من النوع الثاني :

ويحدث هذا النوع الذي كان يسمى سابقاً داء السكري غير المعتمد على الأنسولين أو داء السكري الذي يظهر في مرحلة الكهولة وذلك بسبب عدم فعالية استخدام الجسم للإنسولين ، (أي السكري غير المعتمد على الأنسولين) وتحدث في معظمها نتيجة لفرط الوزن والخمول البدني[54] ، إذ يعاني حوالي 55% من مرضى النوع الثاني من السمنة[51] ، ويتميز النوع الثاني عن الأول من حيث وجود مقاومة مضادة لمفعول الأنسولين فضلاً عن قلة إنتاج الأنسولين، وهذا النوع لم يكن يصادف إلا في البالغين حتى وقت قريب لكنه يحدث الآن في صفوف الأطفال ايضاً[51]. ويقتضي تعاطي الحبوب أو الأنسولين.

وهناك حالات أُخرى من داء السكري مثل داء السكري الحملي الذي يحدث في أثناء فترة الحمل والذي عادة يزول بعد الولادة وقد تحدث مضاعفات تؤدي الى الإصابة بالنوع الثاني من داء السكري ، وهناك أيضاُ مرض السكري الناتج عن اختلال الجلوكوز في أثناء الصيام .

توجد العديد من المسببات النادرة لمرض السكري التي لايمكن تصنيفها كنوع أول أو ثان أو سكري

الحوامل وتثير محاولات تصنيفها الكثير من الجدل ، ويرجع نظام التسمية المستعمل حالياً من قبل منظمة الصحة العالمية الى العام 1999م .

4-1-4 :- أعراض مرض السكري

تشتمل اعراض كلا نوعي مرض السكري على مايلي :

1- زيادة العطش وشرب الكثير من السوائل .

2- كثرة التبول .

3- الشعور بالتعب من دون سبب واضح .

4- نقص الوزن .

5- بطء شفاء الجروح .

6- تشوش الرؤية .

ويلاحظ عند مرضى النوع الأول قد تظهر الأعراض عادة خلال أسابيع قليلة ، ولكن سرعان ما تصبح واضحة جداً ، أما في النوع الثاني فيمكن للأعراض أن تظهر ببطء كبير ، على فترة أشهر ،ويكون لدى بعض المصابين بمرض السكري من النوع الثاني أعراض حقيقية جداً بحيث يعتقدون انها لأسباب أخرى ، وربما لاتوجد أعراض على الإطلاق لدى عدد قليل من الناس ، وقد تكون أعراض هذا النمط مماثلة لأعراض النوع الأول ولكنها قد تكون أقل وضوحاً في كثير من الأحيان ، ولذا فقد يشخص الداء بعد مرور عدة أعوام على بدء الأعراض أي بعد حدوث المضاعفات[54].

5-1-4 :- مضاعفاته

هناك نوعان من المضاعفات الناجمة عن الإصابة بداء السكري هما :

أ- المضاعفات قصيرة الأجل

المضاعفات قصيرة الأجل الناجمة عن النوعين الأول والثاني والتي تتطلب المعالجة الفورية في مثل هذه الحالات التي لاتتم معالجتها فوراً قد تؤدي الى حصول اختلاجات (convulsions)وهي حالة من تبدل الوعي ناجمة عن إطلاق االدماغ شحنات كهربائية والى غيبوبة (Coma) [52].

وأهم تلك المضاعفات

1- هايبوغلايسيميا (Hypoglycaemia) (إنخفاض مستويات الجلوكوز في الدم)

وتحدث الهايبوغليسيميا عندما ينخفض مستوى سكر الجلوكوز في الدم أقل من 4مليمول/لتر . ويمكن أن تحدث للأشخاص الذين يأخـذون ادوية معينة لمرض السكري عن طريق الفم أو أولـئك الـذين يستخدمون الأنسولين .

ممكن ان تنخفض نسبة الجلوكوز في الدم للأسباب التالية :

- تأخر الوجبات أو تفويتها.
- لايوجد ما يكفي من الكربوهيدرات في الوجبة .
- نشاط إضافي أو نشاط متعب أكثر .
- زيادة جرعة الأنسولين أو أدوية السكري .
- الكحول.

2- هايبرغلايسيميا (Hyperglycaemia) (ارتفاع مستويات الجلوكوز)

عندما تكون مستويات الجلوكوز في الدم مرتفعة أي أعلى من المستويات الموصى بها أعلى من 15مليمول / لتر ، والتي قد ترتفع نتيجة للأسباب التالية :

- الإكثار من أكل الكربوهيدرات .
- عدم أخذ المقدار الكافي من الأنسولين وأدوية السكري عن طريق الفم
- المرض أو الإلتهاب .
- الإجهاد العاطفي أو البدني أو العقلي .
- بعض الأقراص أو الأدوية المعينة مثل الكورتيزون أو المنشطات[55].

ب- المضاعفات المزمنة

ممكن أن يتسبب داء السكري مع مرور الوقت في إلحاق الضرر بالقلب والأوعية الدموية والعينين والكلى والأعصاب ، ويزداد خطر تعرض البالغين المصابين بداء السكري للنوبات القلبية والسكتات الدماغية ضعفين أو ثلاثة أضعاف ويؤدي ضعف تدفق الدم والإعتلال العصبي (تلف الأعصاب) في القدمين الى زيادة إحتمالات الإصابة بقرح القدم والعدوى وإلى ضرورة بتر الأطراف في نهاية المطاف .

وقد يحدث ضرر في شبكية العين إذ توجد الأوعية الدموية الرفيعة المهمة للرؤية ، وهذا مايسمى إعتلال الشبكية السكري من الأسباب الرئيسة التي تؤدي الى العمى ويحدث نتيجة لتراكم الضرر الذي

يلحق بالاوعية الدموية الصغيرة في الشبكية على المدى الطويل ، وتعزى نسبة 2.6% من حالات العمى في العالم الى داء السكري . ويعد داء السكري من الأسباب الرئيسة للفشل الكلوي.

وقد تؤدي الى مشكلات في اللثة وتسوس الأسنان ، وايضاً تؤدي إصابة الأشخاص المتقدمين بالعمر بالنوع الثاني من السكري الى الوفاة[54].

وهنالك مضاعفات قد تحدث للأُم بسبب السكر الحملي :

1- تسمم الحمل (pre-eclampsia) وهو إرتفاع ضغط الدم بعد 20 أُسبوع من الحمل .

2- امكانية الاصابة بالسكري الحملي في الحمل التالي .

3- قد تتضاعف مضاعفات مقدمات السكري الى الإصابة بالنوع الثاني من داء السكري في سن متقدم[53].

4- قد تحدث مضاعفات للمولود بسبب السكري الحملي مثل فرط النمو ، نقص السكر في الدم ، اليرقان ، الموت[52].

4-2 جمع البيانات :

يستهدف هذا الفصل تقديم إسهام ما في هذه الدراسة ، وغايته الرئيسة في محاولة تصنيف البيانات من خلال طريقة "SVM" وطريقة "LRM" ، إذ تم جمع البيانات لعينة البحث من مستشفى الموانئ العام في البصرة في مركز الغدد الصم من طبلات المرضى الموجودة في أرشيف المركز المذكور وتم إختيار عينة عشوائية حجمها (216) مريضاً توزعوا على النوعين وكان عدد المرضى من النوع الأول (36) والنوع الثاني (180) .

4-2-1 تعريف متغيرات النموذج :

تم توصيف المتغيرات من خلال مجموعة بيانات العينة وتحويلها الى أرقام وحروف لغرض إدخالها الى الحاسوب لسهولة التعامل معها ، ومن أهم العوامل المساعدة أو المؤثرة في تحديد النوع الذي ينتمي اليه المصاب بالسكري هي :

X_1 : يمثل متغير العمر

X_2 : يمثل متغير الجنس وهو متغير اسمي للذكر = 1 ، وللأُنثى = 0

X_3: يمثل متغير الوزن بالكيلو غرام

X_4 : يمثل متغير الطول بالسنتيمر

X_5 : متغير مدى إستجابة المريض لنوع العلاج وهو متغير اسمي فإذا كان المريض يستجيب للأنسولين = 1 ، للحبوب = 2 ، اذا كان يستجيب للحبوب والأنسولين = 3 .

Y : يمثل متغير الإستجابة فإذا كانت قيمته=1+ في طريقة آلة المتجه الداعم فالمشاهدة تنتمي للمجموعة الأولى ، وإذا كانت =1− فالمشاهدة تنتمي للمجموعة الثانية .أما في طريقة الإنحدار اللوجستي فإذا كانت قيمته =1 فالمشاهدة تنتمي للمجوعة الأولى ، وإذا كانت قيمته = 0 فالمشاهدة تنتمي للمجموعة الثانية .

4-3 : رسم البيانات :

بعد رسم المتغيرات التوضيحية (التفسيرية) الى متغير الإستجابة (المتغير التابع) وبإستخدام البرنامج الإحصائي (R Language) ظهرت النتائج بالشكل التالي :

الشكل (4-1)

إعداد الباحث بالإعتماد على نتائج برنامج (R language)

ويظهر من الشكل السابق كيفية توزيع البيانات الى مجموعتين تشير النقاط أو الاشكال السوداء الى المجموعة الأُولى أو النوع الاول والنقاط أو الأشكال الحمراء الى المجموعة الثانية أو النوع الثاني ، ويظهر الشكل ايضاً أن البيانات ليست متداخلة فيما بينها تداخلاً قوياً وإنما يوجد هناك تداخل طفيف بين المجموعتين ، وسوف يتم استخدام أُسلوب آلة المتجه الداعم ومن ثم سوف نستخدم أُسلوب الإنحدار اللوجستي لنعرف الفرق بين الأُسلوبين ودقة كل منهما في التصنيف وكما يلي :

4-4 : آلة المتجه الداعم Support Vector Machine (SVM)

لقد تمت الإستعانة بالبرنامج الإحصائي الجاهز لغة R من أجل الحصول على النتائج التالية بطريقة SVM وكما يظهر من النتائج الآتية :

جدول (4-1)

ملخص التصنيف الصحيح والخاطئ على وفق طريقة آلة المتجه الداعم "SVM"

Yhat	Type 1	Type 2	Sum
Type 1	31	3	34
Type 2	5	177	182
Sum	36	180	216

إعداد الباحث بالإعتماد على نتائج برنامج (R- language)

يظهر من الجدول المذكور آنفاً المستخلص للتصنيف المبني على أساس أُسلوب آلة المتجه الداعم Support Vector Machine (SVM) للمرضى الـ (216) المصابين بداء السكري من النوع الأول والنوع الثاني ، إذ نلاحظ من خلاله بأنه تم تصنيف مجموعة المصابين بداء السكري من النوع الأول والبالغ عددهم الفعلي (36) كالآتي :-

تم تصنيف(31) مريضاً من مجموع الـ (36) على وفق آلة المتجه الداعم (SVM) على النوع الأول أي ان نسبة التمييز الصحيح كانت 86 % أما الخمسة المتبقين لإكمال الـ (36) المصابين فعلاً بداء السكري من النوع الأول فقد صنفوا للنوع الثاني .

أما بالنسبة للمجموعة الثانية وهي مجموعة المصابين بالنوع الثاني من داء السكري فالعدد الفعلي لهؤلاء هو (180) ، وقد تم تصنيف (177) منهم بشكل صحيح على وفق تقنية آلة المتجه الداعم (SVM)

أما الـ (3) المتبقون فقد تم تصنيفهم الى النوع الأول على وفق تقنية آلة المتجه الداعم (SVM) أي أن نسبة التصنيف الصحيح للمجموعة الثانية وهي مجموعة المصابين بالنوع الثاني من داء السكري كانت% 98 .

أما نسبة التصنيف الكلية لكلا النوعين على وفق أُسلوب أو تقنية آلة المتجه الداعم (SVM) فكانت 96% .

4-4-1 حساب نسب التصنيف الصحيح :

نسبة التصنيف الصحيح للنوع الأول من داء السكري

$$\frac{31}{31+5} * 100 = 86\ \%$$

نسبة التصنيف الصحيح للنوع الثاني من داء السكري

$$\frac{177}{177+3} * 100 = 98\%$$

نسبة التصنيف الصحيح الكلية

$$\frac{31+177}{216} * 100 = 96\ \%$$

أما نسبة التصنيف الخاطئ فيمكن تعريفها على أنها تصنيف مشاهدة الى نوع بينما هي في الحقيقة تعود الى نوع آخر .

4-4-2 ايجاد المتجهات الداعمة Support Vectors

4-4-2-1 المتجهات الداعمة في النوع الأول :

يتم ايجاد أو تحديد قيم المتغيرات التي تمثل المتجهات الداعمة (support vectors) عن طريق مخرجات دالة الهدف $Y= w*x+b$. وكما في الجدول الآتي :-

جدول (4-2)

تحديد قيم المتجهات الداعمة من النوع الأول على وفق آلة المتجه الداعم "SVM"

المتجهات الداعمة للنوع الاول					
obvs	X_1	X_2	X_3	X_4	X_5

1	−2.3842217	−1.2506612	−2.2420756731	−0.71988355	−1.60443221
3	−0.69255	0.795875	−0.58522	3.110039	−1.60443
6	0.838002	0.795875	−0.29284	3.301535	0.014995
7	−1.65922	0.795875	−1.58421	−0.43264	−1.60443
11	−2.22311	0.795875	−1.51111	−0.91138	0.014995
15	−1.98144	0.795875	−1.46238	−0.81563	−1.60443
23	−1.25644	0.795875	−1.31619	−0.81563	1.634422
24	−1.41755	−1.25066	−0.68268	0.716337	1.634422
44	−0.77311	−1.25066	−0.24411	1.09933	−1.60443
62	−1.17589	0.795875	−0.73142	−0.52839	1.634422
66	−1.337	0.795875	−0.92634	−1.10288	1.634422
77	−1.73978	0.795875	−0.29284	0.141849	1.634422
78	−1.09533	0.795875	−0.14664	0.429093	−1.60443
83	−0.612	−1.25066	0.291935	1.290826	−1.60443
113	−0.69255	0.795875	−0.00045	0.141849	−1.60443
122	−0.69255	−1.25066	0.194473	0.333345	−1.60443
124	−1.57867	−1.25066	0.974168	1.769566	0.014995
128	−1.65922	0.795875	0.243204	0.333345	0.014995
130	−0.612	0.795875	−0.00045	−0.1454	−1.60443
149	−1.57867	−1.25066	0.730513	0.907833	1.634422
178	−0.77311	0.795875	0.535589	−0.1454	−1.60443
183	−1.25644	−1.25066	0.438127	−0.43264	−1.60443
186	−1.49811	0.795875	0.681782	−0.24114	0.014995

198	−1.41755	0.795875	0.974168	−0.33689	1.634422
200	−1.57867	0.795875	1.802595	0.716337	−1.60443
202	−1.90089	0.795875	1.266554	−0.33689	0.014995
213	−0.612	−1.25066	3.410716	0.812085	0.014995

إعداد الباحث بالإعتماد على نتائج برنامج (R− language)

ويلاحظ من الجدول السابق أن عدد المتجهات الداعمة للنوع الأول هي (27) متجهاً داعماً ، ويظهر العمود الأول من الجدول السابق المشاهدات التي تمثل المتجهات الداعمة Support Vectors عند إستخدام تقنية Support Vector Machine (SVM) والتي تبدأ من المشاهدة (1,3,6,.....,213) أما الإعمدة من الثاني الى الخامس فتمثل قيم متغيرات تلك المشاهدات مثلاً قيمة (7) في العمود الأول تمثل المشاهدة السابعة والتي كانت قيمتها في العمود الثاني والتي تمثل المتغير الثاني x_2 تساوي (−1.65922) وفي العمود الثالث الذي يمثل المتغير الثالث x_3 تساوي (0.795875) وهكذا بالنسبة لبقية المتغيرات .

علماً أن برنامج R وعند إجراء عملية أو تقنية آلة المتجه الداعم يقوم بشكل أوتوماتيكي بتحويل القيم الأصلية الى قيم معيارية أو بالصيغة القياسية (stander) أي تطرح كل قيمة من متوسطها وتقسمها على الانحراف المعياري لها .

4−4−2−2 المتجهات الداعمة في النوع الثاني :

أما المتجهات الداعمة في النوع الثاني فقد جرى تحديد مشاهداتها في الجدول الآتي:−

جدول (4−3)

مشاهدات المتجهات الداعمة للنوع الثاني على وفق طريقة آلة المتجه الداعم "SVM"

المتجهات الداعمة للنوع الثاني					
obvs	X_1	X_2	X_3	X_4	X_5
4	0.51578	−1.25066	−2.14461	−1.96461	0.014995
8	−1.65922	0.795875	−1.68167	−1.19862	−1.60443
22	−0.20922	0.795875	−1.31619	−0.81563	−1.60443

25	−0.28978	0.795875	−1.16999	−0.62414	1.634422
32	−0.69255	0.795875	−1.12126	−0.71988	0.014995
34	−0.77311	−1.25066	−0.48776	0.812085	−1.60443
50	−0.28978	−1.25066	0.535589	2.631298	0.014995
53	2.529669	0.795875	−0.92634	−0.71988	1.634422
56	−0.37033	−1.25066	0.633051	2.631298	0.014995
67	−0.77311	0.795875	−0.68268	−0.52839	0.014995
71	−0.85367	−1.25066	−0.14664	0.620589	0.014995
73	0.113002	0.795875	0.438127	1.769566	0.014995
75	−0.69255	0.795875	−1.07253	−1.58162	0.014995
76	1.079669	−1.25066	0.097011	1.003581	−1.60443
79	−0.77311	−1.25066	−0.09791	0.524841	1.634422
81	−1.09533	−1.25066	0.04828	0.812085	1.634422
93	−0.12867	0.795875	−0.19538	0.141849	−1.60443
98	−0.69255	0.795875	−0.3903	−0.33689	1.634422
104	0.113002	−1.25066	−0.07355	0.141849	−1.60443
105	−0.69255	0.795875	−0.3903	−0.52839	1.634422
117	−0.28978	0.795875	−0.14664	−0.24114	−1.60443
140	−0.612	0.795875	0.097011	−0.1454	1.634422
147	1.804669	−1.25066	0.779244	1.003581	1.634422
148	−0.85367	−1.25066	−0.00045	−0.43264	0.014995
153	−0.28978	0.795875	0.486858	0.333345	−1.60443
171	−0.69255	0.795875	0.291935	−0.33689	1.634422

176	−0.85367	−1.25066	1.07163	0.812085	0.014995
188	−0.12867	0.795875	0.340666	−0.91138	−1.60443
189	0.113002	0.795875	0.827975	−0.1454	−1.60443
201	−0.69255	0.795875	1.705133	0.524841	0.014995
204	−0.53144	−1.25066	2.631021	1.195078	0.014995
205	−1.49811	0.795875	1.656402	−0.1454	0.014995
209	1.321335	−1.25066	0.535589	−2.06036	1.634422
212	−0.69255	0.795875	2.436097	0.141849	0.014995
214	1.482447	0.795875	0.827975	−2.92209	0.014995
215	1.724113	0.795875	3.410716	−0.04965	0.014995
216	−0.7731099	−1.2506612	6.3345744363	1.29082567	1.63442159

إعداد الباحث بالإعتماد على نتائج برنامج (R– language)

يظهر الجدول المذكور آنفاً أن عدد المتجهات الداعمة للنوع الثاني هي (37) متجهاً داعماً ، ويظهر العمود الأول من الجدول (4–3) المشاهدات التي تمثل المتجهات الداعمة Support Vectors عند إستخدام تقنية Support Vector Machine (SVM) والتي تبدأ من المشاهدة (4,8,22,.....,216) أما الإعمدة من الثاني الى الخامس فتمثل قيم متغيرات تلك المشاهدات .

وبذلك يكون مجموع المتجهات الداعمة (64) لكلا النوعين .

4–4–3 تصنيف المشاهدات :

4–4–3–1 تصنيف مشاهدات النوع الأول :

يتم تصنيف المشاهدات بناءً على دالة القرار أو دالة الهدف y=w*x+b إذ أن مخرجات هذه الدالة تصنف المريض إلى أي نوع ينتمي إليه فإما أن تصنفه الى النوع الاول او الى النوع الثاني وكما في الجدول الآتي :–

جدول (4-4)

تصنيف مشاهدات النوع الأول على وفقة طريقة آلة المتجه الداعم SVM

Observations	Value	observations	Value
1	1.00038248	78	1.00016577
2	1.25366099	83	0.46214169
3	1.00017760	90	1.16814702
6	0.48616126	113	0.22923135
7	0.90015245	122	0.46066452
9	1.19722422	124	-0.10651061
10	1.20753676	128	0.99971162
11	0.99955989	130	-0.05132823
12	1.18270775	131	1.48637576
15	1.00023952	149	-0.11268338
23	0.45815925	178	0.31691988
24	-0.09993102	179	1.52956118
26	1.21866533	183	0.98805132
37	1.72477215	186	0.63394402
44	0.58631967	198	0.24454617
62	0.22042885	200	1.00039856
66	0.51101440	202	0.96905574
77	0.97331877	213	-0.49762038

إعداد الباحث بالإعتماد على نتائج برنامج (R- language)

ويظهر الجدول السابق نتيجة عملية تصنيف مشاهدات النوع الأول إذ تبدو المشاهدات التي نجحت تقنية آلة المتجه الداعم في تصنيفها الى النوع الأول بشكل صحيح وكذلك مشاهدات النوع الأول التي تم

تصنيفها بشكل خاطئ الى النوع الثاني إذ أن العمود الأول والثالث يمثلان المشاهدات والعمود الثاني والرابع يمثلان قيمة المشاهدات وبإستخدام SVM ، إذا علمنا أن العدد الأصلي لمشاهدات النوع الأول في العينة الأصلية كانت (36) وبناءً على ما جاء في الجدول السابق نلاحظ أنه تم تصنيف (5) مشاهدات فقط بشكل خاطئ الى النوع الثاني أي تم تصنيفها الى النوع الثاني علماً أنها في الحقيقية تعود الى النوع الأول وكما مر بنا سابقاً في الجدول (4-1) وكذلك يلاحظ من الجدول (4-4) ان جميع المشاهدات كانت موجبة ما عدا خمسة مشاهدات كانت سالبة وهي كل من المشاهدات (24,124,130,149,213) .

4-4-3-2 تصنيف مشاهدات النوع الثاني :

هنا يتم تصنيف مشاهدات النوع الثاني بموجب معادلة دالة القرار وحسب طريقة "SVM" والتي يظهرها الجدول الآتي :-

جدول (4-5)

تصنيف مشاهدات النوع الثاني على وفق طريقة آلة المتجه الداعم "SVM"

Observations	Value	Observations	Value
4	-0.99956151	112	-2.52230116
5	-1.50239014	114	-1.23606630
8	0.46195618	115	-1.89612426
13	-1.70386587	116	-1.03722273
14	-1.83486065	117	-0.67752213
16	-1.26892548	118	-1.60014796
17	-1.91582048	119	-1.82664491
18	-1.17779755	120	-2.41484750
19	-1.45704089	121	-1.36683150
20	-2.14255333	123	-2.00689860
21	-2.02827005	125	-1.30802209

22	−1.00010913	126	−1.43891184
25	−1.00015554	127	−1.63075516
27	−1.59752112	129	−1.65241340
28	−2.18763272	132	−1.78121144
29	−1.08323762	133	−1.51841526
30	−2.25743152	134	−2.00853961
31	−1.78532095	135	−1.65760117
32	−1.00011582	136	−1.93654056
33	−1.54478316	137	−1.86413602
34	0.50479506	138	−2.40008967
35	−1.17265501	139	−1.08652943
36	−1.51196365	140	−1.00041960
38	−2.01334796	141	−2.30172865
39	−1.35123604	142	−1.47711843
40	−2.30990035	143	−2.48789204
41	−1.73437961	144	−1.15264129
42	−1.60568904	145	−1.01785306
43	−1.46686361	146	−1.94863034
45	−1.63849278	147	−1.00021613
46	−1.27346330	148	−1.00028060
47	−1.79810642	150	−1.77740851
48	−2.06615140	151	−1.39867111
49	−1.53676487	152	−2.16304929

50	−1.00014818	153	−0.31761333
51	−1.99066794	154	−1.34505946
52	−1.71992877	155	−2.42472894
53	−1.00007854	156	−2.09865660
54	−1.78518894	157	−2.48128641
55	−2.06561552	158	−1.62243867
56	−0.98418489	159	−2.51033854
57	−2.18634690	160	−1.21522891
58	−1.90920184	161	−2.16464504
59	−1.83348226	162	−1.58340194
60	−1.25677227	163	−1.82565468
61	−2.10682349	164	−1.82152093
63	−2.21712067	165	−1.96362368
64	−1.73106363	166	−1.93990123
65	−1.93425457	167	−2.48242607
67	−0.89385944	168	−1.63657155
68	−1.54817149	169	−1.13162689
69	−1.25244848	170	−1.98370408
70	−1.87943867	171	−0.86736739
71	−0.98713185	172	−2.24950133
72	−1.49532760	173	−2.24950133
73	−0.99976466	174	−1.96218589
74	−1.52094106	175	−1.90579069

75	−1.00049539	176	−0.99955915
76	−1.00010915	177	−2.36280068
79	−0.99979096	180	−1.81700694
80	−1.71927824	181	−2.12088924
81	−0.62085830	182	−2.47067143
82	−1.58263331	184	−1.88515609
84	−1.98513214	185	−1.93873088
85	−1.27182224	187	−1.01160076
86	−2.10425820	188	−0.99979040
87	−1.07029765	189	−1.00008244
88	−1.83061642	190	−1.65536929
89	−1.89214165	191	−1.91064402
91	−2.49345975	192	−1.40264922
92	−1.52585049	193	−1.49850394
93	−0.74612745	194	−1.52187794
94	−1.76246718	195	−1.30302627
95	−1.56319292	196	−1.32459156
96	−1.91555167	197	−1.47474434
97	−2.12836231	199	−1.87346502
98	−0.74708473	201	−0.99963865
99	−1.90313297	203	−1.20321560
100	−1.08976297	204	−0.88528851
101	−1.73512357	205	0.24016835

102	-2.13764924	206	-1.93348572
103	-1.96351471	207	-1.83007319
104	-0.67371834	208	-1.14929646
105	-0.72550343	209	-0.99995595
106	-1.87908034	210	-1.35241662
107	-1.51674071	211	-1.15243896
108	-2.39373866	212	-1.00009875
109	-1.86698753	214	-1.00024469
110	-2.48610165	215	-0.99982202
111	-1.26975538	216	-1.00019819

الجدول من إعداد الباحث بإستعمال برنامج (R- language)

ويظهر الجدول المذكور آنفاً نتيجة عملية تصنيف مشاهدات النوع الثاني بإستخدام آلة المتجه الداعم إذ يمكن الإستدلال أن المشاهدات التي تم تصنيفها الى النوع الثاني بشكل صحيح من قبل طريقة SVM والمشاهدات التي تم تصنيفها بشكل خاطئ الى النوع الأول إذ أن العمود الأول والثالث يمثلان المشاهدات والعمود الثاني والرابع يمثلان قيمة المشاهدات التي تم تصنيفها الى النوع الثاني أو الأول وبإستخدام طريقة SVM إذا علمنا أن العدد الأصلي لمشاهدات النوع الثاني في العينة الأصلية كانت (180) وبناءً على ما جاء في الجدول السابق نلاحظ أنه تم تصنيف (3) مشاهدات فقط بشكل خاطئ أي تم تصنيفها الى النوع الأول علماً أنها في الحقيقية تتبع النوع الثاني وكما لاحظنا ذلك في الجدول (4-1) ويلاحظ من الجدول (4-5) أن جميع المشاهدات سالبة ما عدا ثلاثة مشاهدات كانت إشارتها موجبة إذ تم تصنيفها بشكل خاطئ الى النوع الأول وهي كل من المشاهدات (8,34,205) .

4-4-4 متجه الأوزان The weights وحد التحيز Bias :

إن الغاية الأساسية لدالة القرار Y= w*x+b هي إيجاد قيم متجه الأوزان (weights) حسب المعادلة (21.2) وقيم حد التحيز (b) أي تمييز الإختلافـات في درجة التحيز، عما إذا كان أي إختلاف يمكن أن يفسر التداخل وكما في الجدول الآتي :-

جدول (4-6)

قيم متجه الأوزان

X_1	X_2	X_3	X_4	X_5
-16.97514	1.847392	-9.201491	5.259065	-6.110011

الجدول من إعداد الباحث بإستعمال برنامج (R- language)

يفسر الجدول السابق قيم متجه الأوزان لكل متغير إذ أن قيمة الوزن للمتغير x_1 كانت (-16.97514) ، وقيمة الوزن للمتغير x_2 كانت (1.847392) ، وقيمة الوزن للمتغير x_3 كانت (-9.201491) ، وقيم الوزن للمتغير x_4 كانت (5.259065) ، وقيمة الوزن للمتغير x_5 كانت (-6.110011) ، علماً بأن قيمة حد التحيز (Bias) كانت تساوي (0.1521142) .

4-5 : أُنموذج الإنحدار اللوجستي (LRM)

يبحث هذا المبحث معالجة البيانات بطريقة LRM وهي الطريقة الثانية التي تم إستخدامها في هذا البحث بعد طريقة SVM، وتم الحصول على النتائج التالية التي يظهرها الجدول الآتي :

جدول (4-7)

ملخص التصنيف الصحيح والخاطئ بموجب أُنموذج الإنحدار اللوجستي "RLM"

Yhat	Type 1	Type 2	Sum
Type 1	27	3	30
Type 2	9	177	186
Sum	36	180	216

إعداد الباحث بالإعتماد على نتائج برنامج (R- language)

يبين الجدول المذكور آنفاً خلاصة للتصنيف المبني على أساس دالة الإنحدار اللوجستي للمرضى الـ (216) المصابين بداء السكري من النوع الأول والنوع الثاني ، إذ نلاحظ من خلاله أنه تم تصنيف مجموعة المصابين بداء السكري من النوع الأول والبالغ عددهم الفعلي (36) كالتالي :

تم تصنيف(27) مريضاً من مجموع الـ (36) على وفق دالة الإنحدار اللوجستي على النوع الأول أي ان نسبة التمييز الصحيح كانت 75% أما التسعة المتبقون لإكمال الـ (36) المصابين فعلاً بداء السكري من النوع الأول فقد صنفوا للنوع الثاني .

أما بالنسبة للمجموعة الثانية وهي مجموعة المصابين بالنوع الثاني من داء السكري فالعدد الفعلي لهؤلاء هو (180) ، وقد تم تصنيف (177) منهم بشكل صحيح على وفق دالة الإنحدار اللوجستي ، أما الـ (3) المتبقون فقد تم تصنيفهم الى النوع الأول على وفق دالة الإنحدار اللوجستي أي أن نسبة التصنيف الصحيح للمجموعة الثانية وهي مجموعة المصابين بالنوع الثاني من داء السكري كانت 98%

أما نسبة التصنيف الكلية لكلا النوعين على وفق دالة الإنحدار اللوجستي فكانت94%

وقد تم حساب نسب التصنيف الصحيح للنوع الأول كالتالي :

نسبة التصنيف الصحيح للنوع الأول من داء السكري

$$\frac{27}{27+9} * 100 = 75\%$$

نسبة التصنيف الصحيح للنوع الثاني من داء السكري

$$\frac{177}{177+3} * 100 = 98\%$$

نسبة التصنيف الصحيح الكلية

$$\frac{27+177}{216} * 100 = 94\ \%$$

4-5-1 تصنيف المشاهدات

يتم تصنيف المشاهدات بناءً على نموذج الإنحدار اللوجستي $f(z)=E(Y/x) = \frac{e^{-z}}{1+e^{-z}}$ فإذا كانت قيمة هذه الدالة تساوي (1) يعني المريض ينتمي إلى النوع الاول أما إذا كانت قيمة الدالة تساوي (صفراً) فالمريض ينتمي الى النوع الثاني وكما يلي :

أولاً : تصنيف مشاهدات النوع الأول

تم تصنيف مشاهدات النوع الأول بإستخدام أُنموذج الإنحدار اللوجستي على وفق الجدول الآتي :-

الجدول (4-8)

تصنيف مشاهدات النوع الأول على وفق طريقة أُنموذج الإنحدار اللوجستيLRM

observations	Classification	Observations	Classification
1	type1	78	Type1
2	type1	83	type2
3	type1	90	Type1
6	type2	113	Type1
7	type1	122	type2
9	type1	124	Type1
10	type1	128	Type1
11	type1	130	type2
12	type1	131	Type1
15	type1	149	Type1
23	type2	178	Type1
24	Type1	179	Type1
26	Type1	183	type2
37	Type1	186	Type1
44	type1	198	Type1
62	type2	200	Type1
66	type2	202	Type1
77	Type1	213	type2

إعداد الباحث بالإعتماد على نتائج برنامج (R– language)

ويظهر الجدول المذكور آنفاً نتيجة عملية تصنيف مشاهدات النوع الأول إذ يظهر المشاهدات التي نجح نموذج الإنحدار اللوجستي في تصنيفها الى النوع الأول بشكل صحيح وكذلك مشاهدات النوع الأول

التي تم تصنيفها بشكل خاطئ الى النوع الثاني إذ أن العمود الأول والثالث يمثلان المشاهدات والعمود الثاني والرابع يمثلان قيمة المشاهدات وبإستخدام نموذج الإنحدار اللوجستي، إذا علمنا أن العدد الأصلي لمشاهدات النوع الأول في العينة الأصلية كانت (36) وبناءً على ما جاء في الجدول المذكور آنفاً نلاحظ أنه تم تصنيف (9) مشاهدات فقط بشكل خاطئ الى النوع الثاني أي تم تصنيفها الى النوع الثاني علماً أنها في الحقيقية تتبع النوع الأول وكما مر بنا سابقاً في الجدول (4-7) وكذلك يمكننا من الجدول السابق تحديد المشاهدات التي تم تصنيفها بشكل خاطئ الى النوع الثاني وهي كل من المشاهدات (6,23,62,66,83,122,130,183,213) .

ثانياً : تصنيف مشاهدات النوع الثاني

الجدول (4-9)

تصنيف مشاهدات النوع الثاني على وفق طريقة أُنموذج الإنحدار اللوجستي LRM

Observations	Classification	Observations	Classification
4	type2	112	type2
5	type2	114	type2
8	type1	115	type2
13	type2	116	type2
14	type2	117	type2
16	type2	118	type2
17	type2	119	type2
18	type2	120	type2
19	type2	121	type2
20	type2	123	type2
21	type2	125	type2
22	type2	126	type2
25	type2	127	type2

27	type2	129	type2
28	type2	132	type2
29	type2	133	type2
30	type2	134	type2
31	type2	135	type2
32	type2	136	type2
33	type2	137	type2
34	type1	138	type2
35	type2	139	type2
36	type2	140	type2
38	type2	141	type2
39	type2	142	type2
40	type2	143	type2
41	type2	144	type2
42	type2	145	type2
43	type2	146	type2
45	type2	147	type2
46	type2	148	type2
47	type2	150	type2
48	type2	151	type2
49	type2	152	type2
50	type2	153	type2
51	type2	154	type2

52	type2	155	type2
53	type2	156	type2
54	type2	157	type2
55	type2	158	type2
56	type2	159	type2
57	type2	160	type2
58	type2	161	type2
59	type2	162	type2
60	type2	163	type2
61	type2	164	type2
63	type2	165	type2
64	type2	166	type2
65	type2	167	type2
67	type2	168	type2
68	type2	169	type2
69	type2	170	type2
70	type2	171	type2
71	type2	172	type2
72	type2	173	type2
73	type2	174	type2
74	type2	175	type2
75	type2	176	type2
76	type2	177	type2

79	type2	180	type2
80	type2	181	type2
81	type2	182	type2
82	type2	184	type2
84	type2	185	type2
85	type2	187	type2
86	type2	188	type2
87	type2	189	type2
88	type2	190	type2
89	type2	191	type2
91	type2	192	type2
92	type2	193	type2
93	type2	194	type2
94	type2	195	type2
95	type2	196	type2
96	type2	197	type2
97	type2	199	type2
98	type2	201	type2
99	type2	203	type2
100	type2	204	type2
101	type2	205	type1
102	type2	206	type2
103	type2	207	type2

104	type2	208	type2
105	type2	209	type2
106	type2	210	type2
107	type2	211	type2
108	type2	212	type2
109	type2	214	type2
110	type2	215	type2
111	type2	216	type2

إعداد الباحث بالإعتماد على نتائج برنامج (R– language)

ويظهر الجدول المذكور آنفاً نتيجة عملية تصنيف مشاهدات النوع الثاني بإستخدام LRM إذ يظهر المشاهدات التي تم تصنيفها الى النوع الثاني بشكل صحيح من قبل LRM والمشاهدات التي تم تصنيفها بشكل خاطئ الى النوع الأول إذ أن العمود الأول والثالث يمثلان المشاهدات والعمود الثاني والرابع يمثلان المخرجات أو عملية التصنيف التي تم تصنيفها الى النوع الثاني أو الأول وبإستخدام LRM إذا علمنا أن العدد الأصلي لمشاهدات النوع الثاني في العينة الأصلية كانت (180) وبناءً على ما جاء في الجدول السابق نلاحظ أنه تم تصنيف (3) مشاهدات فقط بشكل خاطئ أي تم تصنيفها الى النوع الأول علماً أنها في الحقيقية تتبع النوع الثاني وكما لاحظنا ذلك في الجدول (4–7) وكذلك يمكن تحديد المشاهدات التي تم تصنيفها بشكل خاطئ الى النوع الأول من الجدول

المذكور آنفاً وهي كل من المشاهدات (8,34,205) .

4–5–2 تقدير المعلمات :

هنا تم إيجاد قيم المعلمات المقدرة ومعنوية المتغيرات لأُنموذج الإنحدار اللوجستي عن طريق دالة الإمكان الأعظم وكما تم ذكره في الفصل النظري بالمعادلة (32.2) وكما يلي :

جدول (4-10)

نتائج تقدير معلمات أُنموذج الإنحدار اللوجستي LRM

	Estimate	Std. Error	z value	Pr(>|z |)
Intercept	14.03611	7.95916	1.764	0.07781
X_1	0.39944	0.08195	4.874	0.00000109 ***
X_2	-2.15916	0.99600	-2.168	0.03017*
X_3	0.01429	0.01958	0.730	0.46542
X_4	-0.17556	0.05673	-3.094	0.00197**
X_5	1.13805	0.50728	2.243	0.02487*

إعداد الباحث بالإعتماد على نتائج برنامج (R- language)

وتظهر النتائج في الجدول المذكور آنفاً معنوية المتغيرات x_1 الذي يمثل متغير العمر ، والمتغير x_4 الذي يمثل متغير الطول ، والمتغير x_5 الذي يمثل متغير إستجابة المريض لنوع العلاج . ومتغير x_2 الذي يمثل متغير الجنس إذ أن قيمها كانت أقل من مستوى الدلالة 0.05 . إذ كان متغير العمر في المرتبة الأولى من حيث المعنوية وجاء متغير الطول ثانياً ومن ثم متغير نوع العلاج ومتغير االجنس في المرتبة الثالثة في حين كانت قيمة المتغير x_3 الذي يمثل متغير الوزن غير معنوية إذ كانت 0.46542 وهي أقل من مستوى الدلالة 0.05 وبذلك يكون المتغير x_3 غير معنوي . نحصل من الجدول السابق على قيم المقدرات وبذلك تكون معادلة النموذج كالآتي :

$$Y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\beta_5x_5$$

$$Y=14.03611+0.39944\ x_1-2.15916\ x_2+0.01429\ x_3-0.17556\ x_4+1.13805x_5$$

4-5-3 الإختبارات في أُنموذج الإنحدار اللوجستي

4-5-3-1 إختبار والد Wald.Test

أولاً : إختبار والد Wald.Test لكل متغير على حده : إذ تم الحصول على النتائج التالية :

جدول (4-11)

نتائج إختبار والد wald.test الذي يتبع توزيع مربع كاي χ^2 وبدرجة حرية df=1

Variable	Chi-squred	Df	Sig.
X_1 (age)	3.1	1	0.078
X_2 (sex)	23.8	1	0.0000011
X_3(Wieght)	4.7	1	0.03
X_4(hight)	0.53	1	0.47
X_5(therapy)	9.6	1	0.002

إعداد الباحث بالإعتماد على نتائج برنامج (R- language)

ويظهر الجدول المذكور آنفاً نتائج إختبار wald.test الذي نلاحظ من خلاله قيم كل من المتغيرات x_2 وهو متغير الجنس ، والمتغير x_3 وهو متغير الوزن ، ومتغير x_5 وهو متغير استجابة المريض لنوع العلاج إذ إن قيمها جميعاً أقل من مستوى المعنوية 0.05 وبذلك نرفض فرضية العدم أي أن المعلمات $B_2.B_3,B_5$ معنوية ولاتساوي صفراً في المجتمع الذي سحبت منه العينة ، في حين أظهرت قيمة إختبار wald.test للمتغير x_1 الذي يمثل متغير العمر مستوى أكبر من مستوى الدلالة إذ كانت قيمة المتغير السابق 0.078 وهي أكبر من مستوى الدلاله 0.05 أي أن اختبار wald.test لهذا المتغير غير معنوي وكذلك المتغير x_4 الذي يمثل متغير الطول إذ أن إختبار wald.test له غير معنوي لأن قيمته كانت تساوي 0.47 .

ثانياً : إختبار wald.Test لجميع متغيرات النموذج :

أما إختبار wald.test لجميع متغيرات النموذج فكانت نتيجته كما في الجدول الآتي :-

جدول (4-12)

نتيجة إختبار والد wald.test لجميع متغيرات النموذج

Variable	Chi-squred	df	Sig.
All .Var	24	5	0.00021

إعداد الباحث بالإعتماد على نتائج برنامج (R- language)

إذ أظهرت نتيجة إختبار wald.test لجميع متغيرات النموذج من خلال الجدول المذكور آنفاً معنوية المتغيرات مجتمعة إذ بلغت قيمتها المعنوية 0.00021 وهي أقل من 0.05 مما يعني معنوية إختبار wald.test للمتغيرات الكلية .

4-5-3-2 إختبار H&L

إذ تم الحصول على نتائج هوزمر – ليمشو وكما مبينة في الجدول التالي :

جدول (4-13)

نتيجة إختبار هوزمر – ليمشو H&L

Chi-squared	df	Sig
2.596	8	0.9571

إعداد الباحث بالإعتماد على نتائج برنامج (R- language)

يوضح الجدول المذكور آنفاً نتائج إختبار Hosmer and lemshow Test إذ يبين أن إحصاءة H & L التي تتبع توزيع مربع كاي χ^2 تساوي (2.596) وهي غير معنوية لأن sig=0.9571 وهي أكبر من مستوى الدلالة 0.05 لذا نقبل فرضية العدم القائلة بعدم وجود فرق معنوي بين القيم المشاهدة والقيم المتوقعة .

4-5-3-3 إختبارات الجودة $R^2_{cox\&snell}$ و $R^2_{Nagelkerke}$

أما اختبارات جودة النموذج فكانت النتائج كالتالي :

جدول (4-14)

تفسير العلاقة بين متغير الإستجابة والمتغيرات التوضيحية لأنموذج الإنحدار اللوجستي

$R^2_{cox\&snell}$	$R^2_{Nagelkerke}$
0.6358233	0.7077500

إعداد الباحث بالإعتماد على نتائج برنامج (R- language)

يوضح الجدول السابق قيم ($R^2_{cox\&snell}$) و ($R^2_{Nagelkerke}$) إذ كانت قيمة ($R^2_{cox\&snell}$) تساوي 0.6358233 وهذا يشير إلى أن (63%) من التغير في المتغير التابع يتم تفسيره من خلال LRM ، أما قيمة ($R^2_{Nagelkerke}$= 0.7077500) وهذا يشير إلى أن (70%) من التغير في المتغير التابع يتم تفسيره من خلال LRM .

4-5- 4 نسبة الأرجحية Odds Ratio

تم حساب نسب الأرجحية لمتغيرات النموذج وكما في الجدول التالي :

جدول (4-15)

نسب الأرجحية للمتغيرات

Variables	0dds Ratio
Intercept	0.00001
X_1	4.6
X_2	2.5
X_3	6
X_4	0.8
X_5	3.1

إعداد الباحث بالإعتماد على نتائج برنامج (R- language)

يوضح الجدول المذكور آنفاً أن قيم المتغيرات تشير إلى مقدار التغير الحاصل في نسبة أرجحية وقوع الحدث وهو أن يكون المريض مصاب بالنوع الثاني إذ كانت قيمته =1 بينما النوع الأول = 0 (كما تم توضيح ذلك في بداية تعريف المتغيرات في الفصل العملي) عند حدوث تغير في قيمة المتغير التوضيحي (المتغير التفسيري أو المستقل) المرتبط بالمعلمة (B) . ويمكن توضيح ذلك كالتالي :

بالنسبة للمتغير الثاني (X_2 متغير الجنس)

$Exp(X_2)$= 2.5

وهذا يعني أن أرجحية ان يكون الشخص المصاب بداء السكري أُنثى بمقدار (2.5) مرة فيما لو كان المصاب ذكراً .

أما بالنسبة للمتغير الثالث (X_3 متغير الوزن)

$Exp(X_3)= 6$

فيعني عند زيادة وزن المريض بمقدار وحدة واحدة فإن احتمالية أن يكون المريض من النوع الثاني تزداد بمقدار (6) مرات .

أما بالنسبة للمتغير الخامس (X_5 متغير استجابة المريض لنوع العلاج)

$Exp(X_5)= 3.1$

فبالنسبة للمرضى الذين يتناولون الأنسولين فقط كعلاج فإن أرجحية ان يكون المريض من النوع الأول بمقدار (3.1) فيما لو كان من النوع الثاني ، وهو مايدل على أهمية هذا العامل حيث إن إستجابة المرضى من النوع الثاني تكون للحبوب وأحياناً للأنسولين أو الإثنين معاً . في حين تكون استجابة المرضى من النوع الأول للأنسولين وقد يأخذ الحبوب أحياناً لغرض السيطرة على وزنه . ويعد هذا العامل هو العامل الرئيس في تصنيف نوعي مرض داء السكري ومن بعده يأتي العمر والوزن .

4-6 مقارنة النموذجين :

من خلال نتائج التصنيف التي تم الحصول عليها من كلا الطريقتين آلة المتجه الداعم (Support Vector Machine) والإنحدار اللوجستي (Logistic Regression) لوحظ أن آلة المتجه الداعم كانت تعطي نتائج تصنيف أكثر دقة من دالة الإنحدار اللوجستي إذ كانت نتيجة التصنيف بطريقة آلة المتجه الداعم 96 % ونتيجة التصنيف بطريقة الإنحدار اللوجستي كانت 94 % . على الرغم من أن نتيجة التصنيف مقاربة إلا أن آلة المتجه الداعم كانت أفضل ، وكما هو موضح في الجدول الآتي :-

جدول (4-16)

المقارنة بين طريقة آلة المتجه الداعم "SVM" وأُنموذج الإنحدار اللوجستي "LRM"

الطريقة	دقة التصنيف
آلة المتجه الداعم	96%
الإنحدار اللوجستي	94%

إعداد الباحث بالإعتماد على نتائج برنامج (R- language)

وأن السبب في إن النتيجة كانت مقاربة هو أن التداخل لم يكن قوياً جداً في مشاهدات النوع الثاني من المرض بينما كان التداخل قوياً في النوع الأول .

أما عند المقارنة بين نتائج التطبيق العملي والجانب التجريبي عند مستوى (μ =0.7) فإن النتيجة تبدو متقاربة نوعاً ما ، أما عند المستويات الأُخرى فإن SVM تتفوق على LRM بنسب عالية.

في الجانب التجريبي يمكن إيجاز نسبة التعرف الصحيح في الجانب التجريبي كالآتي :

1- في حالة التباين ($\sigma^2 = 1$) ولحجم عينة كبيرة (n=216) وإختيار μ=0.7 لأنها تشبه من خلال الرسم الى حد كبير الرسم في العينة الأصلية في الجانب التطبيقي ، كانت نسبة التعرف الصحيح على وفق المعطيات السابقة في الجانب التجريبي لـ SVM 96% بينما كانت 95% لـ LRM وهي نتيجة مقاربة أيضاً لما تم الحصول عليه في الجانب العملي .

2- في حالة التباين ($\sigma^2 = 1.25$) ولحجم عينة كبيرة (n=216) و μ=0.7 أيضاً كانت نسبة التعرف الصحيح لـ SVM 92% مقابل 90% لـ LRM .

3- في حالة التباين ($\sigma^2 = 1.5$) ولحجم عينة كبيرة (n=216) و μ=0.7 كانت نسبة التعرف الصحيح لـ SVM 89% مقابل 86% لـ LRM .

يلاحظ من النقاط الثلاث الأخيرة بأن تجربة المحاكاة قد أعطت نتائج مقاربة للنتيجة التي تم الحصول عليها في الجانب التطبيقي .

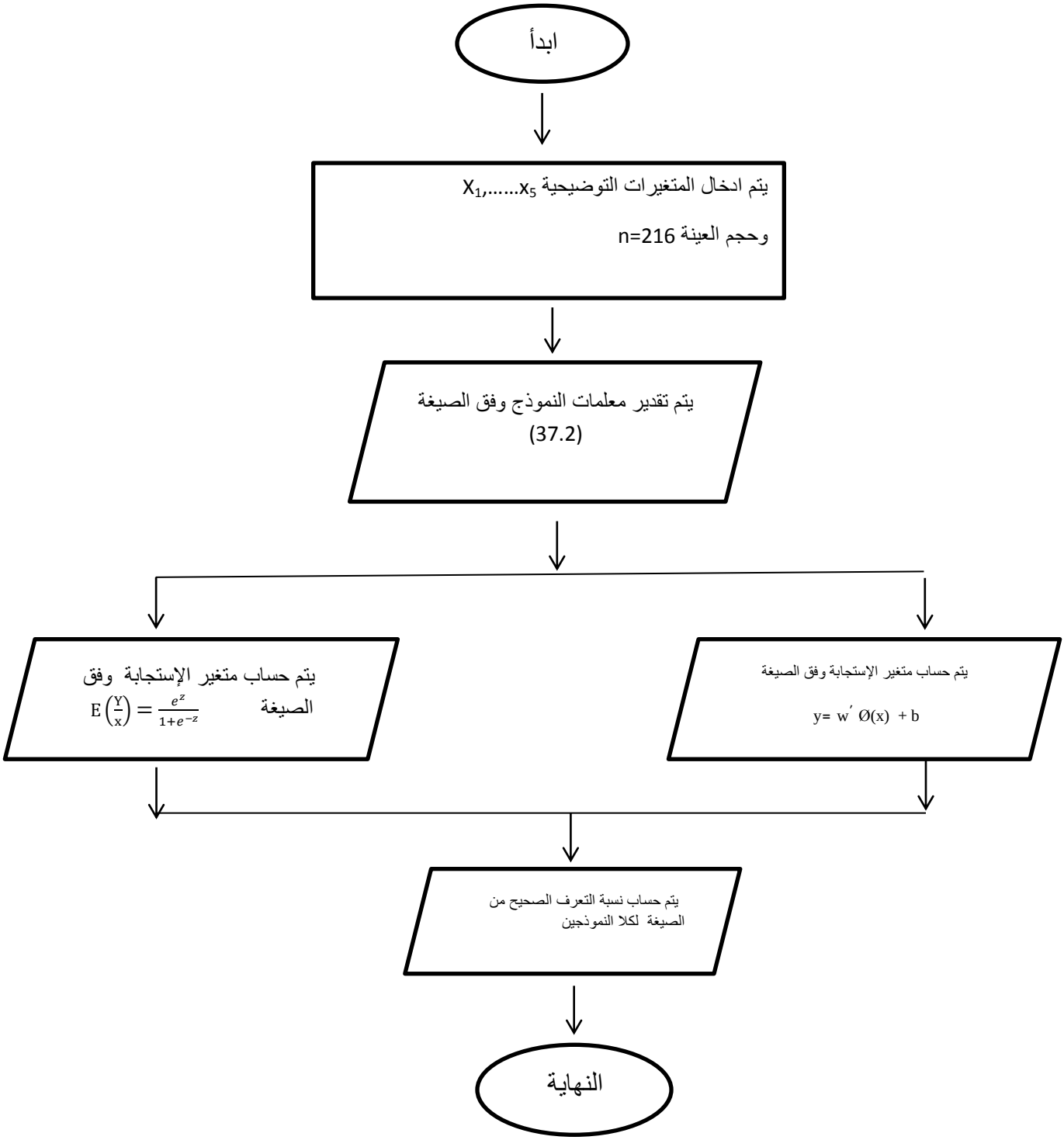
ويستنتج من ذلك أن سبب إنخفاض الدقة في الحالتين الأخيرتين هو بسبب زيادة التداخل بين المشاهدات .

أما في الجانب العملي فقد كانت نسبة التعرف الصحيح لـ SVM ولحجم عينة (n=216) 96% و LRM 94% .

ولتوضيح خطوات البرنامج المستخدم للمقارنة بين طرائق التصنيف وللنموذجين المستعملين يتم اتباع المخطط الإنسيابي الآتي :

شكل (4-1)

المخطط الإنسيابي لعملية التصنيف وفق نموذجي SVM و LRM



الفصل الخامس

الإستنتاجات والتوصيات

(Conclusions & Recommendations)

5-1 : الإستنتاجات : لقد أُستنتج من الدراسة بعض الأبعاد التي ظهرت من التطبيق العملي والتجريبي والتي تساند نظم الآليات المطبقة .

أولاً: أهم الإستنتاجات التي توصل اليها الباحث في الفصل التجريبي (المحاكاة) :

1- لوحظ تفوق SVM على LRM في نسبه التعرف الصحيح عند زيادة حجم العينة .

2- لوحظ تفوق SVM على LRM عند زيادة التباين .

3- لوحظ تفوق SVM على LRM عند زيادة التداخل بين البيانات .

4- لوحظ زيادة الفرق في لطريقة SVM على LRM كلما قل الوسط الحسابي للبيانات .

5- اثبتت النتائج جودة SVM في تصنيف البيانات قياساً الى طريقة LRM في جميع مستويات التباين ولجميع حجوم عينات الدراسة حيث وصل الفارق بين الطريقتين من % 14الى % 18 لصالح طريقة SVM .

ثانياً: أهم الإستنتاجات التي توصل اليها الباحث في الفصل التطبيقي العملي :

1- اظهرت آلة المتجه الداعم تفوقاً واضحاً في تصنيف النوع الأول لمرضى داء السكري بنسبة تعرف صحيح بلغت 86% مقابل % 75 لدالة الإنحدار اللوجستي ، في حين إن كلا الطريقتين قد صنف النوع الثاني لمرضى داء السكري بنسبة % 98 .وذلك بسبب التداخل القليل في متغيرات النوع الثاني في حين التداخل الكبير بين متغيرات النوع الأول هو الذي جعل آلة المتجه الداعم تتفوق على الإنحدار اللوجستي .

2- إن متغير إستجابة المريض لنوع العلاج (X_5) كانت أهم المتغيرات في تحديد إنتماء المريض لنوعي المرض، ويأتي بعده متغير العمر (X_1) ومتغير الوزن (X_3) .

5-2 التوصيات :

يمكن الآن التوصية بعدد من المؤشرات نتيجة الدراسة العملية والتجريبية لإستخدامات المخططين والباحثين وإعطاء أدلة أساسية لأهمية آلة المتجه الداعم .

1- الإستفادة من أسلوب SVM والتوسيع في استخدامه في مجال الدراسات المختلفة(الطبية وغيرها) وذلك لكفاءته ومرونته .

2- يقترح الباحث على الباحثين في مجال الاحصاء وطلبة الدراسات العليا الذين يستخدمون التصنيف في دراساتهم القيام بمزيد من البحث والدراسة في مجال استخدام SVM .

3- يوصي الباحث بضرورة إستخدام SVM عوضاً عن طريقة الانحدار اللوجستي لثبوت كفائتها.

4- إجراء دراسات متقدمة لتقدير أعداد الإصابات الجديدة بمرض داء السكري على مستوى البصرة والعراق لمعرفة إتجاهات إنتشار هذا المرض ووضع استراتيجيات مناسبة لمواجهة مثل هذه الحالات المرضية مستقبلاً .

5- يجب على القائمين في المكتب المركزي للإحصاء أن يعمدوا إلى استخدام الأساليب الحديثة والمتقدمة في الدراسات المستقبلية .

المصـــادر

(References)

أولاً : المصادر العربية

1- ابو شوكان ، محمد و عدلي ، ابراهيم (2014) ، " إستخدام نموذج الإنحدار اللوجستي الثنائي في تفسير المتغيرات التابعة ثنائية القيمة في ميدان الأنشطة البدنية والرياضية "،مجلة علوم وممارسة الأنشطة البدنية الرياضية والفنية رقم 06 (2/2014) ص ص (1-10) .

2- البياتي ، هبة ابراهيم صالح (2005) ، "تحليل المسار في إنموذج الإنحدار اللوجستي مع تطبيق عملي" ، بحث مستل من رسالة ماجستير بالعنوان ذاته ، مجلة الإدارة والإقتصاد ، العدد (70) ص ص (175-194)

3- بابطين ، عادل بن أحمد بن حسن (2010) ، "الإنحدار اللوجستي وكيفية إستخدامه في بناء نماذج التنبؤ للبيانات ذات المتغيرات التابعة ثنائية القيمة " ، أُطروحة دكتوراه تخصص احصاء وبحوث 1430هـ جامعة أم القرى كلية التربية قسم علم النفس المملكة العربية السعودية ص ص 62-66 .

4- رضا، صباح منفي وآخرون (2017) ، " مقارنة بين أُنموذج الإنحدار اللوجستي وانموذج الإنحدار الخطي المميز الخطي بإستعمال المركبات الرئيسية لبيانات البطالة لمحافظة بغداد " ، مجلة العلوم الإدارية ، العدد (95) ، المجلد (23) ، ص ص 367-386 .

5- سعيد، رشا عادل (2015) ، " إستخدام إنموذج الإنحدار اللوجستي في دراسة العوامل المساعدة على تشخيص حالات الإصابة بسرطان المثانة " ، مجلة العلوم الإقتصادية والإدارية المجلد (21) ، العدد (83) . ص ص 344-347 .

6- عباس ، علي خضير(2012) ، " إستخدام الإنحدار اللوجستي في التنبؤ بالدوال ذات المتغيرات الإقتصادية التابعة النوعية " ، مجلة كركوك للعلوم الإدارية والإقتصادية ، المجلد الثاني ، العدد (2) , ص ص 237-238 .

7- عبد الكريم ، أنوار ضياء (2006)، " إستخدام الطرائق التميزية الإحصائية لتشخيص بعض أمراض القلب " ، مجلة جامع كركوك الدراسات العلمية المجلد (1) العدد (2) .

8- عنبر، جنان عبد الله (2010)،"مقارنة بعض طرائق التقدير اللامعلمية لنموذج الإنحدار التجميعي المجزأ بإستعمال المحاكاة مع التطبيق "، رسالة ماجستير مقدمة الى مجلس كلية الإدارة والإقتصاد جامعة بغداد .ص 40 .

9- غانم ، عدنان و الجاعوني ، فريد خليل (2011) ، "إستخدام تقنية الإنموذج اللوجستي الثنائي الإستجابة في دراسة المحددات الإقتصادية والإجتماعية لكفاية دخل الأُسرة (دراسة تطبيقية على عينة عشوائية من الأُسر في محافظة دمشق)" ، مجلة جامعة دمشق للعلوم الإقتصادية والقانونية ، المجلد السابع والعشرون ، العدد (1) , ص ص 123-124 .

10- قاسم ، بهاء عبد الرزاق (2011) ، تحليل أثر بعض المتغيرات في الإصابة بمرض اللثة بإستخدام أُنموذج الإنحدار اللوجستي ، مجلة العلوم الإقتصادية المجلد السابع ، العدد (27) : ص ص 143-144 .

11- محمد ، بشير فيصل (2010) ، "بعض الطرائق المعلمية واللامعلمية لتقدير دالة المعولية مع تطبيق عملي "، رسالة ماجستير مقدمة الى مجلس كلية الإدارة والإقتصاد جامعة بغداد ، ص 42 .

12- نوري، أحمد سامي و عزيز ، سماح فخري (2011) ، " التقصي حول الكشف عن الإخفاء في الصور الملونة " ، مجلة الرافدين لعلوم الحاسبات والرياضيات ، المجلد (8) العدد (2) ص ص 151-167 .

ثانياً : المصادر الأجنبية

13- Acevedo, M. A., Corrada-Bravo, C. J., Corrada-Bravo, H., Villanueva-Rivera, L. J., & Aide, T. M. (2009). Automated classification of bird and amphibian calls using machine learning: A comparison of methods. Ecological Informatics, 4(4),pp 206-214.

14- Ahmad, A. R., Khalid, M., & Yusof, R. (2002). Machine Learning Using Support Vector Machines. Centre Artificial Intelligence and Robotics ,p3.

15- Ali,I.,& Sumaya Saad,S.A. Using One-Class SVM with Spam Classification. Iraqi Journal of Science, , Vol. 57, No.1B, pp. 501-506.

16- Attanayake, A. M. C. H. (2017). Modelling the risk for type 2 diabetes using logistic regression approach,thesis Master in science in business statistics , University of Moratuwa Sri Lanka.pp2-11.

17- Balasubramanian, P. (2014). Automated Classification of EEG Signals Using Component Analysis and Support Vector Machines.pp1-5

18- Cai, C. Z., Wang, W. L., Sun, L. Z., & Chen, Y. Z. (2003). Protein function classification via support vector machine approach. Mathematical biosciences, 185(2),pp. 111-122.

19- Cai, Y. D., Liu, X. J., Xu, X. B., & Chou, K. C. (2002). Support vector machines for predicting the specificity of GalNAc-transferase. Peptides, 23(1),pp. 205-208.

20- Campbell, W. M., Campbell, J. P., Reynolds, D. A., Singer, E., & Torres-Carrasquillo, P. A. (2006). Support vector machines for speaker and language recognition. Computer Speech & Language, 20(2),pp. 210-229.

21- Chinaei, L. (2007). Active Learning with Semi-Supervised Support Vector Machines (Master's thesis, University of Waterloo),pp.10-16.

22- Cramer, J. S. (2002). The origins of logistic regression.pp.3-7.

23- Elhabil, A., & Eljazzar, M. (2014). A comparative study between linear discriminant analysis and multinomial logistic regression. An-Najah university journal research, 28,pp. 1526-1528.

24- Ferrer, A. J. A., & Wang, L. (1999). Comparing the Classification Accuracy among Nonparametric, Parametric Discriminant Analysis and Logistic Regression Methods, pp.5-6

25- Fletcher, T. (2009). Support vector machines explained. University College London, London, pp.2-9

26- Ge, S., Gao, Y., & Wang, R. (2007, August). Least significant bit steganography detection with machine learning techniques. In Proceedings of the 2007 international workshop on Domain driven data mining, pp. 24-32.

27- Geboyts R (2000). « Examples : Binary Logistic Regression » pp.1-2.

28- Guo, Q., Kelly, M., & Graham, C. H. (2005). Support vector machines for predicting distribution of Sudden Oak Death in California. Ecological Modelling, 182(1), pp.80-81.

29- Hosmer, D. W., & Lemeshow, S. (2000). Applied Logistic Regression. 2nd edition. –published by Wiley Series in Probability and Statistics -New York, pp.11-16

https://legacy.wlu.ca/documents/45781/logist.pdf

30- Ivanciuc, O. (2007). Applications of support vector machines in chemistry. Reviews in computational chemistry, 23,p 291.

31- Kester, D. L., Linton, T. H., & Sullivan, L. R. (2002). A Comparison of the Relative Practical Value of a Predictive Discriminant Function Analysis and a Binary Logistic Regression Analysis of Student Success in an Innovative Alternative High School Program in South Texas pp3-13.

32- Kumari, V. A., & Chitra, R. (2013). Classification of diabetes disease using support vector machine. International Journal of Engineering Research and Applications, 3(2), pp.1797-1801.

33- Li, H., Sun, J., & Wu, J. (2010). Predicting business failure using classification and regression tree: An empirical comparison with popular classical statistical methods and top classification mining methods. Expert Systems with Applications, 37(8), pp.5895-5904.

34- Machan, C. M. (2012). Type 2 diabetes mellitus and the prevalence of age-related cataract in a clinic population(Master's thesis, University of Waterloo).pp 3-5 ,

35- Morariu, D., & VINTAN, L. N. (2005). Classification and Clustering using SVM (Doctoral dissertation, Ph. D Report, University of Sibiu).pp.12-15 .

36- Ngassa Piotie, P. (2015). Diabetic nephropathy in a tertiary clinic in South Africa, a cross-sectional study (Doctoral dissertation).p2 .

37- Niyikora,S(2015) , Multiple logistic regression modeling on risk factors of diabetes. Case study of Gitwe Hospital (2011-2013).,thesis Master p 3.

38- Omar, Ibrahim Bashi (2014), "Face Recognation Based On PCA , LBP and SVM Techniques", Eng . & Tech.Journal , Vol(33), part(B),no.(3),pp 384-392 .

39- Qi, Z., Tian, Y., & Shi, Y. (2012). Laplacian twin support vector machine for semi-supervised classification. Neural Networks, 35, pp.46-53.

40- Rumpf, T., Mahlein, A. K., Steiner, U., Oerke, E. C., Dehne, H. W., & Plümer, L. (2010). Early detection and classification of plant diseases with Support Vector Machines based on hyperspectral reflectance. Computers and Electronics in Agriculture, 74(1),pp. 91-99.

41- Shrivastava, N. K., Saurabh, P., & Verma, B. (2011). An efficient approach parallel support vector machine for classification of diabetes dataset. J. Computer Applications, 36(6),pp. 19-24.

42- Stella Appiah(2012), "Multiple Logistic Regression Analysis To Determine Risk Factors For The Clinical Diagnosis Of Diabetes ", Thesis Of Masters Of Philosophy , University Of Science And Technology , Kumasi ,Gana Sylvere Niyikora.p5 .

43- Takeuchi, K., & Collier, N. (2005). Bio-medical entity extraction using support vector machines. Artificial Intelligence in Medicine, 33(2),pp. 125-137.

44- Tamrakar, P. (2014). Prevalence of gestational diabetes mellitus and its associated risk indicators: A hospital based study in Nepal (Msc. thesis).p 2

45- Taylor, M. J. (2013). Risk factors for diabetes mellitus: A comparative analysis of subpopulation differences in a large Canadian sample.p3.

46- Thair A saleh, Mustafa Zuhaer nayef,(2012),"Discrimination Analysis and Support vector Machine ",Eng. & Tech.Journal , Vol(31), part(A),no.(12) , pp.2261-2272

47- Wuensch, K. L. (2014). Binary logistic regression with SPSS. Retrieved March, 18, 2015.pp 2-4

48- Yao, X. J., Panaye, A., Doucet, J. P., Zhang, R. S., Chen, H. F., Liu, M. C., ... & Fan, B. T. (2004). Comparative study of QSAR/QSPR correlations using support vector machines, radial basis function neural networks, and multiple linear regression. Journal of chemical information and computer sciences, 44(4),pp. 1257-1266.

49- Yu, C. N. (2011). Improved learning of structural support vector machines: training with latent variables and nonlinear kernels.p 5 .

ثالثاً : المواقع الألكترونية

50- http//www.annabaa.orgarabichealth5875 موقع شبكة النبأ

51- http//ar.wikipedia.org موقع ويكيبيديا الموسوعة الحرة

52- https//www.webteb.com › موقع ويب طب

53- http//www.mawdoo3.com › موقع موضوع

موقع مقال منظمة الصحة العالمية

الملاحق

(Supplements)

ملحق رقم (1) عينة عشوائية للمصابين بالنوع الأول والثاني وحسب العمر والجنس والوزن والطول ونوع العلاج الذي يأخذه المريض

ت	العمر	الجنس	الوزن بالكيلو غرام	الطول بالسنتيمتر	نوع العلاج	النوع
1	14	ذكر	34	154	انسولين	الاول
2	11	انثى	43	160	انسولين	الاول
3	35	انثى	68	194	انسولين	الاول
4	50	ذكر	36	141	حبوب	الثاني
5	48	ذكر	55	170	حبوب	الثاني
6	54	انثى	74	196	حبوب	الاول
7	23	انثى	47.5	157	انسولين	الاول
8	23	انثى	45.5	149	انسولين	الثاني
9	20	ذكر	63	174	انسولين	الاول
10	17	ذكر	64	175	انسولين	الاول
11	16	انثى	49	152	حبوب	الاول
12	21	ذكر	65	175	انسولين	الاول
13	45	انثى	48	150	حبوب	الثاني
14	54	انثى	48	150	حبوب	الثاني
15	19	انثى	50	153	انسولين	الاول
16	43	انثى	55	160	انسولين وحبوب	الثاني
17	59	انثى	55	160	حبوب	الثاني
18	60	انثى	46.5	147	انسولين وحبوب	الثاني
19	40	انثى	58	162	حبوب	الثاني
20	54	انثى	57	160	حبوب	الثاني
21	50	انثى	52	152	حبوب	الثاني
22	41	انثى	53	153	انسولين	الثاني
23	28	انثى	53	153	انسولين وحبوب	الاول
24	26	ذكر	66	169	انسولين وحبوب	الاول
25	40	انثى	56	155	انسولين وحبوب	الثاني
26	16	انثى	55	153	انسولين	الاول
27	40	انثى	60	159	حبوب	الثاني
28	55	انثى	57	155	حبوب	الثاني
29	68	ذكر	72	174	حبوب	الثاني
30	50	انثى	62	161	حبوب	الثاني

31	43	ذكر	66	166	حبوب	الثاني
32	35	انثى	57	154	حبوب	الثاني
33	60	انثى	62	160	انسولين	الثاني
34	34	ذكر	70	170	انسولين	الثاني
35	55	ذكر	66	165	انسولين	الثاني
36	64	ذكر	66	165	حبوب	الثاني
37	16	انثى	63	161	انسولين	الاول
38	56	انثى	55	150	حبوب	الثاني
39	37	ذكر	65	163	حبوب	الثاني
40	55	انثى	62	158	حبوب	الثاني
41	42	ذكر	71	169	حبوب	الثاني
42	50	انثى	60	155	انسولين	الثاني
43	38	ذكر	69	166	حبوب	الثاني
44	34	ذكر	75	173	انسولين	الاول
45	50	انثى	58	152	انسولين وحبوب	الثاني
46	55	ذكر	85	184	حبوب	الثاني
47	49	ذكر	77	175	حبوب	الثاني
48	51	ذكر	66	162	حبوب	الثاني
49	39	انثى	59	153	حبوب	الثاني
50	40	ذكر	91	189	حبوب	الثاني
51	52	ذكر	74	170	حبوب	الثاني
52	50	انثى	60	153	انسولين وحبوب	الثاني
53	75	انثى	61	154	انسولين وحبوب	الثاني
54	41	انثى	66	160	حبوب	الثاني
55	50	انثى	73	168	حبوب	الثاني
56	39	ذكر	93	189	حبوب	الثاني
57	45	انثى	65	158	حبوب	الثاني
58	52	ذكر	78	173	حبوب	الثاني
59	48	ذكر	80	175	حبوب	الثاني
60	36	ذكر	77	171	حبوب	الثاني
61	47	انثى	73	166	حبوب	الثاني
62	29	انثى	65	156	انسولين وحبوب	الاول
63	45	انثى	66	157	حبوب	الثاني

64	40	انثى	66	157	حبوب	الثاني
65	42	انثى	69	160	حبوب	الثاني
66	27	انثى	61	150	انسولين وحبوب	الاول
67	34	انثى	66	156	حبوب	الثاني
68	38	ذكر	73	164	حبوب	الثاني
69	52	ذكر	89	181	انسولين وحبوب	الثاني
70	42	ذكر	74	165	حبوب	الثاني
71	33	ذكر	77	168	حبوب	الثاني
72	38	ذكر	79	170	حبوب	الثاني
73	45	انثى	89	180	حبوب	الثاني
74	44	ذكر	90	181	حبوب	الثاني
75	35	انثى	58	145	حبوب	الثاني
76	57	ذكر	82	172	انسولين	الثاني
77	22	انثى	74	163	انسولين وحبوب	الاول
78	30	انثى	77	166	انسولين	الاول
79	34	ذكر	78	167	انسولين وحبوب	الثاني
80	40	ذكر	79	168	حبوب	الثاني
81	30	ذكر	81	170	انسولين وحبوب	الثاني
82	50	انثى	55	140	حبوب	الثاني
83	36	ذكر	86	175	انسولين	الاول
84	42	انثى	67.5	155	حبوب	الثاني
85	43	انثى	65	152	انسولين	الثاني
86	50	انثى	61	147	حبوب	الثاني
87	35	انثى	68	155	حبوب	الثاني
88	45	ذكر	87	175	حبوب	الثاني
89	52	ذكر	87	175	حبوب	الثاني
90	16	انثى	64	150	انسولين	الاول
91	50	انثى	73	160	حبوب	الثاني
92	66	ذكر	75	162	حبوب	الثاني
93	42	انثى	76	163	انسولين	الثاني
94	42	انثى	62	147	حبوب	الثاني
95	42	انثى	71	157	انسولين وحبوب	الثاني
96	65	انثى	71	156	حبوب	الثاني

97	48	ذكر	72	158	حبوب	الثاني
98	35	انثى	72	158	انسولين وحبوب	الثاني
99	41	انثى	71	156	حبوب	الثاني
100	35	ذكر	87	172	انسولين وحبوب	الثاني
101	41	ذكر	87	172	حبوب	الثاني
102	44	انثى	68	152	حبوب	الثاني
103	42	انثى	68	152	حبوب	الثاني
104	45	ذكر	78.5	163	انسولين	الثاني
105	35	انثى	72	156	انسولين وحبوب	الثاني
106	60	انثى	74	158	انسولين وحبوب	الثاني
107	47	ذكر	99.5	183	حبوب	الثاني
108	47	انثى	71	154	حبوب	الثاني
109	45	انثى	75	158	انسولين وحبوب	الثاني
110	56	انثى	75	158	حبوب	الثاني
111	55	ذكر	77	160	انسولين	الثاني
112	50	انثى	77	160	حبوب	الثاني
113	35	انثى	80	163	انسولين	الاول
114	38	ذكر	100	182	حبوب	الثاني
115	50	انثى	62	143	حبوب	الثاني
116	35	انثى	64	145	حبوب	الثاني
117	40	انثى	77	159	انسولين	الثاني
118	38	ذكر	82	164	حبوب	الثاني
119	55	ذكر	93.5	175	حبوب	الثاني
120	53	انثى	70	151	حبوب	الثاني
121	36	ذكر	87	168	حبوب	الثاني
122	35	ذكر	84	165	انسولين	الاول
123	57	انثى	79	160	انسولين وحبوب	الثاني
124	24	ذكر	100	180	حبوب	الاول
125	42	ذكر	106	185	حبوب	الثاني
126	65	انثى	68	148	انسولين وحبوب	الثاني
127	39	ذكر	90	170	حبوب	الثاني
128	23	انثى	85	165	حبوب	الاول
129	42	انثى	84	164	انسولين وحبوب	الثاني

130	36	انثى	80	160	انسولين	الاول
131	25	انثى	82	162	انسولين	الاول
132	62	ذكر	86	165	حبوب	الثاني
133	38	انثى	77	156	حبوب	الثاني
134	45	ذكر	92	170	حبوب	الثاني
135	54	ذكر	90	168	انسولين وحبوب	الثاني
136	59	ذكر	88	166	حبوب	الثاني
137	61	ذكر	84	162	حبوب	الثاني
138	55	انثى	73	151	حبوب	الثاني
139	35	انثى	74	152	حبوب	الثاني
140	36	انثى	82	160	انسولين وحبوب	الثاني
141	45	انثى	77	155	حبوب	الثاني
142	54	انثى	87.5	165	انسولين	الثاني
143	51	انثى	84.5	162	حبوب	الثاني
144	43	ذكر	107	182	انسولين وحبوب	الثاني
145	36	انثى	88	165	انسولين وحبوب	الثاني
146	60	انثى	69	146	حبوب	الثاني
147	66	ذكر	96	172	انسولين وحبوب	الثاني
148	33	ذكر	80	157	حبوب	الثاني
149	24	ذكر	95	171	انسولين وحبوب	الاول
150	57	انثى	80	156	انسولين	الثاني
151	38	ذكر	102	176	حبوب	الثاني
152	48	ذكر	91	166	حبوب	الثاني
153	40	انثى	90	165	انسولين	الثاني
154	51	انثى	90	165	انسولين	الثاني
155	47	انثى	83	158	حبوب	الثاني
156	51	ذكر	95	169	حبوب	الثاني
157	50	انثى	87	161	حبوب	الثاني
158	39	ذكر	96	169	حبوب	الثاني
159	55	انثى	81	155	حبوب	الثاني
160	36	انثى	76	150	حبوب	الثاني
161	52	ذكر	92	165	حبوب	الثاني
162	50	ذكر	110	180	حبوب	الثاني

163	42	انثى	92.5	165	حبوب	الثاني
164	54	ذكر	103	174	حبوب	الثاني
165	56	ذكر	83	154	حبوب	الثاني
166	58	انثى	81	154	انسولين وحبوب	الثاني
167	52	انثى	79	152	حبوب	الثاني
168	61	ذكر	100	171	حبوب	الثاني
169	36	انثى	88	160	حبوب	الثاني
170	58	ذكر	87	159	حبوب	الثاني
171	35	انثى	86	158	انسولين وحبوب	الثاني
172	45	انثى	81	153	حبوب	الثاني
173	45	انثى	81	153	حبوب	الثاني
174	48	انثى	92	163	انسولين وحبوب	الثاني
175	46	انثى	92	162	انسولين وحبوب	الثاني
176	33	ذكر	102	170	حبوب	الثاني
177	57	انثى	82	152	حبوب	الثاني
178	34	انثى	91	160	انسولين	الاول
179	19	ذكر	91	160	انسولين	الاول
180	46	انثى	83	152	انسولين وحبوب	الثاني
181	45	انثى	80	149	حبوب	الثاني
182	50	انثى	90	158	حبوب	الثاني
183	28	ذكر	89	157	انسولين	الاول
184	65	انثى	84	152	حبوب	الثاني
185	63	انثى	82	149	حبوب	الثاني
186	25	انثى	94	159	حبوب	الاول
187	35	انثى	84	150	حبوب	الثاني
188	42	انثى	87	152	انسولين	الثاني
189	45	انثى	97	160	انسولين	الثاني
190	40	انثى	95	158	حبوب	الثاني
191	49	انثى	110	170	حبوب	الثاني
192	38	انثى	92	155	حبوب	الثاني
193	45	ذكر	107	167	انسولين وحبوب	الثاني
194	60	ذكر	112	169	حبوب	الثاني
195	47	انثى	94	154	انسولين	الثاني

196	42	ذكر	123	176	حبوب	الثاني
197	60	ذكر	85	146	حبوب	الثاني
198	26	انثى	100	158	انسولين وحبوب	الاول
199	50	انثى	94	153	انسولين وحبوب	الثاني
200	24	انثى	117	169	انسولين	الاول
201	35	انثى	115	167	حبوب	الثاني
202	20	انثى	106	158	حبوب	الاول
203	46	ذكر	120	166	انسولين وحبوب	الثاني
204	37	ذكر	134	174	حبوب	الثاني
205	25	انثى	114	160	حبوب	الثاني
206	55	انثى	96	146	حبوب	الثاني
207	47	انثى	96	146	حبوب	الثاني
208	72	انثى	95	144	حبوب	الثاني
209	60	ذكر	91	140	انسولين وحبوب	الثاني
210	68	انثى	98	145	حبوب	الثاني
211	60	انثى	117	155	انسولين	الثاني
212	35	انثى	130	163	حبوب	الثاني
213	36	ذكر	150	170	حبوب	الاول
214	62	انثى	97	131	حبوب	الثاني
215	65	انثى	150	161	حبوب	الثاني
216	34	ذكر	210	175	انسولين وحبوب	الثاني

المصدر : سجلات مركز الغدد الصم في مستشفى الموانئ العام في البصرة

ملحق رقم (2) برنامج العملي

```
data<-data.frame(x1,x2,x3,x4,x5,y)
data1<-data.frame(x1 ,y)
data2<-data[order(data$y),]
ncol(data)
plot(cmdscale(dist(data2[,-ncol(data2)])),
    col = as.integer(data2[,ncol(data)]),
    pch = c("o","o")[1:216 %in% data2$y + 1])

############### SVM ########################
```

```
fit.svm<-svm(y~.,data=data2,epsilon=0.1)
plot(data2[,5])
points(data[,6],col=2)
names(fit.svm)
fit.svm$x.scale
pred <- predict(fit.svm, data2)
(acc <- table(pred, data2$y))
classAgreement(acc)
fit.svm$sparse
svm.table<-addmargins(acc)
percentage correct for training data
svm.correct.dis<-sum(diag(acc)) / sum(acc)
########### Logistic Regression ############################
glmFit <- glm(y ~ .,data=data, family=binomial(link="logit"))
# predicted probabilities
Yhat <- fitted(glmFit)
# choose a threshold for dichotomizing according to predicted probability
thresh  <- 0.5
YhatFac <- cut(Yhat, breaks=c(-Inf, thresh, Inf), labels=c("type1", "type2"))
# contingency table and marginal sums
cTab <- table(YhatFac,y )
L.table<-addmargins(cTab)
# percentage correct for training data
L.correct.dis<-sum(diag(cTab)) / sum(cTab)
##############  Results ##################################
svm.table
svm.correct.dis
L.table
L.correct.dis
#####################################################################
W= t(fit.svm$coefs)%*%fit.svm$SV

fit.svm$rho

Library(aod)
```

```
Wald.test (b=coef(glmFit),Sigmavcov(glmFit),Terms=1:5)

Library(ResourceSelection)

Hoslem.test(glmFit$y,fitted(glmFit))

Library(BaylorEdPsych)

PseudoR2(glmFit)
```

ملحق رقم (3) برنامج المحاكاة

```
n=          ### يتم إختيار حجم العينة
d=          ### تحديد عدد المتغيرات
μ =         ### يتم تحديد المسافة بين المجموعتين

svm.table.i<-matrix(0,3,3)
svm.correct.dis.i<-0
L.table.i<-matrix(0,3,3)
L.correct.dis.i<-0
no.of.iteration<-1000
for(i in 1:no.of.iteration){
group.x1<-matrix(rnorm((d*n)/2, μ,1),n/2,d)+10
group.x2<-matrix(rnorm((d*n)/2,- μ,1),n/2,d)+10
group.x<-rbind(group.x1,group.x2)
colnames(group.x)=paste("x",1:d,sep="")
y1<-rep("type1",n/2)
y2<-rep("type2",n/2)
y<-c(y1,y2)
data<-data.frame(group.x,y)
data2<-data[order(data$y),]
plot(cmdscale(dist(data2[,-ncol(data2)])),
    col = as.integer(data2[,ncol(data)]),
    pch = c("o","o")[1:n %in% data2$y + 1],xlab="",ylab="")
############## SVM ######################
fit.svm<-svm(y~.,data=data2,epsilon=0.1)
```

```
#plot(data2[,5])
#points(data[,6],col=2)
#names(fit.svm)
#fit.svm$SV
pred <- predict(fit.svm)
acc <- table(pred, data2$y)
#classAgreement(acc)
#fit.svm$sparse
svm.table<-addmargins(acc)
# percentage correct for training data
svm.correct.dis<-sum(diag(acc)) / sum(acc)
########## Logistic Regression ########################
glmFit <- glm(y ~ .,data=data, family=binomial(link="logit"))
# predicted probabilities

Yhat <- fitted(glmFit)
# choose a threshold for dichotomizing according to predicted probability
YhatFac <- cut(Yhat, breaks=c(-Inf, thresh, Inf), labels=c("type1", "type2"))
# contingency table and marginal sums
cTab <- table(YhatFac,y )
L.table<-addmargins(cTab)
# percentage correct for training data
L.correct.dis<-sum(diag(cTab)) / sum(cTab)
############## Results ################################
svm.table
svm.correct.dis
L.table
L.correct.dis
svm.table.i<-svm.table.i+svm.table
svm.correct.dis.i<-svm.correct.dis.i+svm.correct.dis
L.table.i<-L.table.i+L.table
L.correct.dis.i<-L.correct.dis.i+L.correct.dis
} # end of iteration
svm.table.it<-svm.table.i/no.of.iteration
svm.correct.dis.it<-svm.correct.dis.i/no.of.iteration
L.table.it<-L.table.i/no.of.iteration
```

```
L.correct.dis.it<-L.correct.dis.i/no.of.iteration
round(svm.table.it,0)
round(svm.correct.dis.it,2)
round(L.table.it,0)
round(L.correct.dis.it,2)
```

“Abstract “

In this thesis, the process of classification or categorizing statistical data was studied by using the dual-response support vector machine and the logistic regression model based on the correct classification of observations for both methods, The simulation was used in the two methods with different samples and different variations and different arithmetic mean, then the comparison between the two methods, Then applied the two methods in the practical side and with real data for diabetes patients obtained from the Endocrine Center at the General Ports Hospital in Basra , Also the comparison was made between the two methods used in the study. Finally the study found that the supporting vector machine was the best in the classification whether using the real data in the practical side or by using the simulation on the experimental side with different samples of especially when data overlap.

Ministry of Higher Education and Scientific Research

University of Basrah

College of Administration and Economics

Department of Statistics

"Comparison Between The Support vector machine (SVM) and Logistic Regression Model For Classification With application to diabetic patients at Basrah port hospital "

A thesis submitted to the Council of Administration and Economics College , University of Basrah as a partial Fulfillment of the requirements for the degree of M.SC in Statistics

By

Ahmed Abd-Alsamad Habeeb Thamer ALGeBuri

Supervised by

Prof.

Fawzia Ghalip Umar AL-Sadoon

1439 2018