الجمهورية الجزائرية الديمقراطية الشعبية
وزارة التعليم العالي والبحث العلمي
جامعة الجزائر-02- أبو القاسم سعد الله
كلية اللغة العربية وآدابها واللغات الشرقية
قسم علوم اللسان

أطروحة مقدمة لنيل شهادة دكتوراه علوم
تخصص: العلاج الآلي للكلام
بعنوان:

التصنيف الآلي للنصوص الأدبية العربية

من اعداد الطالب : العربي بوعمران بوعلام

الموسم الجامعي 2018 - 2019

الجمهورية الجزائرية الديمقراطية الشعبية
وزارة التعليم العالي والبحث العلمي
جامعة الجزائر-02- أبو القاسم سعد الله
كلية اللغة العربية وآدابها واللغات الشرقية
قسم علوم اللسان

أطروحة مقدمة لنيل شهادة دكتوراه علوم
تخصص: العلاج الآلي للكلام
بعنوان:

التصنيف الآلي للنصوص الأدبية العربية

اعداد الطالب :
العربي بوعمران بوعلام

إشراف:
أ.د/ مراد عباس

الموسم الجامعي 2018 - 2019

الجمهورية الجزائرية الديمقراطية الشعبية

وزارة التعليم العالي و البحث العلمي

جامعة الجزائر (02)

كلية اللغة العربية وآدابها واللغات الشرقية

قسم علوم اللسان

أطروحة مقدمة لنيل درجة دكتوراه العلوم في علوم اللسان تخصص العلاج الآلي للكلام

بعنوان:

التصنيف الآلي للنصوص الأدبية العربية

إعداد الطالب: **إشراف الدكتور :**

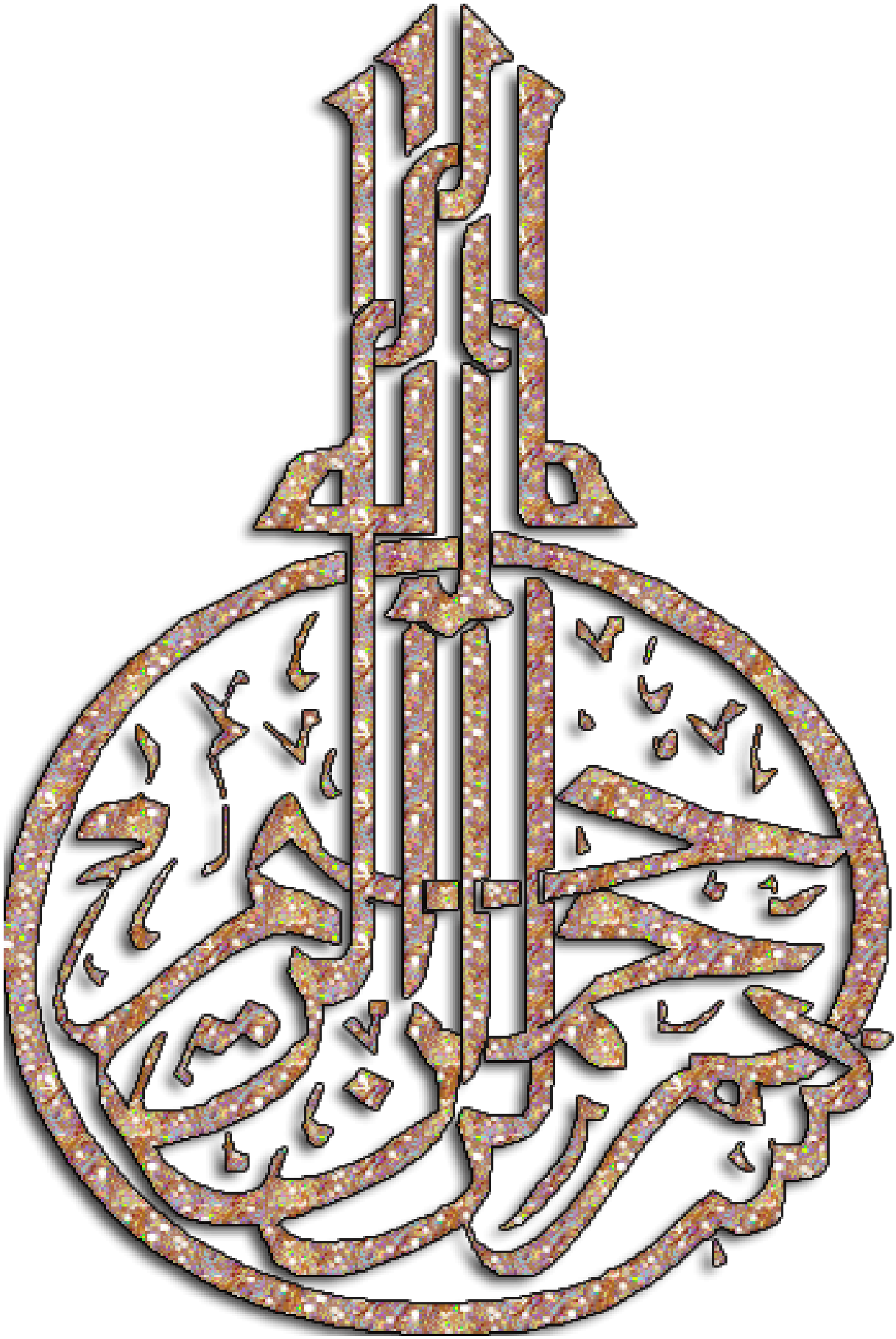
العربي بوعمران بوعلام **مراد عباس**

نوقشت وأجيزت علنا بتاريخ:2019/07/14

أمام اللجنة المكونة من السادة:

1- الاستاذ الدكتور: (عارف غريبي) جامعة الجزائر 02	**رئيسا**
2- الاستاذ الدكتور: (مراد عباس) CRSRDLA	**مشرفا ومقررا**
3- الاستاذ الدكتور: (حسينة عليان) CERIST	**عضوا مناقشا**
4- الاستاذ الدكتور: (كبير بن عيسى) CRSRDLA	**عضوا مناقشا**
5- الاستاذ الدكتور: (محمد قرومي) جامعة باب الزوار	**عضوا مناقشا**
6- الاستاذ الدكتور: (عيسو) المدرسة العليا للأساتذة القبة	**عضوا مناقشا**

الموسم الجامعي: 2018-2019



إهداء

الحمد لله الذي وفقنا لهذا ولم نكن لنصل لولا فضل الله علينا أما بعد:

أهدي هذا العمل المتواضع إلى التي حملتني وهنا على وهن وسهرت الليالي الى جانبي الى والدتي العزيزة ادعو لها بالشفاء العاجل.

الى الذي رباني صغيرا وشجعني كبيرا وكان أبا معينا، والدي الكريم أطال الله في عمره.

الى التي كانت لي سندا في انجاز هذا العمل، رفيقة دربي، زوجتي الكريمة.

إلى فلذة كبدي وقرة عيني ابني محمد أنس حفظه الله ورعاه.

إلى أفراد أسرتي إخوتي وأخواتي ولا احصي لهم فضلا .

إلى كل أقاربي وأصدقائي من دون استثناء أدامهم الله لي من المقربين.

شكر وتقدير

اشكر الله عز وجل على توفيقي في انجاز هذا البحث المتواضع
كما أتوجه بجزيل الشكر والامتنان إلى أستاذي الكريم " مراد عباس" الذي شجعني للبحث في هذا المجال، والذي لم يبخل عليا بتوجيهاته العلمية الصائبة، ونصائحه القيمة، أدامه الله في خدمة اللغة العربية.
وإلى الأساتذة الأفاضل أعضاء لجنة المناقشة لما تكبدوه من عناء تقييم هذه الدراسة فلهم مني أسمى عبارات التقدير والاحترام.
كما أشكر كل من ساعدني من قريب أو من بعيد على انجاز هذا العمل وفي تذليل ما واجهته من صعوبات.

المقدمة

المقدمة

أصبحنا في عصرنا هذا محاطين بكم هائل من المعلومات وهذا راجع إلى التدفق السريع للبيانات على نطاق واسع، بسبب التطور الحاصل وظهور أجهزة الاتصال الذكية، مما ولد الحاجة إلى ضرورة إدارة هذه المعلومات قصد تسهيل الوصول إليها، لذا كان لابد من إنتاج برمجيات فعالة تؤدي وظائفها على نحو مرض بتكلفة أقل وجهد محدود ومن بين هذه البرامج برنامج (WEKA) الذي يعمل على تصنيف المستندات النصية بالاعتماد على خوارزميات متطورة.

وكنتيجة لتراكم البيانات النصية باللغة العربية صار التصنيف الآلي هو أحد أكثر الحلول الفعالة من قبل الباحثين للوصول إلى المعلومة بسرعة وسهولة، لذا اخترت في هذا البحث تقنيتين مهمتين في علم التنقيب عن البيانات وهما التصنيف والتنبؤ، كمحاولة لتصنيف البيانات النصية الأدبية والتعرف على أساليبها الخبرية والإنشائية.

أهمية البحث :

لهذا البحث أهمية وقيمة علمية، كونه يتطرق إلى حقل معرفي مهم وهو التصنيف الآلي للنصوص بالاعتماد على برنامج جد متطور وجملة من الخوارزميات ذات الكفاءة العالية .

- كما يعطي هذا البحث صورة للدارسين في مجال اللغة العربية تصور عن كيفية تألية اللغة العربية وتطويع البرامج الغربية واعتماد الخوارزميات الذكية.
- هذا البحث يتوخى دعم الدراسات العربية التي تعنى بالتصنيف الآلي وفق خوارزميات التعليم والبرامج مفتوحة المصدر.
- يقدم البحث للدارسين في هذا المجال تصور حول تقنيات التنقيب في البيانات، وكيفية

تصنيف النصوص وطريقة العمل بالخوارزميات وكيفية تقييم أدائها ومعرفة كفاءتها.

أهداف البحث:

يقدم هذا البحث دراسة تطبيقية في مجال التصنيف الآلي للنصوص العربية قصد استخلاص الأساليب الخبرية والإنشائية وبناء نموذج تصنيفي تنبؤي للحالات الجديدة وأهداف أخرى يمكن عرضها فيما يلي:

- يهدف هذا البحث إلى تقديم أداة متطورة وفعالة تسمح للدارسين في مجال اللغة العربية باستنباط الأساليب الخبرية والإنشائية من كم هائل من البيانات النصية بشكل مرن وديناميكي.
- دراسة وتحليل آليات عمل برنامج weka في عملية التصنيف الآلي للنصوص العربية والتعرف على مجموعة من الخوارزميات الذكية التي يتيحها لنا.
- تطبيق أهم خوارزميات التصنيف الآلي واختيار الأنسب منها لتصنيف النصوص الأدبية العربية.
- اقتراح نموذج لاستكشاف المعرفة من البيانات النصية الأدبية العربية.

إشكاليات البحث:

يطرح هذا البحث مجموعة من التساؤلات نعرضها فيما يلي:

✓ إلى أي مدى يمكن تطبيق آليات التصنيف الآلي على النصوص الأدبية العربية ؟

- ✓ هل يمكن تطويع برامج التصنيف الآلي (غربية المنشأ) لتتناسب مع خصائص اللغة العربية؟
- ✓ ما هي أهم الفروق بين خوارزميات التصنيف الآلي؟
- ✓ ما هي الخوارزمية التي تمتاز بأداء جيد وكفاءة عالية في عملية التصنيف الآلي للنصوص الأدبية؟
- ✓ هل يمكن الاستفادة من خوارزميات التصنيف الآلي لبناء نموذج تنبؤي قادر على معرفة الأساليب الخبرية والإنشائية في النصوص العربية؟

خطة البحث:

للإجابة عن التساؤلات المطروحة في هذا البحث اتبعنا الخطة التالية:

- **مقدمة:** تناول المقدمة الإطار المنهجي لهذا البحث إذا وضحنا من خلالها الأهمية التي تنطوي عليها هذه الدراسة والأهداف المرجو تحقيقها، واهم التساؤلات المطروحة كما وضحنا المنهج المعتمد واهم الدراسات السابقة المعتمدة.
- **الفصل الأول:** يتم في هذا الفصل إعطاء نبذة موجزة عن علم التنقيب في البيانات كون أن التصنيف الآلي يعد إحدى التقنيات المهمة في هذا الحقل المعرفي، كما يتم تسليط الضوء على مختلف تقنيات التنقيب وعرض مجموعة من البرامج الآلية، كما يقدم هذا الفصل لمحة عن كيفية معالجة البيانات، ارتأيت كذلك أن أشير إلى فرع معرفي مهم وهو التنقيب في النصوص وذكر أهم التقنيات التي يتيحها لمعالجة المستندات النصية.
- **الفصل الثاني:** خصص هذا الفصل للحديث عن تقنية التصنيف الآلي للنصوص العربية، إذ حاولنا أن نلم بمختلف التعريفات لهذه الآلية وإعطاء نبذة عن النشأة وأسباب الظهور واهم

الخوارزميات التي تعتمدها، يتطرق هذا الفصل أيضا إلى المعالجة الآلية للغات الطبيعية وأهم معوقات التصنيف الآلي للمستندات النصية العربية.

- **الفصل الثالث:** ارتأينا أن يختص هذا الفصل بالحديث عن مختلف خوارزميات التصنيف الآلي بدء بإعطائها تعريفا محددا ونشأتها وكيفية العمل بها مع التمثيل لها، ومن أهم هذه الخوارزميات: Bayes Naïve ، أشجار القرار، J48 ، الشبكات العصبونية MLP ، خوارزمية KNN ، SVM ، خوارزميات العنقدة، انتقاء المعايير. يتطرق أيضا هذا الفصل إلى الحديث عن الأداة المستخدمة في هذه الدراسة وهي برنامج WEKA ، حاولنا إعطاء لمحة عن كيفية تصميم هذا البرنامج وشرح كيفية العمل به واهم الآليات والخوارزميات التي يتيحها والمزايا التي يتسم بها.

- **الفصل الرابع:** خصص هذا الفصل للجانب التطبيقي من البحث عنون بـ: التصنيف الآلي للنصوص العربية باستخدام برنامج (WEKA) اعتمدنا في هذا الفصل على تطبيق أربع خوارزميات يتيحها لنا هذا البرنامج على مدونة نصوص عربية أدبية، بغية تصنيفها وفق فئتين نصوص ذات أساليب خبرية وإنشائية الهدف من هذا التصنيف بناء نموذج تدريبي للتعرف على مختلف مؤشرات الأسلوبين، وكذا بناء نموذج تنبؤي يتعرف على مختلف الحالات الجديدة، قمنا باختبار هذه الخوارزميات وتقييم أداءها كما تم تمثيل مختلف النتائج المتوصل إليها.

- **خاتمة** انتهينا بخاتمة لهذا البحث والتي كانت عبارة عن مجموعة من النتائج المتوصل إليها، تعتبر بمثابة إجابات لمختلف التساؤلات المطروحة، كما ذيلنا هذه الخاتمة بمجموعة من التوصيات التي لابد من العمل عليها مستقبلا.

منهج البحث:

تقوم هذه الدراسة على ثلاثة مناهج:

1. **المنهج الوصفي التحليلي** المبني على الملاحظة والاستقراء والاستنباط في دارسة كيفية تصنيف النصوص الأدبية العربية بالاعتماد على أربع خوارزميات يتيحها البرنامج المعتمد في هذا البحث، واستنباط أهم النتائج المتوصل إليها والاحتساب الرياضي لأهم نسب المقاييس المتوصل إليها مثل مقياس كابا الإحصائي ومصفوفات التعارض....
2. **المنهج المقارن** للمقارنة بين الخوارزميات المعتمدة في البحث وتحديد الفارق بينها بالاعتماد على جملة من اختبارات الأداء ومقاييس احتساب الدقة، وأيضا التعرف على قدرة كل خوارزمية في تنبؤها بالحالات الجديدة من خلال النموذج التنبؤي.
3. **المنهج التاريخي** لدراسة تصنيف النصوص وعلم التنقيب في البيانات والتنقيب في النصوص وذلك عند تحديد النشأة والتطور واهم الأدوات والتطبيقات.

الدراسات السابقة والمعتمدة:

- دراسة **مراد عباس وآخرون (2011)** بعنوان(**تقييم طرق التعرف الموضوعي للنصوص العربية**) تهدف هذه الدراسة إلى عرض أهم طرق التعرف الموضوعي من بينها خوارزمية الجار الأقرب و TF-IDF كما تم عرض طريقة جديدة وهي مصنف الزناد الذي يعتمد على حساب الزنادات أو المعلومة المتبادلة المتوسطة لكل زوج من الكلمات، بالإضافة إلى ذلك قام الباحثون بمحاولة ربط هذه الطرق المستعملة بغية الحصول على نتائج جيدة بالاعتماد على مجموعة من الأساليب وهي تصويت الأغلبية، تصويت الأغلبية المحسن والترابط الخطي.

- دراسة **محمد سعيد الدسوقي (2014)** بعنوان **(تطبيق العنقدة المتعددة المستويات على نص القرآن الكريم)** تهدف هذه الدراسة إلى تطبيق تقنية العنقدة على مجموعة من النصوص القرآنية واستخراج التشابه بين الآيات اعتمادا على الكلمات وتصنيفها ضمن مجموعات ثم تم القيام بحساب التشابهات بين المجموعات وإنشاء شجرة تصنيف هرمية تعبر عن توزع الآيات القرآنية وتجميعها بحسب كلماتها.
- دراسة **خلوف وآخرون (2009)** بعنوان **(استخدام آليات التنقيب في المعطيات للمساعدة في اكتشاف عمليات الاحتيال في البيئة المصرفية)** تهدف هذه الدراسة إلى استعراض آلية جديدة لاكتشاف الاحتيال المصرفي من خلال مقارنة نسبة التحذيرات الخاطئة لكل من خوارزمية Bayes Naïve وخوارزمية SVM لمساعدة المصارف في اتخاذ القرار.

الصعوبات والقيود:

إن تطبيق هذه التقنيات على اللغة العربية يعتبر تحديا كبيرا بحد ذاته بدء بإدخال اللغة العربية لبرنامج ويكا ثم جمع البيانات النصية الهائلة وتنظيفها وتصميم نموذج تدريبي وآخر تنبؤي واختبار أداء مختلف الخوارزميات وتمثيل النتائج وتقييمها.

لذا علينا أن نجزم بالقول أن محاولة قولبة اللغة العربية في الحاسوب تعتبر من أهم المشاكل التي تعترض طريق تطبيق البرامج الآلية التي يتيحها التصنيف الآلي، والتي تعتمد على لغة الجافا، ولعل المعوق الأساسي هو أن معظم البرامج الآلية تبنى على خوارزميات هي في الأصل غربية المنشأ، بالإضافة إلى ما تتميز به اللغة العربية عن بقية اللغات في العديد من الخصائص.

الفصل الأول :

مدخل الى التنقيب في البيانات

تمهيد

مع تطور العلوم وتفرعها وظهور تكنولوجيات متعددة تميز عصرنا الحالي بتدفق كم هائل من المعلومات في شتى المجالات، مما أدى إلى صعوبة الوصول إلى المعرفة إذ لم تعد الطرق التقليدية قادرة على استخلاص المعلومات نظرا لضخامتها، هذه الأسباب وغيرها ولدت الحاجة إلى ايجاد تقنيات ذات فعالية تعمل على استكشاف المعرفة وإدارة هذه المعلومات، نتيجة لذلك ظهرت العديد من العلوم من بينها علم التنقيب في البيانات الذي عمل على ايجاد تقنيات وأدوات وخوارزميات تسهل إجراءات الوصول للمعلومات، من خلال كمية هائلة من البيانات بطريقة سهلة وسريعة ودقيقة جدا، كما تعمل على بناء تنبؤات مستقبلية لحل المشكلات.

1.1 تعريف التنقيب في البيانات:

يعتبر التنقيب عن البيانات علما قائما بذاته نظرا لمجالاته الواسعة ونتائجه المتقدمة والناجعة وأساليبه الدقيقة في التعامل مع المعلومات وقواعد البيانات، يختص هذا الحقل المعرفي بتنظيم المعلومات وتخزينها تنقيتها وتصنيفها والعمل على استرجاعها ومعالجتها بالاعتماد على برامج آلية متطورة.

هناك العديد من التعريفات التي أطلقت على هذا العلم وان كانت الدراسات التي تطرقت إليه شحيحة جدا، إذ نلمح قلة من الدارسين الذين يغامرون بالبحث في هذا الميدان إلا أصحاب التخصص الذين يولون أهمية بالغة للجانب التطبيقي فغض الطرف عن الجوانب التنظيرية، هذه الدراسات جلها أعطت تعريفات تصب في قالب واحد وهو معالجة المعلومات واسترجاعها نذكر بعض هذه التعريفات:

تنقيب البيانات $(DataMining)$ هي عملية تحليل البيانات بأنماط مختلفة من منظور مختلف ثم جمعها

وتلخيصها لإعطاء معلومات مفيدة للاستخدام أو هي عملية استخدام مجموعة متنوعة من أدوات تحليل البيانات لاكتشاف الأنماط والعلاقات بين البيانات المختلفة والتي يمكن استخدامها لجعل التنبؤات صحيح، وهي خليط من علم الإحصاء وعلم الرياضيات وبرمجيات الحاسب الآلي [1].

(هي عملية بحث محوسب ويدوي عن معرفة من البيانات دون فرضيات مسبقة عما يمكن أن تكون هذه المعرفة) [2].

وتعرف أيضا بأنها " الاستكشاف الآلي أو المؤتمت لأنماط شائعة ومخفية في قاعدة بيانات معينة " [3] .

قد يستخدم المصطلح بالتساوي مع مصطلح اكتشاف المعرفة knowledge discovery أي أنه عملية مرتبطة بتحليل البيانات من وجهات نظر مختلفة ودمجها في معلومات مفيدة، ومن ثم يمكن استخدام المعلومات في زيادة الإيرادات والتقليل من التكاليف أو الاثنين معا.

ولقد تم تصميم برامج خاصة بالتنقيب عن البيانات تتميز باحتوائها على مجموعة من الأدوات التحليل تستخدم في تحليل البيانات، حيث تمكن المستفيد من تحليل البيانات من أبعاد وزوايا مختلفة مع إمكانية تصنيفها وتلخيص العلاقة بينهم [4].

أخذ هذا المصطلح العديد من الترجمات في اللغة العربية من بينها التنقيب في البيانات، التنقيب في المعطيات، استنباط البيانات، التنقيب عن البيانات، تنقيب البيانات.

تجمع جل التعريفات على أن التنقيب في البيانات يتعامل مع الكم الهائل من المعلومات والبيانات الضخمة بالاعتماد على الحاسوب وبرامج الكترونية تعمل وفق خوارزميات رياضياتية.

[1] ايهاب عثمان، عبدالرحمن عثمان التنبؤات الانتخابات الامريكية كدراسة حالة الانتخابية باستخدام تنقيب البيانات aamosman7@gmail.com hobatonga@gmail,

[2] محاضرات في تنقيب البيانات، نبيل محمد لطف مصلي، جامعة المستقبل لعلوم الادارة و تكنولوجيا المعلومات

[3] قندليجي، عامر ابراهيم، عبد الستاؤ العلي، غسان العمري، المدخل إلى إدارة المعرفة، دار المسيرة، عمان، 2006،

[4] رحاب فايز احمد، التنقيب عن بيانات مؤسسات العمل التطوعي على الويب، مجلة كلية الآداب، جامعة بني سويف، المجلد الأول، العدد 27، 2013، ص 373.

ويتحدد هدف هذا العلم من خلال ايجاد حلول لبعض المشاكل فيما يخص التعامل مع المعلومات في شتى المجالات المعرفية والخروج بقرارات دقيقة وفعالة والتنبؤ المستقبلي بها، أصبح هذا العلم ضرورة حتمية في ظل التنامي المتسارع للمعلومات وصار لزاما على الدارس الأكاديمي وغيره من الباحثين أن يستعينوا بالأدوات والمنهجيات التي يتيحها علم التنقيب في البيانات.

2.1 نشأة علم التنقيب في البيانات:

بالنسبة إلى النشأة الأولى لتنقيب البيانات يمكن إرجاعها إلى خمسينيات القرن العشرين حيث ساهمت مجموعة من علماء الرياضيات والحاسوب بمجالات وضع مرتكزات الذكاء الصناعي إذ بدأ الاهتمام الفعلي بهذا المجال عام 1989 أثناء انعقاد ورشة عمل حول اكتشاف المعرفة في قواعد البيانات ومنذ ذلك الحين تم عقد هذه الورشة بصفة مستمرة سنويا حتى عام 1994 أما في 1995 أصبح المؤتمر الدولي لاكتشاف المعرفة و التنقيب عن البيانات من أهم الأحداث السنوية ومن ثم بدأ التخطيط العملي للتنقيب عن البيانات واكتشاف المعرفة من خلال كتابين " اكتشاف المعرفة في قواعد البيانات " [5] و " التقدم في اكتشاف المعرفة والتنقيب عن البيانات " [6].

لذا يعتبر علم التنقيب عن البيانات وليد القرن العشرين استطاع أن يثبت وجوده كأحد الحلول الناجحة لتحليل كميات ضخمة من البيانات وذلك بتحويلها من مجرد معلومات متراكمة وغير مفهومة إلى معلومات قيمة يمكن استغلالها و الاستفادة منها .

إذ أصبح علما قائما بذاته لنضوج مواضيعه مما يجعله العلم المسؤول عن أساليب وطرق إنتاج

[5] Piatetsky-Shapiro, G., and Frawley,W., (Eds) .(1991) Knowledge Discovery in Databases, AAAI/MIT Press, .1991

[6] Fayyad,U., Piatesky-Shapiro,G., Smyth, P., and Uthurusamy, R., (1996) Advances in Knowledge Discovery and Data Mining, AAAi/MIT Press, .1996

المعلومات وقواعد المعرفة من خلال كم كبير من البيانات التي يتم التنقيب عنها.

3.1 أهداف التنقيب في البيانات:

عادة ما نقصد بالتنقيب البحث في كميات ضخمة من البيانات من اجل الوصول إلى تحقيق غرض ما ونتلخص أهداف هذا العلم فيما يلي :

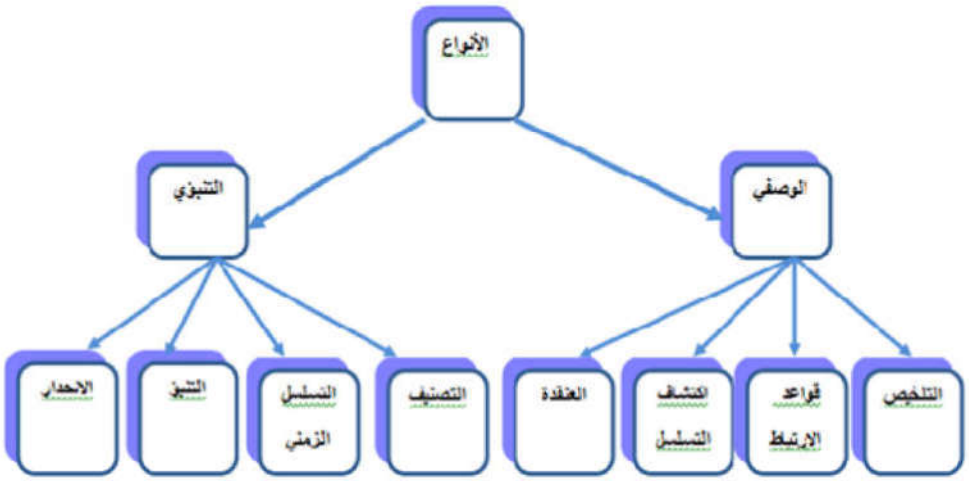
- من اجل تعليل بعض الظواهر المرئية
- من أجل التثبت من نظرية ما مثل التثبت من النظرية التي تقول بان الأسرة الكبيرة تهتم بالضمان الصحي أكثر من الأسر الصغيرة عددا.
- من اجل تحليل البيانات الحصول على علاقات جديدة وغير متوقعة مثل كيف سيكون الإنفاق العام إن كان ملازما لعملية خداع واسعة من قبل البطاقات الائتمانية [7].

4.1 تقنيات تنقيب البيانات:

توجد العديد من التقنيات التي تستخدم في عملية التنقيب عن البيانات، تختص كل تقنية بأداء أهداف معينة وذلك حسب طبيعة البيانات وحجمها تعمد أغلب هذه التقنيات على أدوات وخوارزميات متعددة من اجل الوصول إلى الهدف المحدد وعادة ما تكون هذه الأهداف لغرض تنبؤي أو وصفي حسب ما تقتضيه الدراسة وعليه يتحدد نموذجان لتنقيب البيانات هما التنقيب التنبؤي والوصفي فالتنقيب الوصفي في المعطيات يمكن أن يفسر كعملية للتنقيب في قاعدة المعطيات لاستخراج النماذج المخفية أو فرضيات للنماذج التي لم تحدد مسبقا أما التنقيب

[7] هالة حسن،: تعدين بينات التمويل الاصغر باستخدام تقنيات التصنيف والعنقدة اشراف طارق عبد الكريم، مذكرو ماجستير، تخصص تقانة المعلومات، كلية علوم الحاسوب وتقانة المعلومات. جامعة النيلين، مارس 2013. ص 42

التنبؤي في المعطيات فهو عملية استخدام النماذج المكتشفة لتوقع المستقبل من خلال النموذج التنبؤي [8].



شكل 1.1: تقنيات تنقيب البيانات

1.4.1 التنقيب التنبؤي:

تعتمد تنقيب البيانات في تنبؤاتها على ما يعرف بالنموذج التنبؤي وهو يرتكز على نموذج إحصائي يعتمد على عوامل متغيرة وهي غالبا ما تؤثر على السلوك المستقبلي والنتائج يعتمد هذا النموذج على النتائج المتوصل إليها ويتم بناء هذا النموذج بعدة طرق والتي تقوم على نظريات الإحصاء والمعادلات الرياضية [9] من اجل التنبؤ بقيم جديدة، أي أنه يستخدم نماذج محددة لتوقع ما سيحدث مستقبلا إذ يسعى هذا النموذج للتوصل إلى أحسن التنبؤات من خلال اعتماده على

[8] فادي خلوف، تطوير آليات جديدة للتنقيب في المعطيات لإدارة علاقات الزبائن في بيئة مصرفية، مجلة جامعة دمشق للعلوم الهندسية، مج:26، العدد 01/2010، ص 87.

[9] إيهاب عثمان ، عبد الرحمن عثمان، التنبؤات الانتخابية باستخدام تنقيب البيانات،سبتمبر 2016، https://www.researchgate.net/publication/307925413

مجموعة من التقنيات كالتصنيف والانحدار وتحليل السلاسل الزمنية.

- **التصنيف:** يعتبر التصنيف أكثر التقنيات استعمالا في وقتنا الحالي نظرا لأهميته في عملية إدارة المعلومات، تقوم هذه التقنية بمعالجة كم هائل من البيانات إذ تعمل على تصنيفها وفق فئات محددة مسبقا وذلك بالاعتماد على مجموعة من الخوارزميات.

- **الانحدار:** تعمل هذه التقنية على معالجة البيانات من أجل تحديد العلاقة التي تربط بين متغيرين أو أكثر وكذلك من أجل تقدير قيمة أحد المتغيرين وهو نموذج تنبؤي "يلجأ إليه المحللون عند الرغبة في تقييم العلاقة السببية بين أحد المتغيرات الكمية وعدة متغيرات أخرى ويطلق على المتغير الذي نريد التغيير فيه أو التنبؤ بقيمه في المستقبل عدة أسماء المتغير التابع، متغير الاستجابة أو المتغير المفسر، ويطلق على المتغيرات الأخرى بالمتغيرات المستقلة والمتنبئات" [10].

- **تحليل السلاسل الزمنية:** : تقوم هذه التقنية بمعالجة البيانات عبر أزمنة متعددة وملاحظة أهم التغيرات و الربط بينها وهذا يعني أن القيم المتوصل إليها تكون متعلقة ببيانات متغيرة عبر الزمن والهدف من ذلك هو " تحليل الأسباب والنتائج وتحديد الاتجاهات حتى يمكن استخدامها للتقدير والتنبؤ بالمستقبل" [11]

- **التنبؤ:** يعتبر التنبؤ تقنية مهمة في تنقيب البيانات وعادة ما تكون الهدف الرئيسي في عملية التنقيب وذلك من اجل تحديد قيم الحالات الجديدة أو البيانات المستقبلية بناءا على نماذج سابقة ومن الأدوات التقليدية المستخدمة في التنبؤ: الانحدارات بأنواعها والتحليل التمييزي

[10] عبد الحميد محمد العباسي : التنقيب في البيانات ، دراسات معهد الدراسات والبحوث الاحصائية، جامعة القاهرة، 2013.

[11] سيف الدين عثمان،الشفيع جعفر: التنقيب في البيانات واتخاذ القرارات، مجاة النيل الابيض للدراسات والبحوث ، العدد الثالث، مارس 2014، ص06

أما بالنسبة للأدوات الجديدة فنجد الخوارزميات، وأشجار القرار.

2.4.1 التنقيب الوصفي:

: هو نموذج يتم من خلاله معرفة الأنماط والعلاقات في البيانات يكمن الاختلاف بينه وبين النموذج السابق كون أن هذا النموذج يعمل على تحديد خصائص البيانات المعالجة ووصفها وليس لتحديد قيم حالات جديدة أو التنبؤ بها، ومن أهم أدواته:

- **قواعد الارتباط:** لدى هذه التقنية قدرة على تحليل كميات هائلة من البيانات وهي تسمح بمعرفة الصفات الموجودة اعتمادا على وجود صفات أخرى بمعنى آخر الكشف على مجموعة من القواعد المشتركة بين نسبة كبيرة من البيانات هذه التقنية تعطي لنا مجموعة من الحقول وكل حقل يحتوي على مجموعة من العناصر إذ أن أكثر العناصر ارتباطا مع العناصر الأخرى نعتبرها هي قاعدة الارتباط أو هي البحث بين المتغيرات عن علاقة تربطهم.
- **العنقدة:** تقوم هذه التقنية على وضع الحالات المتشابهة داخل نفس المجموعة بالاعتماد على صفات الحالات المتشابهة في حين توضع الحالات المختلفة في مجموعات منفصلة فالعنقدة هي التصنيف ضمن فئات غير محددة مسبقا إذ يقاس التشابه بواسطة تابع قياس المسافة، إذ يشتمل العنقود الواحد على مجموعة من الحالات المتشابهة والتي تختلف عن العنقود الآخر، بالإضافة إلى أن التشابه الكلي بين حالات العنقود الواحد يؤدي إلى أفضل عنقدة إذ أن تحديد مدى التشابه هو مقياس جودة العنقدة وتشمل تقنيات العنقدة: طريقة الجار الأقرب، الخرائط الذاتية التنظيم، نوع خاص من الشبكات العصبونية [12]

[12] فادي خلوف: تطوير أليات جديدة للتنقيب في المعطيات لادارة علاقات الزبائن في بيئة مصرفية، مجلة جامعة دمشق للعلوم الهندسية ،المجلد 26، العدد الاول، 2010، ص88.

- **اكتشاف التسلسل:** : تستخدم هذه التقنية لتحديد أنماط متسلسلة في البيانات وهذه الأنماط معتمدة على تسلسل زمن التأثيرات عادة ما تكون الحالات عبارة عن بيانات تشكل مجموعة متسلسلة " إذ أنها تستخدم التحليل القائم على الوقت لانتزاع معلومات مفيدة وهي مماثلة للتجميع في أنها تستخدم لتحديد العناصر التي تحدث معا لكن الأهم من ذلك أنها تستخدم لتحديد أي من العناصر يحدث أولا" [13] .

- **التلخيص:** تعمل هذه التقنية على تفتيت كل البيانات الكبيرة إلى مقاييس موجزة وتوفر وصفا عاما للمتغيرات وعلاقاتها ومن أمثلة أساليب التلخيص نذكر: المتوسطات المجاميع الإحصائيات الوصفية مثل المتوسط الحسابي والوسيط والمنوال ومقاييس التشتت مثل الانحراف المعياري [14].

5.1 مراحل عملية التنقيب في البيانات:

إن عملية التنقيب في البيانات ليس أمرا هينا بل تحتاج إلى كفاءة وخبرة وإلمام شامل بكل التقنيات والبرمجيات التي يتيحها هذا العلم وكذا معرفة الخطوات اللازمة التي بد من إتباعها أثناء الدراسة بغية الوصول إلى نتائج جيدة :

- قبل الشروع في العملية لا بد من تحديد مجال الدراسة والمشكلة المراد بحثها وإيجاد حلول لها.

- **جمع البيانات :** وهي مرحلة تجميع البيانات التي تمت بصلة للدراسة وذلك بشكل عشوائي

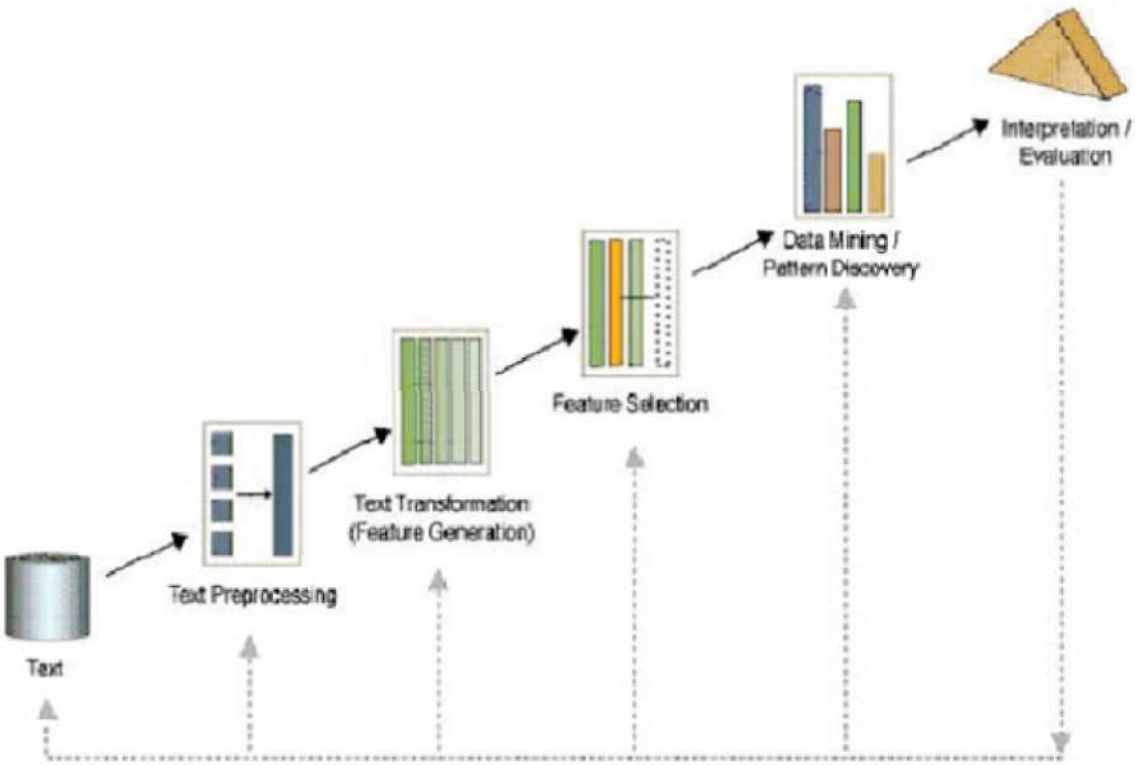
[13] سميرة محمد علي القدم، تطبيق تقنيات التنقيب في البيانات لتقييم أداء طلاب قيم الحاسوب ،كلية العلوم،مذكرة بكالوريوس، إشراف محمود حفص الدين/ قسم الحاسوب، كلية العلوم ،جامعة سبها،2018، ص 6,

[14] بشير عباس، العلاق: الإدارة الرقمية المجالات والتطبيقات، مركز الإمارات للدراسات والبحوث الاساراتيجية، أبو ظبي، 2005

والشروع في بناء قاعدة بيانات التنقيب.

- **اختيار البيانات Data Selection :** يتم في هذه المرحلة تعيين واختيار البيانات الملائمة من مجموع البيانات قصد استرجاعها.
- **تصفية البيانات وتنقيتها Data Cleaning :** يتم في هذه المرحلة حذف البيانات الزائدة التي لا تشكل أهمية أثناء الدراسة وتشتمل على التخلص من الحقول المتكررة ، إزالة البيانات المزعجة التي تعيق عملية التنقيب، تعيين البيانات غير المكتملة، تحديد الفراغات وإزالتها.
- **تحويل البيانات Data Transformation :** يتم في هذه المرحلة تحويل البيانات إلى صيغ أو نماذج تتلاءم مع إجراءات البحث قصد تحضيرها لعملية التنقيب.
- **التنقيب في البيانات Data Mining :** تعتبر هذه المرحلة الأهم حيث يتم فيها تنفيذ العمل وبناء النماذج أي اختيار النموذج الملائم لتمثيل البيانات المستكشفة باستخدام آليات ذكية والاستعانة بخوارزميات متطورة.
- **التقييم Pattern Evaluation :** يتم في هذه المرحلة تحديد النموذج النهائي وتطبيقه واستخراج النتائج.
- **تمثيل البيانات وتقديمها Knowledge Representatio :** وهي المرحلة الأخيرة من الدراسة نقوم في هذه المرحلة بتمثيل الخطوات المتبعة والنتائج المتوصل إليها بصور مرئية عن طريق استخدام التقنيات المصورةVisualization قصد تسهيل الفهم وتقريب المعلومة.

يمثل الشكل(2) المراحل المختلفة لتنقيب البيانات:



شكل 2.1: مراحل التنقيب في البيانات

6.1 تطبيقات تنقيب البيانات:

لقد أصبح التنقيب في البيانات علما قائما على أسس ممنهجة يعمل على إدارة مختلف البيانات لذا أصبح يتخلخل جميع الحقول المعرفية يمكن أن نجمل بخض هذه الميادين فيما يلي:

- **إدارة الأعمال التجارية:** كمجال التسويق والمبيعات: تحديد الزبائن والعملاء، الكشف عن حملات المبيعات وتفسيرها.

- معرفة عائدات الاستثمار والتنبؤ بها.

- الكشف عن حالات الاحتيال.
- تحديد سمات الاستفادة من القروض.

- **العلوم الطبية:**
  - تصنيف الحالات المرضية وتحديد أسبابها.
  - تحليل الوصفات الطبية لإرسال المواد الترويجية للزبائن المستهدفين.
  - إحصاء المرضى وتحديد الحالات.
  - تحديد العلاقات السببية بين الأمراض.
- **التصنيع:** مثل تحديد أسباب مشاكل التصنيع وبناء نماذج تنبؤية بها
- **اللغات الطبيعية:** ظهرت العديد من البحوث في هذا المجال كتصنيف الظواهر اللغوية آليا، أو التعرف الموضوعي، أو الترجمة الآلية.
- **علم المكتبات:** التصنيف الآلي للكتب والدوريات ومختلف المقتنيات والعمل على التعرف عليها.
- **مجال الاتصالات:** مثل تحديد مناطق تدفق شبكة الانترنيت وإحصاء الزبائن.
- **تطبيقات تكنولوجيا المعلومات:** يساعد التنقيب في البيانات في التأكد من جودة البيانات

7.1 أهمية التنقيب في البيانات:

- تعمل على بناء التنبؤات المستقبلية واستكشاف السلوك والاتجاهات مما يسمح بتقدير القرارات الصحيحة واتخاذها في الوقت المناسب.

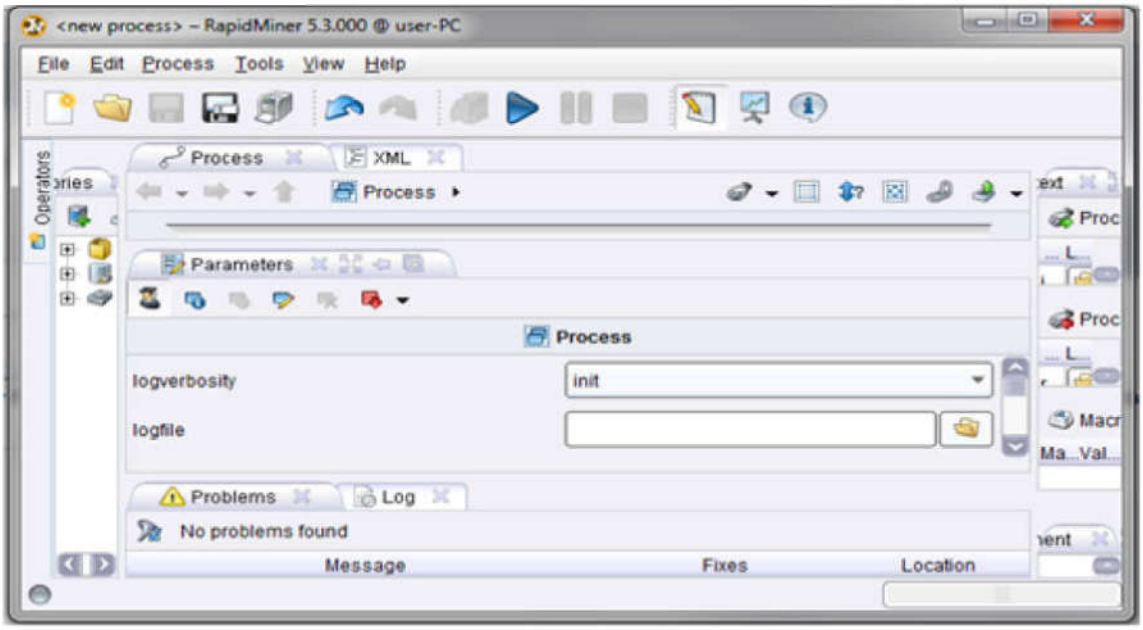
- معالجة كم هائل من البيانات في وقت قياسي وبدقة متناهية عكس تقنيات الإحصاء الكلاسيكية التي تستغرق وقتا طويلا مع عدم ضمان خلوها من الأخطاء.
- سهولة المشاركة في استخدام البيانات نظرا لشمولية البيانات وحفظها في مكان واحد.
- العمل على استخلاص المعلومات و القواعد من البيانات.
- تعميم النتائج المستخلصة من مجموعة من البيانات على كامل البيانات الأخرى.
- تعمل على إدارة البيانات وتنظيمها وتصغيرها من دون ضياع المعلومة وكذا تجنب التكرار غير اللازم للبيانات.
- تعمل على المساهمة في تحسين أداء المؤسسات لما لها من قدرة على استكشاف المعلومات الموجودة في قواعد البيانات.

8.1 برامج التنقيب في البيانات:

ظهرت العديد من البرامج والأدوات التي تقوم بعملية التنقيب في البيانات إذ تتيح مجموعة من الخوارزميات المتطورة التي تعمل على تحليل كميات ضخمة من البيانات بسرعة فائقة ومن هذه البرامج:

1.8.1 برنامج Rapidminer :

يعتبر من البرامج المجانية مفتوحة المصدر صمم من قبل شركة Germany Rapid-I يعمل بلغة الجافا، يتوفر هذا البرنامج على واجهة رسومية سهلة الاستخدام مقارنة ببرامج أخرى إذ لا يستلزم الأمر صعوبة في التعامل هذا البرنامج يتيح هذا البرنامج جملة من الخوارزميات المعروفة لمعالجة كميات ضخمة من البيانات.



شكل 3.1: الواجهة الرسومية لبرنامج Rapidminer

2.8.1 برنامج Clementine :

صمم هذا البرنامج من قبل شركة (SPSS) يتوفر هذا البرنامج على مكتبات كاملة لتنقيب البيانات بواسطة مختلف خوارزميات التصنيف والتحليل العنقودي وقواعد اكتشاف العلاقات والارتباطات يتصف هذا البرنامج بسهولة الاستخدام والتعلم.

3.8.1 برنامج WEKA :

يعتبر من البرامج المجانية مفتوحة المصدر، تم تصميم هذا البرنامج في جامعة ويكاتو بنيوزلندا جاء بهذا الاسم اختصارا لـ Wekato Envernoment for the Knoweledge Analysis يعمل بلغة الجافا، يتميز بقدرته على معالجة كمية هائلة من البيانات، يمدنا بمجموعة كاملة لمختلف الخوارزميات المعروفة في هذا المجال.

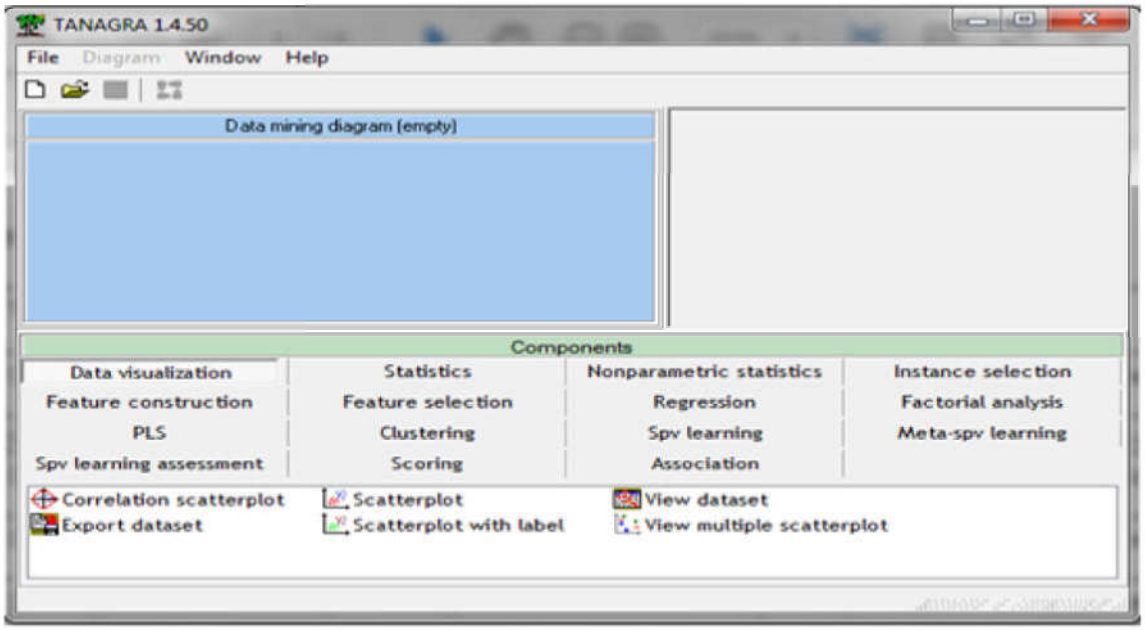
شكل 4.1: الواجهة الرسومية لبرنامج WEKA

4.8.1 برنامج Rattle :

يعتبر هذا البرنامج من البرامج مفتوحة المصدر صمم من قبل شركة Australia Togaware يعمل بلغة (R) تنفرد هذه الأداة بتضمينها حجم كبير من البيانات ما يأخذ على هذا البرنامج عدم مرونته في التعامل مع البيانات.

5.8.1 برنامج Tanagra :

يعتبر من البرامج مفتوحة المصدر صمم من قبل شركة Lumière University Lyon - France يعمل بلغة C++ سهل الاستخدام يتوفر على مجموعة من الخوارزميات إلا أن ما يعاب على هذا البرنامج هو أنها تعرض البيانات والنموذج بشكل ضعيف.



شكل 5.1: الواجهة الرسومية لبرنامج Tanagra

9.1 التنقيب في النصوص :

أولت البحوث في السنوات الأخيرة الكثير من الاهتمام لمعالجة البيانات النصية وهذا عائد لعدة أسباب من بينها تزايد مجموعة البيانات على شبكات التواصل وتطوير البنية التحتية للاتصالات والانترنيت مما أدى إلى الحاجة الماسة لتنظيم ومعالجة كميات ضخمة من البيانات إذ أن المعالجة اليدوية لهذه البيانات مكلفة للغاية في الوقت والأفراد كما أنها ليست مرنة وتعميمها إلى ميادين أخرى مستحيلة عمليا لذلك كان لابد من تطوير أساليب آلية.

إن الكميات الضخمة من البيانات المتوفرة حاليا هي عبارة عن بيانات نصية والتي تتألف بدورها من مجموعة كبيرة من المستندات من مصادر مختلفة مثل مقالات من صحف إخبارية، بحوث الكترونية، رسائل البريد الالكتروني، صفحات الويب إذ نجد هذه البيانات النصية تنمو بشكل كبير نظرا للتزايد الهائل من البيانات الرقمية على شبكات الانترنيت وبالتالي أصبح هناك

ثغرة واسعة بين البحث والمعالجة لاسترجاع معلومة من هذه المستندات وبين كمية البيانات الهائلة لذا تولدت الحاجة إلى التنقيب في البيانات النصية (النصوص) فظهر ما يسمى التنقيب في النصوص.

يعد التنقيب في النصوص أحد فروع المعالجة الآلية للغة وقد تزايد الاهتمام بت في الآونة الأخيرة نظرا لتزايد حجم البيانات ذات المحتوى النصي لذا ظهرت العديد من التقنيات والأدوات والخوارزميات التي تعمل على معالجة النصوص آليا منها الربط بين الكلمات والمقاطع في النصوص وتصنيف النصوص ضمن موضوعات محددة مسبقا.

10.1 تطبيقات التنقيب في النصوص :

- **المعالجة التلقائية لرسائل البريد الالكتروني:** ويشمل القيام بالفلترة الآلية للبريد الالكتروني غير المرغوب فيه بالاستناد إلى تحديد بعض المدخلات التي لايمكن أن نجدها في الرسائل المرغوب فيها بهذه الطريقة يمكن تفادي هذه الرسائل تلقائيا،
- **التصنيف التلقائي للمستندات النصية:** هي عملية التعرف الآلي على صنف أو موضوع المستند(رياضة، اقتصاد، أدب....) ويعتبر التعرف الموضوعي لنص ما العملية الآلية التي تتيح إرفاق الموضوع الصحيح الذي يتسم به ذلك النص أما التصنيف فهو تجميع النصوص التي تعالج موضوعا مشتركا في فئة واحدة ونظرا للتشابه الكبير بين هذين التعريفين فإنه يمكن استعمال طرق التصنيف بهدف التعرف الموضوعي والعكس صحيح. [15]
- **المعالجة الآلية لمحتويات صفحات الويب:** من خلال هذا التطبيق يمكن استكشاف المواقع الالكترونية للمستخدمين والتوغل للصفحات وإنشاء قائمة الملفات المتاحة في ذلك

[15] مراد عباس، تقييم طرق التعرف الموضوعي للنصوص العربية، مجلة الخليج العربي للبحوث العلمية، 29(3/4)، 2011، ص 184.

الموقع وبالتالي الحصول على معلومات قيمة عن المستخدمين بشكل آلي.

- **معالجة أوراق التأمينات وعقود البيع والمقابلات التشخيصية ومختلف الصيغ التجارية:** إدارة هذه المعلومات والعمل على تحليلها يقلص من إمكانية حدوث الأخطاء وتحديد مجموعة المشاكل والشكاوي الشائعة في هذا المجال.

11.1 التنقيب في النصوص:

يتم التعامل مع النصوص من خلال لأدوات وطرق خاصة لاستخلاص معلومات قيمة من النصوص المكتوبة كأسماء الأشخاص والشركات والتواريخ الموجودة في نص معين وبين وضع ملخص عن هذا النص أو تصنيف مجموعة من النصوص بحسب محددات معينة تستخلص منها.

تبدأ مراحل عملية التنقيب في النصوص بعملية استكشاف المعارف التي تسمح باستخلاص المحددات وهي المصطلحات والعناصر الهامة في النص والتي يمكن أن تفيد باعتبارها كلمات جوهرية حيث يتم تحليلها والربط بينها ودراسة توزعها في النص أو في مجموعة النصوص ضمن إطار التنقيب في النصوص للبحث عن معالم وظواهر وتوجهات.

- **استخراج المحددات:** تتضمن هذه الخطوة العديد من الإجراءات المهمة للتعامل مع النص:

- **التصفية:** هي العملية التي يتم فيها حذف الحروف الخاصة وعلامات الترقيم التي لا تعطي أي قوة تمييزية للوثيقة.
- **التقطيع:** هي عملية تجزيء الوثائق إلى كلمات.
- **التجذيع:** وهو عملية إعادة الكلمات إلى جذوعها باستخدام آلية للتحليل الصرفي

- **حذف الكلمات الزائدة:** ونقصد بها الكلمات التي لا تعطي معنى تمييزي للنص مثل الروابط بين الكلمات التي لا تحمل معنى في ذاتها بل تأخذ معناها بمجاورتها لكلمات أخرى.
- **حذف التشكيل:** يتم في هذه المرحلة حذف الحركات مثل الفتحة والضمة والسكون والتنوين.
- **التقليم:** هي عملية حذف الكلمات التي تظهر بتردد أقل أو أكثر في النصوص، فالكلمات الأقل ترددا تكون لنا عناقيد صغيرة غير مفيدة حتى وإن كان لديها قدرة تمييزية للنص، أما الكلمات الأكثر ترددا تعتبر كلمات غير تمييزية لأنها موجودة في معظم الوثائق.
- **حساب تردد الكلمات:** تجري هذه العملية بالاعتماد على طرق تحدد نسبة تردد الكلمة في الوثيقة الواحدة ونسبة ترددها في جميع الوثائق الأخرى وتسمى هذه الطريقة ب TFIDF [16].

12.1 أدوات تنقيب النصوص:

إن أية معالجة لنص لغوي تعتمد اعتمادا كليا على لغته لذا فمن الضروري أن تكون أدوات البحث قادرة على التعرف آليا على لغات الوثائق أو النصوص التي تتعامل معها لذا ظهرت العديد من التقنيات في هذا المجال.

[16] كادي زين الدين،خديم خديجة: التنقيب المعلوماتي ودوره في تحليل احتياجات مستعملي المكتبات ومراكز المعلومات،ص 149.

1.12.1 التجريد:

يستخدم التجريد في معالجة اللغات الطبيعية واسترجاع البيانات وتصنيف المستندات إذ يعمل على تقليل عدد المزايا وبالتالي تقليل حجم المستند وزيادة سرعة التعلم والتصنيف خاصة فيما يتعلق بالمصنفات التي تعتمد على مجموعة بيانات نصية ضخمة.
هناك أربع طرق مختلفة للتجريد في اللغة العربية وهي

- التجريد الخفيف الذي يعمل على إزالة السوابق واللواحق من دون المساس بالجذع آو تحديد الوزن آو استنتاج الجذر.
- التحليل المورفولوجي لإيجاد الجذور: تعمل هذه المحللات الصرفية على ايجاد الجذور المحتملة للكلمة استنادا إلى مجموعة من القواعد التحليلية.
- التجريد الإحصائي : يحاول أن يجمع مختلف تشكيلات الكلمة باستخدام تقنيات العنقدة التي تعتمد بشكل أساسي على مبدأ التشابه الصرفي بين الكلمات
- بناء قواميس بشكل يدوي: تعطي لنا معلومات عن التجريد وهي عبارة عن قواميس للجذور والأوزان [17]

2.12.1 مصنف الزناد:

تعرف مجموعة زنادات كلمة ما بأنها الكلمات التي لديها ترابط قوي بينها وبين تلك الكلمة ويتم حساب هذه الزنادات باستعمال المعلومة المتبادلة المتوسطة لكل زوج من الكلمات التي تنتمي لمجموعة المفردات والزنادات الأكثر أهمية تلك التي تملك قيم كبرى للمعلومة المتبادلة المتوسطة

[17] مصعب شاهين، شادي صالح:مشروع تخرج بعنوان: تطوير نظام لتصنيف المستندات العربية،اشراف الدكتورناصر ناصر، جامعة تشرين سوريا، قسم البرمجيات ونظم المعلومات، 2011/2012، ص 10.

الموافقة لكل زوج من الكلمات وبالتالي يصبح كل موضوع ممثلا بعدد من الزنادات المستخرجة من مدونة التدريب [18]

3.12.1 تقنية TF-IDF :

تقوم هذه التقنية على تمثيل كل وثيقة d بمتجه $D = (d1, d1, ...d|V|)$ حيث يرمز |V| إلى حجم مجموعة المفردات ويتم حساب مركبات المتجه عن طريق ضرب تكرار اللفظة $TF(w, d)$ الذي هو عبارة عن عدد المرات التي تظهر فيها اللفظة w في الوثيقة d بعكس تكرار الوثيقة IDF(w) ويمثل تكرار الوثيقة DF(w) عدد الوثائق التي تظهر فيها اللفظة w مرة واحدة على الأقل .

وتعرف القيمة di بوزن اللفظة wi في الوثيقة d وتعطى كالآتي:

$di = TF(w, d) * IDF(w)$ مع $IDF(w) = log(N/DF(w))$

و N هو عدد الوثائق.

ولحساب التشابه $sim(Dj, Di)$ التشابه الموجود بين الوثيقة Di والموضوع Dj تستعمل المعادلة وتنسب الوثيقة إلى الموضوع الذي يحصل على أكبر قيمة تشابه $sim(Dj, D)$

$$sim(Dj, Di) = \frac{\sum_{k=1}^{|v|} d_{jk} d_{ik}}{\sqrt{\sum_{k=1}^{|v|} (d_{jk})^2 \sum_{k=1}^{|v|} (d_i k)^2}}$$

4.12.1 تقنية ن. غرام:

يستخدم هذا المصطلح في العديد من المجالات مثل أنظمة التعرف الآلي على الكلام بقيم نموذجية لـ ن تساوي 3 آو 4 وتستخدم أيضا في أنظمة المعالجة الآلية للغة في ميدان البحث عن المعلومات ويمكن لـ ن. غرام أن تشير إلى حد سواء إلى ن. غرام حروف أو ن. غرام كلمات في هذا

[18] مراد عباس: تقييم طرق التعرف الموضوعي للنصوص العربية، مرجع سابق، ص185.

النموذج يتم تمثيل النصوص بمتجهات الـ ن. غرام عوضا عن متجهات الألفاظ عناصرها تواترا ت الـ ن. غرام في النصوص المطابقة لها.

و الـ ن. غرام هو سلسلة ألفاظ متتالية بالنسبة لأي نص مجموعة الـ ن. غرام التي يمكن استخراجها هي النتيجة التي نحصل عليها عن طريق تحريك خلال النص نافذة متكونة من ن مربعات وتتم هذه الحركة على مراحل مرحلة واحدة تقابل حرف واحد لكل ن. غرام حروف وكلمة لكل ن. غرام كلمات ثم نحسب ن. غرام المتحصل عليها [19].

اقتصرنا على ذكر بعض التقنيات إلا أن هناك العديد من الدراسات التي بحثت في هذا المجال ووضعت تقنيات جديدة تتعامل بدقة مع النصوص بناء على لغتها.

خلاصة:

ما نخلص إليه أن علم التنقيب في البيانات هو أحدى الطرق الحديثة والعلوم المتطورة التي استثمرت المنجزات التكنولوجية من برامج حاسوبية، وعملت على تطوير مختلف الخوارزميات الذكية لتنشأ أدوات ذات كفاءة عالية، تعمل على إدارة البيانات لاستكشاف المعرفة وتقديم أفضل ما يمكن الوصول إليه، وقد أصبح حاليا من بين الضروريات اللازمة في كل حقل معرفي .

[19] عبد المالك أمين، مقاربة لتحديد اللغات تلقائيا في مدينة نصوص متعددة اللغات، المجلة العربية الدولية للمعلوماتية، المجلد الثاني،العدد الرابع،2013، ص 37.

الفصل الثاني : دراسة فنية حول التصنيف الآلي للنصوص

تمهيد

إن التطورات التي عرفها العصر الحالي، على المستوى العلمي والتقني تشهد أن العالم يعرف ثورة علمية وتكنولوجية لها أبعادها الكبرى في شتى ميادين الحياة،وتقنية المعلومات أحد المحاور المهمة في هذا التطور إن لم تكن المحور الأساس للثورة العلمية المعاصرة. وقد اتسع مجال التقنية المعلوماتية ليشمل العديد من المجالات الحيوية، ومن ضمنها اللغة التي تعتبر الوسيلة الطبيعية التي يستخدمها الإنسان لاستمرار الحضارة،فإنها تمكنه من نقل المعلومات وتساعده على حفظها وتوارثها جيلا بعد جيل،وقد اعتمد التقدم في عصر المعلومات بشكل أساس على التحام اللغة بالحاسوب،تجلى هذا الالتحام في الثورة التي حدثت على مستوى التنظيم اللغوي الذي صاحبته تكنولوجيا متقدمة. فالتطور التكنولوجي الصناعي وأدوات البحث (العلمي) قربت المسافة بين الروحاني والجسماني، وبين الروحاني والمادة، وجعلت الآلة، وكأنها أصبحت إنسانا جديداً وكشفت على وجود كوجيطو صنعي يقف مع الكوجيطو الذاتي على قدم المساواة من حيث المعرفة، والقدرة على الكشف والاختراع [1].

كما يؤدي بنا النمو السريع لكمية المعلومات المتوفرة هذه الأيام إلى عصر المعلومات،وقد سبب التوسع في المعلومات بعض الصعوبات في البحث عن معلومات محددة. وتم تطوير العديد من الأدوات للمساعدة في إدارة المعلومات والتحكم بها خلال العقود القليلة الأخيرة،كما كان لابد من تطوير نظام لاسترجاع المعلومات التي تعمل وفق مجموعة من الأدوات التي تهتم بتمثيل وتخزين وترتيب والوصول إلى مواد المعلومات لذلك كان لا بد من إيجاد تقنيات وأساليب فعالة تعمل على تصنيف هذه الوثائق وتجميعها بطريقة منظمة، وقد ظهرت العديد من مناهج التصنيف الآلي

[1] سامي أدهم، الذكاء الاصطناعي، ثنائية الآلة والدماغ، مجلة كتابات معاصرة،ع 29-28، دجنبر 1996، يناير 1997، ص.35.

في الآونة الأخيرة التي تعتمد على خوارزميات التعلم الآلي وإن كانت عملية التصنيف معروفة منذ القديم يدويا إلىأن تطور الوسائل التكنولوجية تولدت عنها تقنية التصنيف الآلي للنصوص.

وتصنيف النصوص هو عبارة عن عملية تجميع آلي لمجموعة من الوثائق في صنف أو عدة أصناف، وفقا لمعايير مختلفة كمحتواها النصي ونوع الوثيقة، ويعتبر التصنيف الآلي للنصوص من أكثر العمليات التي لاقت تطبيقا واسعا في الآونة الأخيرة ويرجع ذلك أساسا إلى التزايد الكبير والهائل للوثائق الرقمية المتاحة خاصة على الشبكات الالكترونية لذلك كان لا بد من تصنيفها بطريقة منظمة.

وقد أصبح حاليا من بين الميادين الأكثر فعالية ونشاطا ولاقى اهتماما واسعا من قبل المتخصصين و الدارسين بمختلف المجالات العلمية، كما أن اعتماد التشغيل الآلي كان تحديا للتطور العلمي الحاصل في المجتمع بشكل متسارع وهائل خلال السنوات العشرينية الماضية، لذا ظهرت العديد من تقنيات التعلم الآلي في مجال تصنيف النصوص منها: تقنية روشيو،تقنية المكائن ذات الدعم ألاتجاهي، وتقنية الاحتمالات المسماة بآيس وتقنية أشجار القرار.

كما ظهرت العديد من النماذج كالتصفية (إشراف من قسمين)، التوجيه(إشراف متعدد الأقسام) أو التصنيف العادي (تصنيف النصوص المرتبة وفق فئات)، ومن خلال هذه النماذج تم تطوير منهجيات وأساليب الاختبار وأدوات التقييم وطرق التمثيل الموافقة للمعالجة المسبقة.

إلا أنه رغم تعدد البحوث في هذا المجال، هناك العديد من العوائق التي تحول دون إمكانية تصنيف آلي دقيق لان خوارزميات التصنيف حتى وإن عملت بشكل دقيق وصحيح، إلا أنه من الصعب جدا تحديد الأسبقيات دون بعضها البعض أو حتى تحسين طريقة التصنيف من خلال دمج نماذج أخرى وهذا ما سنحاول تبيينه في هذه الدراسة.

يشمل مجال المعالجة الذكية للبيانات النصية جميع الأدوات والأساليب الفعالة التي تعمل على استخراج المعلومات من النصوص المكتوبة باللغة الطبيعية، إذ يوجد مجالين مهمين لمعالجة هذه

المشكلة ولكل مجال أساليبه الخاصة به.

أولا تعمل المناهج على تحليل البيانات وإعداد دراسة إحصائية خاصة كما أنها تسعى لتقديم أدوات للإحصائيين واللغويين لتمكينهم من تحليل قواعد البيانات النصية ذات الحجم الكبير، من خلال توفير معلومات موجزة عن المدونة، كما تقدم برامج تحليل المدونات قوائم تردد الكلمة وتمثيلها البياني من خلال تحليل جزء من هذه المصنفات، كما تقدم هذه المناهج أنظمة و طرق تعالج الوثائق بصفة آلية تلقائية وتحقق في كثير من الأحيانوظائف منخفضة المستوى: التحليل المعجمي، التحليل النحوي والتركيبي والبحث عن المعلومات عن طريق الكلمات الرئيسية المفتاحية.

1.2 تعريف تصنيف النصوص:

التصنيف وفق ما ورد في المعاجم هو تمييز الأشياء بعضها عن بعض، وصنف الأشياء أي قسمها وفق تشابهها إلى مجموعات تضم كل مجموعة وحدات تشترك في صفة أو خاصية واحدة على الأقل. أو يمكن إعطاءه تعريف آخر أكثر تحديدا إذ أن تصنيف النصوص هو تعيين النص بعلامة أو أكثر لفهرسة هذا المستند في مجموعة من الفئات محددة مسبقا، صممت في الأصل للمساعدة في أعمال الترتيب الوثائقي أو المقالات في المجالات التقنية أو العلمية،لذا فعملية تصنيف النصوص هي ضم وثيقة إلى فئة أو فئات محددة مسبقا، والهدف من هذه العملية هو القدرة على انجاز تصنيف آلي لمجموعة جديدة من النصوص [2] .

التصنيف الآلي للنصوص (Automatic Text Categorization) هي مهمة تصنيف المستندات النصية الإلكترونية اتوماتيكيا إلى أصنافها المعرفة مسبقا بحسب محتوياتها.

[2] (Brown & Chong , 1998) G.Brown, H.A.Chong « The Guru System in TREC-6 »

بمعنى آخر تحديد الصنف الرئيسي الذي يندرج تحته النص أو المستند "سياسة ، اقتصاد ، رياضة، ... الخ" [3].

يعلمك تصنيف النصوص كيفية التحكم في مجموعة من مواصفات التمييز للسماح بتخزين الوثيقة المعطاة في طبقات أو فئات موافقة لمحتواها.

أما خوارزميات التصنيف تعتمد أساسا على مناهج التعلم من خلال المدونات التي تسمح بتصنيف النصوص الجديدة، هذا النوع من المناهج يعمل على معرفة البيانات من خلال المدخلات التي هي النصوص والمخرجات والتي هي الفئات، إن الأعمال المتعددة في هذا المجال تتطلع إلى إيجاد خوارزمية تعمل على ضم النص إلى الفئة التي ينتمي إليها فعليا وتحقق نسبة نجاح كبيرة من دون ضم النص إلى الكثير من الفئات.

وفي هذا السياق يسمح التشابه النصي بتحديد الفئات الأقرب إلى الوثيقة المصنفة، وان كانت فكرة التشابه النصي في الكثير من الأحيان تبدو بديهية بالنسبة للإنسان فإنه قد ينتج عنها عمليات معقدة وغير مفهومة من قبل العقل.

ويمكن تلخيص مشكلة التصنيف من الاستياء من فكرة التشابه النصي، لذا لا بد من السعي لإيجاد نموذج رياضي قادر على تمثيل وظيفة تقرير تحدد انتمائية وعضوية النصوص في الفئات، مع ذلك فإن نظام التصنيف الآلي يساعد على ربط كل وثيقة بمجموعة من المصنفات.

كما تقوم أنظمة التصنيف على مجموعة من المناهج التعليمية تعمل على المعالجة ووظيفة التقرير باستخدام مدونة تدريبية، هذه الوظيفة تستطيع التدخل بعدد هائل من القيم الرقمية التي لا يستطيع الإنسان القيام بها.

[3] بسام محمد واحمد السالمي، التصنيف الآلي للنصوص العربيةبإستخدام تعليم بايزيناالإحتمالي،2011.

وهناك العديد من التقنيات الحاسوبية المعروفة بما يسمى "التعليم الآلي المسبق " (Learning Supervised Machine) والتي تم استخدامها بغرض حل مشكلة التصنيف الآلي للمستندات. "التعليم الإحصائي" (Statistical learning) هو إحدى تلك التقنيات المستخدمة لذات الغرض والتي تعتمد في أداءها تخمين الصنف الذي يندرج إليه النص بطرق احتمالية،واحد من أكثر التقنيات الإحصائية المستخدمة يسمى " Bayesian Learning " والذي يعتمد على نظرية بايزيان الاحتمالية،يوجد هنالك العديد من موديلات " Bayesian " والتي لم يتم استخدامها مسبقا لغرض تصنيف النصوص العربية آليا [4].

كما نجد أن التصنيف الآلي ارتبطت في أولا لأمر بعملية استرجاع المعلومات إذ نجد الدارسين الذين اهتموا بهذا المجال حاولوا إيجاد العديد من طرق التصنيف التي تسهل عملية استرجاع البيانات المخزنة، وفي علم استرجاع المعلومات يتم تصنيف الوثائق في نفس المجموعة إذا كان لها نفس التصرف تجاه طلب المعلومات [5].

مما يعني أنه إذا كانت إحدى الوثائق في مجموعة معينة ذات صلة باستعلام معين فإن احتمالية كون بقية الوثائق في تلك المجموعة ذات صلة بذلك الاستعلام ستكون عالية أيضاً.

وللتصنيف الآلي عدة تطبيقات في علم استرجاع المعلومات يمكن تقسيمها إلى نوعين وفقاً لمجموعة الوثائق التي تسعى لتصنيفها ولجوانب استرجاع المعلومات التي تحاول تحسينها، ففي النوع الأول يمكن إجراء التصنيف على نتائج البحث، أو على جزئية من مجموعة الوثائق، أو على كامل مجموعة الوثائق،أما في النوع الثاني، فإن التصنيف يستخدم لتحسين واجهة المستخدم، أو خبرات المستخدم، أو فاعلية وأداء نظام البحث.

[4] بسام محمد احمد السالمي، التصنيف الآلي للنصوص العربيةبإستخدام تعليم بايزينالإحتمالي، 2011

[5] Manning, P. Raghavan, and H. Schütze, An Introduction to Information Retrieval, Cambridge University Press, .2009

وقد وجدنا عددا قليلا فقط من الأبحاث التي اهتمت بتصنيف الوثائق العربية لغرض استرجاع المعلومات ففي أحد الدراسات [6]. قام الباحثون ببناء مصنف مبني على خوارزمية (Bayes Naive) والغرض منه توفير فهرس بالمواضيع يسهل عملية البحث ويعمل هذا المصنف على تصنيف الوثائق إلى خمس مواضيع أساسية وهي: الرياضة، الأعمال، الثقافة والفن، والعلوم، والصحة،ويتم قبيل عملية التصنيف إزالة التشكيل واستخراج جذور الكلمات، وقد بلغ معدل نسبة دقة التصنيف 78.68 % . كما قدم باحثون آخرون، خوارزمية للتصنيف. الآلي للوثائق العربية مبنية على استخراج الكلمات التي تغطي المفهوم الأساسي لموضوع كل وثيقة، بحيث يتم حساب وزن كل كلمة بناءً على مدى تكرار هذه الكلمة في الوثيقة وأماكن تواجدها،وقد وجد الباحثون أن استخدام خوارزمية التصنيف هذه قد زاد من كفاءة نظام استرجاع المعلومات [7].

2.2 الإرهاصات الأولى لتصنيف النصوص:

إذا أردنا تحديد تاريخ فعلي لبداية العمل بالتصنيف فلا بد من البحث عن جذوره التاريخية فهو نظام قديم إلى حد ما ظهر عام 1627م مع (Gabriel Naudè) حيث اقترح ترتيبا وفق خمس مواضيع مهمة : علم الديانات اللاهوتيات، الفقه الحقوق و القوانين، العلوم، الفنون، الأدب الجميل، كما رتبت موسوعة ديدروت(نشرت في الفترة مابين 1751- 1772) وفق الترتيب الأبجدي مع وجود إحالات ترابطية، كما نشرت مؤسسة بانكوك(مابين 1776-

[6] M. Elkourdi, A. Bensaid, and T . Rachidi , "Automatic Arabic Document Categorization Based on the Naïve Bayes Algorithm", in Proc. of COLING 20th Workshop on Computational Approaches to Arabic Script-based Languages, .2004

[7] S. Ghwanmeh, G. Kanaan, R. Al-Shalabi and A. Ababneh, "Enhanced Arabic Information Retrieval System based on Arabic Text Classfiication", 4th International Conference on Innovations in Information Technology, pp.461 - ,465 .2007

1780) تنظيما وفق منهج شجرة القرار [8]

إن نظام تصنيف الموضوعات ظهر مع الأيام الأولى للتدوين من اجل إضفاء الطابع المؤسساتي للإسكندرية، مما استلزم ذلك إلى إنشاء نظام التصنيف العالمي (ديوي) عام 1876م الذي يتعلق بتصنيف الوثائق الشبيهة بالموسوعات والكتب.

إلا أن فكرة إجراء تصنيف النصوص عن طريق الأجهزة يعود إلى أوائل الستينات،وقد حقق تقدما كبيرا مع بداية التسعينات مع ظهور العديد من الخوارزميات التي أصبحت أكثر كفاءة من ذي قبل.

لذا حتى أوائل الثمانينات كان الواجب علينا لبناء مصنف تكريس الكثير من الموارد البشرية لهذه المهمة، فالعديد من الخبراء قاموا بنشر قواعد يدوية ووضعوا لها اختبارات ومع ظهور التعلم الآلي تم توفير الكثير من الجهد و الوقت، هذه التطورات التكنولوجية والخوارزميات المتقدمة جعلت التصنيف اليوم أداة يمكن الاعتماد عليها بشكل كبير.

في بداية التسعينات بدأ البحث في المجمعات حول استرجاع المعلومات وتم وضع مناهج الرقمنة وخوارزميات التصنيف خاصة في المؤتمرات (مؤتمر استرجاع المعلومات).

كما اهتم مجمع التعليم الآلي بهذه المشكلة منذ أكثر من عشر سنوات مثل النظر في الخوارزميات للتعرف على الأشكال، وحاليا لا تزال مناهج رقمنة النصوص موجودة ومستوحاة إلى حد كبير من قبل مجمع استرجاع المعلومات في حين أن المصنفات الأكثر أداء هي المتعلقة بالتعلم الآلي.

مجمع آخر يتكون أساسا من اللغويين والإحصائيين عملوا على حل مشكلة تصنيف النصوص استنادا على مناهج تحليل البيانات، والهدف من ذلك ليس لخلق نظام يصنف الوثائق آليا دون تدخل الإنسان، ولكن لاسترجاع المعلومات المصنفة في المدونات ولعلاج مشاكلها على سبيل

[8] (Fayet-Scribe, 1997) S.Fayet-Scribe « Chronologie des supports, des dispositfis et des outils de repérage de l'information »

المثال دراسة الأنواع الأدبية أو تحديد مؤلف النص.

3.2 الحاجة إلى تصنيف النصوص:

لقد اهتمت في السنوات الأخيرة الكثير من البحوث بتألية النصوص متعددة اللغات وذلك لعدة أسباب نذكر منها: تزايد مجموعة البيانات على الشبكة العنكبوتية وانتشارها على نطاق واسع بمختلف الأصناف و اللغات إذ يمكن الحديث حسب الإحصائيات الأخيرة عن أكثر من 03 مليارات من الصفحات بالإضافة إلى تزايد حجم المدونات المستخدمة.

لذا تولدت الحاجة إلى ضرورة تصنيفها آليا لأن المعالجة اليدوية لهذه البيانات قد تكلف الكثير من الوقت و الأفراد كما أنها غير مرنة وتعميمها على كل الميادين أمر شبه مستحيل . [9]

إذ أن عملية التصنيف اليدوي للنصوص والوثائق تكلف الكثير من الجهد فقط من اجل ربط نص بمجموعات أو فئات مختلفة كذلك يحتاج المصنف من أجل تصنيف وثيقة إلى الكثير من الوقت فقط في القراءة الأولى وقد يحتاج المصنف إلى قراءة ثانية، من أجل استيعاب مضمون النص مع أن هذه العملية تحتاج لوقت أطول خاصة وان سرعة التصنيف تختلف من شخص لآخر نظرا لتفاوت سرعة القراءات بالإضافة إلى ذلك قد يحتاج المصنف اليدوي إلى زيادة وقت للتأكد و التأمل وذلك بالرجوع إلى النظر في نصوص أخرى مصنفة للتحقق من صحة التصنيف.

وهناك عوامل مشتركة في عدد من المصنفات التي لا بد التفريق فيما بينها مثلا هناك فئات مختلفة عن بعضها،إذ نجد إمكانية إعطاء أكثر من عنوان لنص ما فيصعب الاختيار فيما بينها. أيضا المصنفات ذات العلامات الكثيرة تحتاج إلى اهتمام كبير قبل ربط أي وثيقة أو ضمها

[9]R. Carnap (The LogicalSyntax of Language) 5th edition, Routledge&Kegan Paul Ltd., London, UK, .1959

من خلال هذا يمكن القول أن انجاز هذه المهمة بخبير يدوي مكلفة للغاية من حيث الوقت و الموظفين نظرا للكم الهائل من النصوص المتاحة اليوم خاصة على شبكة الانترنيت [10] .

المعالجات اليدوية تتسم بمرونة قليلة إذ يستحيل تعميمها على كل الميادين لذا لابد من السعي لتطوير الأساليب الآلية [11] .

كما أن العملية اليدوية تكون ذاتية أكثر منها موضوعية عند تفسير الوثيقة إذ يمكن لخبيرين أن يفسرا الوثيقة الواحدة بطريقة مختلفة أو يمكن للخبير نفسه أن يصنف الوثيقة المعطاة له في وقتين مختلفين بطريقة مختلفة عن بعضها [12] .

وبالتالي لا بد من استبدال العملية اليدوية بالتصنيف الآلي للنصوص من اجل تجنب العديد من النقائص التي يمكن أن يحدثها التصنيف اليدوي وعليه فقدت ركزت البحوث في السنوات الأخيرة على التصنيف الآلي.

و المصنفات واحدة من التطبيقات المهمة في الوقت المعاصر تعرف باستخلاص المعلومات ومجالاتها المعتبرة تعتمد على معالجة كميات كبيرة من النصوص و التنقيب فيها بحثا عن قوالب و سمات محددة وهذا لا يكون إلا بتحليل اللغة إلى أجزائها الأولية و هذه وظيفة المصنفات بالأساس.

وبالإضافة هناك وظائف أخرى للمصنفات نرصد منها ما يلي:

- توضيح أجزاء النص بالإضافة السمات الدالة على وظائف مفردات الجملة.
- اقتراح المعلومات الناقصة في النصوص برصد المعلومات المفقودة مثل علامات التشكيل.

[10] (Moulinier, 1996) I.Moulinier « Une approche de la catégorisation de textes par l'apprentissagesymbolique .»

[11] (Sebastiani, (2002 F.Sebastiani « Machine learning in automatedtextcategorization »

[12] (Clech&Zighed, 2004) J.Clech, D.A.Zighed « Une technique de réétiquetage dans un contextede catégorisation de textes »

وبمقدور هذه الوظائف تجهيز المدونات المذخرة هذا يعني إبقاء النص الأصلي على حالة مع إضافة علامات تميز أجزاء الكلام وميزات أخرى مهمة بغية تجهيزه لمراحل لاحقة من المعالجة فالمدونة غير المذخرة لا فائدة منها تقريبا سوى في النواحي الإحصائية الخاصة بمجالات التحليل الصرفي، هذا بالنسبة للغة العربية ومعظم اللغات الطبيعية ، فالمدونات ليست هدفا في ذاتها بل تستعمل مدخلات لنوع آخر من التطبيقات.

4.2 أسس عمل المصنفات:

المصنف يتعامل مع النصوص على حسب طبيعة اللغة فهو مع اللغات المرسلة (الصينية و اليابانية) يتعامل بطريقة تبدأ بمحاولة التعرف على حدود الجملة ومع اللغة العربية يتعامل على مستوى المفردة كما يمكن جعل المصنف يتعامل مع اللغة العربية على حسب الجملة و لكن هنالك بشكل عام أسلوبين لعمل المصنف والثالث مجرد موائمة منها:

أسلوب إحصائي يعتمد على نظرية المعلومات وهي تنجح سريعا ولكن بنمط انتقائي بسبب أن المدخلات غالبا ما تكون من شاكلة عينة التدريب لذا سرعان ما يلحظ مطوروها أنها تفشل في الأمثلة البسيطة غير التقليدية فالناس يستعملون في حياتهم اليومية نسبة ضئيلة من جملة مفردات اللغة.

أسلوب تحليلي ،و هو أصعب أنواع الطرق المستخدمة في تطوير تطبيقات معالجة اللغات وذلك لاعتماده على المعرفة العميقة بمكونات اللغة قيد التحليل ألا الأسلوب الأول الإحصائي يتطلب وجود عينات ضخمة من النصوص هي بذاتها تتطلب إجراء التصنيف بطريقة مباشرة وهذه نقطة ضعف إذ أن مهمة تحضير العينات عملية يقوم بها البشر.

5.2 كيفية إجراء عملية التصنيف:

لتنفيذ عملية التصنيف الآلي للنصوص كما عرفناها سابقا لا بد من إتباع الطريقة التالية:
المرحلة الأولى: تتعلق بإضفاء الصبغة الرسمية على النصوص و تحضيرها بحيث تكون مفهومة من قبل الآلة واستخدامها من قبل خوارزميات التعلم.
إما المرحلة الثانية تتعلق بتصنيف الوثائق: هذه المرحلة هي الحاسمة لأنها تسمح أو لا تسمح باستخدام تقنيات التعليم لإنتاج تعميم جديد لثنائيات الوثائق والمصنفات و لتحسين أداء النماذج يتم تقييم جودة المصنفات ومقارنة النتائج المقدمة من طرف مختلف النماذج التي أجريت في نهاية الدورة.
ولوضع طريقة ثابتة للتصنيف الآلي للنصوص يمكن أن نلخصها بالطريقة التالية:

- إزالة الأحرف الفاصلة: كعلامات الترقيم و الكلمات الفارغة الزائدة...
- الكلمات الباقية هي التي تختص بالتصنيف.
- تصبح الوثيقة متجه (تكرار الكلمة).
- تدريب نموذج التصنيف على ثنائيات(الوثيقة- المصنف).
- تقييم نتائج المصنف.

6.2 صعوبات التصنيف الآلي للنصوص:

قد يتعرض تصنيف النصوص إلى العديد من الصعوبات، قد تكون صعوبات متعلقة بالتعلم الآلي بالإشراف مثل القرارات الذاتية التي يصدرها الخبراء، إلا أن هناك مشاكل خاصة تتعلق بطبيعة البيانات المعالجة للنصوص مثل تعدد معاني الكلمة، التكرار، التغيرات المورفولوجية...وغيرها من

الصعوبات، وهنا نشير إلى أهم عشرة قواعد تعيق تصنيف النصوص و هي:

1.6.2 الاشتراك في المعنى (الترادف):

التكرار و المترادفات تستخدم للتعبير عن نفس المفهوم أو المعنى بعبارات مختلفة وطرق مختلفة للتعبير عن نفس الشيء، و ترتبط هذه الصعوبة بطبيعة المستندات المعالجة على عكس البيانات الرقمية المعبر عنها باللغات الطبيعية.

ويوضح (Lefévre) الصعوبة هذه في مثال القط و العصفور:

- أكل قطي عصفورا.
- التهم قطي الكبير عصفورا.
- التهم قطي المفضل ريش حيوان صغير. [13]

إذ نجد نفس الفكرة مثلت بثلاثة طرق مختلفة، فكل عبارة استخدمت بطريقة مغايرة للتعبير عن شيء معين ولكن في نهاية الأمر هي تصب في مفهوم واحد وهو أن الطائر هو من أكل من قبل القط.

و لتصنيفها لا بد من تمثيل ناقلات الوثيقة، وتمثل فيها العبارات على حدا أي انه من المهم جدا جمع هذه العبارات في مجموعة دلالية مشتركة، وهذا بطبيعة الحال يولد تكاليف أو جهد إضافي من اجل التدقيق و التصرف فيها.

[13] (Lefévre, 2000)P. Lefévre « La recherche d'information - du texte intégral au thésaurus »

2.6.2 تعدد المعاني(الغموض):

يشير الغموض إلى المعنى غير الواضح، والتعبير غير التام وهذا ما ينتج عنه الالتباس والخلط، ومن أهم مظاهر هذا الالتباس ما نسميه الالتباس اللفظي (Ambiguity) والذي غالباً ما ينتج عن تعدد معاني اللفظ الواحد وتشاركها وعدم وضوح معناها المخصص من خلال السياق .

فعلى عكس البيانات الرقمية نجد أن البيانات النصية غنية بالدلالات لأنها مستوحاة و مصممة من قبل الفكر الإنساني.

خلافا للغات الكمبيوتر اللغات الطبيعية تسمح بانتهاك القواعد النحوية فتؤدي إلى عدة تغيرات، فالكلمة نجد إن لديها أكثر من معنى واحد أو العديد من التعاريف المرتبطة بها ولذلك فان تعدد معاني الكلمات قد تحمل في بعض الأحيان واصفات سيئة

3.6.2 التجانس اللفظي:

هو أن تتجانس الكلمات في رسمها الإملائي وتختلف في معناها و يطلق على كلمتين مكتوبتين بنفس الطريقة دون أن يكون لهما بالضرورة نفس النطق بالتجانس اللفظي وهو نوع من الغموض يولد التجانس و الغموض نوع من التعمية من شأنها أن تتسبب في تدهور دقة المؤشر الضروري لقياس أداء المصنف، و سيكون من الأفضل حينذاك إزالة هذه الالتباسات.

4.6.2 الشكل الخطي للكلمات:

قد يحتوي مصطلح ما على أخطاء إملائية أو كتابية كما يمكن أن تكون مكتوبة بعدة طرق وهذا ما يؤثر بشكل كبير على نوعية النتائج لأنه إذا وردت كلمة بطريقتين في نفس الوثيقة، فان البحث البسيط لهذا المصطلح مع شكل خطي واحد يهمل وجود نفس المصطلح بشكل خطي آخر ولان هذا الأمر يؤثر على عملية التصنيف فان الأشكال الخطية سيتم التعامل معها بشكل

منفصل.

ومع ذلك فإنه من وجهة نظر عملية أن الكلمة غير المعروفة تكون قريبة من الكلمة الأخرى هذا ما يثبت أنها ربما تحتوي على أخطاء إملائية.

ودائما في هذا السياق اثبت بأن الشكل الخطي يمكن أن يوفر معلومات متعلقة بمعنى الكلمة المستعملة إلا انه لابد من مراعاة هذه الاختلافات من اجل التصنيف الآلي للنص [14] .

5.6.2 التغيرات الصرفية :

التعريف و الجمع قد يؤثر سلبا على جودة النتائج، فالتغيرات الصرفية المختلفة تدرس بشكل منفصل مثل الكلمات: كتب، كتاب، مكتب، مكتبة يتم معالجتها بشكل مستقل على الرغم من أن الحقيقة أن هذه المتغيرات تدور حول نفس الفكرة.

6.6.2 الكلام المركب:

الكلمات المركبة قد تشكل صعوبة جد كبيرة عند التصنيف مثل قوس قزح

إذ نجدها متواجدة بعدد جد كبير في كل اللغات و لمعالجة هذه الكلمة لابد أن يكون بكلمتين منفصلتين وهذا ما يقلل بشكل ملحوظ من أداء نظام التصنيف، إلا انه باستخدام تقنية لترميز النص يقلل بشكل ملحوظ من مشكلة الكلمات المركبة.

[14] (Loupy& El-Bèze, 2000)C.de Loupy, M.El-Bèze « Using few cuescancompensate the small amount of resourcesavailable for WSD »

7.6.2 حضور وغياب الكلمات:

حضور الكلمة في النص يشير إلى ما أراد المؤلف التعبير عنه، لذلك هناك علاقة ضمنية بين الكلمة و المفهوم المرتبط بها.
على الرغم من أننا عرفنا انه يوجد الكثير من الطرق التي تعبر عن نفس الأشياء ، مع ذلك غياب الكلمة لا يعني بالضرورة أن هذا المفهوم الذي يرتبط معها مفقود من المستند ، وهذه الملاحظة الخاصة تقودنا إلى توخي الحذر في استخدام تقنيا التعلم القائمة على استبعاد كلمة معينة.

8.6.2 تعقيد خوارزمية التعلم:

نتحدث هنا عن تمثيل و ترميز الوثائق، حيث أن النص عادة ما يمثل في عمومه كثاقل للميزات يحتوي على العديد من الكلمات المتمظهرة فيه، أو عدد النصوص التي سيتم معالجتها، فمن المهم جدا أن نمهل العبارات المكونة لنفس النص لذا يمكن للمرء أن يتصور حجم الجدول (النص والعبارات)المعالج الذي سوف يعقد بشكل كبير مهمة التصنيف وتخفيض أداء النظام وبالتالي التقليل من حجم الجدول.

9.6.2 الذاتية في اتخاذ القرارات:

من بين المشاكل القديمة و المألوفة في مجال التعلم بالإشراف هو الذاتية في اتخاذ القرارات من قبل الخبراء الذين يقررون الفئة التي ينتمي إليها النص بالتأكيد بعد قراءة النص المصنف فإن الخبراء وضحوا كيفية تقرير الفئة التي ينتمي إليها النص استنادا إلى محتواه الدلالي و سياق النص وكذا استثارة نصوص أخرى مرتبطة مسبقا مع فئات معينة الذي لا يمكنه إلا أن يكون ذاتيا فالخبراء لا يقرءون بنفس الطريقة ولا يفكرون بنفس الطريقة وبالتالي التصنيف لا يكون بنفس الطريقة وهكذا يمكن تصنيف نفس الوثيقة بشكل مختلف من قبل خبيرين أو حتى تصنيف

وثيقة بشكل مختلف من قبل الخبير نفسه في وقتين مختلفين [15].

ومن خلال التجارب عندما يقوم خبيرين بتحديد فئات مجموعة من النصوص غالبا ما يكون هناك اختلاف في أكثر من 5 بالمائة من النصوص لذلك من المهم جدا البحث عن التصنيف الآلي الأمثل.

7.2 التعرف الآلي على اللغات الطبيعية وتصنيف بياناتها:

وباعتبار اللغة وسيط إشباعي عاطفي للتواصل فهي تتفاوت في قدراتها في التعبير عن محمولات الكلام المنطوق أو النص المكتوب وما يريده المؤلف وهذا يعتمد على مهارات أخرى من بينها الأسلوب و السياق يكون في مقدور السامع بعدها أو القارئ معرفة قصد المتكلم أو الكاتب . المصنف أدنى من ذلك وهو ليس هدفا في ذاته وتنطبق عليها ذوات الظروف لأدوات معالجة اللغات الطبيعية الأخرى، و الحاجة إليه تكمن في ضرورة تفنيد الجمل إلى مكونات وظيفية حتى يمكن معاملة كل منها حسب أهميتها في النسيج العام للنص قيد التصنيف.

لقد كانت محاولات ربط اللغة بالوسائل التكنولوجية الحديثة ليس قصد تطويعها وإخضاعها وإنما لتطويرها ولعل تزايد البيانات المكتوبة باللغات الطبيعية ومحاولة رقمنتها والاستعانة بالحواسيب لتنظيمها ولد العديد من ميادين البحث كعلم الحاسوبيات الذكاء الاصطناعي وادي الى ظهور العديد من خوارزميات التعلم و التقنيات الحديثة والتصنيف الآلي من بين تلك التقنيات الهامة و الضرورية التي لا يمكن الاستغناء عنها في أي ميدان علمي أو مجال بحث متطور.

إن اية تألية لنص لغوي ما تعتمد اعتمادا كليا على لغته لذا فمن الضروري ان تكون أدوات البحث قادرة على التعرف آليا على لغات الوثائق أو النصوص و ايجاد تقنيات ناجعة للتعامل

[15] (Clech&Zighed, 2004)J.Clech, D.A.Zighed « Une technique de réétiquetage dans un contexte de catégorisation de textes »

معها.

وعندما نتكلم عن التصنيف لا بد أن نجزم بإمكانية التعرف الآلي على اللغة لأن جميع اللغات تمتلك انتظاما في استخدام الحروف وتسلسلها لذا ظهرت العديد من خوارزميات التعرف على لغات النصوص قصد تصنيفها بالعديد من الطرق مثل الطريقة اللغوية و طريقة الاحتمالات والإحصاء [16].

8.2 تطبيق تقنيات التصنيف الآلي للنصوص العربية:

ونظراً للنمو المتزايد للمحتوى العربي الرقمي سواءً على الإنترنت أو الوسائط الإلكترونية الأخرى، فإن الحاجة أصبحت أشد إلحاحاً لإيجاد أنظمة استرجاع معلومات ومحركات بحث تعتني بالخصائص الفريدة للغة العربية وتحسن التعامل معها. فالعربية هي لغة النوحيين القرآن والسنة وتعد أكثر اللغات السامية الحية من حيث عدد المتكلمين بها. وتتميز العربية عن بقية اللغات الجرمانية بأنها تكتب وتقرأ من اليمين إلى اليسار، كما أن حروفها تكتب بأشكال مختلفة تبعاً لموقعها والحروف المجاورة لها، وتختلف طريقة نطق الحرف وبالتالي معنى الكلمة وموقعها الإعرابي بناءً على حركة التشكيل الموجودة عليه ، بالإضافة إلى أن العربية لغة اشتقاقية وليست إلصاقية، حيث يعد نظامها الصرفي من أكثر النظم الصرفية تقدماً، فهو مبني على تصريف الجذور وفقاً لمجموعة محددة من الأوزان للحصول على كلمات ذات دلالات مختلفة من نفس الجذر. وكل ما سبق ذكره يمثل تحديات لمقننة التحليل الصرفي والإعرابي والدلالي للغة العربية ومن ثم التصنيف الآلي لمجمل النصوص العربية.

[16] P. F. Strawson (Introduction to LogicalTheory) Methuen & Co. Ltd., London, UK, 1960.

9.2 أوجه صعوبة التصنيف الآلي للنصوص العربية:

مما لا شك فيه أن محاولة إخضاع اللغة للحاسوب لا بد وأن يعترضها العديد من الإشكاليات والعقبات. وعندما تتشابه العقبات في لغات عديدة فإنه بلا شك تتشابه طرق حلها. غير أن تحليل اللغة العربية بوساطة الحاسوب يكتنفه عقبات كثيرة، أكثر من أي لغة أخرى. ومعظم هذه المشاكل متعلقة بالجوانب التي تختلف فيها العربية عن اللغات الأوروبية، تلك اللغات التي صممت معظم البرامج الحاسوبية أصلاً لتحليلها.

ولا شك أن محاولة قولبة اللغة العربية في الحاسوب من أهم المشاكل التي تعترض طريق وضع المصطلحات العربية. ولذلك يقترح (Slocum & Aristar) وجوب حصر الأوزان العربية حصراً دقيقاً وتحليلها وفق نظام تصنيفي معين، وهو ما سيمكن من وضع رموز رياضية لها في الحاسوب [17] . ويجب أن يكون هذا التصنيف من الشمولية بحيث يستوعب الفروق بين الكلمات الناتجة عن اختلاف التشكيل. ثم يتم بعد ذلك تطوير برامج آلية يمكنها استيعاب القواعد النحوية العربية، بحيث يتمكن الحاسوب من تصويب الجمل الخاطئة عند قراءتها. ومن الجانب النظري، فإن تطوير مثل هذه البرامج ليس مستحيلاً.

وقد ظهرت محاولات كثيرة لوضع برامج آلية لتصنيف بيانات اللغة العربية بحيثُ يمكنها التعامل مع مختلف المشاكل و العوائق، خاصة فيما يتعلق التركيب الصرفي للغة العربية، بالإضافة إلى بعض هذه العوائق الناشئة عن عملية التعريب نفسها، أي تعدد الطرق المستخدمة في التعريب وتباينها فيما بينهابالإضافةإلى مشاكل أخرى نحصرها فيما يلي:

[17] سعد بن هادي قحطاني، تحليل اللغة العربية بواسطة الحاسوب، مركز اللغة الانجليزية، معهد الادارة ، الرياض.

1.9.2 تعدد الطبقات:

بعض اللغات ثنائية الطبقات أو ازدواجية النطق مما يعني أن الناطقين باللغة يعتمدون على نوعين نوع تستخدمه النخبة أما النوع الثاني تستخدمه العامة، كما هو الحال بالنسبة للغة العربية إذ نجد فيها طبقات الطبقة الأولى هي اللغة المنضبطة المكتوب بها القرآن و أسمى الأعمال الأدبية كالشعر الجاهلي ثم لغة ابسط منها و هي اللغة الفصيحة نجدها عند الصحفيين ووسائل الإعلام، ثم نجد لغة العامية تختلف بحسب العوامل الجغرافية و الثقافية و الاجتماعية،إلا أن الصعوبة تكمن في التداخل بين هذه الطبقات الثلاثة إذ نجد مثلا لغة الصحفيين تستعمل العامية بدل الفصحى و نجد أيضا كتابات رسمية توظف بعض الكلمات العامية هذا ما يشكل صعوبة أثناء تصنيف النصوص لأنه لابد الأخذ بعين الاعتبار تداخل هذه الطبقات الثلاثة مع بعضها البعض.

2.9.2 تأثير اللغات الأخرى:

هذه مشكلة لا تتعلق باللغة العربية فقط إذ نجدها في كل اللغات التي تتداخل طبيعيا إذ نجدها تتداخل في مجموعة من المستويات مثل التداخل المعجمي بتوظيف ألفاظ و مصطلحات مثلما الحال في اللغة العربية إذ نجد أغلبية الدارسين يوظفون مصطلحات غربية أثناء حديثهم عن علوم غربية المنشأ أو مصطلحات لمستحدثات العلوم الجديدة أو الترجمة الحرفية لبعض المصطلحات الغربية،أما التداخل النحوي بتبني القواعد النحوية كما الحال في العامية السورية و اللبنانية التي استعارت بعض القواعد الآرامية و عادة ما تكون مسببات التداخل اللغوي إلي التداخل بين منطقتين جغرافيتين و كذا الهجرات السكانية المتبادلة إذ نجد في هذه الحالة الكثير من التأثر في مختلف مناحي الحياة خاصة اللغة وهذا ما ينتج عنه لغة هجينة وفي هذه الحالة عملية التصنيف الآلي تتعر قل بوجود ازدواجية لغوية أو لغة هجينة لا تتقيد بقواعد لغوية محددة.

3.9.2 تعدد الاختلافات الإقليمية:

اللغة العربية تتوسع على نطاق واسع من الدول الناطقة بها و التعامل معها يختلف من إقليم إلى آخر كما هو الحال بالنسبة للإنجليزية التي تختلف بين بريطانيا و أمريكا ، فاللغة العربية الفصحى دخلت عليها العديد من الاختلافات النطقية ونجد ذلك يتجلى في وسائل الإعلام الصحفية إذ أنها تمثل نسبة كبيرة من النصوص العربية المتوفرة اليوم على هيئة مدونات و التي لابد من تصنيفها إلا أن هذا الاختلاف في اللغة على حسب الأقاليم يعتبر معوق جد كبير في عملية التصنيف الآلي إذ لا بد من حصر مختلف اللهجات التي لها صلة باللغة العربية الفصحى و نجد أهم التغييرات التي تدخل عليها هي التغيرات الصرفية التي تخالف القاعدة اللغوية الصحيحة لذا أثناء التصنيف لا نجد أنفسنا أمام لغة أو لغتان وإنما العديد من اللغات.

4.9.2 عدم وجود فوارق شكلية واضحة بين مكونات النص:

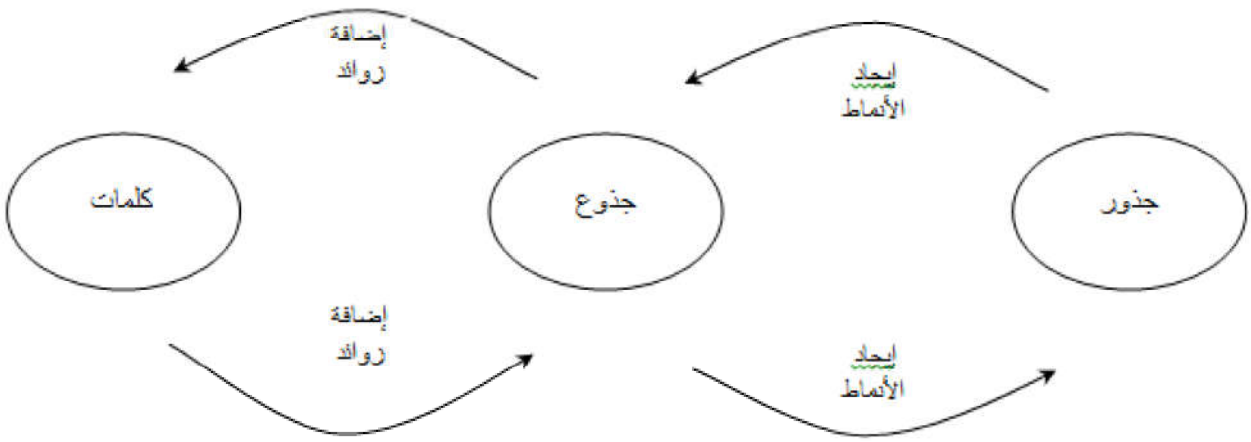
بعض اللغات تميز أجزاء الكلام بأنواع مختلفة من الحروف و اللواصق إلا أن اللغة العربية لا نجد فيها أدوات تفرق بين الكلمات إذ أنها تعتمد على الأوزان و الحركات الإعرابية لذا لا بد من العودة إلى الحركات التي تعمل على إحداث فوارق بين الكلمات المختلفة.

5.9.2 افتقار اللغة العربية لمبدأ الوحدة الدلالية:

تقوم اللغة العربية على عكس اللغات الأخرى على مبدأ الاشتقاق الذي ينبني أساسا على الجذر إذ يمكن لجذر ثلاثي أن تشتق منه العديد من المفردات التي تختلف في أكثرها دلاليا و هذا ما يعيق عملية التصنيف إذا إن الكلمات المشتقة من جذر واحد يمكن ان تحمل معان مختلفة عن بعضها البعض و هذا ما يمكنا تسميته بالتشتت الدلالي مثل : الجذر بلغ نشتق منه الكلمات التالية و التي تختلف دلاليا فيما بينها: بلوغ- مبالغة- بليغ- بلاغة- مبلغ- بلاغ.

ثم إن معظم مفردات اللغات الأخرى المصنفة بالحاسوب مصنفة حسب الصيغة المبنية(أي صيغة الفعل، أو الاسم المجرد) وليس بالجذر كما في اللغة العربية. وتصنَّف السوابق واللواحق كمدخلات أساسية عكس اللغة العربية. وبما أن اللغة العربية تحتوي على صيغ صرفية داخلية(تحدث داخل الكلمة نفسها) وليست سوابق أو لواحق؛ فإنه يتحتم التعامل مع الجذر وليس مع كل صيغة على حدة. وهذه الخصائص التي تختص بها العربية تجعل من الصعب استقطاب البرامج الآلية الحديثة التي صممت أصلاً للتعامل مع الإنجليزية [18].

والفرق الرئيسي بين اللغة العربية وغيرها من اللغات هي أنها اشتقاقية أما اللغات الأخرى فهي لصقية. الشكلالتالي يوضح رسم تخطيطي ومثال على النظام العربي للاشتقاق :



شكل 1.2: يوضح نظام الاشتقاق العربي

المشكلة الرئيسية للخوارزمية المعتمدة على الجذر في عملية التصنيف الآلي للنصوص هي أن العديد من التهجئات المختلفة للكلمة ليس لديها تفسيرات دلالية متشابهة. أي بالرغم من أن هذه الكلمات تنشأ وتنتج من نفس الجذر, إلا أنها مختلفة في المعنى. لذا، استخدام الخوارزميات المعتمدة على الجذر في التصنيف تزيد من غموض الكلمة [19].

[18]سعد بن هادي قحطاني، تحليل اللغة العربية بواسطة الحاسوب، مركز اللغة الانجليزية، معهد الادارة ، الرياض.
[19]Kareem Darwish. "Building Shallow Arabic Morphological Analyzer in One Day", Associ-

إضافة إلى مشكل آخر وهو ارتباط بعض المفردات بالسياقات الزمنية فمصطلحات معاصرة مثل (عربة حافلة سيارة) لا يجوز الحكم بمدلولاتها بمعزل عن سياقها الزمني للنصوص التي وردت فيها فكلمة حافلة في الشعر الجاهلي لم يكن أبدا يقصد بها المركبة ذات الأربع عجلات .

6.9.2 الاستخدام المفرط للأساليب البيانية (المجاز- الكناية – الاستعارات):

تستخدم اللغة العربية هذه الأساليب بشكل واسع جدا بغية تحسين الخصائص البلاغية للنص أو التأثير العاطفي للمتلقي و بإمكان المتلقي أن يستوعب مضمون الأساليب البيانية بالاعتماد على خبراته المرجعية إلا أن جهاز الكمبيوتر لن يكون بالإمكان استعابهالأنه يعتمد في مجمل عملياته على التفسير المنطقي المباشر و بالتالي الأساليب البيانية تشكل صعوبة أمام التصنيف الآلي للنصوص العربية.

7.9.2 عدم وجود علامات التشكيل:

تعتمد اللغة العربية بالأساس على التشكيل و التنقيط و عادة ما تسمى تلك الحركات بالصوائت ، ونجد أغلبية الوثائق أو النصوص الالكترونية لا تتضمن تشكيلا و ربما السبب يعود في ذلك إلي عدم مرونة لوحة مفاتيح عند استخدام علامات التشكيل إذ نجد مفتاحين يعملان معا لكل علامة تشكيل،ونجد أغلبية النصوص المتعلقة بالأخبار و الروايات لا تعتمد على التشكيل.

ation for Computational Linguistics. 40th Anniversary Meeting,July 6-12, 2002 pp. 47-54.
University of Pennsylvania.

8.9.2 الأخطاء اللغوية الشائعة:

يتأثر التصنيف الآلي بهذه الأخطاء التي تخرج عن القاعدة اللغوية فالأفعال السداسية و الخماسية في اللغة العربية تبدأ بهمزة وصل إلا أن بعض النصوص نجدها مكتوبة بهمزة قطع مثل : استخرج و إستخرج، وحتى بالنسبة للكلمات المبدوءة بهمزة قطع تكتب بهمزة وصل مثل:أنباء و انباء، أيضا بالنسبة للكلمات التي النصوص تنتهي بياء نجدها في بعض الأحيان تكتب بياء مقصورة ، مثل:على وعلي، وبعض اللهجات تضيف همزة وصل في بداية أسماء الأعلام مثلا محمد تكتب امحمد.

فهذه الأخطاء تمثل عائق كبير جدا في عملية التصنيف الآلي إذ لا بد من تصحيحها يدويا أو بمصحح آلي بالإشراف للتأكد من أخذ الكلمة الصحيحة ثم تصنيفها.

بالرغم من هذه الأخطاء التي تعيق عملية تطبيق أهم التقنيات التكنولوجية على اللغة العربية أهمها عملية التصنيف الآلي للنصوص أو البيانات إلا أن البحوث مستمرة و هناك العديد من البحوث التي قدمت تقنيات حاسوبية (آلية) حاولت أن تعطي حلولا قيمة لعملية حوسبة اللغة العربية وكذلك تم تطويع العديد من المناهج الغربية و الخوارزميات حتى تناسب اللغة العربية .

خلاصة:

إن التعامل اليدوي مع هذا الكم الهائل من البيانات دون استخدام تقنيات حديثة يبعدنا عن التطور و الارتقاء إلى مستويات أداء أفضل وإدخال الآلات إلى العمل فقط، بل من الأفضل استخدام تقنيات وبرمجيات تخدم آلية تصنيف البيانات وتقدم لها ما يمكن أن تستفيد منه دون إضاعة الوقت و الجهد، لذا فإن هذا المجال تطور بشكل كبير في العشر سنوات الأخيرة و لعل ذلك يعود إلى الطلب الواسع لمستخدمي هذه التكنولوجيا.

إن التصنيف الآلي لهذه للنصوص وفق تقنيات التعلم والخوارزميات يقدم الحل الأمثل لمشكلة التزايد الهائل للبيانات النصية فهي تكنولوجيا جديدة تهدف إلى تنظيم وتصنيف النصوص المتراكمة التي لا يمكن بأي حال من الأحوال معالجتها يدويا.

الفصل الثالث : خوارزميات التصنيف الآلي

تمهيد

كلما تزايدت كمية البيانات أصبحت الحاجة ملحة لإيجاد تقنيات ذات كفاءة عالية، من أجل القيام بعملية تحليل هذه البيانات وتعد تقنية التصنيف من أهم تقنيات التنقيب في البيانات.

تعتبر تقنية التصنيف أو المصنف (classfier) هي طريقة منظمة systematic لبناء نماذج تصنيف، من خلال ادخال مجموعة من البيانات ولها عدة تقنيات من أهمها مصنفات أشجار القرار والمصنفات القاعدية المعتمدة على القاعدة (rule-based) ، والشبكات العصبية (network neural) ، ومكائن الإسناد الموجه (support vector machines) ، ومصنفات بيز البسيطة (bues classfier).

تستخدم كل تقنية من التقنيات السابقة لخوارزمية تعلم algorthm learning لتحديد نموذج يلاءم العلاقة بين مجموعة الصفات ومؤشر الصنف لبيانات الإدخال حيث يتم توليد النموذج من خلال خوارزمية تعلم ويجب على كل من النموذج والخوارزمية أن يتلاءما مع البيانات المدخلة بصورة جيدة والتنبؤ بصورة دقيقة لمؤشرات الصنف، لذلك فإن الهدف الرئيسي لخوارزمية التعلم هو بناء نماذج يمكن تعميمها، أي نماذج تتنبأ بشكل دقيق بتسميات أصناف سجلات غير معروفة مسبقا [1]

1.3 خوارزميات التصنيف:

يمكن تعريف خوارزميات التصنيف ضمن بيئة التعلم الآلي بأنها عملية توزيع البيانات والتي تدعى بيانات التدريب، ضمن فئات مختلفة حسب خاصياتها المشتركة بالاعتماد على آليات محددة

[1] محمد حسن عبد الله، تنقيب بيانات نتيجة التعليم الأساسي، مذكرة ماجستير في تقانة المعلومات، كلية الدراسات العليا، جامعة النيلين، 2016، ص53.

مسبقا، هذا وتعتبر عملية التصنيف أساسا لعملية التنبؤ من خلال النماذج التي يتم بناؤها أثناء عملية التصنيف والمرتبط بنوع المصنف المستخدم وفيما يلي نستعرض أهم المصنفات والخوارزميات الرائجة في هذا المجال:

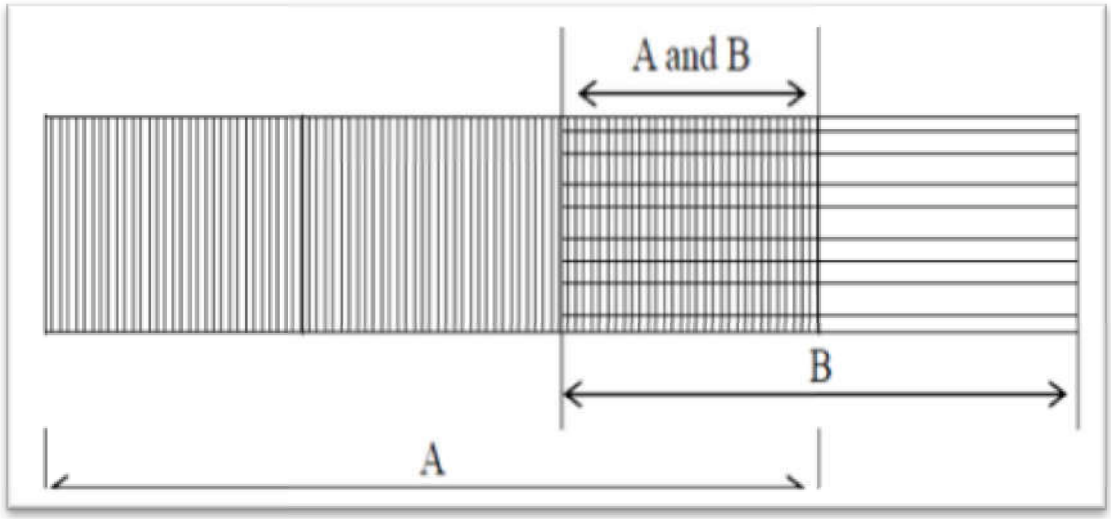
1.1.3 المصنف Naïve Bayes :

يستند هذا المصنف إلى نظرية بايز الاحتمالية، القائمة على مبدأ الاحتمال الشرطي الذي يعمد لحساب احتمال وقوع أحد الأحداث الاحتمالية بناء على وقوع حدث آخر وفق المعادلة التالية:

$$Prob(B\,gaiven\,A) = Prob(A\,and\,B)/Prob(A)$$

بحيث أن:

- وقوع الحدث B بناء على وقوع الحدث A : $Prob(B\,gaiven\,A)$
- احتمال وقوع الحدثين A و B معا : $Prob(A\,and\,B)$
- احتمال وقوع الحدثA : $prob(A)$



شكل 1.3: آلية عمل الاحتمال الشرطي بييز الاحتمالية.

يمتاز هذا التصنيف بالسرعة في بناء النماذج كما أنه يمتاز بأنه قابل للتوسع مع ازدياد بيانات التدريب ثنائية الفئات أو متعدد الفئات

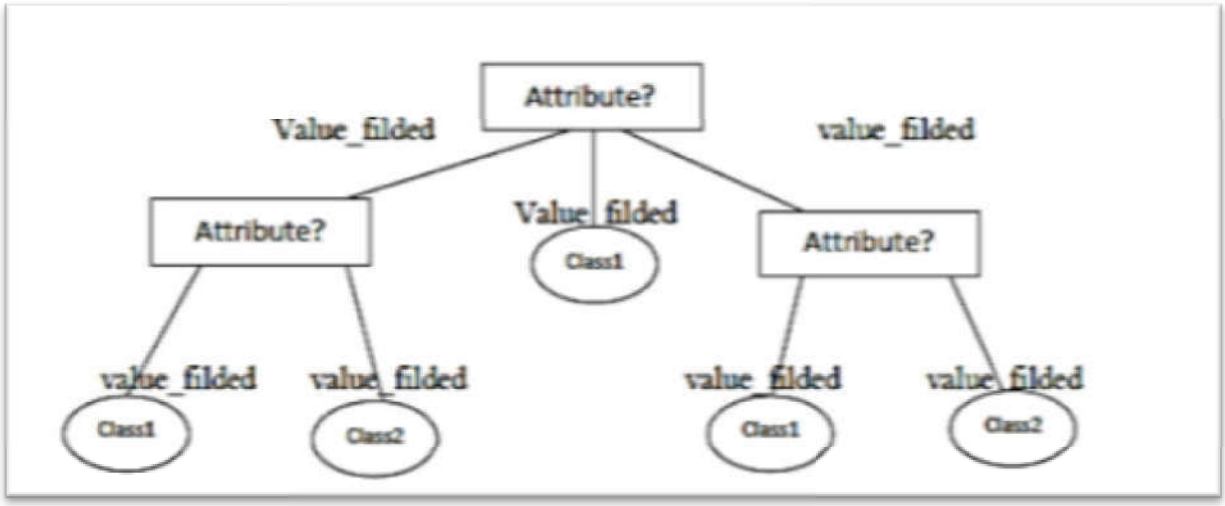
2.1.3 مصنف أشجار القرار:

تعرف شجرات القرار decision tree ببساطة على أنها طريقة بيانية أو نموذجية لتمثيل سلسلة من القواعد تقودنا إلى فئة أو قيمة، وهي من أهم النماذج التنبؤية للتنقيب في البيانات تتكون أشجار القرار من العقد nodes وهي الجذور أما عقدة الجذر هي قمة الشجرة التي تتفرع منها جميع التصنيفات، يتفرع من كل عقدة مجموعة من الفروع branches كل فرع يعبر عن أحد الإجابات الممكنة [2].

تصنف الحالات في أشجار القرار عن طريق فرزها على أساس قيم الصفة attribute كل عقدة

[2]عبد الحميد محمد العباسي، التنقيب في البيانات مجموعة محاضرات، قسم الإحصاء الحيوي والسكاني، معهد الدراسات والبحوث الإحصائية، جامعة القاهرة، مصر، 2013. ص 13.

داخلية في شجرة القرار تمثل صفة اختبار attribute test ، كل فرع يمثل قيمة العقدة وكل عقدة طرفية تمثل قيمة الاختبار ،value أعلى عقدة في الشجرة هي جذر الشجرة وتصنف الحالات انطلاقا من عقدة الجذر في شجرة القرار يتم تعيين كل عقدة ورقية كفئة class كما يتضح من خلال الشكل التالي:



شكل 2.3: شجرة قرار التصنيف

ومن أهم خوارزميات أشجار القرار نجد J48 ، ID3 ، CART ، CHAID ، QUEST .

3.1.3 المصنف J48 :

يندرج هذا المصنف ضمن خوارزميات أشجار القرار، والتي على اختلاف أنواعها تشابه إلى حد ما خوارزمية Naïve Bayes من حيث اعتمادها على الاحتمالات الشرطية مع اختلاف رئيسي، يكمن في أن هذه الخوارزمية تقوم بتوليد قواعد لاستخدامها كجمل شرطية لتحديد السجلات والأحداث الاحتمالية في شكل عبارة شرطية (IF....THEN) .

يستند هذا النوع من التصنيفات إلى هيكلة شجرية مؤلفة من عقد رئيسية تدعى الجذر(Root) ومجموعة عقد داخلية ،(Nodes) ومجموعة عقد نهائية (Terminals) بحيث يتضمن كل من الجذر ومجموعة العقد الداخلية القاعدة ،Rule التي تحدد المسار للفروع المرتبطة بما يسمح في النهاية بالوصول إلى النتيجة النهائية.

تعمد هذه الخوارزمية إلى تقسيم مجموعة بيانات التدريب المراد تصنيفها إلى مجالات متقاطعة ذات تسمية أو قيمة أو عملية لتوضيح وشرح البيانات داخل هذا المجال وذلك بالاعتماد على معيار يستخدم لحساب أو تعيين أفضل المعايير لتجزئة هذا المجال من البيانات التي يتم تدريبها والذي يدعى التابع الإحصائي والمعرف بالمعادلة التالية

بحيث: $Gain(S,A) = Entropy(S) - \sum_{v \in Value(A)} \frac{|S_v|}{|S|} Entropy(S_v)$

- مجموع بيانات التدريب (S) ومجموعة المعايير (A)
- values (A) جميع القيم الممكنة للمعيار A
- Sv: مجموعة جزئية من المجموعة (s) المنتمية للمعيار (A) وذات قيمة (v)
- تابع العشوائي (Entropy) يعبر هذا التابع عن عشوائية المعطيات وتتراوح قيمته بين[0-1] ويعبر عنه بالمعادلة $Entropy(s) = \sum_{i=1}^{c} -Pi \log^2 li$

حيث يعبر المتغير (pi) عن احتمالي انتماء مجموعة البيانات (S) الى الفئة (i) إن بناء الهيكلية الشجرية لأغلب أشجار التصنيف يتم من الأعلى إلى الأدنى، والاستفادة من طريقة البحث لتحديد قيم التابع (Gain) لمختلف المعايير، واختيار المعايير ذات القيم الأعلى للتابع ،(Gain) ومن ثم إعادة العملية لباقي المعايير وصولا لمجموعات جزئي متجانسة.

تتصف النماذج التي يولدها هذا النوع من التصنيفات بالدقة العالية والسرعة في بناء النموذج، كما يمكن تطبيقها على البيانات متعددة الفئات، وبأنها قابلة للتأويل والفهم من خلال تحليل شجرة

شجرة القرار ومعاينة الرسم البياني المولد عن بناء النموذج.

4.1.3 المصنف MLP (Multi Layer Perception)

تعتبر الشبكات العصبونية من أكثر الخوارزميات تعقيدا إذ تتطلب الكثير من البيانات لتدريبها، ووقتا إضافيا إلا أنها تتصف بميزة هي توقع الحالات الجديدة بسرعة فائقة، كما أنها لا تعمل إلا مع المعطيات الرقمية أي لا بد من تحويل البيانات إلى أرقام.

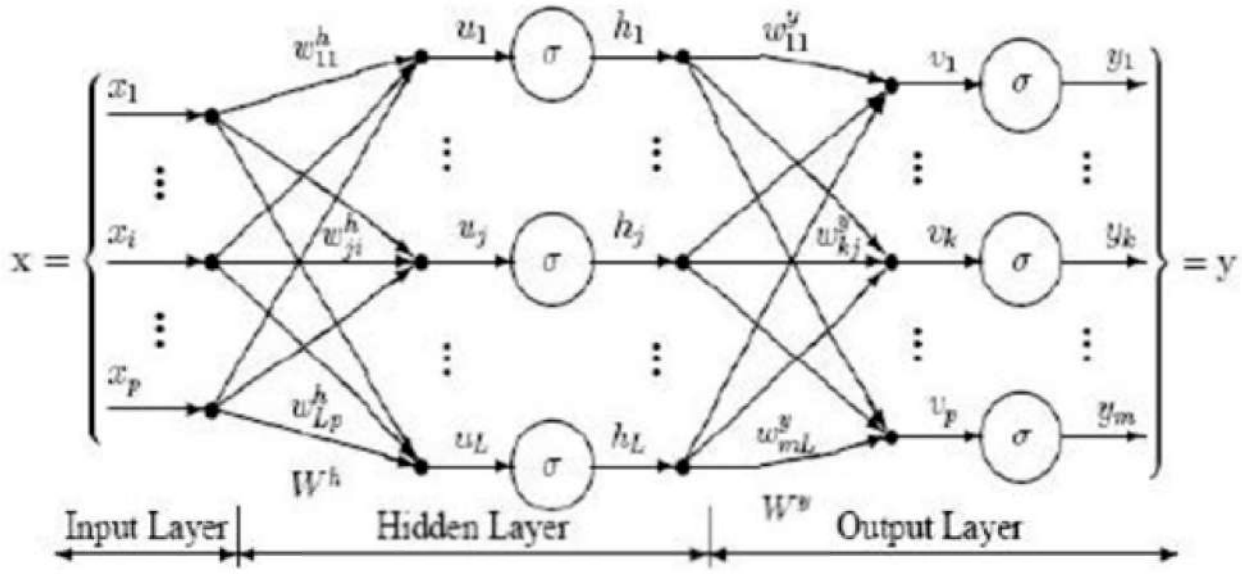
تتألف الشبكة العصبونية الأكثر شيوعا من ثلاثة وحدات طبقة لوحدات الدخل مرتبطة بطبقة من الوحدات المخفية التي ترتبط بدورها بطبقة لوحدات الخرج [3].

تكمن الفكرة الأساسية لآلية عمل هذه الشبكة في خلق آلية عمل مشابهة للنظام العصبي العضوي المشكل من خلايا عصبية مترابطة، تعمل مجتمعة لحل المشكلات التي تواجه المستخدم تمتاز بقدرتها على تعلم الاستجابة الصحيحة للمتغيرات المختلفة.

ومع الاستفادة من القدرات البرمجية يمكن محاكاة عمل الخلية العصبية العضوية، ومن ثم ايجاد شبكة من هذه العصبونات لتشكيل نظام عصبي صناعي متكامل، عند تدريب البيانات يتم ادخال البيانات عبر طبقة الإدخال (Input Layer) وتتم معالجتها ضمن الطبقات المخفية (Hidden Layers) وعرضها بالنهاية عبر طبقات الخرج (Output Layer) [4].

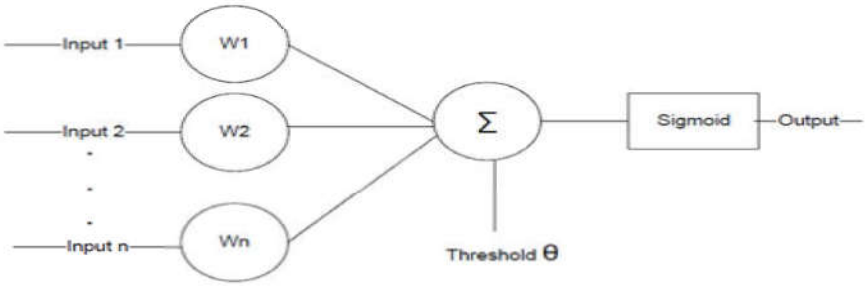
[3] فادي خلوف، تطوير آليات جديدة للتنقيب في المعطيات لإدارة علاقات الزبائن في بيئة مصرفية، مجلة جامعة دمشق للعلوم الهندسية، المجلد 26، العدد 1، 2010، ص 90.

[4] جلال الضاهر، تصميم نموذج نظام دعم القرار لإدارة الموارد البشرية بالاعتماد على تقنيات الذكاء الصنعي، مذكرة ماجستير، الجامعة الافتراضية السورية، 2013/ 2014، ص 18.



شكل 3.3: الطبقات الثلاث لخوارزمية Multi Layer Perceptron

يظهر في الشكل (3) الطبقات الثلاثة المتعارف عليها في بنية الشبكات العصبونية حيث تتألف من طبقة واحدة أو أكثر من العصبونات الصنعية المتوازية لكل عصبون كما يظهر في الشكل (4) عدد N من المدخلات ذات الوزن لكلW منها بالإضافة لمخرج واحد فقط يقوم كل عصبون بدمج المدخلات مختلفة الأوزان من خلال جمعهم سوية و بالاستناد إلى حد العتبة θ يقوم بتحديد قيمة الخرج (Output) .



شكل 4.3: بنية العصبون الصنعي الواحد

لشرح آلية عمل هذه الخوارزمية مبسطة لا بد من تعريف المتغيرات التالية:

- المدخلات(x1 , x2 , ... , xn) ذات الأوزان (w1 , ... , wn)
- التابع μ تابع يعبر عن احتمالية التنشيط (activation potential).
- تابع حد العتبة θ (threshold)
- تابع الخرج y (output)
- تابع التنشيط f (activation function)

وعليه يمكن تعريف تابع احتمالية التنشيط بالمعادلة :

$$u = \sum_{i=1}^{N}(w_i x_i) \qquad (1)$$

$$y = f(u - \theta) \qquad (2)$$

$$y = f\left(\sum_{i=1}^{N}(w_i x_i)\right) : w_0 = \theta, \quad x_0 = -1 \qquad (3)$$

تعبر المعادلة (3) عن الشرط الخاص بكل عصبون صنعي لجميع الطبقات مع اختلاف المتغيرات لدى الانتقال من طبقة إلى أخرى أو اختلاف الصف ضمن الطبقة الواحدة.

يتم تدريب شبكة البيانات بداية باختيار مجموعة أوزان بشكل عشوائي وحدود عتبات داخلية ومن ثم تعديل قيمة الأوزان بعد كل محاولة بشكل تكراري وصولا للأوزان المناسبة بحيث تصبح ضمن حد مقبول [5].

إن قوة تصنيف هذه الخوارزمية غير الخطية مهدت لاستخدامها بشكل واسع في عدة مجالات، كتشخيص النطق تشخيص الصور إضافة لبرامج البرمجة، كما أثبتت هذه الخوارزمية قدرتها العالية

[5] جلال الضاهر، تصميم نموذج نظام دعم القرار لإدارة الموارد البشرية بالاعتماد على تقنيات الذكاء الصنعي، ص 20.

في عمليات التخمين والتقريب ومعالجة الإشارات البيولوجية والاتصالات والالكترونية وصولا إلى علوم الفضاء.

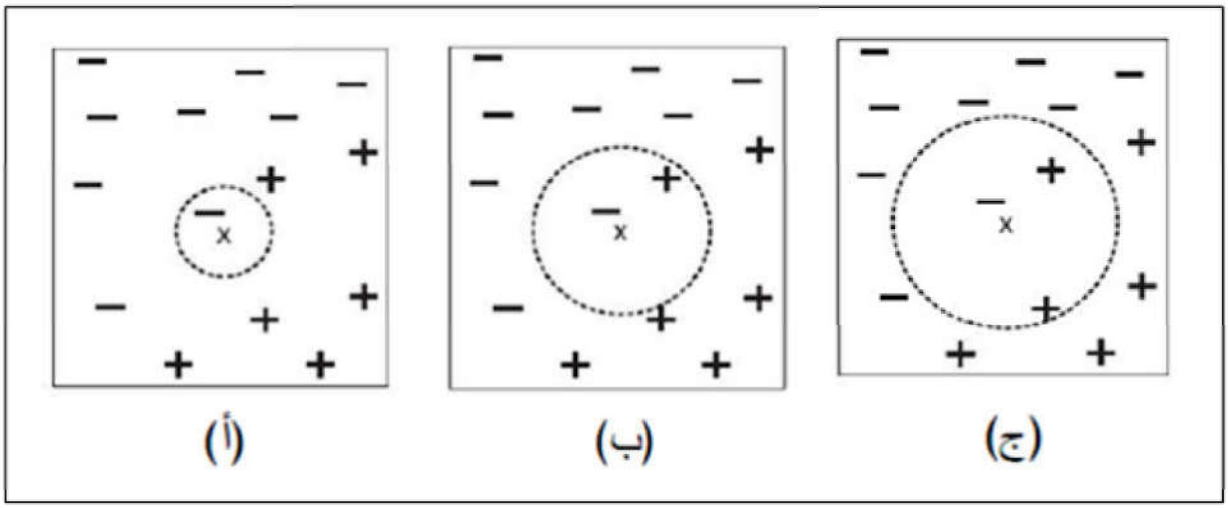
5.1.3 خوارزمية KNN :

طريقة الجار الأقرب(K-Nearest Neighbor) هي تقنية تنبؤية مناسبة لنماذج التصنيف إذ تمثل (K) عدد الحالات المتشابهة أو عدد العناصر(items) في المجموعة، وتعد معطيات التدريب في طريقة الجار الأقرب هي النموذج فلا يتم بناؤه، عندما يتم تقديم حالة جديدة للنموذج تبحث الخوارزمية في المعطيات كلها لإيجاد مجموعة جزئية (subset) من الحالات التي هي أكثر تشابها وتستخدمها لتوقع الخرج، هناك محددان أساسيان في خوارزمية الجار الأقرب:

- عدد الحالات الأقرب ليتم استخدامها .(K)
- وحدة قياس (metric) لقياس التشابه [6].

والشكل التالي يوضح عمل هذه الخوارزمية حيث تظهر النقطة المجاورة الأقرب لإحدى نقاط البيانات المراد تصنيفها (X) ضمن الشكل (أ) أما الشكل (ب) يظهر النقطتين المجاورتين للنقطة (X) ويظهر الشكل (ج) النقاط الثلاثة المجاورة للنقطة (X) .

[6] فادي خلوف، تطوير أليات جديدة للتنقيب في المعطيات لإدارة علاقات الزبائن في بيئة مصرفية، مرجع سابق، ص 88



شكل 5.3: مثال على توزع القيم عند استخدام التصنيف KNN

إن النقطة (X) في حالة الشكل (أ) تنتمي إلى الصف السالب، و في حالة الشكل (ج) تنتمي إلى الصف الموجب وذلك حسب نظام التصويت للأغلبية (Majority Voting Scheme)، أما في حالة الشكل (ب) فإنه يتم اختيار الصف بناء على وحدة القياس (metric) ليتم تصنيف النقطة على أساسه ، يتم اختيار العدد (K) بشكل مناسب مع عدد البيانات بحيث يتم التغلب على التراكب الناتج عن عملية التصنيف والتي تزداد مع ازدياد شذوذ البيانات وعدم تناسقها.

يمكن تلخيص خطوات خوارزمية الجار الأقرب كالتالي:

1. تحديد عدد الجيران الأقرب ولتكن (k) .
2. حساب المسافة التقليدية بين السجل المستكشف وأقرب الجار.
3. ترتيب المسافات بإعطاء الرتب لها من اصغر مسافة إلى أعلى مسافة، ثم تحديد الجيران الأقرب بالاستناد إلى مسافة حد أدنى (k-th) .

4. جمع الصنف للجار الأقرب.

5. حساب الوسط الحسابي للجيران الأقرب كقيمة تنبؤ لحالة السجل المستكشف.

6. نستمر بتقدير دالة الهدف للسجلات المستكشفة.

7. نحسب قيمة RMSE (جذر معدل الخطأ تربيع) لكل قيمة K .

6.1.3 المصنف SVM :

تعد SVM من أشهر طرق التصنيف الآلي والتي تعتمد على ايجاد منحنى أو مستوى فاصل، يفصل العينات المدخلة عن بعضها البعض وتتميز باستخدامها في تصنيف المسائل ذات الفئات الثنائية حصرا، حيث ترمز للعينات الايجابية بـ (1) وللعينات السلبية بـ (-1)، تقوم الخوارزمية بحساب المستوى الفاصل أو مجموعة المستويات الفاصلة في بعد يختلف طوله عن طول بعد متجه خصائص البيانات المدروسة، وتحدد دقة الخوارزمية بقدرتها على الفصل بين النوعين، بحيث تكون أقرب عينتين من كلا النوعين أبعد ما يكون عن بعضهما البعض وندعو هذا المستوى الفاصل بالهامش الفصل فكلما زاد هامش الفصل كلما قل الخطأ عند التعميم على مجموعة بيانات جديدة [7].

يعتبر هذا المصنف أحد أقوى المصنفات التقليدية لامتلاكه آلية عمل تدمج كلا من خوارزمية الشبكات العصبونية مع خوارزمية الشعاع الأساسي لإيجاد أفضل سطح فاصل بين بيانات التدريب، يمتاز هذا المصنف بالمرونة، قابلية التوسع والسرعة في الأداء مما يعطيه الأفضلية في معالجة مسائل التشخيص المتنوعة كما يتميز هذا المصنف بقدرته على معالجة المعطيات ذات عدد

[7] بسام الديب، تصنيف النصوص العربية باستخدام الخصائص الغرضية في قواعد البيانات، مجلة جامعة البعث، المجلد 38، العدد 15، 2016، ص 116.

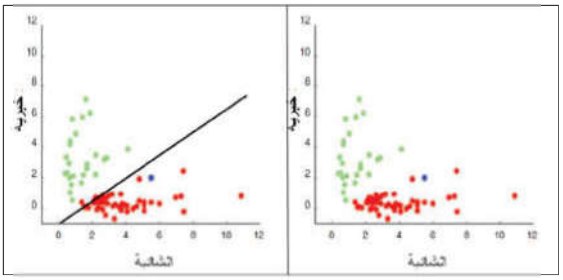
كبير من المعايير مقارنة بعدد سجلات البيانات المتواجدة [8].

تحوي SVM على أربع مفاهيم أساسية:

***السطح الفائق Hyper plane :**

على سبيل الافتراض لدينا مجموعة من بيانات التدريب Training Dataset كل حالة منها تمثل ملفات نصية، يمكن أن تكون إما أساليب إنشائية أو خبرية هذه الحالات تحوي على واصفتين عدديتين الأولى تمثل تكرار الأداة (يا) والثانية تمثل تكرار الأداة (لم) صنف كل حالة يمكن أن يكون أما أن يكون خبري أو إنشائي وكما يتضح من خلال الشكل (6.3) النقاط الخضراء هي أشعة تمثل نصوص الأساليب الخبرية والحمراء تمثل والحمراء هي أشعة تمثل نصوص الأساليب الإنشائية يمكن ملاحظة أن نصوص الأساليب الخبرية تجتمع في الجهة العلوية اليسرى ونصوص الأساليب الإنشائية تجتمع في الجهة السفلية اليمنى، يمكن الفصل بين هذه الأشعة بواسطة خط فاصل يفصل بينهما.

بهذه الحالة يمكن التنبؤ بصنف نص جديد مجهول الصنف بمعرفة موقعه من هذا الفضاء على أي جهة من الخط الفاصل اصطلح على تسمية هذا الخط بالسطح الفائق ومهمة SVM ايجاد هذا السطح ويتضح ذلك من خلال الشكل التالي:

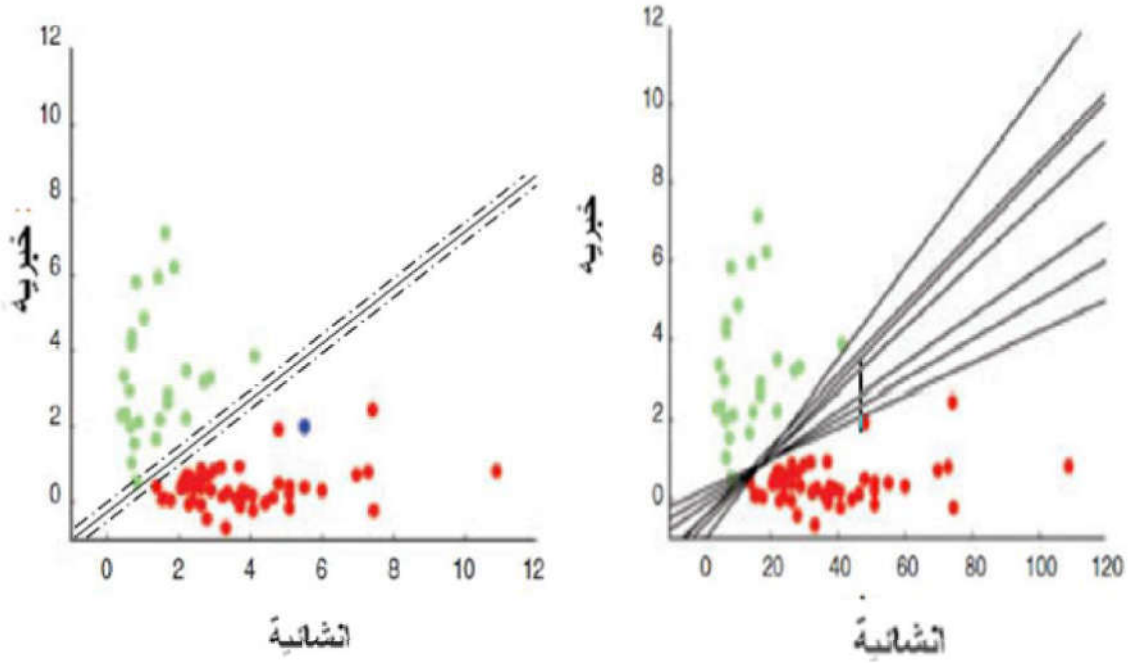


شكل 6.3: يوضح السطح الفائق Hyper plane

[8] جلال الضاهر، تصميم نموذج نظام دعم القرار لإدارة الموارد البشرية بالاعتماد على تقنيات الذكاء الصنعي، مذكرة ماجستير، اشراف طاهر رجب قدار، 2013/2014 ، ص22.

***الهامش الأكبر للسطح The maximum-margin hyper plane :**

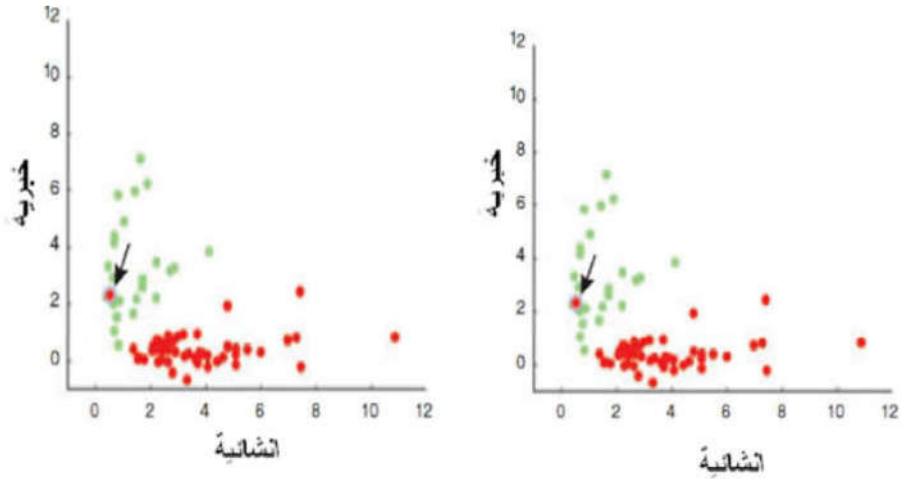
يمكننا رسم عدد لامتناهي من السطوح الفائقة تقوم باختيار السطح الذي يكون في المنتصف تماما من أجل الحصول على أفضل النتائج في عملية التصنيف ويتم اختيار السطح الذي يملك أكبر هامش بينه وبين اقرب شعاع في هذا الفضاء كما يتضح من خلال الشكل التالي: ***الهامش**



شكل 7.3: الهامش الأكبر للسطح The maximum-margin

المرن Soft Margin :

في أغلب الحالات لا تتوضع الأشعة في الفضاء بحيث يمكن فصلها بواسطة سطح فائق لذلك تم تعديل خوارزمية SVM لحل هذه المشكلة بإضافة ما يعرف بالهامش المرن بحيث تسمح لبعض الأشعة بأن تتوضع في المكان الخاطئ دون أن تؤثر على النتيجة النهائية كما يتضح من خلال الشكل التالي:

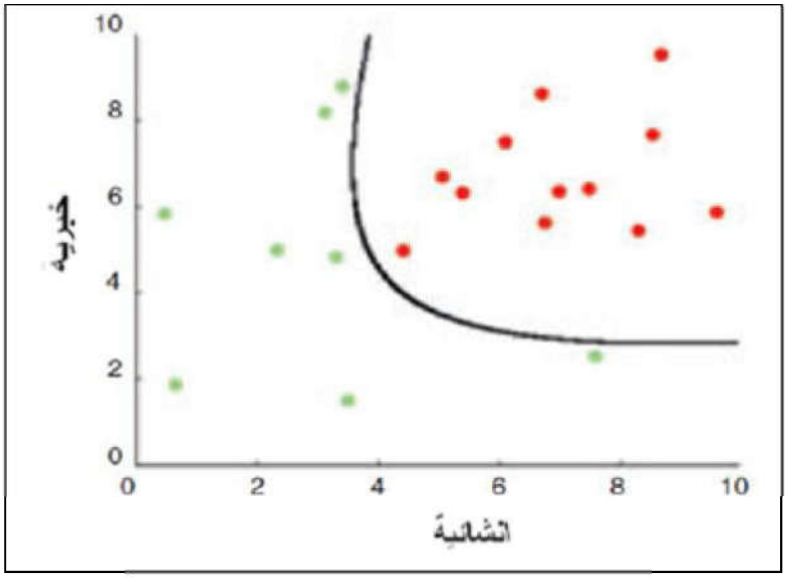


شكل 8.3: الهامش المرن Soft Margin

***تابع النواة Kernel Function :**

في بعض الحالات لا يمكن إن يتم فصل الأشعة بواسطة سطح فائق لكن يمكن حل هذه المشكلة بتربيع هذه القيم وإضافة القيم المربعة كأبعاد جديدة في الفضاء وفي هذا الفضاء الجديد تتوضع النقاط بحيث يمكن فصل الأشعة بواسطة سطح فائق وعملية التربيع هي ابسط أنواع توابع النواة لكن غير مستخدم لأنه يضاعف أبعاد الفضاء ويجعل تعقيد عملية التصنيف كبيرة جدا حيث تزداد درجة التعقيد مع زيادة أبعاد الفضاء فدور توابع النواة هو القيام بعملية التوفيق بين الأداء والدقة لعملية التصنيف والشكل (9.3) يوضح الإسقاط للسطح الفائق في الفضاء ذو الأبعاد العالية على الفضاء ذو البعدين [9] حيث يظهر بشكل منحنى:

[9] مصعب شاهين، شادي صالح، تطوير نظام لتصنيف المستندات العربية، مشروع تخرج، إشراف ناصر ناصر، فسم البرمجيات ونظم المعلومات، كلية الهندسة المعلوماتية، 2011/2012، ص 41.



شكل 9.3: تابع النواة Kernel Function

7.1.3 خوارزميات العنقدة (clustrering algorithms) :

يمكن تعريف العنقدة بأنها التقنية التي تضع الكيانات المتشابهة داخل المجموعة نفسها بالاعتماد على صفات المعطيات المتشابهة في حين توضع الكيانات المختلفة في مجموعات منفصلة، يقاس التشابه (similarity) بواسطة تابع قياس المسافة (function distance measure) لذلك فإن معنى العناقيد (clusters) يعتمد على تابع المسافة المستخدم [10].

وتتحقق جودة العنقدة إذا كانت عناصر العنقود الواحد متشابهة وذات علاقة قوية مع بعضها البعض، وإذا بعدت المسافة بين العنقود والآخر.

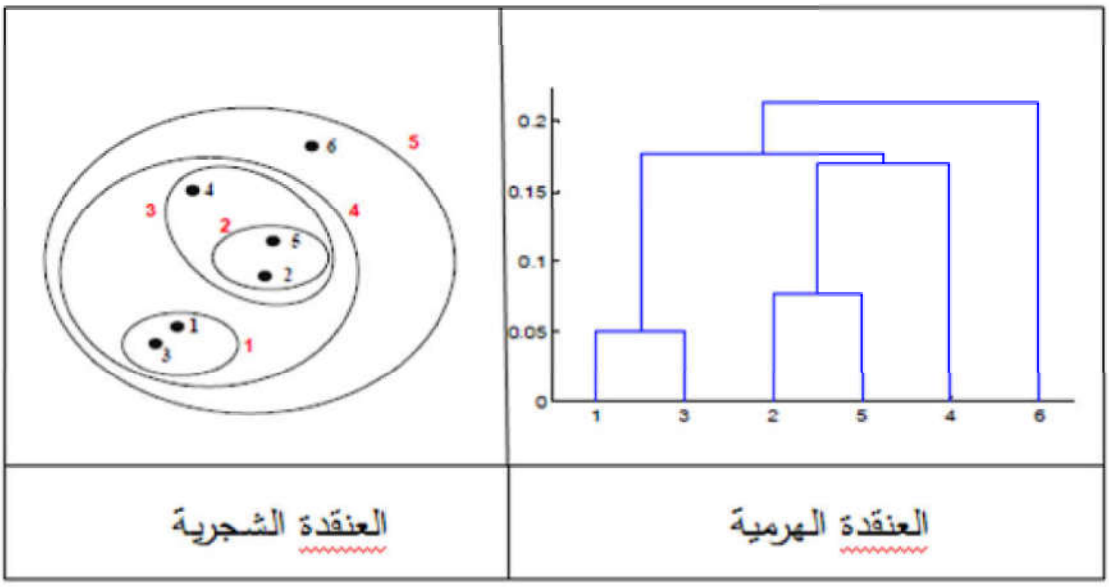
إن عملية العنقدة بحد ذاتها ليست عبارة عن خوارزمية خاصة يمكن إجراءها لمختلف أنواع المعطيات وإنما عبارة عن عملية يمكن إنجازها من خلال عدة خوارزميات متفاوتة بشكل كبير

[10] فادي خلوف، تطوير آليات جديدة للتنقيب في المعطيات لإدارة علاقات الزبائن في بيئة مصرفية، مرجع سابق، ص 88.

لاسيما في طريقة بناء مجموعة وتشكيلها، ولا يمكن الفصل بشكل كامل بين أنواع وتقسيمات خوارزميات العنقدة نظرا لوجود تقاطع وتداخل فيما بينها ولكن يمكن تصنيف عمليات العنقدة حسب التقسيمات التالية:

***العنقدة القائمة على ارتباط البيانات (Connectivity-based Clustering):**

يطلق عليها أيضا العنقدة الهرمية وهي قائمة على إن النقاط الإحصائية تكون أكثر ارتباطا مع النقاط القريبة وتتناقص قوة الارتباط مع ازدياد بعد النقاط عن بعضها، خوارزميات هذه العنقدة تقوم بتشكيل المجموعة من خلال ربط نقاط البيانات بعضها ببعض بالقدر الذي يسمح به الحد الأقصى للمسافة، ويمكن أن تأتي هذه الخوارزمية على شكل بنية شجرية واسعة وممتدة من أجزاء المجموعات ومن ثم الدمج بينها بناءا على المسافة المسموح بها كما يتضح من خلال الشكل (10).[11]

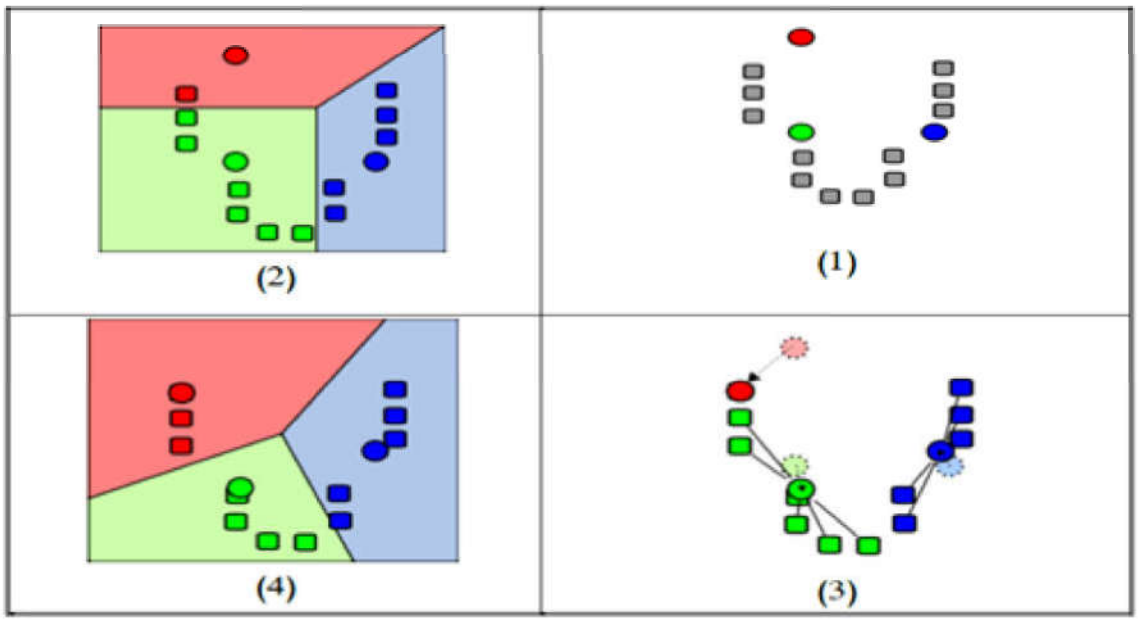


شكل 10.3: أنواع العنقدة القائمة على ارتباط البيانات.

[11] هالة حسن محمود، تعدين بيانات التمويل الأصغر باستخدام تقنيات التصنيف والعنقدة، مرجع سابق، ص 78.

***العنقدة القائمة على النقاط المركزية (Centroid-based Clustering):**

يعمل هذا النوع من العنقدة إلى تقسيم نقاط المعطيات إلى عدد ثابت(K) من مجموعات وتعتبر خوارزمية simple K-means هي المثال النموذجي وأكثر تطبيقا لهذا النوع من العنقدة، إذ تقوم بتقسيم البيانات إلى مجموعات تحوي نقاطا مركزية تنسب لها والشكل (11.3) يوضح خطوات عمل هذه الخوارزمية.



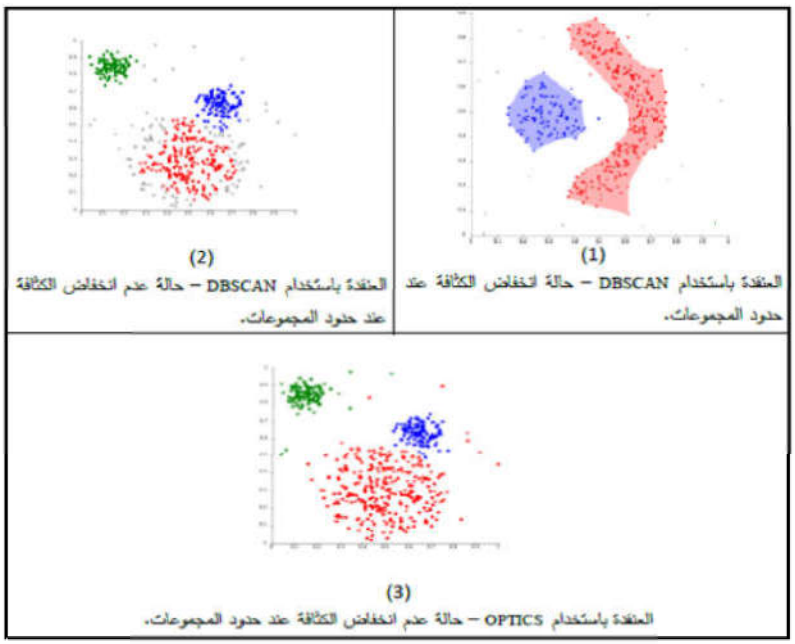
شكل 11.3: خطوات عمل خوارزميات العنقدة القائمة على النقاط المركزية

يوضح الشكل(3-11) طريقة عمل هذه الخوارزمية أولا تم توليد ثلاثة نقاط من مجموعة البيانات التي تم تقسيمها إلى ثلاثة مجموعات كما يتضح من خلال الصورة رقم (1)، ثم تم نسب جميع نقاط البيانات إلى المراكز الثلاثة حسب قربها منها كما هو مبين في الصورة رقم (2)، يتم بعد ذلك إعادة حساب نقاط مركزية للمجموعات بناء على نقاط البيانات الموجودة في المجموعة بحيث تصبح النقاط المركزية الجديدة نقاطا متوسطة كما هو بارز في الصورة رقم (3)، في المرحلة

الأخيرة تعاد الخطوة (2) و(3) حتى يتم إثبات النقاط المركزية ويظهر ذلك الصورة رقم (4).

***العنقدة القائمة على كثافة المعطيات(Density-based Clustering):**

في هذا النوع من العنقدة تعرف المجموعة كنطاقات من نقاط المعطيات ذات كثافات متفاوتة لتوزيع المعطيات، وبالتالي فإن نقاط المعطيات المتناثرة يتم اعتبارها عادة كنقاط شاذة ،(noise) إن المبدأ الأساسي الذي يقوم عليه هذا النوع من العنقدة هو ربط نقاط المعطيات ضمن عتبات مسافات وأبعاد معينة وذلك بشكل مشابه للعنقدة الهرمية إلا انه يربط النقاط التي تلبي معيار الكثافة (Density criterion) الذي يعرف بأنه الحد الأدنى لعدد نقاط البيانات ضمن نصف القطر [12]. أشهر خوارزميات هذا النوع من العنقدة خوارزمية OPTICS وخوارزمية DBSCAN، والشكل (12.3) يوضح طريقة العمل بكلتا الخوارزميتين:



شكل 12.3: حالات عمل خوارزميات العنقدة القائمة على كثافة المعطيات

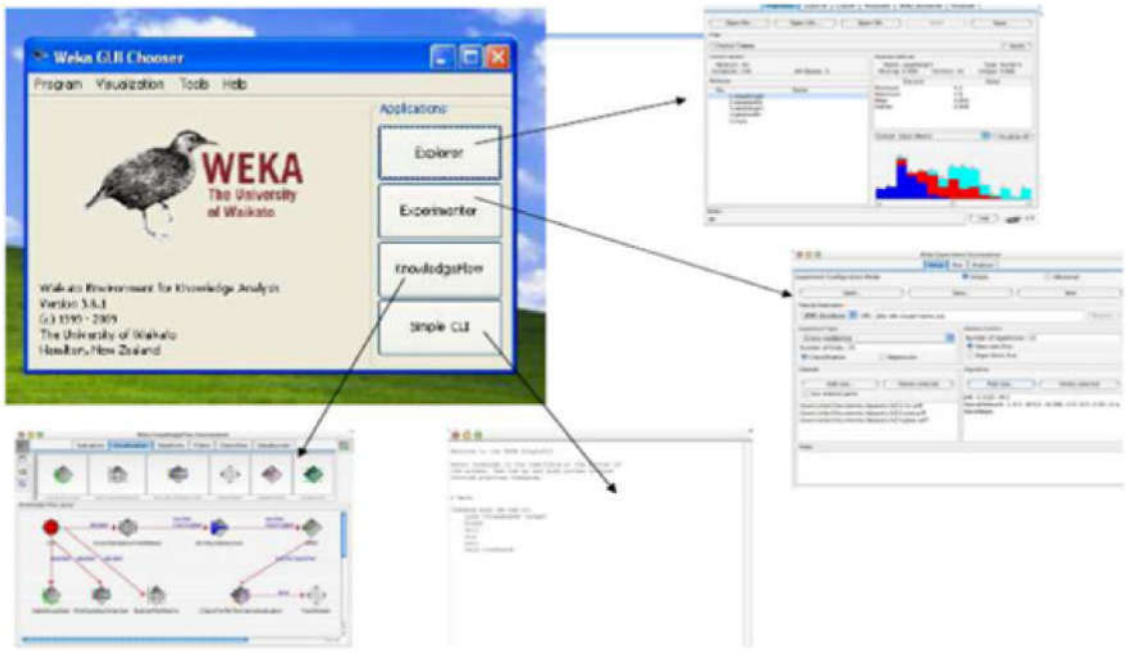
[12] جلال الضاهر، تصميم نموذج نظام لدعم القرار لإدارة الموارد البشرية بالاعتماد على تقنيات الذكاء الصنعي، مرجع سابق، ص 30.

اقتصرنا على ذكر أشهر الخوارزميات في مجال التصنيف الآلي، إلا أن هناك العديد منها والتي قد تنطوي ضمن أشجار القرار واهم خوارزميات انتقاء المعايير وخوارزميات أخرى نجدها كلها متاحة ضمن برامج التصنيف الآلي مثل برنامج WEKA الذي اعتمدناه في هذه الدراسة نظرا لاحتوائه أشهر خوارزميات التصنيف وقدرته على معالجة كميات هائلة من البيانات بسرعة ودقة في النتائج المتوصل لها.

2.3 تطبيق WEKA (Environment for Knowledge Analysis Waikato) :

تطبيق حديث ومتجدد يؤمن إجراء عمليات التنقيب في البيانات بسهولة، له العديد من الميزات التي تجعله أحد أهم التطبيقات المعروفة للتنقيب في البيانات في الوقت الحاضر من خلال تزويده بمجموعة ضخمة من خوارزميات التعلم الآلي التي يمكن أن تطبق مباشرة على مجموعة البيانات أو يتم استدعاؤها من التعليمات البرمجية الخاصة بالجافا، ويحتوي أيضا على أدوات تجهيز البيانات وتحويلها.

تم تطوير تطبيق WEKA في مخابر جامعة (Waikato) في نيوزلندا بالاعتماد على لغة البرمجة (JAVA) وكانت النسخة الأولى عام 1996 ويعتبر أحد أفضل برمجيات التنقيب في المعطيات المعروفة حتى الآن حيث يعمل بالتوافق مع كافة أنظمة التشغيل المعروفة (Windows , Macintosh , Linux) ويمكن الوصول والتعامل مع كافة عناصر هذا التطبيق من خلال لغة البرمجة جافا.



شكل 13.3: صورة توضيحية لواجهة عمل تطبيق WEKA

1.2.3 طريقة استخدام برنامج WEKA :

أسهل طريقة لاستخدام (WEKA) هو من خلال انتقاء إحدى الواجهات الأربعة التي يتيحها البرنامج وهي:

- **المستكشف (Explorer)** : وهذا الأخير يتيح الوصول إلى جميع منشآت البرنامج إذ تقوم بعرض جملة من الخيارات على شكل قوائم تضم كل من: تحضير البيانات ،(Preprocess) التصنيف (Classfiication)، التجزئة (Clustering)، التجميع ،(Association) اختيار المعايير(Attribute Selection)، التمثيل البياني (Visualization).
- **المختبر (Experimental)** : يتضمن عمليات اختبار وتقييم خوارزميات التعلم الآلي.

- **تدفق المعرفة (Knowledge Flow)**: يتضمن إمكانية التصميم المرئي لعمليات استكشاف المعرفة.
- **واجهة بسيطة غير رسومية (Simple CLI)**: يمكن من خلالها كتابة الأوامر التنفيذية وهي بدورها تسمح بالوصول لكل المميزات بالنظام.

2.2.3 كيفية ادخال البيانات إلى برنامج WEKA :

يتم تصدير البيانات إلى البرنامج على شكل اكسل أو ملفات نصية بشرط أن لا تحتوي هذه البيانات على فواصل منقوطة أو علامات التنصيص أو فراغات وحتى يتقبلها البرنامج لا بد من إضافة أوامر خاصة به لتعريف حقول البيانات(Attributes) في رأس الملف وذلك لتعريف كل عمود من الأعمدة في هذه البيانات ومن ثم حفظه بصيغة (arff) .
يمكن تحميل البيانات إلى التطبيق من عدة مصادر تتضمن:

1. الملفات ذات الصيغ المعرفة لهذا التطبيق(bsi , names , csv , data , xrff , arff)
2. مختلف قواعد البيانات.
3. مسارات ونطاقات عناوين الحواسب ضمن الشبكات الحاسوبية(URLs).

3.2.3 المصطلحات الرئيسية للبرنامج (Basic Terms):

فيما يلي تعريف وتوضيح لبعض المصطلحات التي سيتم استخدامها والمرتبطة بتطبيق WEKA علما أن اغلب هذه المصطلحات ذات تعاريف رياضية وإحصائية منفصلة عن هذا التطبيق ولكن ضمن هذه الدراسة نعتمد على تعريفها وبيان أهميتها ضمن التطبيق المستخدم.

- مصفوفة الشك (Confusion Matrix):

مصفوفة الشك أو كما تدعى أيضا مصفوفة الخطأ (Error Matrix) هي عبارة عن جدول يسمح بتشخيص مدى أداء خوارزمية التصنيف من خلال توزع الصفوف داخل هذه المصفوفة بحيث يمثل العمود الفئات المتوقعة (Predicted class) ويمثل السطر النتيجة الحقيقية (class Actual) كما في الشكل(14.3):

		التصنيف المتوقع	
		A	B
التصنيف الحقيقي	A	2321	572
	B	245	1248

شكل 14.3: مصفوفة الشك خاصة بتطبيق WEKA

إن قطر المصفوفة يمثل القيم المتوقعة بشكل صحيح ومطابقة للواقع فالخلية $([A, A] = 2321)$ تمثل عدد الحالات التي توقع خلالها التصنيف القيمة (A) متغير قيمته الأصلية هي (A) أما في الحالة $([A, B] = 572)$ فإن التصنيف توقع القيمة (A) بمتغير قيمته الأصلية هي ،(B) وبالتالي فإن دقة التصنيف ناتجة من مجموع قيم القطر بالنسبة للعدد الكلي.

- معامل الإحصاء كابا (Kappa Static) :

يعتبر هذا المعامل أحد أهم المؤشرات لقوة التصنيف من خلال قراءته لمصفوفة الشك واختصار دقة التصنيف الموجود داخلها إلى رقم يتراوح مابين (0 إلى 1) من خلال تطبيق المعادلة التالية:

بحيث: $k = \frac{P(a)-P(e)}{1-P(e)}$

P (a) : نسبة عدد حالات التطابق

P (e) : نسبة عدد الحالات التصادفية

- **مقاييس الدقة:** وهي من بين المقاييس التي يتيحها البرنامج حيث تقوم باحتساب دقة المصنفات بالنسبة المئوية للحالات التي صنفت بشكل صحيح مع تحديد نسبة الحالات الخاطئة وارتباطها بالصف الخاطئ ونستعرضها فيما يلي:

- **مقياس Precision:** : هو النسبة المئوية التي يتنبأ النموذج فيها بشكل صحيح عند اتخاذ القرار وتعطى بالمعادلة التالية:

$$Precision = \frac{TP}{TP+FP}$$

- **مقياس Recall :** هو النسبة المئوية التي تم تحديدها بشكل صحيح من كل الايجابيات الموجودة ويعطى بالمعادلة التالية: : هو مقياس توافقي من Precision و فيRecall مقياس واحد وتتراوح قيم مجال المقياس بين (0 و1) ويكون المصنف جيد كلما اقترب هذا المقياس من الواحد ويعطى بالمعادلة التالية:

$$Recall = \frac{TP}{TP+FP}$$

- **مقياس F-measure :** : هو مقياس توافقي من Precision و فيRecall مقياس واحد وتتراوح قيم مجال المقياس بين (0 و1) ويكون المصنف جيد كلما اقترب هذا المقياس من الواحد ويعطى بالمعادلة التالية:

$$F - measure = \frac{2(Precision)(Recall)}{Precision+Recall}$$

- **مقياس ROC :** وهو اختصار لـ Receiver Operating Characteristic لقي اهتماما متزايدا في تقييم أداء الخوارزميات وهو عبارة هن منحنى حيث يمثل المحور Y معدل القيم الايجابية الصحيحة والمعطى بالمعادلة التالية

$$TPR = \frac{TP}{TP+FN}$$

بالمقابل المحور X معدل القيم الإيجابية الخاطئة والمعطى بالمعادلة التالية:

$$FPR = \frac{FP}{TN+FP}$$

يظهر المنحنى كفاءة وفعالية الخوارزمية التي يتم اختبارها في ترتيب الحالات الإيجابية بالنسبة للحالات السلبية، فكلما اقترب المصنف من النقطة (1) كان المصنف مثاليا والعكس كلما اقترب من النقطة(0) قل أداؤه.

خلاصة:

هناك العديد من المقاييس الأخرى والتي سيتم إيرادها بالتفصيل في الجانب التطبيقي من هذا البحث مع التمثيل لها بنتائج حقيقية، قمنا بعرض هذه المصطلحات المتداولة حتى نبرز كفاءة برنامج WEKA وإلمامه الواسع بكم هائل من الخوارزميات، مع تقديم اختبارات ومقاييس لتقييمها ومعرفة مدى دقتها وكفاءتها، كما يتضمن هذا البرنامج بيئة عمل مرئية ذات مرونة جد عالية تسمح بمشاهدة النتائج من خلال رسوم وأشكال، كما يسمح أيضا بمعالجة كم هائل من البيانات مع اختزال في الجهد والوقت.

الفصل الرابع : تصنيف النصوص الأدبية لأساليب إنشائية وخبرية باستخدام برنامج (Weka)

خصص هذا الفصل للدراسة التطبيقية التي استلزمت جهدا وعملا جادا بدء من المعالجة المسبقة للبيانات حتى تكون جاهزة للتنقيب وهي المرحلة الأكثر جهدا والأطول وقتا حيث يتم تحويل البيانات إلى صيغة (ARFF) وتحميلها في أداة Weka ثم يأتي دور استخدام المصفيات في المعالجة المسبقة للبيانات، وهذه المرحلة تتضمن تغيير نوع البيانات وتقسيمها، سنوضح هذه المراحل وطريقة تنفيذها فيما يلي:

المرحلة الأولى: وهي تمثل عملية اختيار البيانات وتجميعها إذ أن أي دراسة بحثية تأخذ بعين الاعتبار هذه الخطوة التي تعتبر مهمة جدا لأنها تستلزم ضرورة إنشاء مدونة تضم مجموعة من النصوص التي تعتبر نماذج لعملية التطبيق وهذه النصوص عادة ما تكون موثقة المصدر.

قمنا في هذه الخطوة بجمع نصوص عربية مختلفة المصادر قسمتها إلى ثلاث مجالات نصوص قرآنية تحوي شواهد لآيات طويلة وقصيرة تتلاءم مع متطلبات الدراسة، أخذت هذه النصوص القرآنية من مصحف المدينة المنورة للنشر الحاسوبي حيث يتيح هذا المصحف مجموعة من الإمكانيات إذ يسمح للمستخدم من إضافة النص القرآني إلى الوثائق وملفات النصوص مع احتفاظها بخصائصها، كما يمكننا من البحث عن كلمة أو أكثر في نص القرآن الكريم لاستعراض المواضيع المختلفة التي جاءت فيها هذه الكلمة أو الجملة، وأيضا قابلية نسخ الآيات القرآنية إلى برنامج وورد، وسبب اعتمادي على النصوص القرآنية بالدرجة الأولى يرجع إلى كونه زاخر بالأساليب الإنشائية والخبرية.

المجال الثاني خصصته للشواهد الشعرية لكون أن الأمة العربية هي أمة شاعرة بالدرجة الأولى، يتيح لنا هذا المصدر كم هائل من البيانات حاولت أن اختار ما يتناسب مع الدراسة مع الأخذ بعين الاعتبار الاستشهاد بنصوص من عصور مختلفة بدء بالعصر الجاهلي حتى المعاصر وكذا من دول عربية مختلفة، اقتبست مجمل الشواهد الشعرية من موسوعة الشعر العربي الالكترونية (ديوان) والتي تضم حوالي 25693 قصيدة لـ 449 شاعر، نجد في هذه الموسوعة تسهيلات

كثيرة للإبحار فيها بعد عملية الاشتراك والانضمام إليها [1].

أما المجال الثالث من المدونة خصصته لمقتطفات من النصوص الروائية لكون أن كل من الشعر والرواية هي نصوص أدبية تحفل بمختلف الأساليب الفنية، مصدر هذه النصوص كان من الموقع الالكتروني حكايا للرواية العربية [2] الذي يتيح العديد من الروايات والقصص والحكايات كما يسمح بتحميلها وطبعها ونسخها من دون قاعدة الاشتراك.

الطريقة التي اعتمدتها في جمع النصوص على شكل مدونة لها العديد من المزايا من بينها:

- أنها أكثر عملية كونها تحوي نصوص حقيقية.
- تحتوي العديد من النصوص التي تخضع للتطبيق مما يساعد في ظهور نتائج يقينية.
- تنوع النصوص وفقا لأسس علمية لتمثل كل مؤشرات الأساليب الخبرية والإنشائية مع مراعاة مختلف الاستعمالات اللغوية.
- الاستخدام الحاسوبي في التعامل مع نصوص المدونة "وهذا هو الغرض من الدراسة".
- الاعتماد على أساليب التحليل الكمي (إحصاء التكرارات).

المرحلة الثانية: تتعلق هذه المرحلة بعملية تنظيف البيانات وتنقيتها من الشوائب فبعض البيانات تكون فارغة كعلامات الترقيم " النقطة ، الفاصلة، علامة الاستفهام والتعجب، الأقواس، المزدوجتان، والأرقام فكلها ليست لها أهمية أثناء التصنيف وقد تؤدي إلى تدني كفاءة المصنفات، لذا يتم حذفها إما يدويا أو بطريقة آلية فمثلا في البرنامج الذي اعتمدته في هذه الدراسة weka تكون عملية تصفية البيانات بطريقة آلية مباشرة بعد إدخالها عليه.

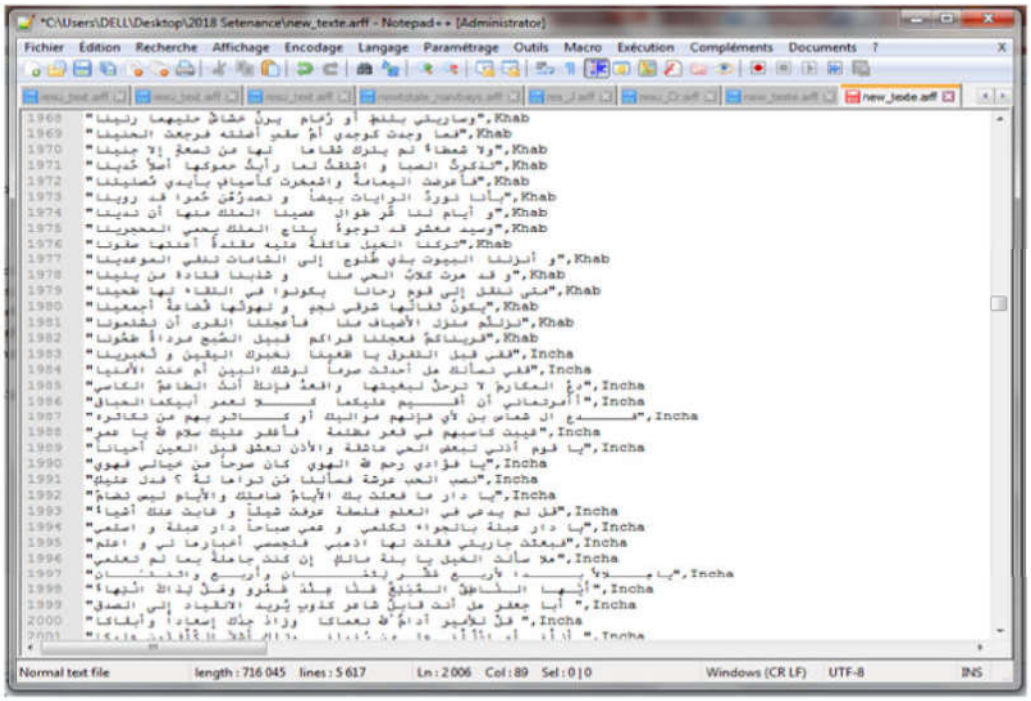
ومن مزايا هذه المرحلة:

[1] موقع موسوعة الشعر العربي (الديوان): https://www.aldiwan.net

[2] موقع حكايا للرواية العربية https://hakaya.com

- أنها تعمل على تخفيض كمية البيانات إلا أنها لا تمس بالبيانات الأساسية.
- إزالة الشوائب التي قد تعيق عملية تصنيف البيانات

المرحلة الثالثة: تتعلق بتحويل البيانات فالبرامج الآلية المعتمدة في عملية معالجة البيانات تستدعي التعامل مع نوع معين من الملفات ، مما يقتضي ضرورة تحويل شكل البيانات بهدف إدخالها إلى البرنامج و طبعا مما هو معروف فإن برنامج weka يتعامل مع ملفات (arff) Attribute Relation File Format وبالتالي كان لابد من تحويل جميع البيانات من صيغة (txt) إلى صيغة (arff) والبرنامج بدوره يوفر مجموعة من الصفوف أو الأعمدة تسهل عملية إدخال هذا الملف، إلا انه أحيانا أخرى يتم الاستعانة ببرامج أخرى تعمل على تحويل البيانات مثل برنامج Notepad ++ . الشكل (1.4) يوضح نص بصيغة arff :



شكل 1.4: تحويل البيانات إلى صيغة ARFF

المرحلة الرابعة: وهي مرحلة تنقيب البيانات والتي تأتي بعد إدخال البيانات إلى البرنامج لإيجاد آلية قادرة على تصنيف النصوص، إذ تم تغيير نوع البيانات واستخدام أكبر قدر من الخوارزميات المتاحة وقسمت البيانات إلى حزمتين بيانيتين الحزمة البيانية الأولى للتدريب وتحتوي على نصوص قرآنية وشعرية وروائية أي ما يعادل نسبة80 % من البيانات، أما الحزمة البيانية الثانية فتحتوي على ما يعادل نسبة 20 % من البيانات، هذه الحزمة ستترك جانبا لاستخدامها في مرحلة التنبؤ. بعد تحديد مجموعة التدريب وتحديد الفئات يأتي دور تقييم واختبار الفئة وهو مهم جدا لمعرفة هل هذه الفئة التي تم تدريبها يمكن استخدامها في تصنيف بيانات جديدة أم لا، ويمكن تحديد إمكانية إعادة استخدام الفئة التي تم تدريبها من عدمه من خلال نسبة تصنيفه في مرحلة الاختبار، فكلما كانت عالية كانت إمكانية استخدامه أكبر.

1.4 أدوات اختبار التصنيف:

حتى يمكننا تقييم أداء مصنف لا بد من إجراء اختبار بإعطائه مجموعة من البيانات التي تم تدريبه عليها مسبقا ويتحدد ذلك من خلال النسبة المئوية للتصنيف، فكلما كانت النسبة عالية كانت دقة تصنيفه جيدة وتتم عملية التقييم واحتساب الدقة باعتماد أربعة طرق:

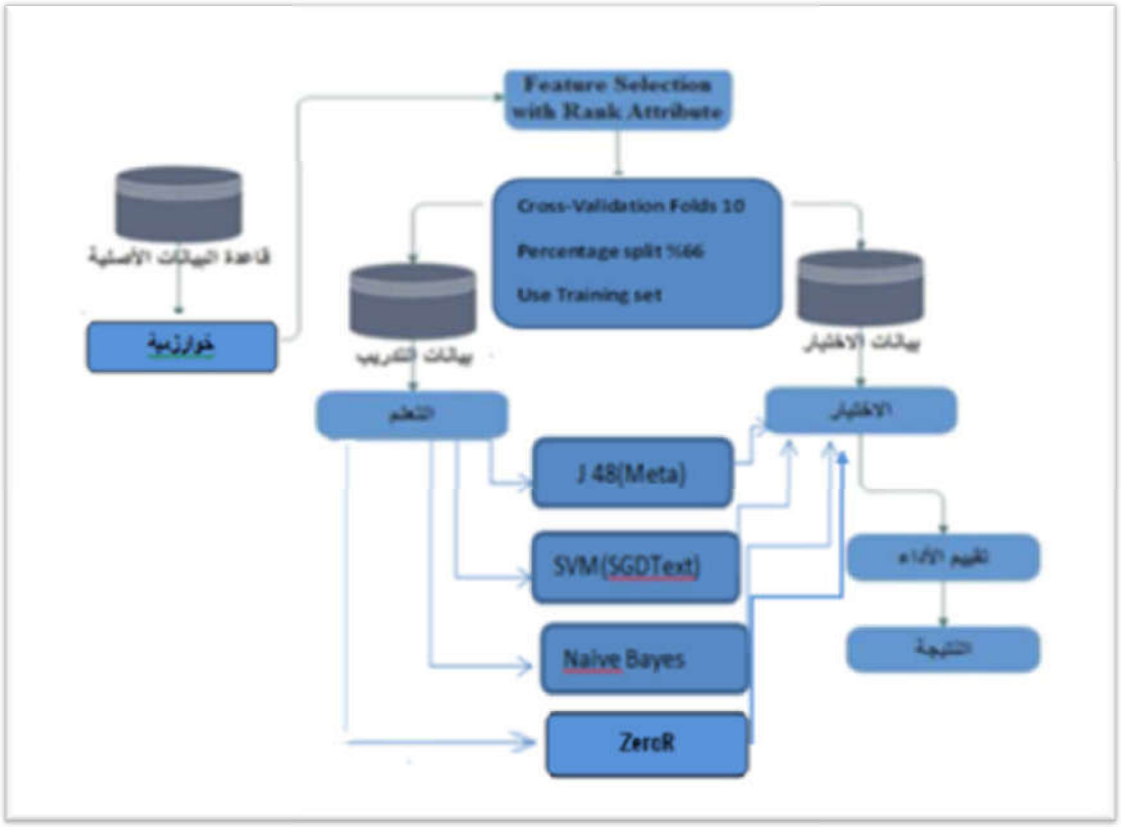
- UsingTrainig set : تعتمد هذه الطريقة في عملية اختبار المصنف على البيانات التدريبية وتغطية وتخمين(label) الذي تم إعطائه للبيانات أي التسمية التي أعطيت للبيانات المرجو تصنيفها.

- Cross Validation : تقوم هذه الطريقة على تقسيم البيانات حسب الرقم الذي يعطى (fold)ومن ثم يتدرب على 90% من هذه البيانات المقسمة و10% يتم اختبار المصنف عليها، مثلا لدينا مجموعة تدريب عددها 100 يتم أولا تقسيم البيانات حسب (fold) إلى

عشرة مجموعات إذ يتم تدريب المصنف على تسع مجموعات والعاشرة يختبر عليها، ومن ثم يتم إجراء الاختبار على مجموعة أخرى وإدخال المجموعة السابقة التي اجري عليها الاختبار ضمن مجموعة التدريب وهكذا حتى يتم اختبار المصنف عشر مرات.

- Percentage Split : يتم في هذه الطريقة تقسيم مجموعة البيانات إلى قسمين مجموعة للتدريب ومجموعة للاختبار.
- Supplied test Set : تقوم هذه الطريقة على إدخال بيانات جديدة على المصنف ويتم الاختبار عليها وعادة ما تستخدم هذه الطريقة في عملية التنبؤ.

وفي عملنا هذا تم تجربة أربع خوارزميات لتحديد أفضلها بغية استخدامها في عملية تصنيف النصوص الأدبية (الخبرية والإنشائية)، ومن أجل معرفة وتحديد دقة تصنيف البيانات تم اعتماد ثلاث طرق للاختبار Using Trainigset ، Cross-Validation ، Percentage Split .



شكل 2.4: النموذج المقترح لمعالجة المعطيات التي تم جمعها

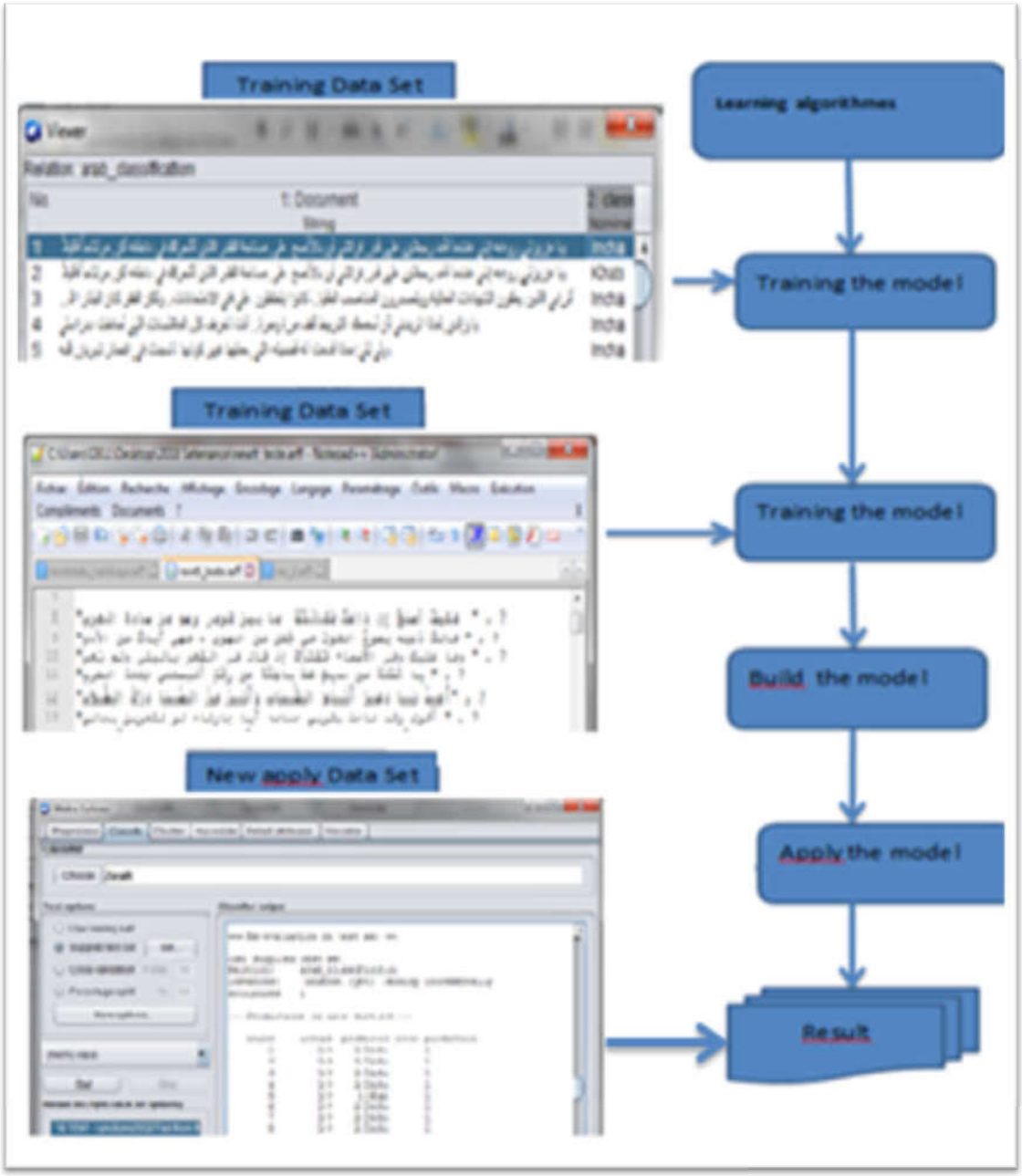
2.4 التنقيب في المعطيات (Data Manning):

للاستفادة من المعطيات التي تم جمعها ضمن بيئة التنقيب في البيانات يستلزم بناء النماذج التنبؤية بالاعتماد على خوارزميات التصنيف، ومن ثم انتقاء أهم معايير تقييم الأداء استنادا لخوارزميات انتقاء المعايير وأخيرا تحليل البيانات للاستفادة من الخوارزميات .

3.4 بناء نموذج التصنيف:

تتم عملية بناء نموذج التصنيف بخطوتين رئيسيتين هما: التصنيف وهو خطوة بناء واختبار نموذج التصنيف، والتنبؤ وهو استخدام النموذج ليتنبأ بفئات البيانات غير معروفة الفئة وبذلك تكون خطوات التصنيف كما يلي :

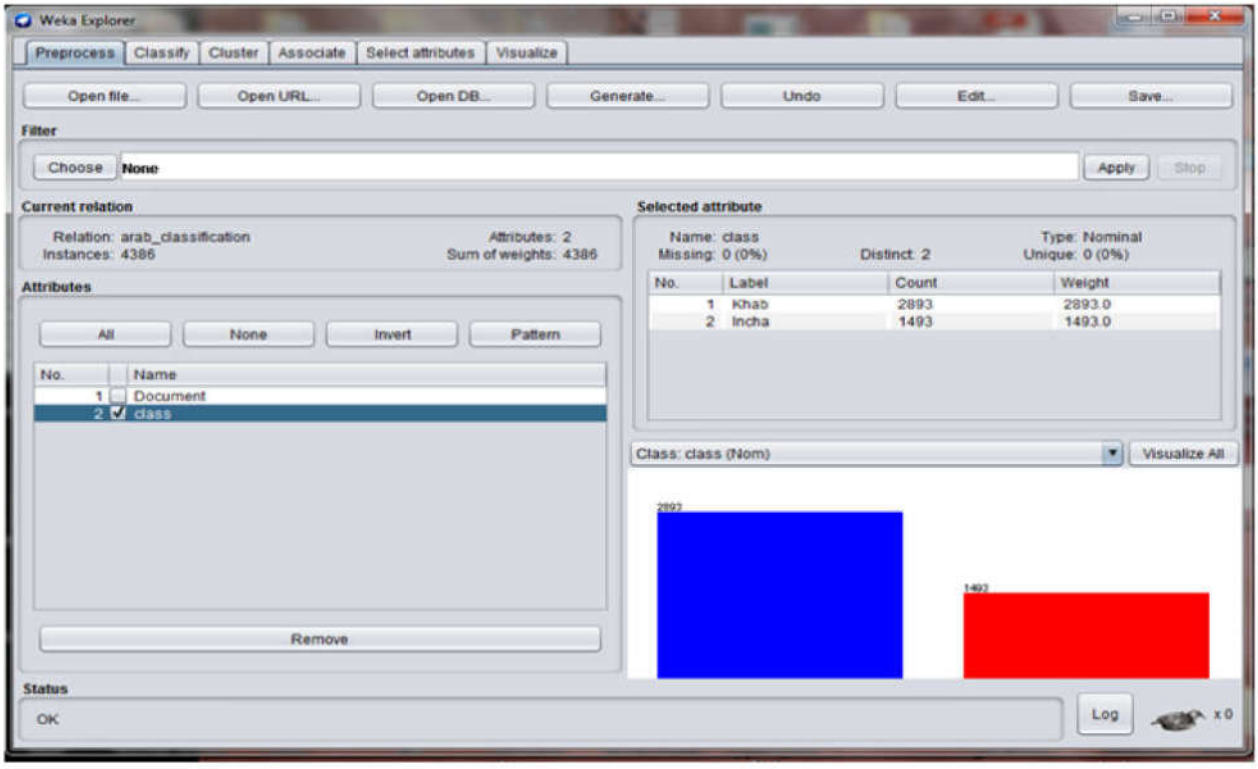
- يتم تدريب خوارزمية التصنيف (التعلم) على بيانات التدريب Training Data Set المحتوية على سجلات معروفة الفئة لبناء النموذج الذي يستخدم لاختبار حزمة بيانات التي تحتوي على سجلات غير معروفة الفئة.
- يتم تقييم أداء النماذج باستخدام خوارزميات التصنيف والهدف هو الحصول على أعلى دقة وأقل نسبة خطأ عند تطبيقها على بيانات الاختبار ومن خلال ذلك يتم بناء أفضل نموذج (Model) ويتم تنفيذ النموذج الذي تم بنائه على حزمة بيانات التطبيق Apply Data Set والتي تحتوي على لنصوص غير معروفة الفئة وبعد إظهار النتائج يتم الاستفادة منها في التنبؤ المستقبلي والشكل (3.4) يوضح خطوات بناء النموذج.



شكل 3.4: خطوات بناء النموذج والتنبؤ

4.4 التنقيب في المعطيات بالاستعمال مصنف (J48) :

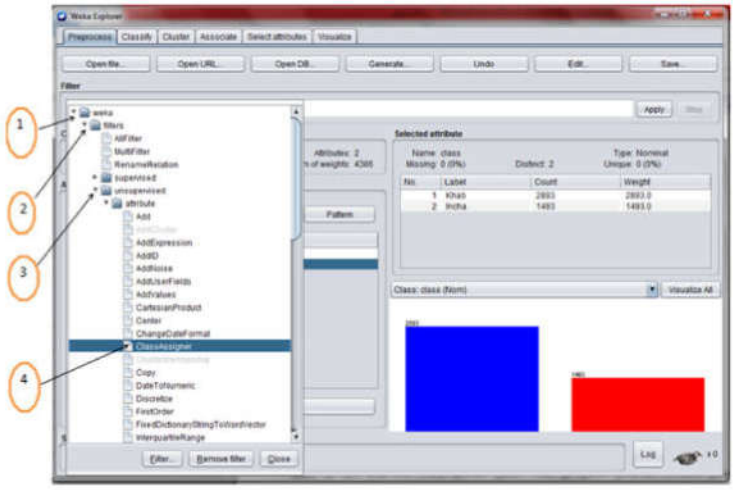
حتى يمكننا استعمال المصنف (J48) لابد من إتباع مجموعة من الخطوات، أولا لابد من الدخول إلى برنامج weka فتتاح أمامك مجموعة من التعليمات ابدأ بالضغط على زرfilters (ترشيح) تظهر لك مجموعة من المؤشرات اختر unsupervised (بدون إشراف) ثم اكبس على تعليمة class assigner (الفئة المعينة)، بعد اجتياز هذه المراحل تظهر أمامك تعليمة Apply قم بالضغط عليها فتظهر أمامك الشاشة الموضحة في الشكل (4-4).



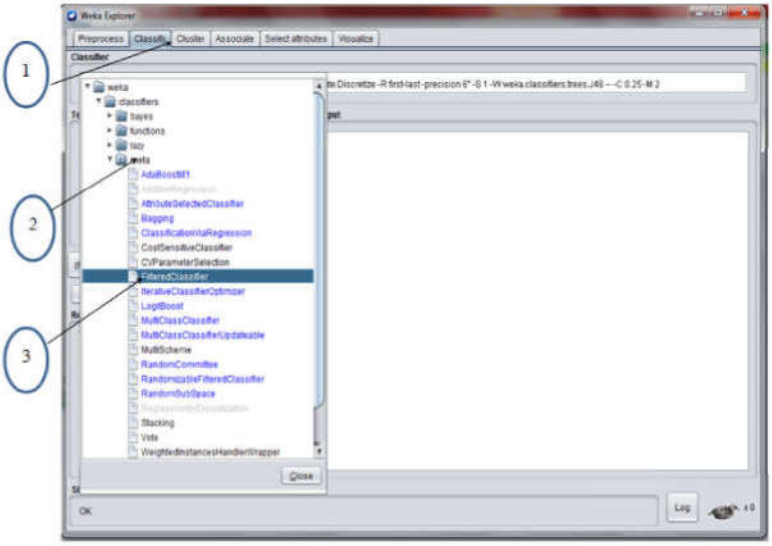
شكل 4.4: قاعدة بيانات مفتوحة Data set

لا يمكن لمصنفات النصوص التعامل مع السمة " النص" كخاصية واحدة، لذلك يجب علينا أولا تحويل السمة " النص" إلى فئات لإجراء ذلك ما عليك سوى النقر على الزر " اختيار" "Choose" ضمن مربع مجموعة التصفية "Filter" ، سوف يعرض برنامج Weka شجرة من الفلاتر المتاحة، كما

هو مبين في الشكل (5.4).



شكل 5.4: فتح ملفarff



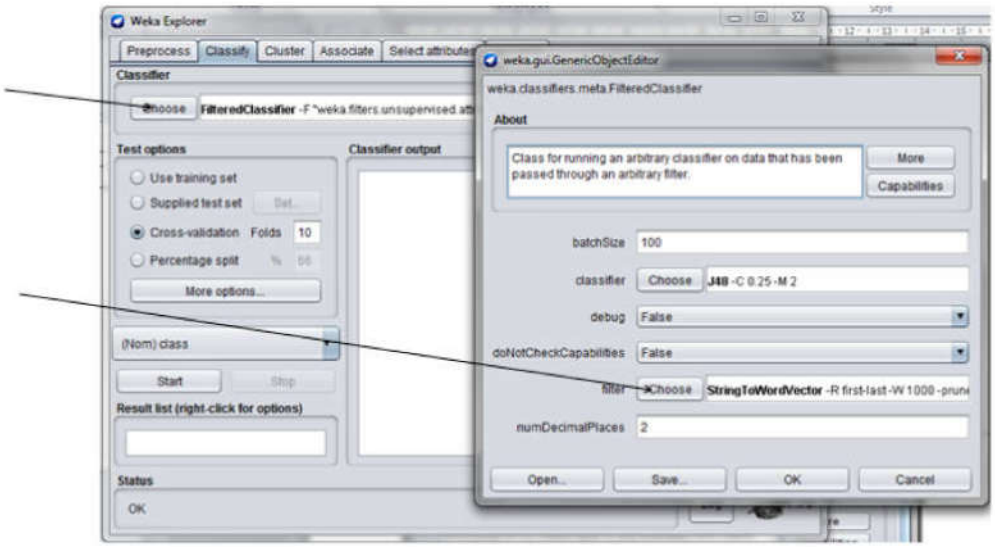
شكل 6.4: تحويل السمة " النص" إلى فئات

وبما أن فلتر "SteingToWordVector" يعتمد أساس على معامل تردد المصطلح- معكوس تردد الوثيقة(TF-IDF) هو معامل غالبا ما يستخدم في استرجاع المعلومات وتعدين النصوص، هذا المعامل هو مقياس إحصائي يستخدم لتقييم مدى أهمية وجود كلمة في مستند معين في

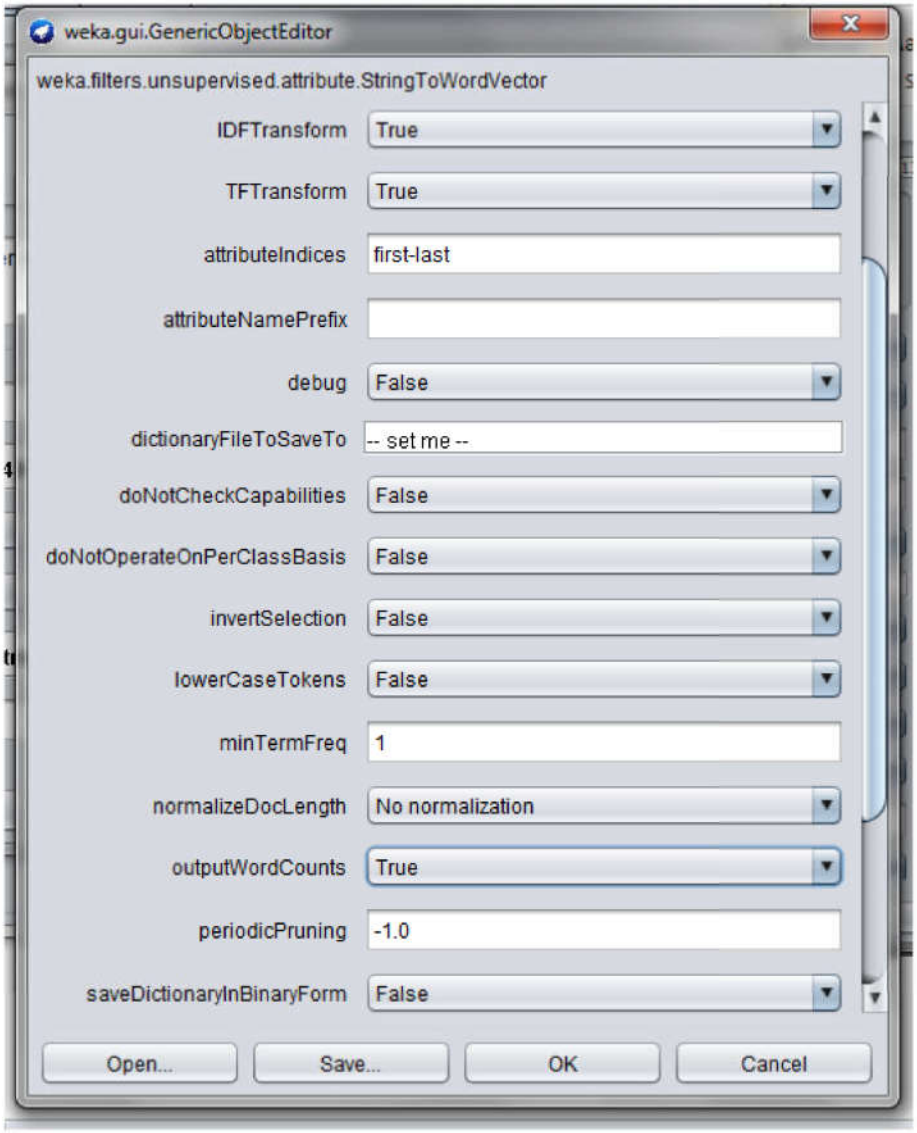
مدونة النصوص.

الأهمية تزيد نسبيا بزيادة عدد مرات ظهور الكلمة في المستند ولكن تقابل بتردد الكلمة في المدونة بشكل عام ويتم ذلك بالاعتماد على مدونات التدريب الخاصة لكل موضوع، والتي تمثل خصائصه.

بحيث يتم مقارنة هذه الخصائص مع تلك المتعلقة بالنص وتقوم فكرة معامل TF-IDF أساسا على تمثيل كل وثيقة d بمتجه $D = (d_1, d_2, ..., d_{|v|})$ في فضاء المتجهات، حيث يرمز |v| إلى حجم مجموعة المفردات، ويتم حساب مركبات المتجهة عن طريق ضرب تكرار الكلمة TF(w ,d) الذي هو عبارة عن عدد المرات التي تظهر فيها الكلمة w في الوثيقة d بعكس تكرار الوثيقة IDF(w) ، ويمثل تكرار الوثيقة DF(w) عدد الوثائق التي تظهر فيها الكلمة w مرة واحدة على الأقل. ومن اجل استعمال (TF-IDF) في برنامج Weka يتم إعداد"IDFTransform" "TFTransform" و"OutputWordCounst" كما هو مبين في الشكل(4-7) والشكل (8.4).



شكل 7.4: اختيار فلتر "SteingToWordVector"



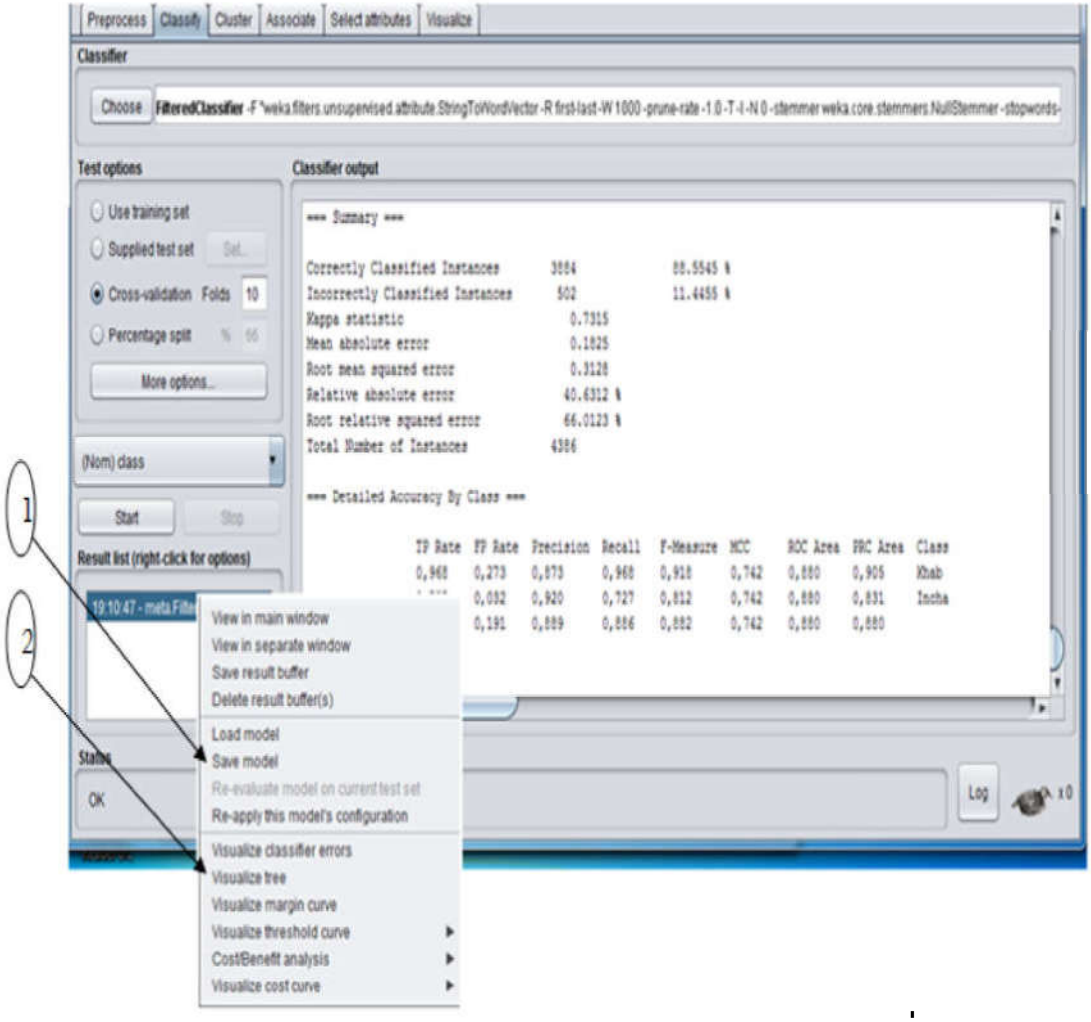
شكل 8.4: إعداد"IDFTransform" "TFTransform"

1.4.4 اختبار المصنف (J48) :

- **اختبار (Cross validation):**

من أجل تحديد دقة تصنيف البيانات التي تم تدريبها تم اعتماد أسلوب اختبار (validation Cross) بنسبة تقطيع مساوية عشرة Folds=10 وذلك لجميع حالات التصنيف في هذا البحث، ويمكن تلخيص عمل هذه الطريقة بالخطوات التالية:

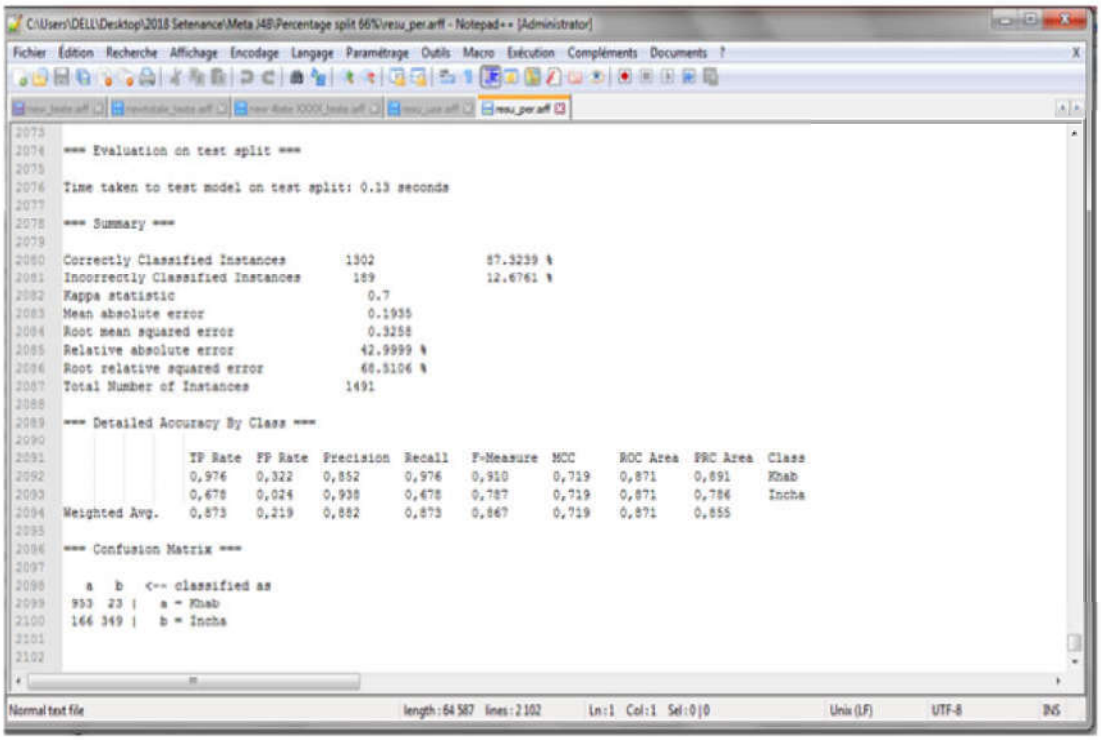
- تدريب نسبة 100 %من بيانات التدريب.
- تقسيم بيانات التدريب إلى /10 / أقسام بشكل عشوائي.
- تنفيذ عملية تخمين المقادير لكل من هذه الأقسام العشرة وحساب نسبة تطابق القيم الحقيقية لهذه البيانات مع القيم التي تم تخمينها.
- حساب هذه النسبة لكامل مجموعة البيانات لتكون نسبة الدقة المطلوبة.
- يجب التنويه هنا أن مقدار نسبة التقسيم مرتبط بعدد بيانات التدريب ولا يوجد نسبة ثابتة تحدد الخيار الأفضل للتقسيم ولكن الافتراض المستخدم ضمن تطبيق WEKA هو10 حيث أشارت العديد من الدراسات إلى أن هذه النسبة شكلت القيمة الأمثل في العديد من الاختبارات.الشكل (9.4) يوضح ذلك



شكل 9.4: الاحتفاظ بالنموذج الناتج عن اختبار Cross validation للخوارزمية J48 Meta المراد تطبقيها للتنبؤ.

- **اختبار (Percentage split 66%)**

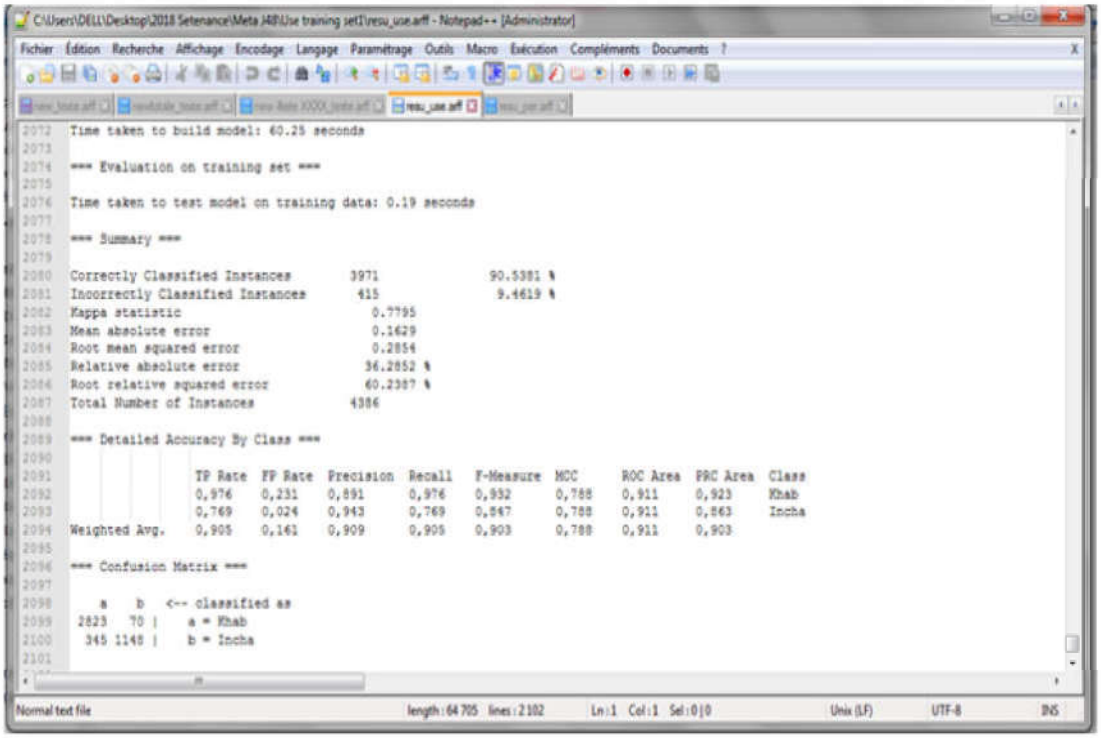
يتم في هذه الطريقة تقسيم مجموعة البيانات إلى قسمين مجموعة للتدريب (66%) ومجموعة للاختبار(34%).



شكل 10.4: اختبارsplit Percentage للمصنف J48

- اختبار(Using training set)

تم اختبار هذا المصنف على مجموعات التدريب التي تم تدريبه عليها



```
Time taken to build model: 60.25 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.19 seconds

=== Summary ===

Correctly Classified Instances        3971               90.5381 %
Incorrectly Classified Instances       415                9.4619 %
Kappa statistic                          0.7795
Mean absolute error                      0.1629
Root mean squared error                  0.2854
Relative absolute error                 36.2852 %
Root relative squared error             60.2387 %
Total Number of Instances             4386

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,976    0,231    0,891      0,976    0,932      0,788    0,911     0,923     Khab
               0,769    0,024    0,943      0,769    0,847      0,788    0,911     0,863     Incha
Weighted Avg.  0,905    0,161    0,909      0,905    0,903      0,788    0,911     0,903

=== Confusion Matrix ===

    a    b   <-- classified as
 2823   70 |    a = Khab
  345 1148 |    b = Incha
```

شكل 11.4: اختبار set training Using للمصنف J48

2.4.4 استظهار النتائج:

من خلال الاختبارات المجراة على المصنف J 48 والمبينة في كل من الشكل (9.4)،(10.4) و(4-11) والتي تظهر لنا مجموعة من الأسطر والقيم العددية إذ نجد في الأسطر الأولى من كل شكل مجموع الحالات التي صنفتها الخوارزمية بشكل صحيح Correctly classfiied instances مع ذكر نسبتها المئوية وأما السطر الثاني يبين لنا مجموع الحالات (instances) التي أخطأت الخوارزمية في تصنيفها ونسبتها المئوية incorrectly classfiied instance .

حيث وجدنا عدد الحالات المصنفة لاختبار Cross Validation للمصنف Meta(J48) بشكل صحيح هو 3884 حالة بنسبة مئوية مقدارها 88,5545% وعدد الحالات المصنفة بشكل غير صحيح هو 502 حالة بنسبة 11,4455% .

أما بالنسبة لاختبار Percentage split 66% وجدنا عدد الحالات المصنفة بشكل صحيح هو 1302 حالة بنسبة مئوية قدرت بـ 87,3239 % أما عدد الحالات المصنفة بشكل غير صحيح هو 189 حالة بنسبة مئوية مقدارها 12,6761% .

أما فيما يخص اختبار Using training set لاحظنا عدد الحالات المصنفة بشكل صحيح هو 3971 حالة بنسبة مئوية مقدارها 90,5381% أما بالنسبة للحالات المصنفة بشكل غير صحيح هي 415 حالة بنسبة مئوية مقدارها 9,4619%.

أما السطر الثالث يمثل مقياس لتصحيح احتمال الاتفاق بين التصنيفات الحقيقية إحصاء الكابا (Kappa Statistiques) والتي كان مقدارها بالنسبة لاختبار Cross Validation 0,7315

حيث $K = \frac{P0-Pe}{1-Pe}$

حيث $P0 = \frac{(2799+1085)}{(2799+1085+408+94)}$ و $Pe = \frac{(3207*2893)+(1179*1493)}{4386*4386}$

ومنه نحصل على $K = \frac{0,8855-0,5737}{1-0,5737} = 0,7315$

ويتعلق الأمر نفسه باختبار Percentage split 66% والتي قدرت نسبتها بـ 0,7 أما بالنسبة لاختبار Using training set فنسبتها 0,77 .

أما السطر الرابع نجد Mean absolute error (الخطأ المطلق في المتوسط) ويستخدم معدلات الخطأ للتنبؤ الرقمي بدلا من التصنيف حيث أن التنبؤات ليست فقط الصحيحة و الخاطئة بالنسبة لاختبار Cross Validation تساوي 0,1825, بالنسبة لاختبار Percentage split 66% تساوي0,1935 , أما بالنسبة لاختبار Using training set قيمتها 0,1629.

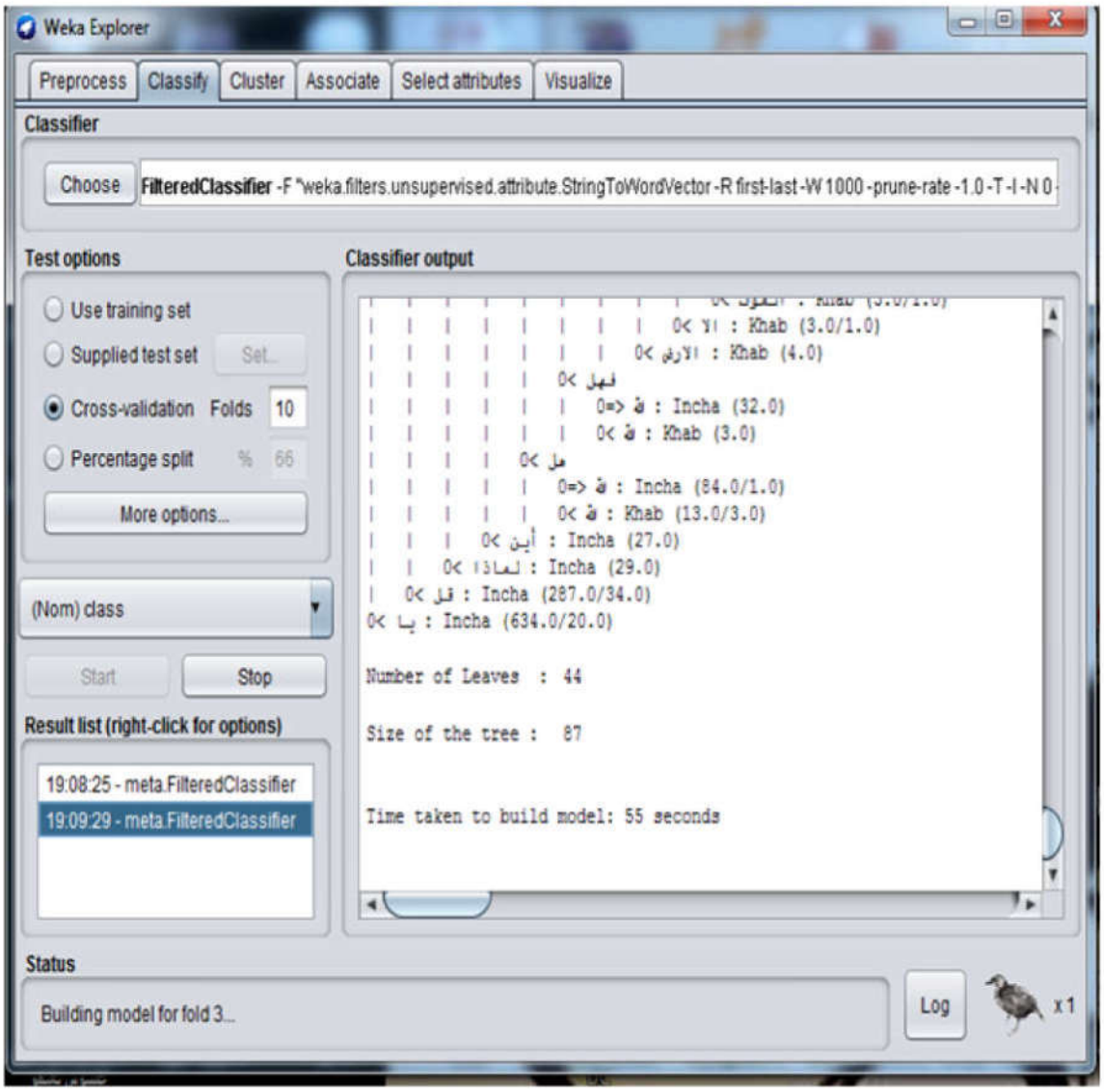
أما السطر الخامس Root mean squared error جذر مربع متوسط الخطأ بالنسبة لاختبار

Cross Validation تساوي 0,3128, ويتعلق الأمر نفسه باختبار 66% Percentage split والتي قدرت قيمتها بـ 0,3258 , وبالنسبة لاختبار Using training set قيمتها 0,2854, ويليه السطر السادس Relative absolud error الخطأ المطلق النسبي بالنسبة لاختبار Validation Cross تساوي 40,6312%, بالنسبة لاختبار Percentage split 66 % تساوي42,9999%, أما بالنسبة لاختبار Using training set فنسبتها 36,2852%.

أما السطر السابع Root relative squared error هو الجذر التربيعي النسبي لخطأ يساوي 66,0123% بالنسبة لاختبار Cross Validation , بالنسبة لاختبار 66% Percentage split فنسبتها تساوي68,5106%, أما بالنسبة لاختبار Using training set فنسبتها 60,2387%.

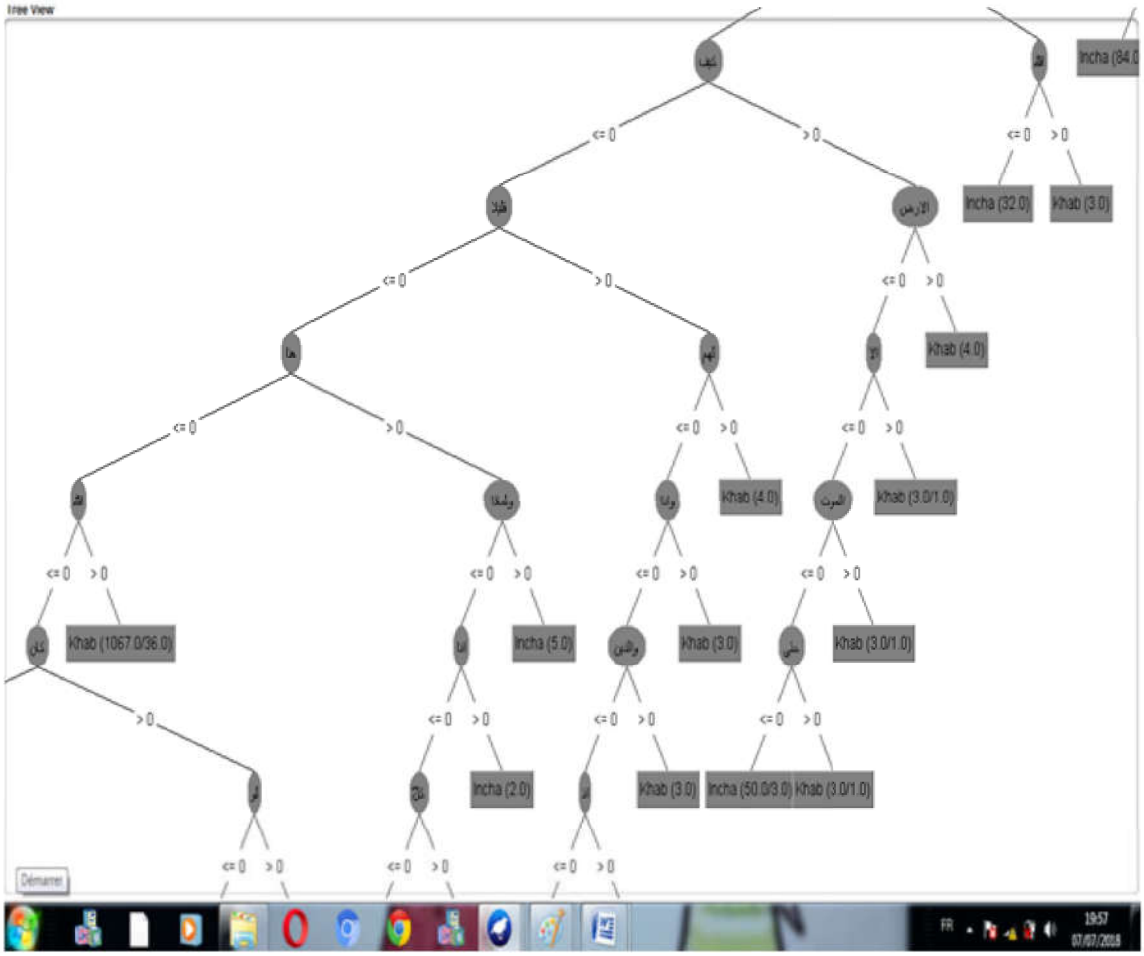
3.4.4 تصنيف النصوص بأشجار القرار J48 :

تعتمد خوارزمية J48 أساسا على أشجار القرار ، إذ تعتبر كأدوات متداولة للتصنيف والتنبؤ حيث تظهر فيها عملية تصنيف النصوص على شكل شجرة لها عقد وفروع التي تظهر بصورة دوائر ومربعات تمثل لنا القرارات والحالات المتوقعة والشكل(4-12) و(4-13) يوضح لنا شجرة القرار المعتمدة في هذه الخوارزمية والتي تتكون من 44 عقدة تمثل لنا مخرجات الاختبار إذ تعين كل عقدة على أنها فئة و87 فرع (أوزان) اللاتي تمثل لنا التنبؤ بالحالات أو قيمة العقدة.



شكل 12.4: بناء شجرة قرار بالتدريب علي مجموعة بيانات

بالرغم من أن شجرة القرار تستخدم للتدريب على البيانات إلا أنها أيضا تستخدم وبشكل كبير للتنبؤ .



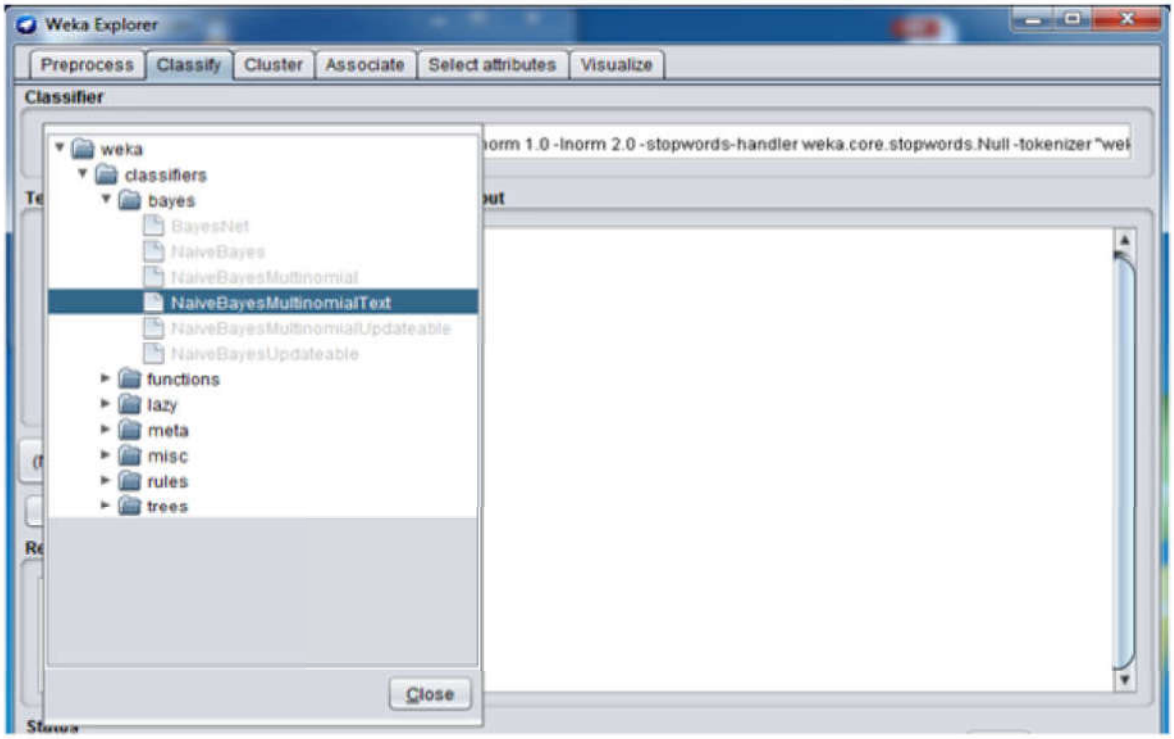
شكل 13.4: مصنف شجرة القرار الناتج لقاعدة بيانات

5.4 التنقيب في المعطيات بالاستعمال مصنف (Naïve Bayse) :

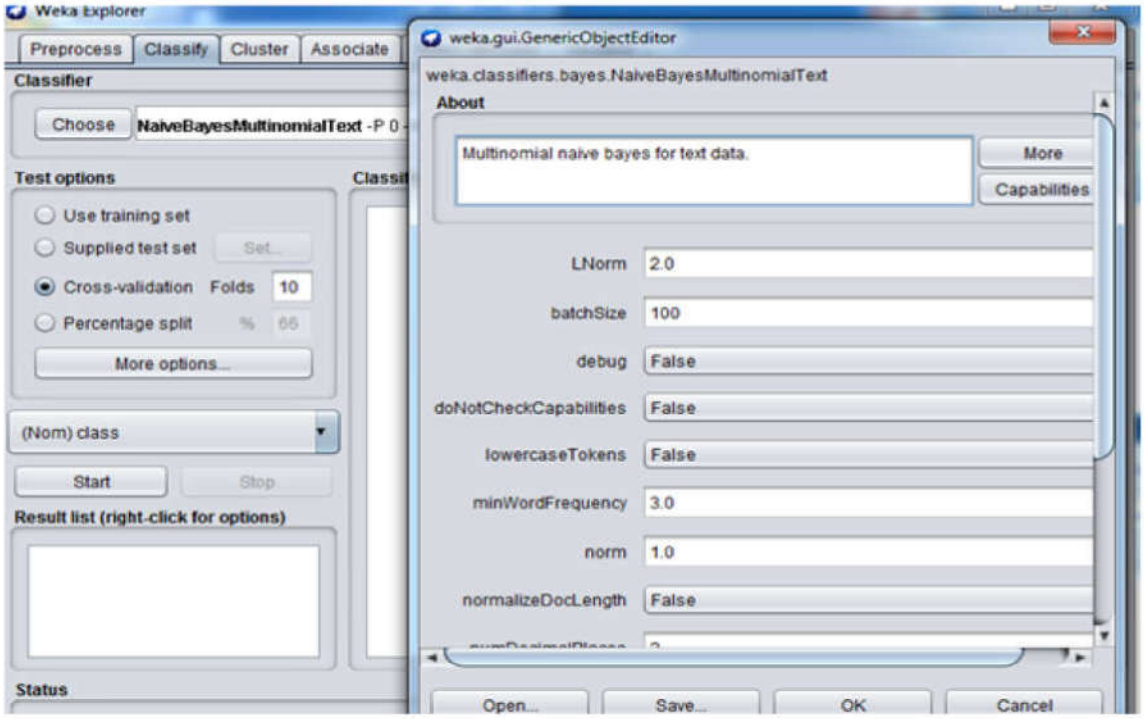
من أجل الوصول إلي تطبيق خوارزمية Naive bayse في برنامج Weka يتم المرور بمجموعة من المراحل المتمثلة في كل من الشكل (14.4) إلى غاية الشكل (15.4).

ويتم الاستعانة بهذه الخوارزمية بعد تعديلها من أجل الوصول إلى استظهار النتائج مع مختلف الاختبارات Using training set , Percentage split 66% , Cross validation ، وذلك

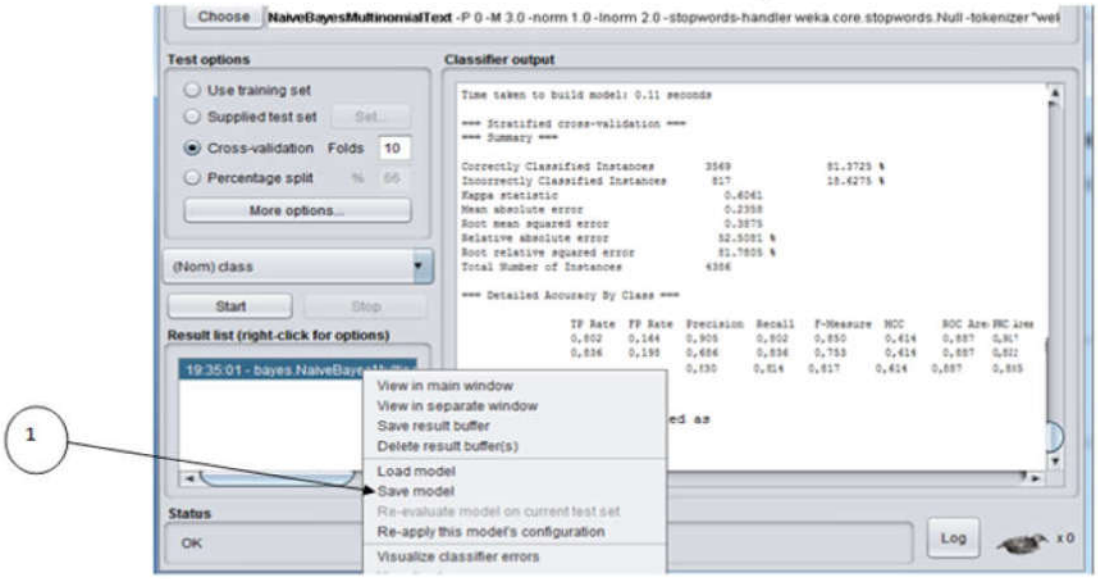
بغية التوصل إلى النموذج الناتج عن اختبار هذه الخوارزمية والاحتفاظ به لتطبيقه في عملية التنبؤ ويتضح ذلك من خلال الشكل (16.4).



شكل 14.4: اختيار خوارزمية Naïve Bayes



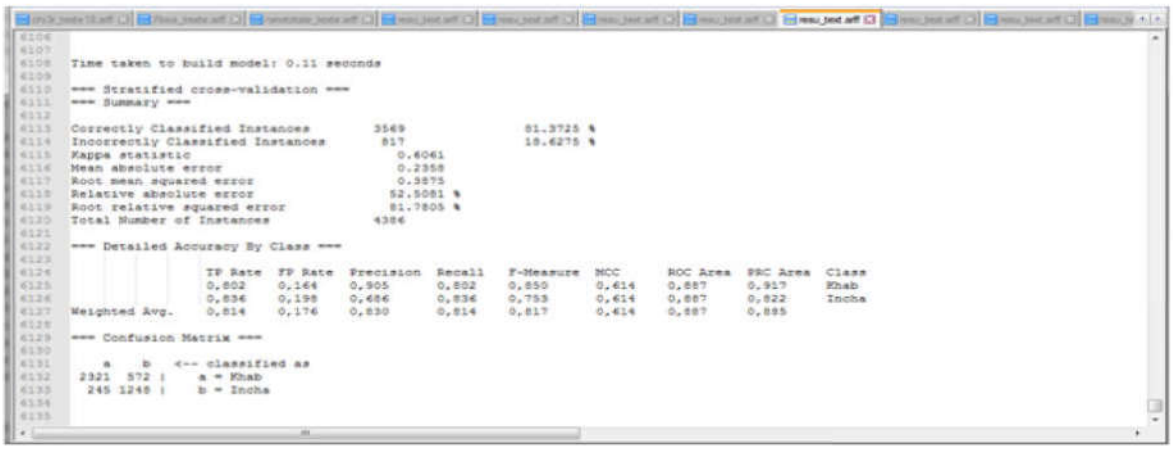
شكل 15.4: إعداد خوارزمية Naïve Bayes



شكل 16.4: الاحتفاظ بالنموذج ناتج عن اختبار Cross validation للخوارزمية Naïve Bayes المراد تطبيقها للتنبؤ.

1.5.4 اختبار المصنف Bayes Naïve :

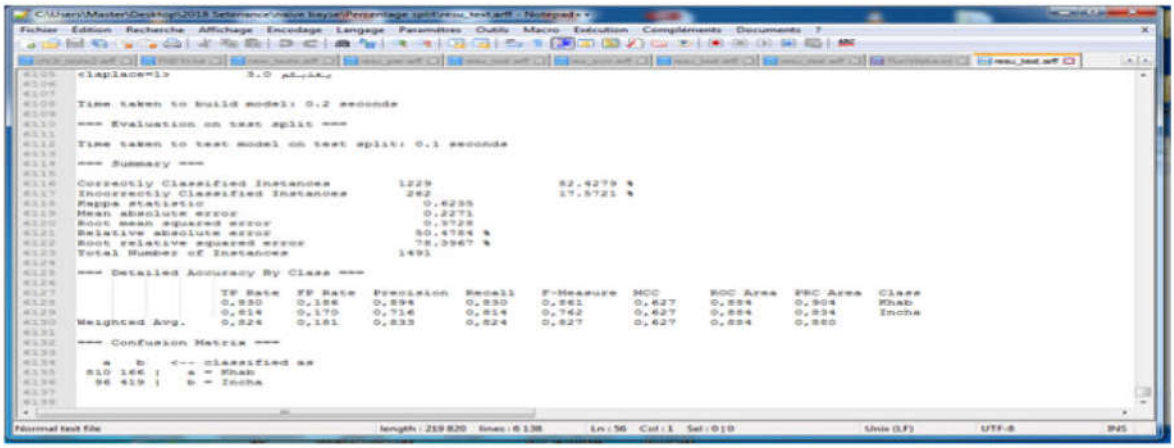
- اختبار (Cross validation)



```
Time taken to build model: 0.11 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        3569               81.3725 %
Incorrectly Classified Instances       817               18.6275 %
Kappa statistic                          0.6061
Mean absolute error                      0.2358
Root mean squared error                  0.3875
Relative absolute error                 52.5081 %
Root relative squared error             81.7805 %
Total Number of Instances             4386

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0,802    0,164    0,905      0,802   0,850      0,614    0,887     0,917     Khab
                 0,836    0,198    0,686      0,836   0,753      0,614    0,887     0,822     Incha
Weighted Avg.    0,814    0,176    0,830      0,814   0,817      0,614    0,887     0,885

=== Confusion Matrix ===

    a    b   <-- classified as
 2321  572 |    a = Khab
  245 1248 |    b = Incha
```

شكل 17.4: اختبار Cross validation للمصنف Naïve Bayes

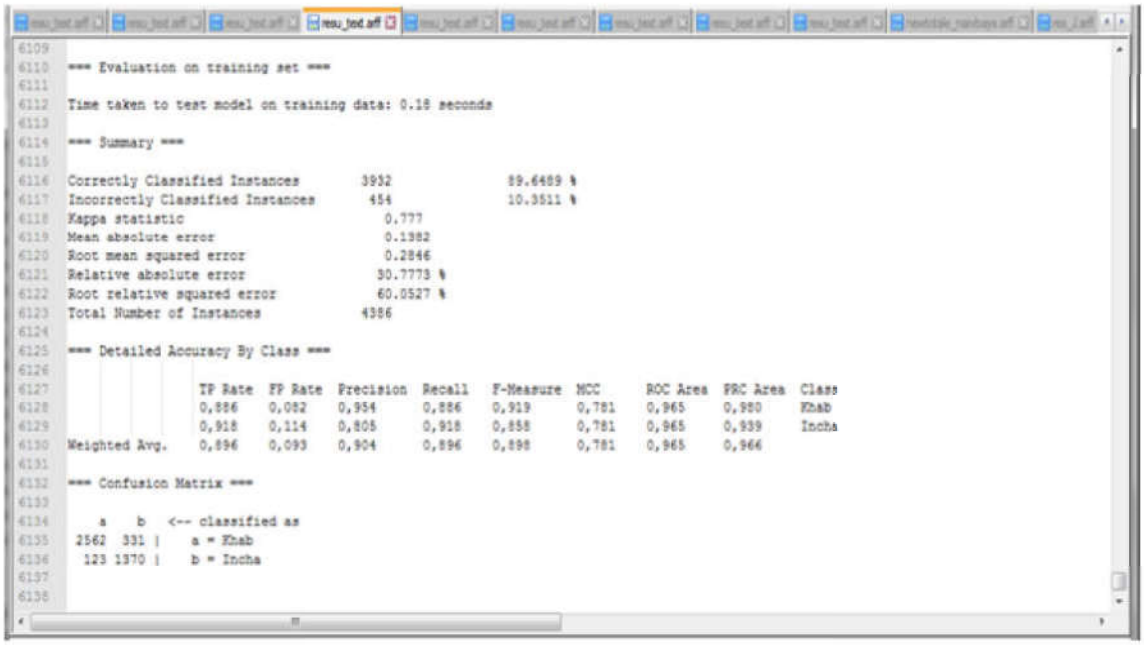
- اختبار(Percentage split 66%)



```
<laplace=1>                3.0 بكتيريا

Time taken to build model: 0.2 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.1 seconds

=== Summary ===

Correctly Classified Instances        1229               82.4279 %
Incorrectly Classified Instances       262               17.5721 %
Kappa statistic                          0.6235
Mean absolute error                      0.2271
Root mean squared error                  0.3728
Relative absolute error                 50.4784 %
Root relative squared error             78.3967 %
Total Number of Instances             1491

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0,830    0,186    0,894      0,830   0,861      0,627    0,884     0,904     Khab
                 0,814    0,170    0,716      0,814   0,762      0,627    0,884     0,834     Incha
Weighted Avg.    0,824    0,181    0,833      0,824   0,827      0,627    0,884     0,880

=== Confusion Matrix ===

   a   b   <-- classified as
 810 166 |   a = Khab
  96 419 |   b = Incha
```

شكل 18.4: اختبار Percentage split للمصنف Naïve Bayes

• اختبار (Using training set)



```
=== Evaluation on training set ===

Time taken to test model on training data: 0.18 seconds

=== Summary ===

Correctly Classified Instances        3932               89.6669 %
Incorrectly Classified Instances       454               10.3511 %
Kappa statistic                          0.777
Mean absolute error                      0.1382
Root mean squared error                  0.2846
Relative absolute error                 30.7773 %
Root relative squared error             60.0527 %
Total Number of Instances             4386

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,886    0,082    0,954      0,886    0,919      0,781    0,965     0,980     Khab
                 0,918    0,114    0,805      0,918    0,858      0,781    0,965     0,939     Incha
Weighted Avg.    0,896    0,093    0,904      0,896    0,898      0,781    0,965     0,966

=== Confusion Matrix ===

    a    b   <-- classified as
 2562  331 |    a = Khab
  123 1370 |    b = Incha
```

شكل 19.4: اختبار Using training set للمصنف Naïve Bayes

2.5.4 استظهار النتائج:

اكتفينا باستظهار نتائج اختبار Cross Validation بينما الاختبارات الأخرى ونتائجها تظهرها الأشكال المبينة أعلاه.

أولا فيما يتعلق بـ Correctly classfiied instances و incorrectly classfiied instances تظهر عدد الحالات المصنفة بشكل صحيح هو3567 حالة بنسبة مئوية مقدارها 81,3269% وعدد الحالات المصنفة بشكل غير صحيح هو 819 حالة بنسبة 18,6731% .

- السطر الثالث يمثل مقياس لتصحيح احتمال الاتفاق بين التصنيفات الحقيقية إحصاء كابا (Kappa Statistiques) والتي كان مقدارها 0,6051 حيث $K = \frac{P0-Pe}{1-Pe}$

- أما السطر الرابع نجد (Mean absolut error) الخطأ المطلق في المتوسط ويستخدم معدلات الخطأ للتنبؤ الرقمي بدلا من التصنيف حيث أن التنبؤات ليست فقط الصحيحة والخاطئة Mean absolut error =0,2354

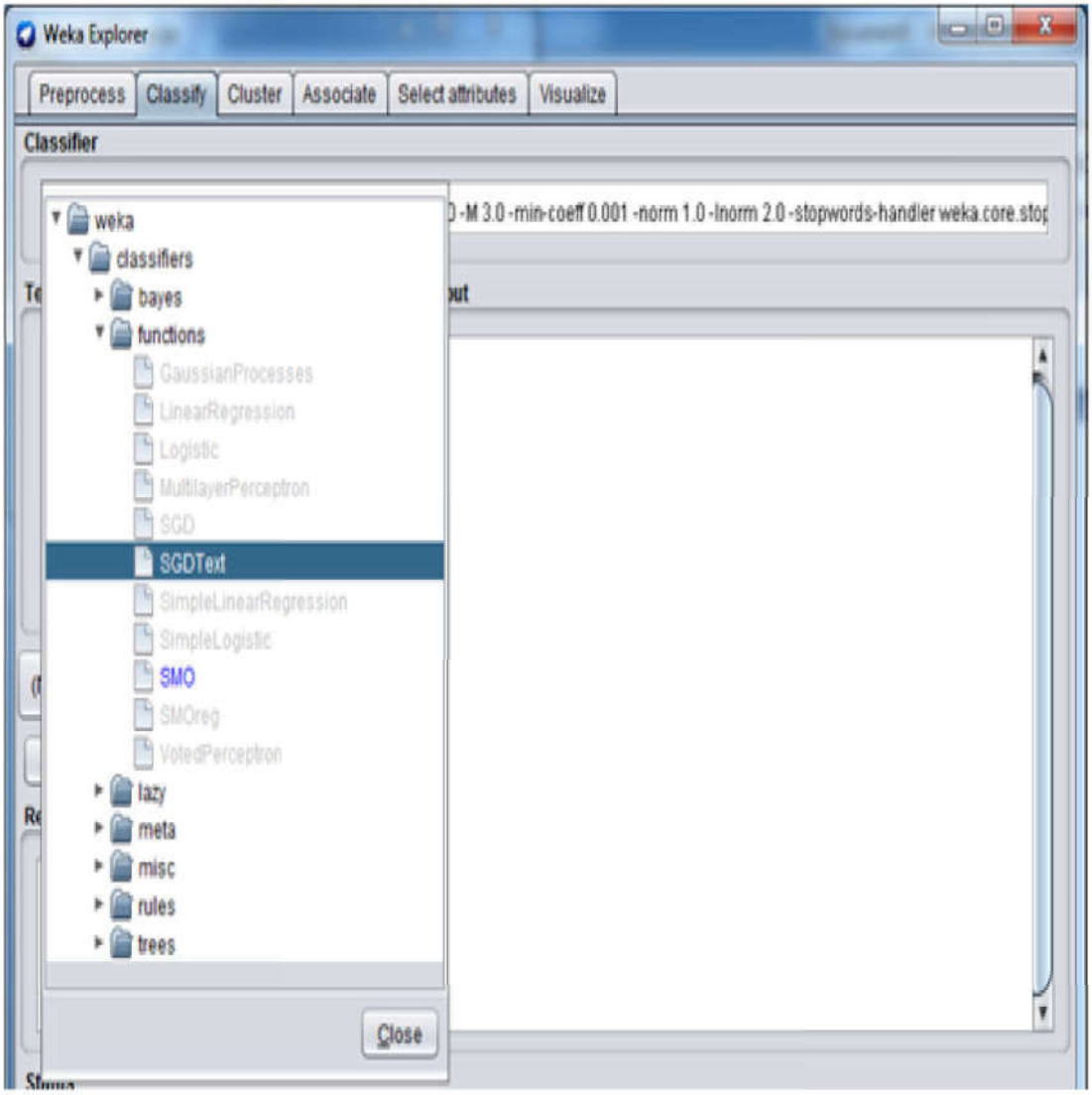
- أما السطر الخامس Root mean squared error جذر متوسط مربع الخطأ يساوي 0,3876

- السطر السادس Relative absolut error الخطأ المطلق النسبي يساوي 52,4192%

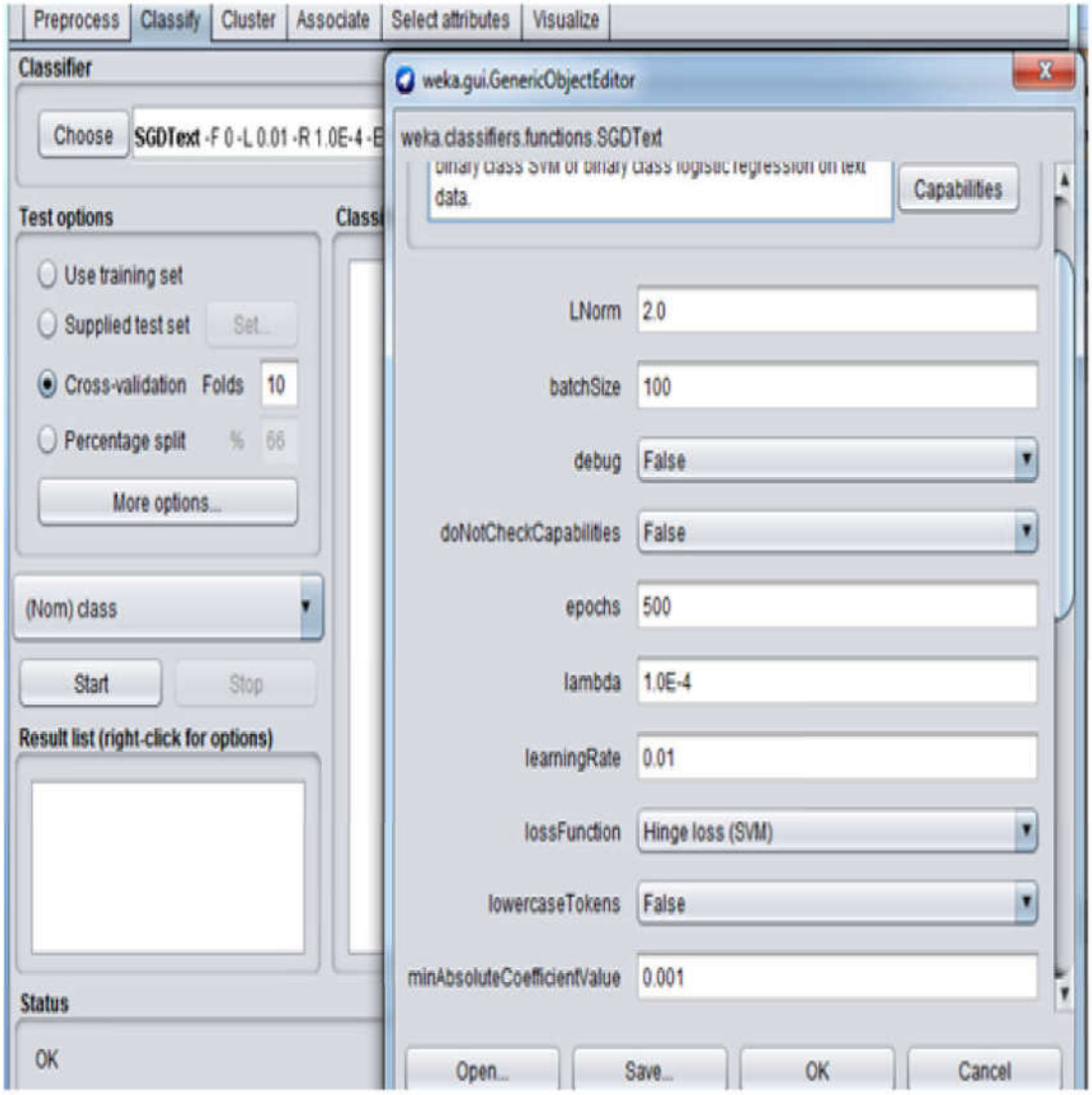
- السطر السابع Root relative squared error هو الجذر التربيعي النسبي الخطأ يساوي 81,789%

6.4 التنقيب في المعطيات بالاستعمال خوارزمية التصنيف (SVM) :

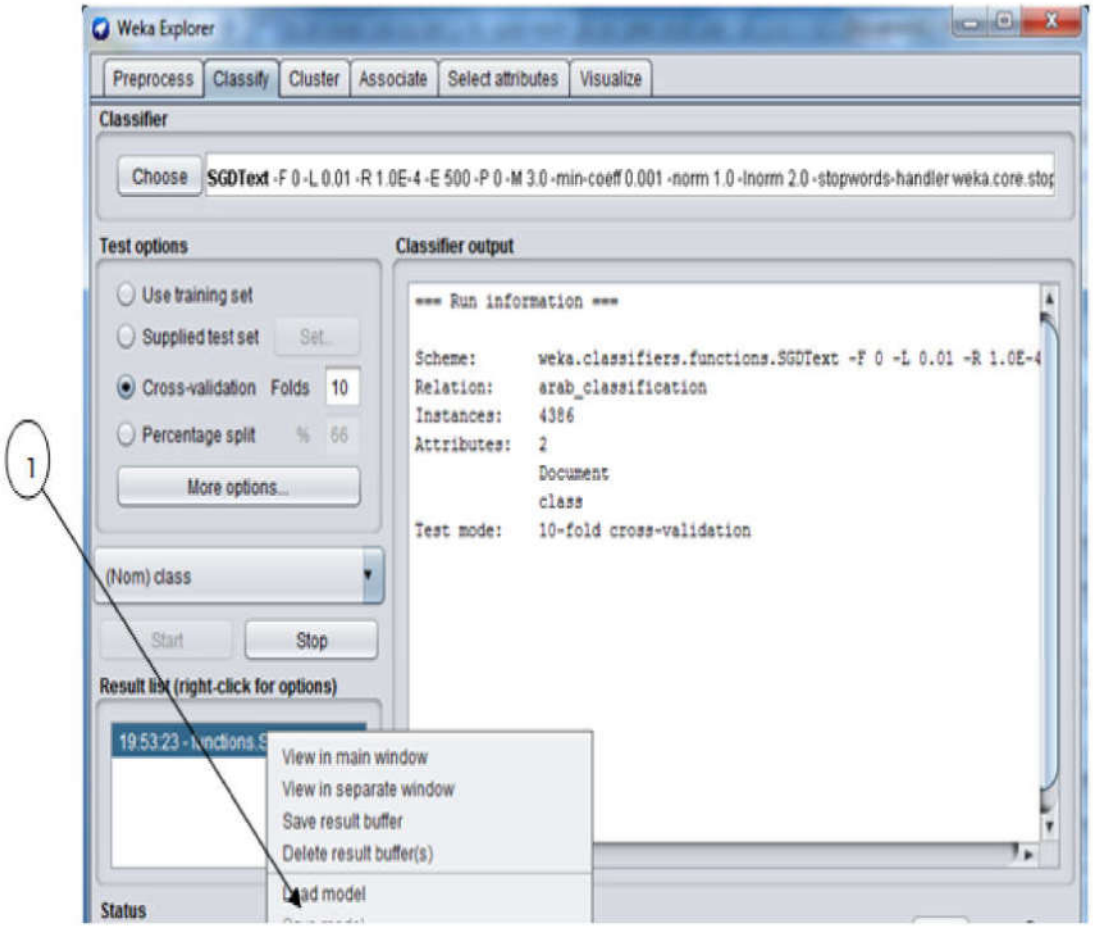
لتطبيق خوارزمية SVM على بيانات التدريب لا بد من بناء نموذج قادر على تصنيف النصوص وفق فئتين أساليب خبرية وإنشائية، وبعد بناء النموذج يتم اختبار دقته بتطبيقه على معطيات الاختبار وذلك باختبارات الأداء Cross validation , split Percentage 66% , Using training set والأشكال التالية توضح أهم المراحل المتبعة للعمل بهذه الخوارزمية التي يتيحها لنا برنامج ويكا.



شكل 20.4: اختيار خوارزمية SVM



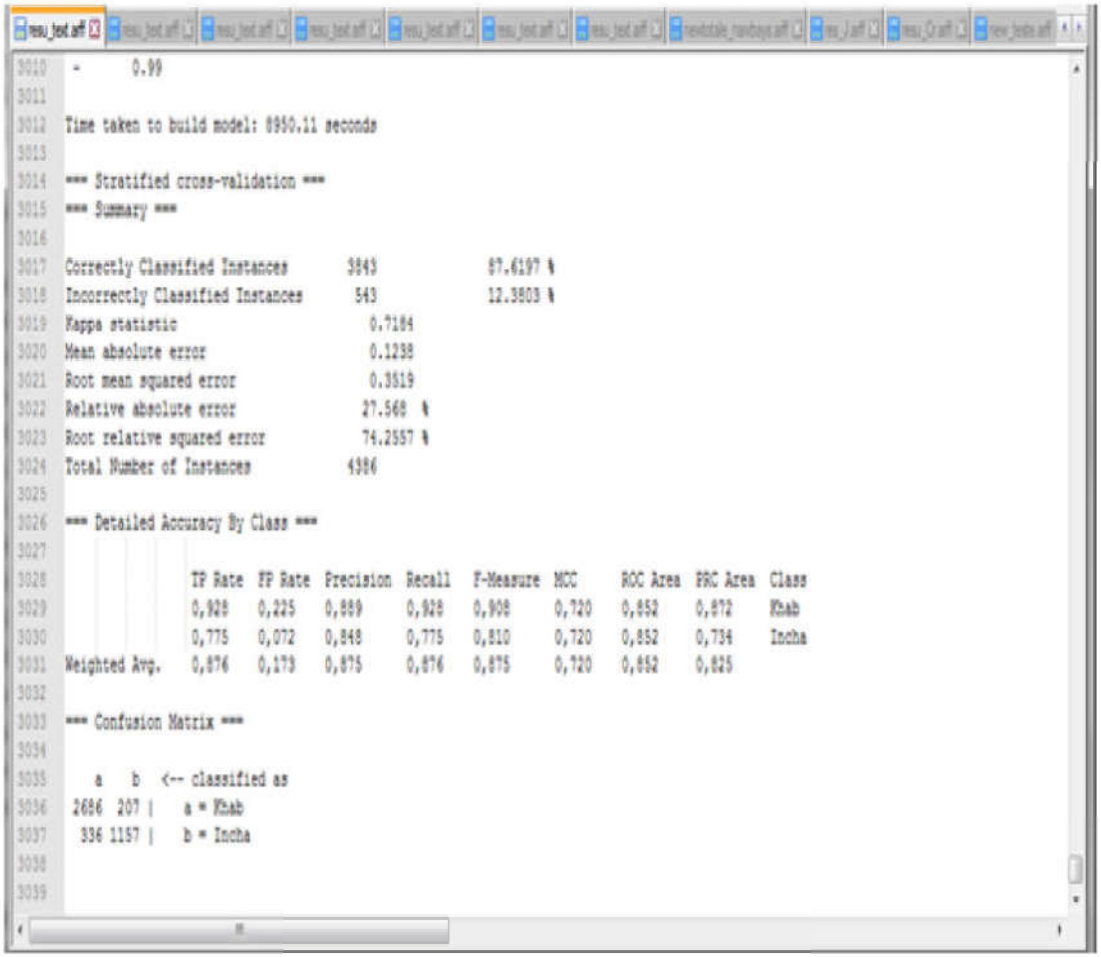
شكل 21.4: إعداد خوارزمية SVM



شكل 22.4: الاحتفاظ بالنموذج ناتج عن اختبار Cross validation للخوارزمية SVM المراد تطبقها للتنبؤ.

1.6.4 اختبار المصنف (SVM) :

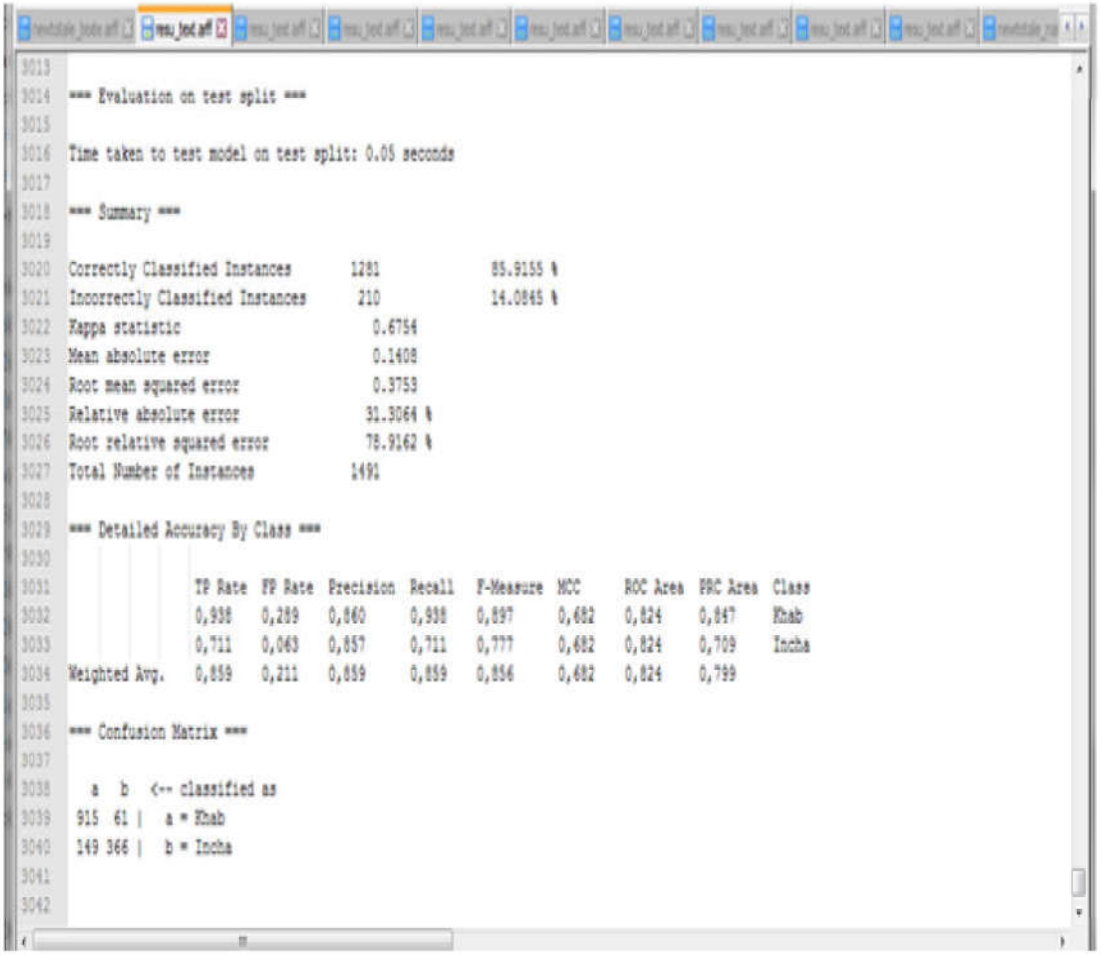
- اختبار (Cross validation)



```
-       0.99

Time taken to build model: 8950.11 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        3843               87.6197 %
Incorrectly Classified Instances       543               12.3803 %
Kappa statistic                          0.7184
Mean absolute error                      0.1238
Root mean squared error                  0.3519
Relative absolute error                 27.568  %
Root relative squared error             74.2557 %
Total Number of Instances             4386

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0,928     0,225     0,889     0,928     0,908      0,720     0,852     0,872     Khab
               0,775     0,072     0,849     0,775     0,810      0,720     0,852     0,734     Incha
Weighted Avg.  0,876     0,173     0,875     0,876     0,875      0,720     0,852     0,825

=== Confusion Matrix ===

   a    b   <-- classified as
 2686  207 |    a = Khab
  336 1157 |    b = Incha
```

شكل 23.4: اختبار Cross validation للمصنف SVM

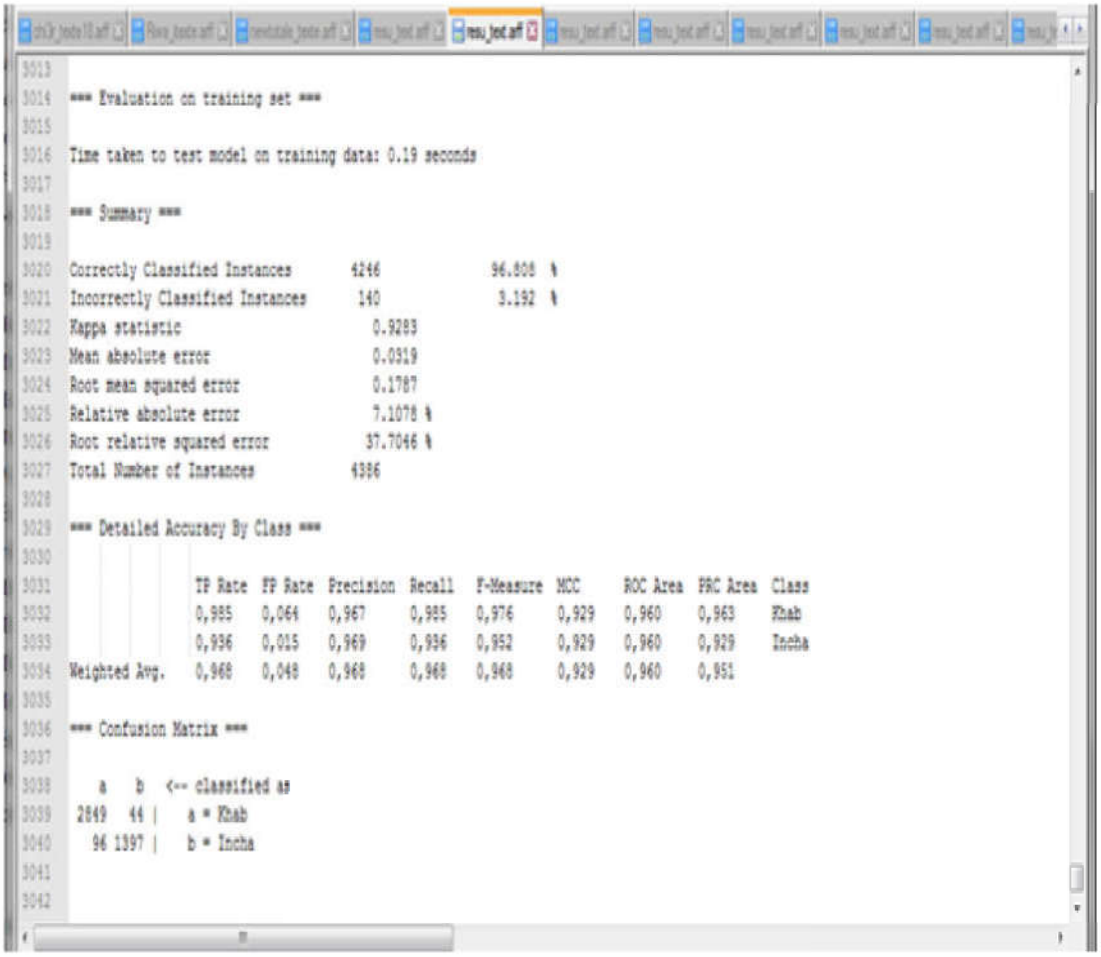
• اختبار(Percentage split 66%)



```
=== Evaluation on test split ===

Time taken to test model on test split: 0.05 seconds

=== Summary ===

Correctly Classified Instances        1281               85.9155 %
Incorrectly Classified Instances       210               14.0845 %
Kappa statistic                          0.6754
Mean absolute error                      0.1408
Root mean squared error                  0.3753
Relative absolute error                 31.3064 %
Root relative squared error             78.9162 %
Total Number of Instances             1491

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0,938    0,289    0,860      0,938   0,897      0,682    0,824     0,847     Khab
               0,711    0,063    0,857      0,711   0,777      0,682    0,824     0,709     Incha
Weighted Avg.  0,859    0,211    0,859      0,859   0,856      0,682    0,824     0,799

=== Confusion Matrix ===

   a   b   <-- classified as
 915  61 |   a = Khab
 149 366 |   b = Incha
```

شكل 24.4: اختبار Percentage split 66% للمصنف SVM

• اختبار (Using training set)



```
=== Evaluation on training set ===

Time taken to test model on training data: 0.19 seconds

=== Summary ===

Correctly Classified Instances        4246               96.808 %
Incorrectly Classified Instances       140                3.192 %
Kappa statistic                          0.9283
Mean absolute error                      0.0319
Root mean squared error                  0.1787
Relative absolute error                  7.1078 %
Root relative squared error             37.7046 %
Total Number of Instances             4386

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0,985    0,064    0,967      0,985   0,976      0,929    0,960     0,963     Khab
               0,936    0,015    0,969      0,936   0,952      0,929    0,960     0,929     Incha
Weighted Avg.  0,968    0,048    0,968      0,968   0,968      0,929    0,960     0,951

=== Confusion Matrix ===

    a    b   <-- classified as
 2849   44 |    a = Khab
   96 1397 |    b = Incha
```

شكل 25.4: اختبار Using training set للمصنفSVM

2.6.4 استظهار النتائج:

نشاهد عدد الحالات المصنفة في اختبار (Cross Validation) للمصنف SVM بشكل صحيح هو 3843 حالة بنسبة مئوية مقدارها 87,6197% وعدد الحالات المصنفة بشكل غير صحيح هو 543 حالة بنسبة 12,3803% .

- السطر الثالث يمثل مقياس لتصحيح احتمال الاتفاق بين التصنيفات الحقيقية كابا إحصاء

(Kappa Statistiques) والتي كان مقدارها 0,7184 حيث $K = \frac{P0-Pe}{1-Pe}$

- السطر الرابع نجد (Mean absolut error) الخطأ المطلق في المتوسط ويستخدم معدلات الخطأ للتنبؤ الرقمي بدلا من التصنيف حيث ام التنبؤات ليست فقط الصحيحة و الخاطئة Mean absolut error =0,1238

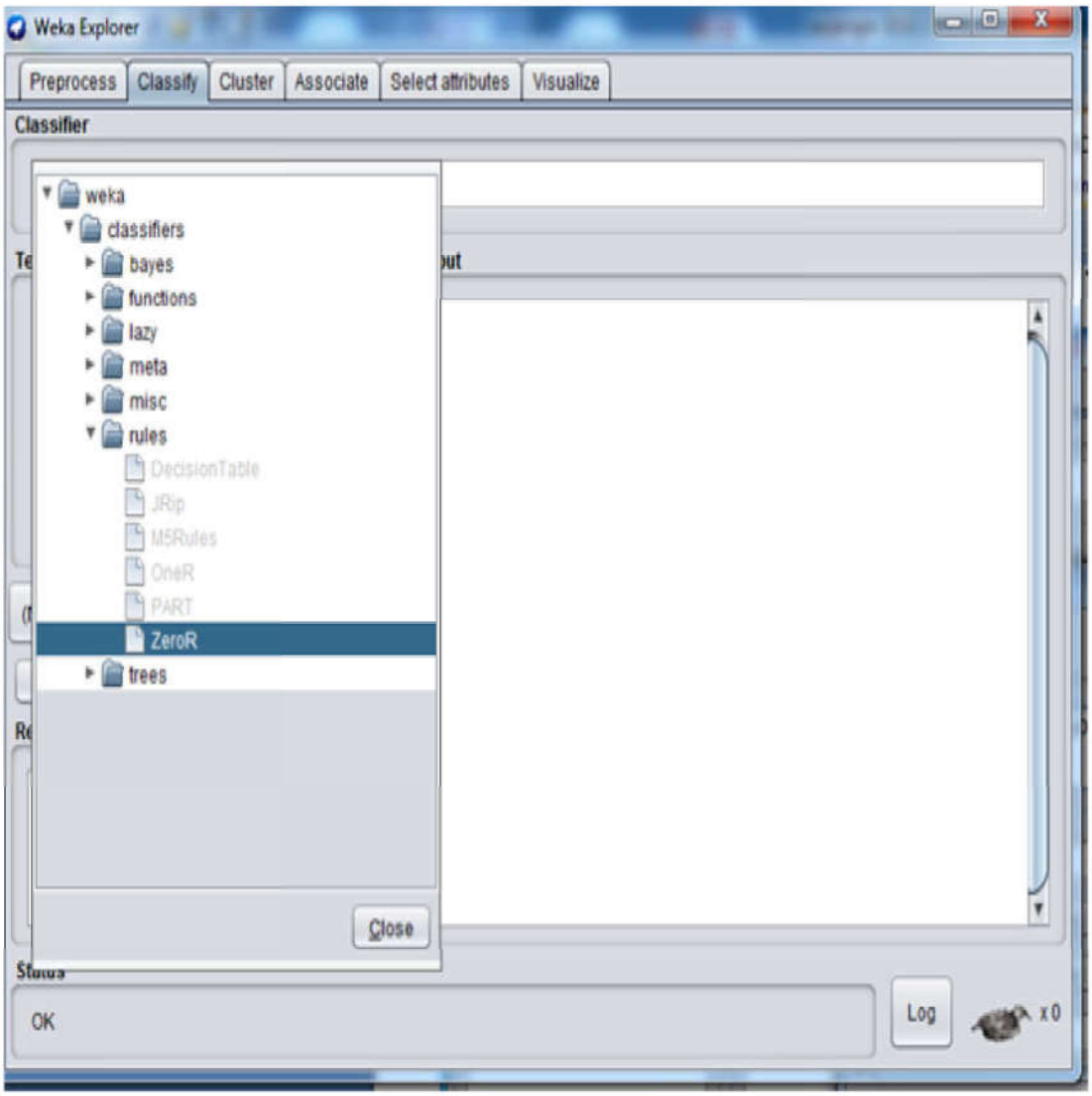
-السطر الخامس Root mean squared error جذر متوسط مربع الخطأ يساوي 0,3519

- السطر السادس Relative absolut error الخطأ المطلق النسبي يساوي 27,568%

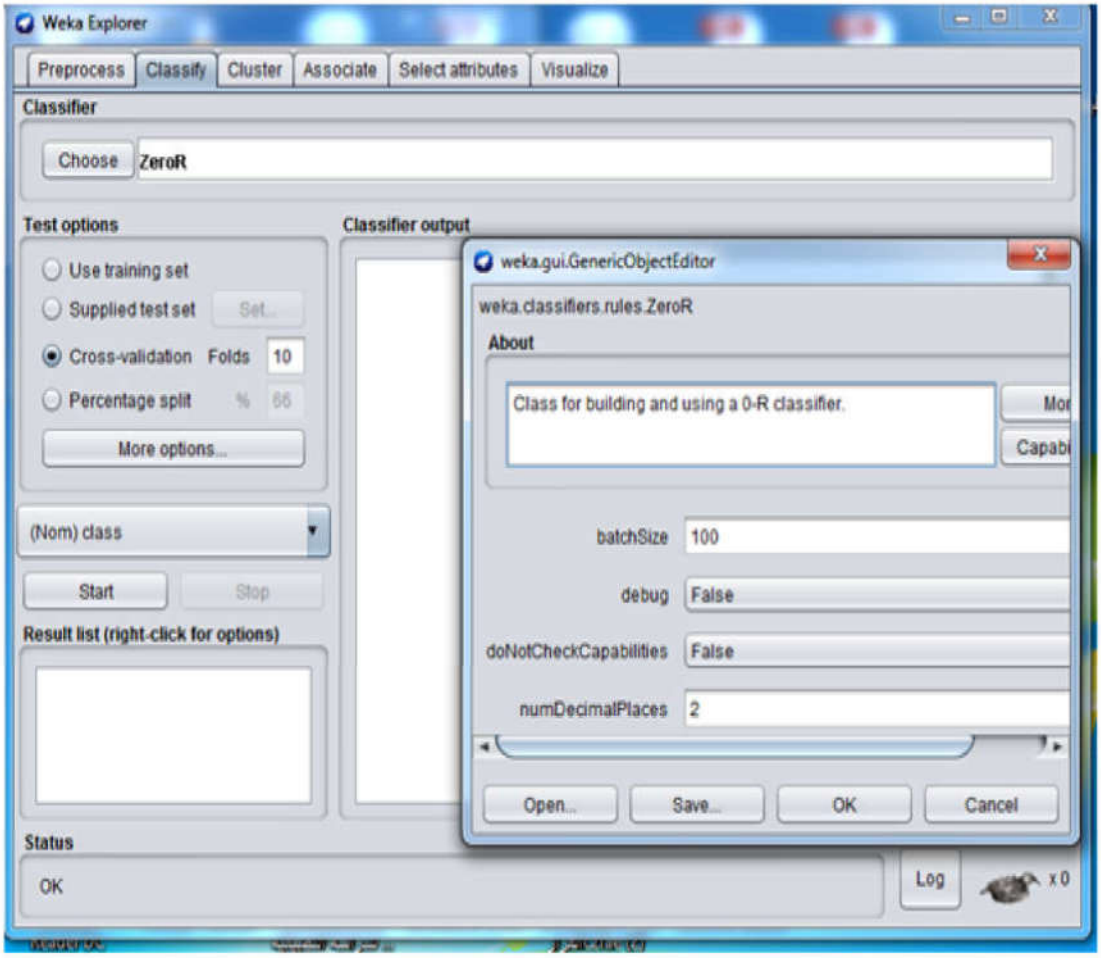
- السطر السابع Root relative squared error هو الجذر التربيعي النسبي الخطأ يساوي 74,2557% .

7.4 التنقيب في المعطيات بالاستعمال خوارزمية التصنيف (Zero R)

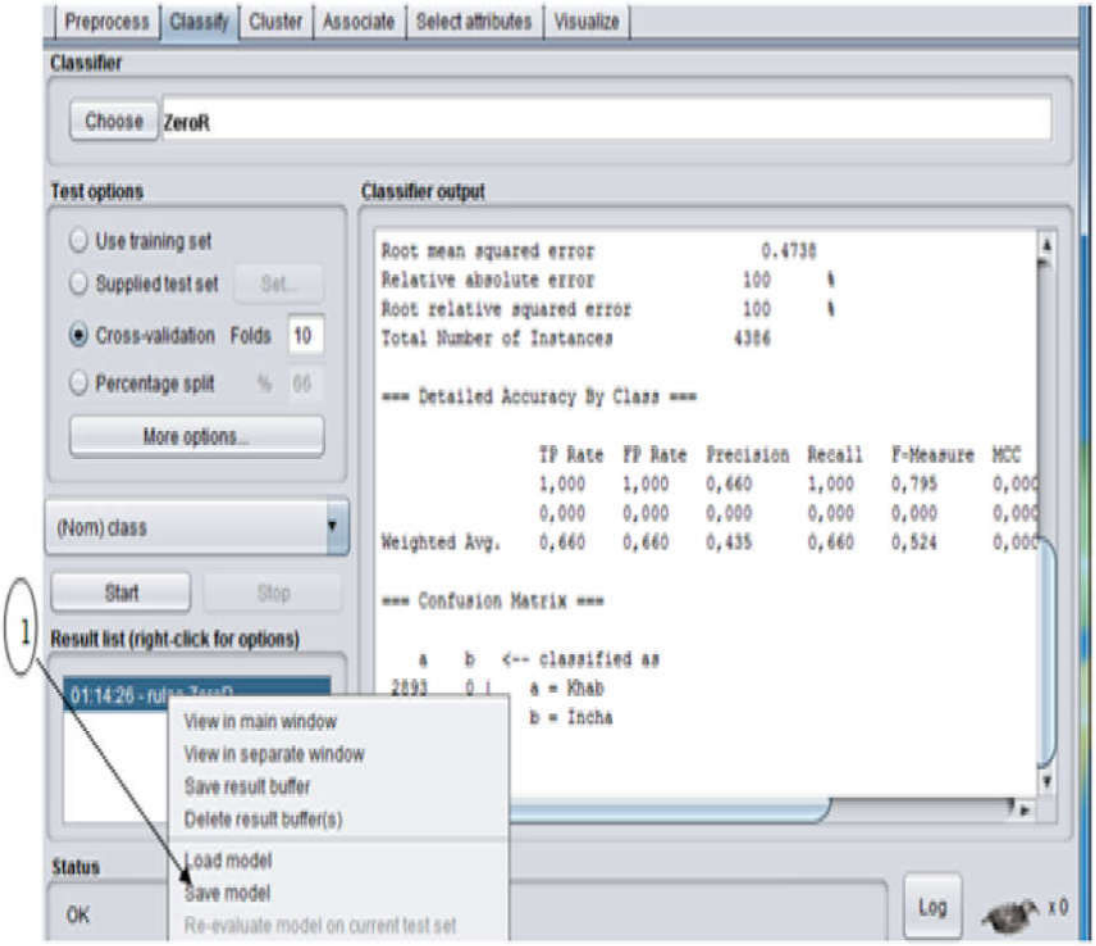
تعد Zero R من ابسط الخوارزميات تستخدم عادة في عملية التصنيف تعمل على تحديد الفئات المطلوبة، إلا أنها ليست لديها قدرة تنبؤية، طريقة العمل بهذه الخوارزمية يكون أسهل مقارنة بالخوارزميات الأخرى يتم تقييم أداء هذه الخوارزمية من خلال مجموعة من الاختبارات هي Using training set , Percentage split 66% , Cross validation ويتضح ذلك من خلال الاشكال التالية:



شكل 26.4: اختيار خوارزميةZeroR



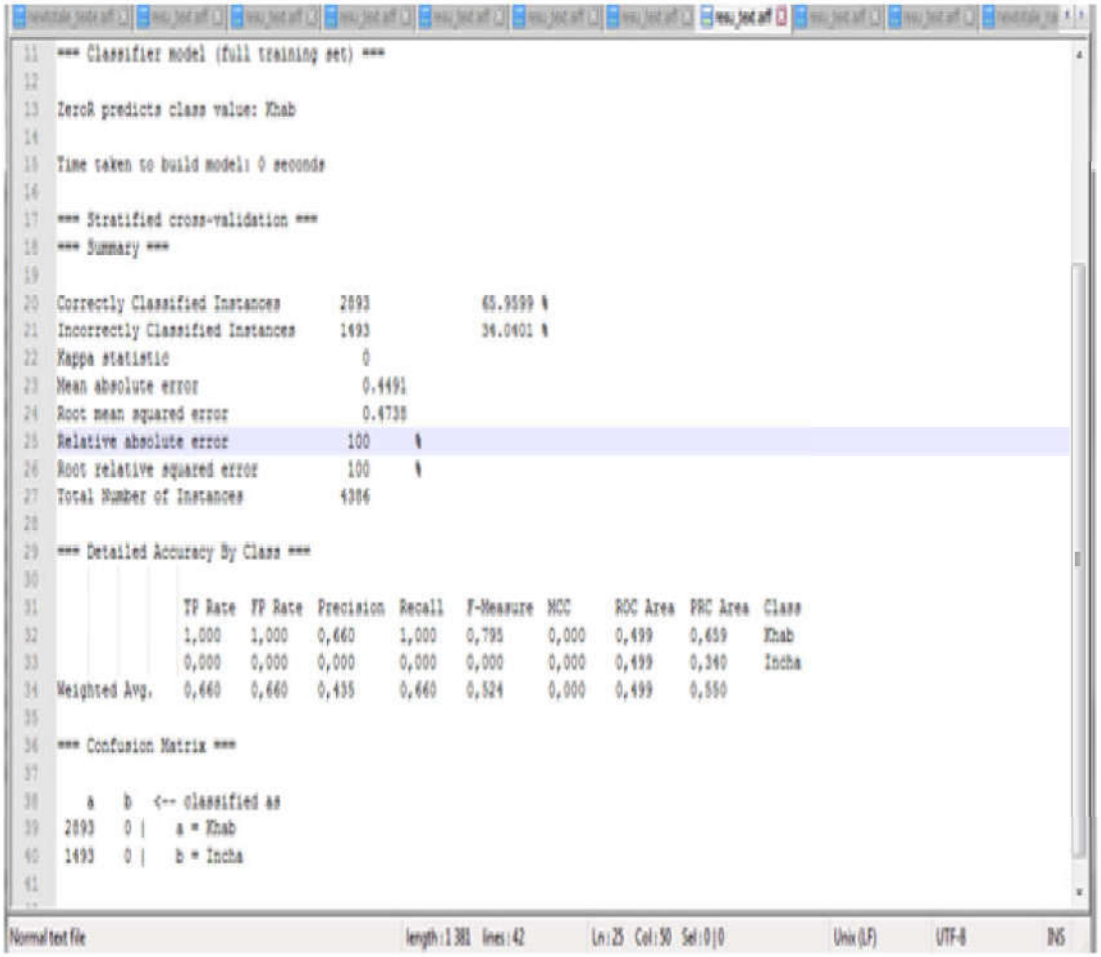
شكل 27.4: إعداد خوارزمية ZeroR



شكل 28.4: الاحتفاظ بالنموذج ناتج عن اختبار Cross validation للخوارزمية ZeroR المراد تطبقيها للتنبؤ.

1.7.4 اختبار المصنف ZeroR :

- اختبار (validation Cross)

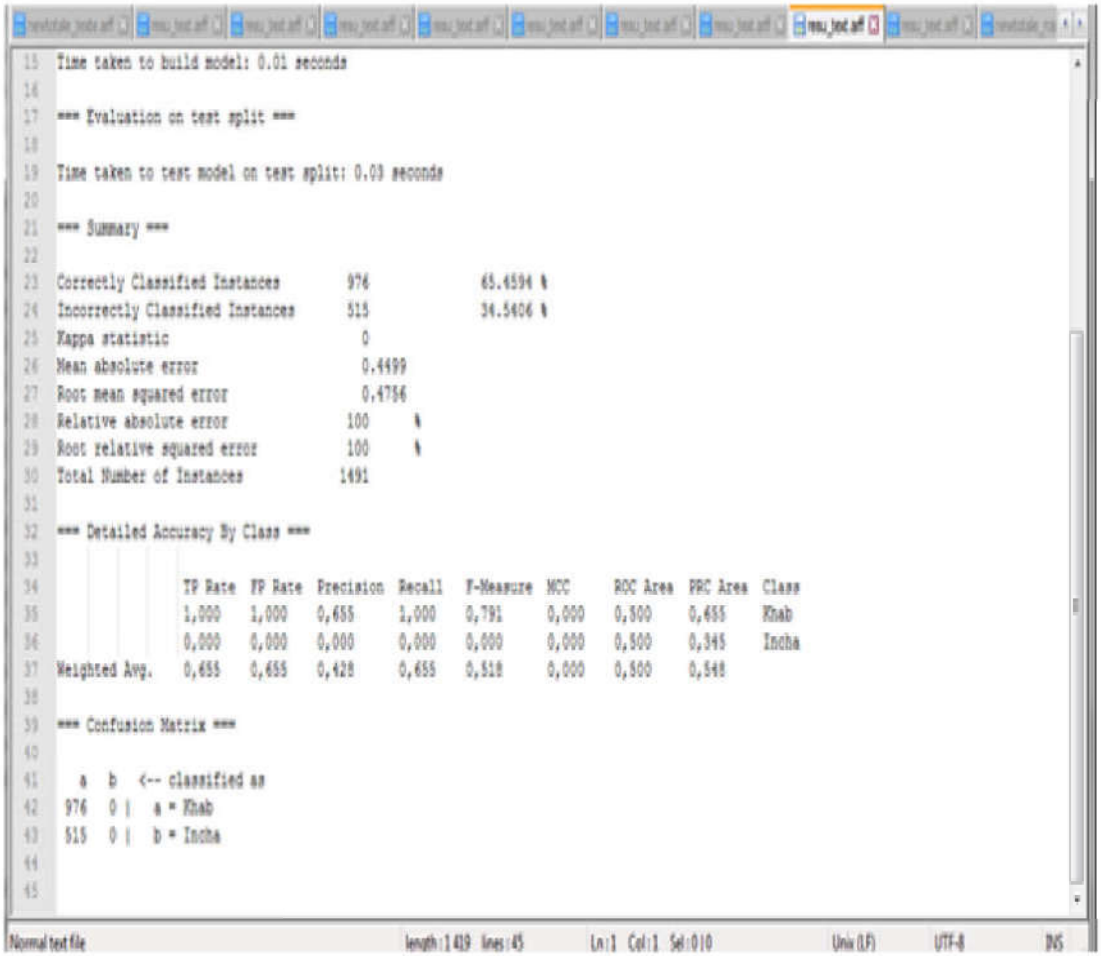


```
=== Classifier model (full training set) ===

ZeroR predicts class value: Khab

Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        2893               65.9599 %
Incorrectly Classified Instances      1493               34.0401 %
Kappa statistic                          0
Mean absolute error                      0.4491
Root mean squared error                  0.4738
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances             4386

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               1,000    1,000    0,660      1,000   0,795      0,000    0,499     0,659     Khab
               0,000    0,000    0,000      0,000   0,000      0,000    0,499     0,340     Incha
Weighted Avg.  0,660    0,660    0,435      0,660   0,524      0,000    0,499     0,550

=== Confusion Matrix ===

   a   b   <-- classified as
2893   0 |   a = Khab
1493   0 |   b = Incha
```

Normal text file | length : 1 381 lines : 42 | Ln : 25 Col : 50 Sel : 0 | 0 | Unix (LF) | UTF-8 | INS

شكل 29.4: اختبار validation Cross للمصنف(ZeroR)

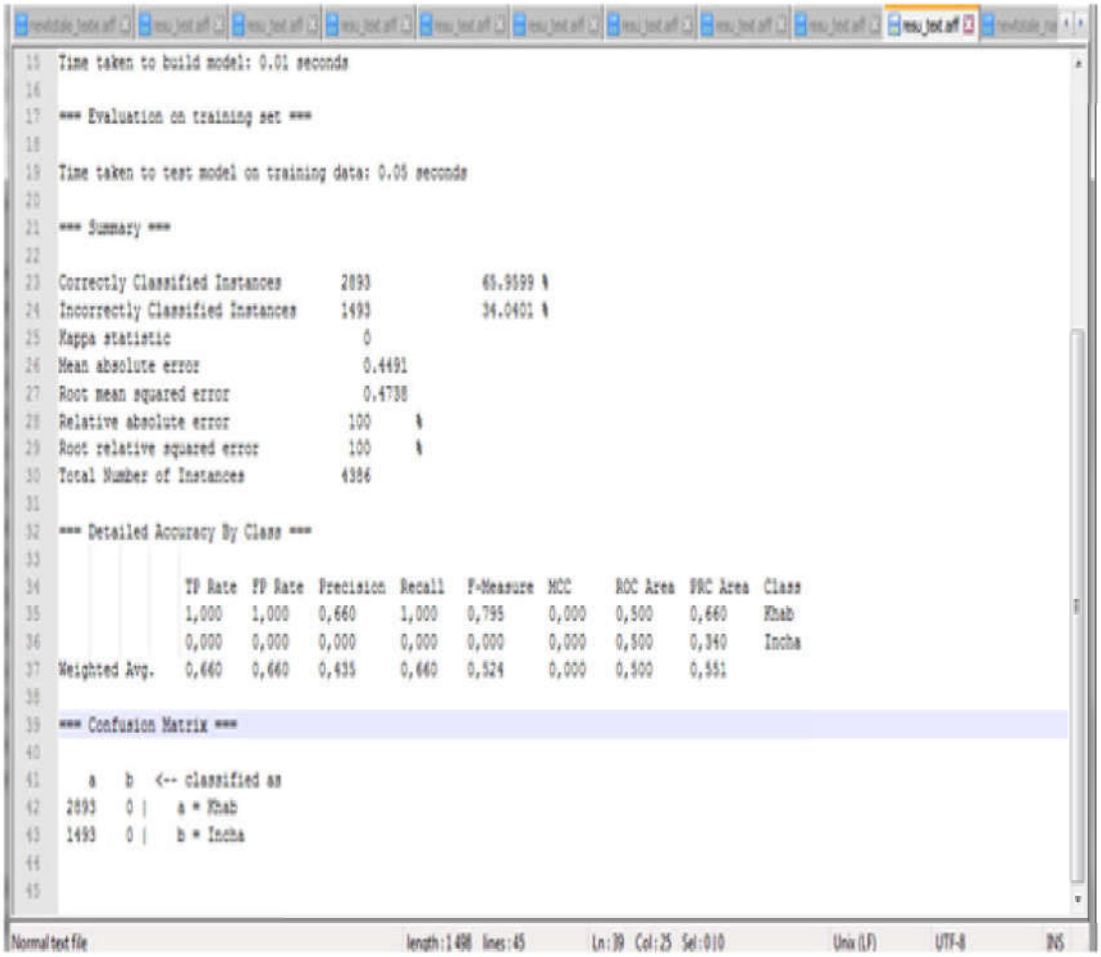
• اختبار(Percentage split 66%)



```
Time taken to build model: 0.01 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.03 seconds

=== Summary ===

Correctly Classified Instances         976               65.4594 %
Incorrectly Classified Instances       515               34.5406 %
Kappa statistic                          0
Mean absolute error                      0.4499
Root mean squared error                  0.4756
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances             1491

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               1,000    1,000    0,655      1,000    0,791      0,000    0,500     0,655     Khab
               0,000    0,000    0,000      0,000    0,000      0,000    0,500     0,345     Incha
Weighted Avg.  0,655    0,655    0,428      0,655    0,518      0,000    0,500     0,548

=== Confusion Matrix ===

   a   b   <-- classified as
 976   0 |   a = Khab
 515   0 |   b = Incha
```

شكل 30.4: اختبار Percentage split 66% للمصنف (ZeroR)

• اختبار (Using training set)



```
Time taken to build model: 0.01 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.05 seconds

=== Summary ===

Correctly Classified Instances        2893               65.9599 %
Incorrectly Classified Instances      1493               34.0401 %
Kappa statistic                          0
Mean absolute error                      0.4491
Root mean squared error                  0.4738
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances             4386

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               1,000    1,000    0,660      1,000    0,795      0,000    0,500     0,660     Khab
               0,000    0,000    0,000      0,000    0,000      0,000    0,500     0,340     Incha
Weighted Avg.  0,660    0,660    0,435      0,660    0,524      0,000    0,500     0,551

=== Confusion Matrix ===

   a   b   <-- classified as
2893   0 |   a = Khab
1493   0 |   b = Incha
```

شكل 31.4: اختبار Using training set للمصنف (ZeroR)

2.7.4 استظهار النتائج:

من خلال اعتمادنا على اختبار Cross Validation للمصنف ZeroR نشاهد عدد الحالات المصنفة بشكل صحيح هو 2893 حالة بنسبة مئوية مقدارها 65,9599% وعدد الحالات المصنفة بشكل غير صحيح هو 1493 حالة بنسبة 34,0401% .

أما السطر الثالث يمثل مقياس لتصحيح احتمال الاتفاق بين التصنيفات الحقيقية إحصاء كابا

(Kappa Statistiques) و التي كان مقدارها 0 حيث $K = \frac{P0-Pe}{1-Pe}$

أما السطر الرابع نجد Mean absolut error (الخطأ المطلق في المتوسط) ويستخدم معدلات الخطأ للتنبؤ الرقمي بدلا من التصنيف حيث ان التنبؤات ليست فقط الصحيحة والخاطئة

Mean absolut error =0,4491

أما السطر الخامس Root mean squared error جذر متوسط مربع الخطأ يساوي 0,4738 ويليه السطر السادس Relative absolut error الخطأ المطلق النسبي يساوي 100% أما السطر السابع Root relative squared error هو الجذر التربيعي النسبي الخطأ يساوي100% .

8.4 التنفيذ والمقارنة بين الخوارزميات:

تم تجربة هذه الخوارزميات لتحديد الأفضل بالنظر إلى المعايير التي تم توضيحها، واختيار الأمثل لاستخدامها في حل مشكل تصنيف النصوص والجدول (2-4) يوضح المقارنة بين أفضل الخوارزميات التي تم اختيارها وذلك بعد تجربة العديد منها والتي يوفرها برنامج Weka وقصد معرفة أداءها ركزنا على اختبارين هما Cross-validation و Percentage split 66% ، أما استخدام طريقة الاختبار use training set لاحظنا أن جميع النتائج تعطي تقريب 100%أي أن هنالك إشكالية في استخدام هذا النوع من الاختبار، حيث يتم تدريب البيانات واختبارها على نفسها وليس هناك تحديد لنسبة الاختبار لمعرفة مدى دقة الخوارزمية عند اختبارها على الحزم البيانية وبالتالي لا نعتمد كثيرا على هذا النوع من الاختبارات.

توصلنا إلى أن أفضل نسبة كانت لاختبار Cross-validation فقد تحصلنا على أفضل نتيجة بتقسيم الحزمة إلى عدد folds 10 ويليه اختبار Percentage split 66%

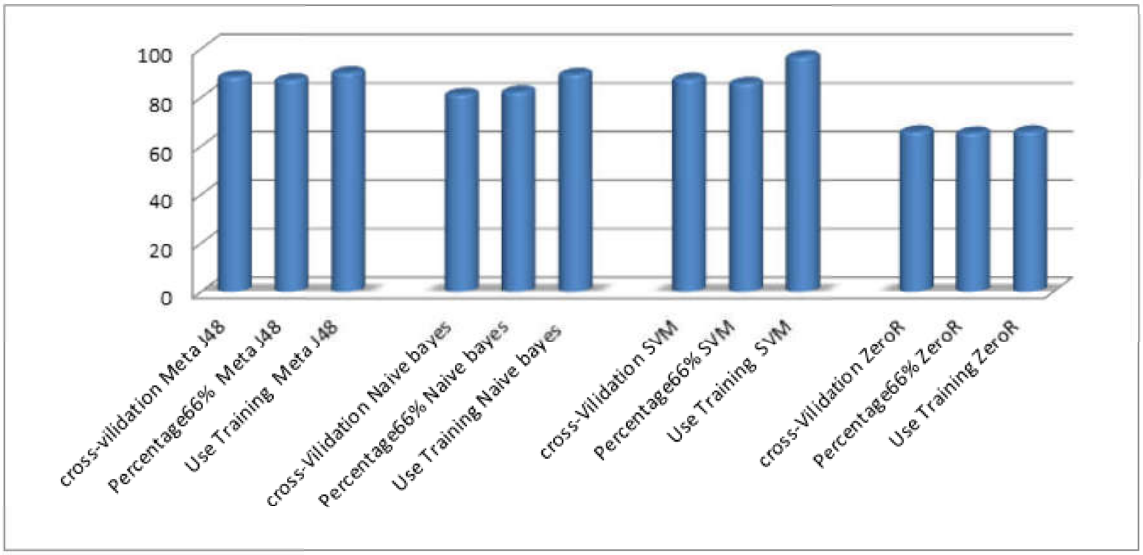
Algorithms	Test option	Correctly	Incorrectly
Meta J48	Cross-validation (10) folds	88,5545%	11,4455%
	Percentage split 66%	87,3239%	12,6761%
	Use Training	90,2873%	9,7127%
Naïve Bayes	Cross-validation (10) folds	81,3725%	18,6275%
	Percentage split 66%	82,4279%	17,5721%
	Use Training	89,6489%	10,3511%
SVM	Cross-validation (10) folds	87,6197%	12,3803%
	Percentage split 66%	85,9155%	14,0845%
	Use Training	96,808%	3,192%
ZeroR	Cross-validation (10) folds	65,9599%	34,0401%
	Percentage split 66%	65,4594%	34,5406%
	Use Training	65,9599%	34,0401%

جدول 1.4: النسب الصحيحة والخاطئة لخوارزميات التصنيف

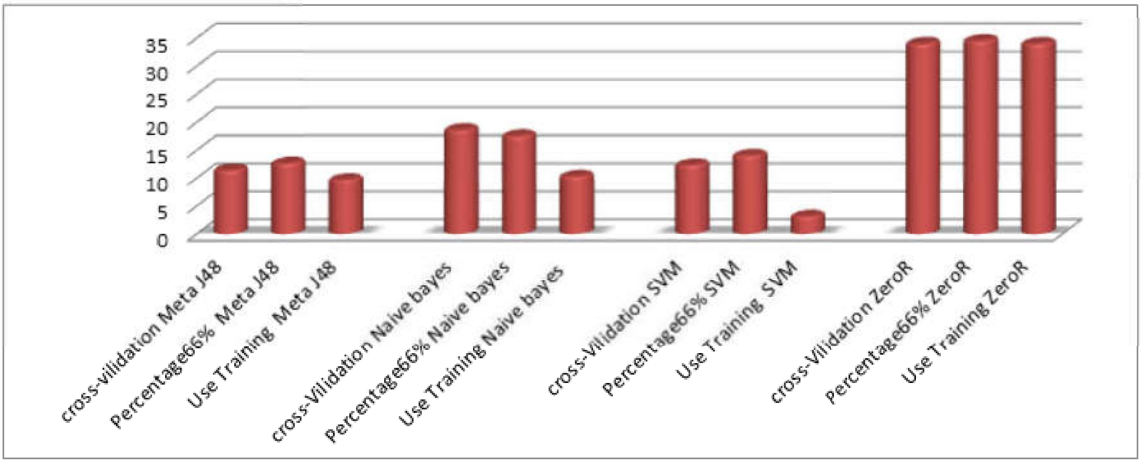
بناءا على الاختبارات التي أجريناها على الخوارزميات المستخدمة في عملية التصنيف وبالأخص اختبار Cross-validation لكون هذا الأخير تحصلنا من خلاله على أفضل النتائج نظرا لتقسيم حزمة البيانات إلى عدد 10 folds وبعد المقارنة بين النتائج المتوصل إليها، تبين أن النسب كانت متقاربة إلى حد ما حيث كانت نسبة دقة خوارزمية J48 تعادل (88,5545%) بينما تليها خوارزمية SVM بنسبة (87,6197%) أما خوارزمية Naive Bayes ظهرت بنسبة (81,3725%)

وأخيرا خوارزمية ZeroR بنسبة (65,9599%).

بعد استظهار هذه النتائج التي يوضحها الشكل (4- 32) تبين أن دقة خوارزمية J48 كانت أفضل من ناحية إبراز الحالات المصنفة بشكل صحيح أما بالنسبة للحالات المصنفة بشكل خاطئ فانه يظهر من خلال قراءة النتائج بشكل عكسي وهذا ما يوضحه الشكل (33.4).



شكل 32.4: نسبة الحالات المصنفة بشكل صحيح



شكل 33.4: نسبة الحالات المصنفة بشكل خاطئ

9.4 مقاييس تقييم أداء الخوارزميات

يتم معرفة أداء الخوارزميات من خلال مجموعة من مقاييس الأداء التي تعمل على تحديد النسبة المئوية للحالات المصنفة بشكل صحيح مع توضيح نسبة الحالات المصنفة بشكل خاطئ، تظهر مقاييس الأداء في شاشة Accuracy class التي توضح نتائج دقة الفئات المصنفة لكلتا الخوارزميتين وهي فئة الخبري والإنشائي ومن بين هذه المقاييس نجد مقياسRoc مصفوفة التشويش(الشك) Confusion matrix ، F-Meausure ، recall ، precision سنكتفي بعرض نتائج F-Meausure ، مصفوفة التشويش(الشك) Confusion matrix ، مقياس الأداء ROC ، recall للمصنفات المعتمدة في الدراسة ونستعرض فيما يلي المقاييس مع ذكر أهم النتائج المتوصل إليها .

1.9.4 مصفوفة الشك (التشويش) Confusion matrix :

تعتبر من أهم مقاييس الأداء، يكمن دورها في تقييم أداء المصنف بحساب عدد الحالات المتوقعة المصنفة بشكل صحيح والمصنفة بشكل خاطئ، وهي عبارة عن جداول تحتوي على قيم التصنيفات الحقيقية والخاطئة ومن خلال مصفوفات الشك المبينة في الأشكال التالية نستخلص مجموعة مختلفة من مقاييس الدقة مع توضيح قيمها لكل مصنف.

- **مصفوفة الشك بالاستعمال خوارزمية J48 :**

		التصنيف المتوقع	
		A	B
التصنيف الحقيقي	A	2799	94
	B	408	1085

مصفوفة الشك Confusion Matrix

Meta(J48)Cross-validation(10)

		التصنيف المتوقع	
		A	B
التصنيف الحقيقي	A	953	23
	B	166	349

مصفوفة الشك Confusion Matrix

Meta(J48)Percentage Split66%

		التصنيف المتوقع	
		A	B
التصنيف الحقيقي	A	2823	70
	B	345	1148

مصفوفة الشك Confusion Matrix

Meta(J48)Use training

شكل 34.4: مصفوفة الشك بالاستعمال خوارزمية J48

إذ نجد أن هذه المصفوفات في اختبار Cross-validation(10) تتوقع عدد الأصناف الايجابية 2799 والخاطئة 94 وتتوقع عدد الأصناف السلبية الصحيحة 1085 والخاطئة 408 ويمكننا أيضا من حساب القيم التالية:

-التوقعات الصحيحة 3884=2799+1085

-االتوقعات الخاطئة المقدمة من النموذج 502=408+94

-معدل الخطأ هو $0.114 = \frac{(408+94)}{408+2799+94+1083}$

-معدل الدقة الإجمالي هو Accuracy Overall $0.904 = \frac{3884}{408+2799+94+1083}$

-الدقة المتوسطة هي Average Accuracy $0.904 = (\frac{2779}{2779+94} + \frac{1085}{1085+408})/2$

- Kappa statistics إحصائية كابا K = 0,731

بالنسبة لاختبار Percentage Split 66%

-التوقعات الصحيحة 1302=953+349

-التوقعات الخاطئة المقدمة من النموذج 189=166+23

-معدل الخطأ هو $0.0946 = \frac{166+23}{953+349+166+23}$

-معدل الدقة الإجمالي هو Accuracy Overall $0,873 = \frac{1302}{953+349+166+23}$

-الدقة المتوسطة هي Average Accuracy $0,8270 = (\frac{953}{953+23} + \frac{349}{166+349})/2$

- Kappa statistics إحصائية كابا K = 0,699

بالنسبة لاختبار Use training

-التوقعات الصحيحة 3971=2823+1148

-التوقعات الخاطئة المقدمة من النموذج 415=345+70

-معدل الخطأ هو $0,126 = \frac{(345+70)}{2823+70+345+1148}$

-معدل الدقة الإجمالي هو Accuracy Overall $0,9053 = \frac{3971}{2823+70+345+1148}$

-الدقة المتوسطة هي Average Accuracy $0,8723 = (\frac{2823}{2823+70} + \frac{1148}{345+1148})/2$

- Kappa statistics إحصائية كابا K = 0,779

• مصفوفة الشك بالاستعمال خوارزمية **Naive bayse** :

		التصنيف المتوقع	
		A	B
التصنيف الحقيقي	A	2321	572
	B	245	1248

Matrix Confusion مصفوفة الشك Naïve Bayés Cross-validation(10)

		التصنيف المتوقع	
		A	B
التصنيف الحقيقي	A	2562	331
	B	123	1370

Matrix Confusion مصفوفة الشك Naïve Bayés Use training

		التصنيف المتوقع	
		A	B
التصنيف الحقيقي	A	810	166
	B	96	419

Matrix Confusion مصفوفة الشك Naïve Bayés Percentage Split66%

شكل 35.4: مصفوفة الشك بالاستعمال خوارزمية Naïve bayes

بالنسبة لاختبار Cross-validation(10) :

-التوقعات الصحيحة 3569=2321+1248

-التوقعات الخاطئة المقدمة من النموذج 817=572+245

-معدل الخطأ هو $0,1862 = \frac{(572+245)}{2321+572+245+1248}$

-معدل الدقة الإجمالي هو Accuracy Overall $0,8137 = \frac{3569}{2321+572+245+1248}$

-الدقة المتوسطة هي Average Accuracy $0,8190 = (\frac{2321}{2321+572} + \frac{1248}{245+1248})/2$

- Kappa statistics إحصائية كابا K = 0,604

بالنسبة لاختبار **Percentage Split 66%**

-االتوقعات الصحيحة 1229=810+419

-التوقعات الخاطئة المقدمة من النموذج 262=166+96

-معدل الخطأ هو $\frac{(166+96)}{810+419+166+96} = 0,1757$

-معدل الدقة الإجمالي هو Accuracy Overall $\frac{810+419}{810+419+166+96} = 0,824$

-الدقة المتوسطة هي Average Accuracy $(\frac{419}{419+96} + \frac{810}{166+810})/2 = 0,821$

- Kappa statistics إحصائية كابا K = 0,604

بالنسبة لاختبار Use training

-االتوقعات الصحيحة 3932=2562+1370

-التوقعات الخاطئة المقدمة من النموذج 454=123+331

-معدل الخطأ هو $\frac{(331+123)}{331+123+2562+1370} = 0,1035$

-معدل الدقة الإجمالي هو Accuracy Overall $\frac{3932}{331+123+2562+1370} = 0,896$

-الدقة المتوسطة هي Average Accuracy $(\frac{1370}{1370+123} + \frac{2562}{2562+331})/2 = 0,942$

- Kappa statistics إحصائية كابا K = 0,776

- **مصفوفة الشك بالاستعمال خوارزميةSVM :**

		التصنيف المتوقع	
		A	B
التصنيف الحقيقي	A	2686	207
	B	336	1157

Matrix Confusionمصفوفة الشكSVM Cross-validation(10)

		التصنيف المتوقع	
		A	B
التصنيف الحقيقي	A	915	61
	B	149	366

Matrix Confusionمصفوفة الشكSVM Percentage Split66%

		التصنيف المتوقع	
		A	B
التصنيف الحقيقي	A	2849	44
	B	96	1397

Matrix Confusionمصفوفة الشكSVM Use training

شكل 36.4: مصفوفة الشك بالاستعمال خوارزميةSVM

بالنسبة لاختبار Cross-validation(10) :

- التوقعات الصحيحة 3843=2686+1157

-التوقعات الخاطئة المقدمة من النموذج 543=207+336

-معدل الخطأ هو $0,1238 = \frac{(336+207)}{1157+2686+336+207}$

-معدل الدقة الإجمالي هو Accuracy Overall $0,876 = \frac{3843}{1157+2686+336+207}$

-الدقة المتوسطة هي Average Accuracy $0,8516 = (\frac{2686}{2686+207} + \frac{1157}{336+1157})/2$

- Kappa statistics إحصائية كابا K = 0,717

بالنسبة لاختبار Percentage Split 66%

-االتوقعات الصحيحة 1281=915+336

-التوقعات الخاطئة المقدمة من النموذج 210=61+149

-معدل الخطأ هو $\frac{(149+61)}{149+366+61+915} = 0,140$

-معدل الدقة الإجمالي هو Accuracy Overall $\frac{366+915}{149+366+61+915} = 0,859$

-الدقة المتوسطة هي Average Accuracy $(\frac{915}{915+61} + \frac{366}{149+366})/2 = 0,824$

- Kappa statistics إحصائية كابا K = 0,678

بالنسبة لاختبار Use training

-التوقعات الصحيحة 3932=2562+1370

-التوقعات الخاطئة المقدمة من النموذج 454=123+331

-معدل الخطأ هو $\frac{(331+123)}{331+123+2562+1370} = 0,1035$

-معدل الدقة الإجمالي هو Accuracy Overall $\frac{3932}{331+123+2562+1370} = 0,896$

-الدقة المتوسطة هي Average Accuracy $(\frac{1370}{1370-123} + \frac{2562}{2562+331})/2 = 0,942$

- Kappa statistics إحصائية كابا K = 0,776

- **مصفوفة الشك بالاستعمال خوارزميةZeroR**

		التصنيف المتوقع	
		A	B
التصنيف الحقيقي	A	2893	0
	B	1493	0

Matrix Confusionالشك مصفوفة ZeroRCross-validation(10)

		التصنيف المتوقع	
		A	B
التصنيف الحقيقي	A	976	0
	B	515	0

Matrix Confusion مصفوفةالشك
ZeroRPercentage Split66%

		التصنيف المتوقع	
		A	B
التصنيف الحقيقي	A	2893	0
	B	1493	0

Matrix Confusionالشك مصفوفة ZeroR
Use training

شكل 37.4: مصفوفة الشك بالاستعمال خوارزميةZeroR

بالنسبة لاختبار Cross-validation(10) :

- االتوقعات الصحيحة 0+2893=2893

-التوقعات الخاطئة المقدمة من النموذج 0+1493=1493

-معدل الخطأ هو $\frac{(0+1493)}{1403+2893} = 0,326$

-معدل الدقة الإجمالي هو Accuracy Overall $\frac{2893}{1403+2893} = 0,673$

-الدقة المتوسطة هي Average Accuracy $(\frac{2893}{2893} + 0)/2 = 0,5$

- Kappa statistics إحصائية كابا K = 0

بالنسبة لاختبار Percentage Split 66%

-االتوقعات الصحيحة 0+976=976

-التوقعات الخاطئة المقدمة من النموذج 0+515=515

-معدل الخطأ هو $\frac{515}{976+515} = 0,345$

-معدل الدقة الإجمالي هو Accuracy Overall $\frac{976}{976+515} = 0,654$

-الدقة المتوسطة هي Average Accuracy $(\frac{976}{976+0} + 0)/2 = 0,5$

- Kappa statistics إحصائية كابا K = 0

بالنسبة لاختبار Use training

-التوقعات الصحيحة 0+2893=2893

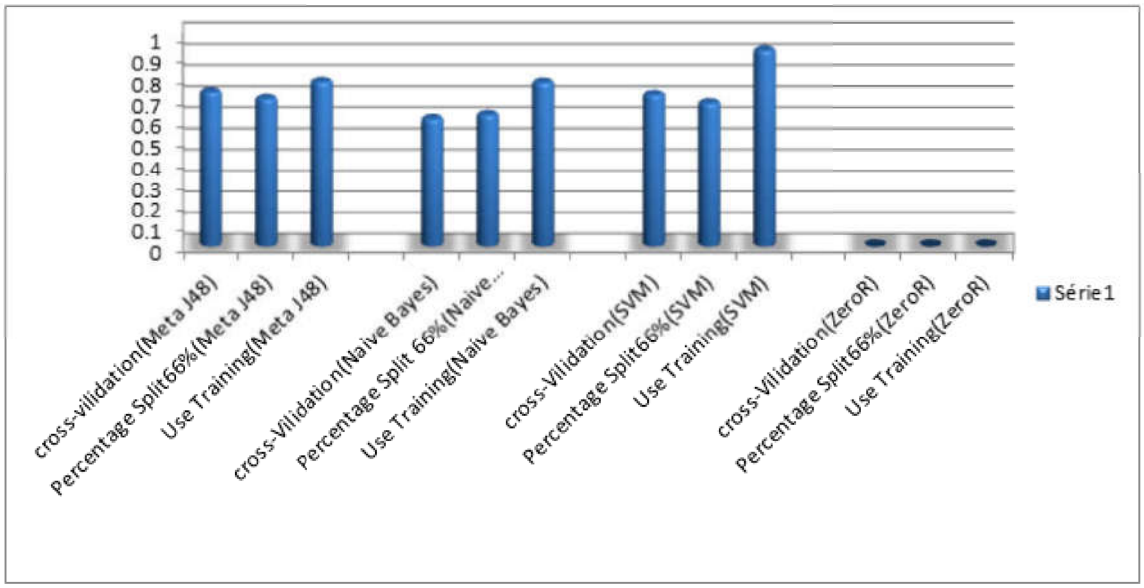
-التوقعات الخاطئة المقدمة من النموذج 0+1493=1493

-معدل الخطأ هو $\frac{(0+1493)}{1403+2893} = 0,326$

-معدل الدقة الإجمالي هو Accuracy Overall $\frac{2893}{1403+2893} = 0,673$

-الدقة المتوسطة هي Average Accuracy $(\frac{2893}{2893} + 0)/2 = 0,5$

- Kappa statistics إحصائية كابا K = 0



شكل 38.4: قيم مقياس كابا (Kappa statistics) للمصنفات

تبين جداول المصفوفات مجموعة من القيم للخوارزميات المعتمدة بعد اخضاعها لاختبارات الاداء، ومن خلال مجموعة مقاييس الدقة التي تم عرضها سابقا نكتفي بالتعليق على مقياس كابا Kappa statistics وهو مقياس لتصحيح احتمال الاتفاق بين التصنيفات والطبقات الحقيقية. $K = \frac{P0-Pe}{1-Pe}$ حيث نلاحظ من خلال الشكل(4-38) قيم مقياس كابا للمصنفات بغية المقارنة بين النتائج المتوصل لها إذ أنه بعد تقييم هذه النتائج اتضح أن خوارزمية J48 حققت دقة عالية في التصنيف أما بالنسبة لخوارزمية SVM و Naive bayse حققتا قيم متقاربة إلى حد ما، ناتج القيم لهذه الخوازميات كانت فوق 0,5 ما يؤكد دقتها في التصنيف، إلا أن خوارزمية ZeroR تحصلت على قيمة 0 كون هذا المصنف كان غير جيد في عملية التصنيف.

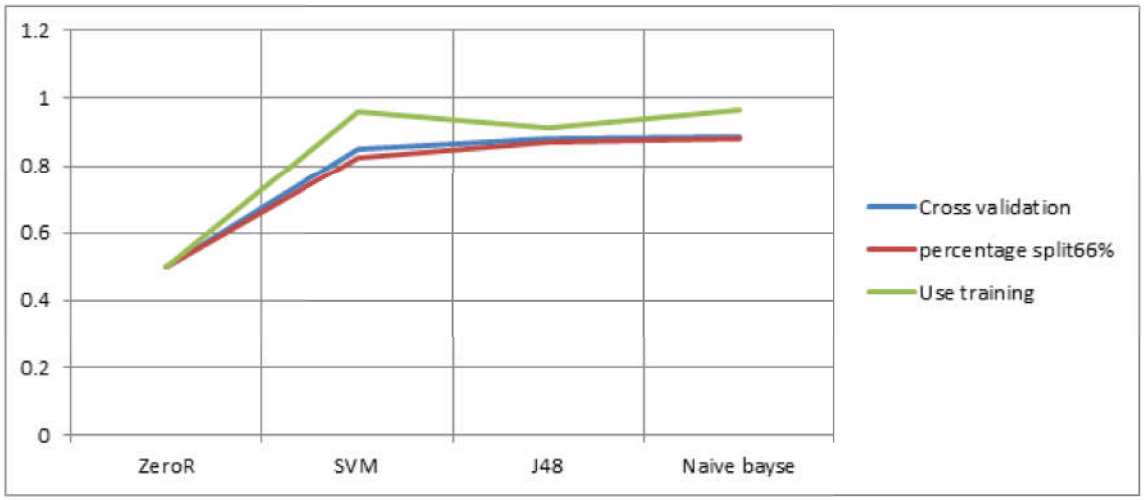
2.9.4 مقياس الأداء (receiver operating characteristic) ROC (الخصائص التشغيلية الاستقبال):

يعتبر مقياس ROC من بين المقاييس المستخدمة بكثرة لمعرفة فعالية أداء المصنف من خلال مخطط يظهر معدل القيم الايجابية الصحيحة والخاطئة، بحيث يحوي المخطط على نقطة (0-1) كلما اقترب منحنى الحالات من النقطة 1 كان أداء المصنف مثالي وكلما اقترب من 0 كان أداء المصنف ضعيف ومن خلال الدراسة التي أجريناها حاولنا استخراج معدلات قيم ROC لخوارزميات التصنيف من أجل معرفة الخوارزمية الأمثل والأكثر فعالية في عملية تصنيف النصوص وهذا ما يوضحه الجدول التالي:

ZeroR	Naïve baye	SVM	J48	اختبارات
0,4989	0,8874	0,8517	0,880	Cross validation
0,5	0,8838	0,8241	0,8711	Percentage split 66%
0,5	0,9647	0,9602	0,9113	Use training

جدول 2.4: النسب ROC لخوارزميات التصنيف

يبين هذا الجدول القيم الناتجة عن اختبارات أجريت على الخوارزميات وهي تعبر عن منحنياتROC (الأساليب الخبرية والإنشائية) ويمكن الرجوع للملحق للاطلاع على هذه المنحنيات بدء من الشكل 01 إلى غاية الشكل 12.



شكل 39.4: منحنى لقياس نسب ROC لخوارزميات التصنيف

قصد توضيح الفارق النسبي بين أداء هذه المصنفات أجرينا منحنى تجميعي لقياس نسب ROC للخوارزميات المستخدمة وهذا ما يبنه الشكل (39.4) حيث لوحظ أن المنحنيات تقترب من المستوى (1) في كل من الخوارزمية J48 ، SVM ، Naïve Bayes مما يدل على فعالية أداء هذه المصنفات إلا أنه في خوارزمية ZeroR لوحظ أن المنحنى أقل من (0,5) أي أنه يقترب من النقطة (0) مما يدل على ضعف هذا المصنف.

ما نخلص إليه بعد استعراض نتائج مقاييس الأداء أن مصنف J48 اثبت فعالية جيدة وأداء مثالي أثناء عملية تصنيف النصوص وهذا على غرار المصنفات الأخرى.

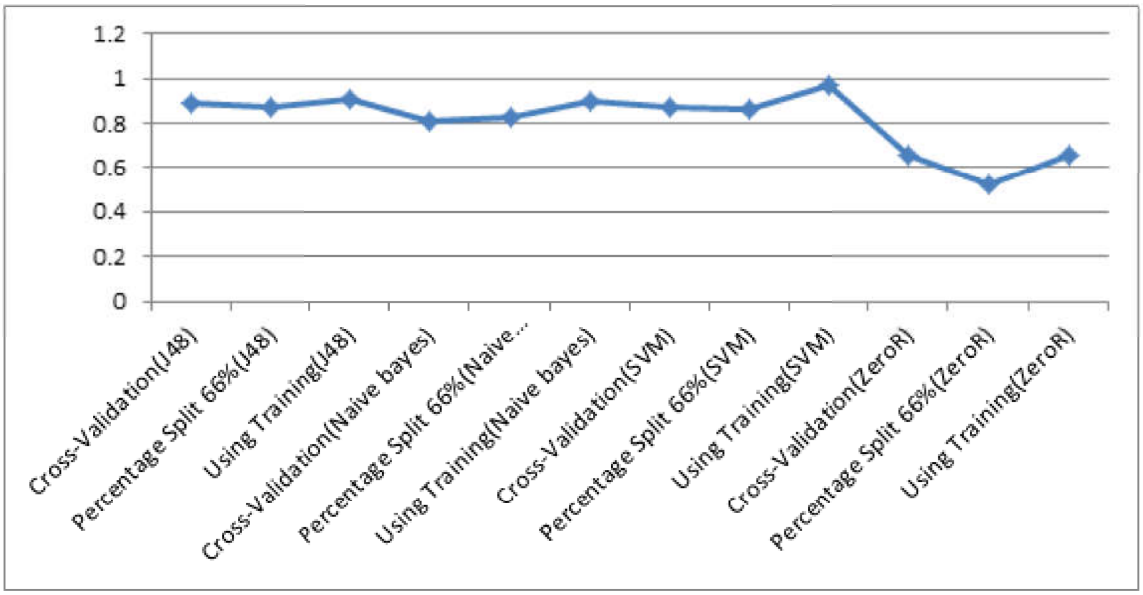
3.9.4 مقياس الدقة Recall

وهو تحديد النسبة المئوية للحالات الايجابية التي تم تصنيفها بشكل صحيح ويتحقق من خلال المعادلة التالية: $Recall = \frac{TP+FN}{TP}$

الخوارزميات	نوع الاختبار	Recall
Meta J48	Cross -validation (10) folds	0,886
	Percentage split 66%	0,873
	Use Training	0,905
Naïve Bayes	Cross-validation 10% folds	0,814
	Percentage split 66%	0,824
	Use Training	0,896
SVM	Cross-validation 10% folds	0,876
	Percentage split 66%	0,859
	Use Training	0,968
ZeroR	Cross-validation 10% folds	0,660
	Percentage split 66%	0,524
	Use Training	0,660

جدول 3.4: مقياس Recall

يوضح لنا الجدول (3.4) أهم قيم Recall المتحصل عليها من قبل الخوارزميات إذ تحصلت خوارزمية J48 على نسب عالية مقارنة بخوارزميات الاخرى ، مثلنا هذه القيم بواسطة منحنى تجميعي لإظهار الفارق النسبي بين المصنفات كما يتضح من خلال الشكل(4-40).



شكل 40.4: مقياس Recall

يبين الشكل (4-40) أهم قيم مقياس recall حيث لوحظ أن منحنى خوارزمية J48 كان في ارتفاع مستمر مقارنة بمنحنى الخوارزميات الأخرى التي تراجعت بشكل طفيف وذلك باختبار Cross validation ، أما في اختبار Percentage split 66% نجد منحنيات الخوارزميات الأخرى منخفض على عكس خوارزمية J48 التي حققت نسب عالية ، ونظرا لأننا اعتمدنا ناتج اختبار Cross validation لفعاليته فإننا نجزم القول أن خوارزمية J48 حققت دقة جد عالية في تصنيف الفئات.

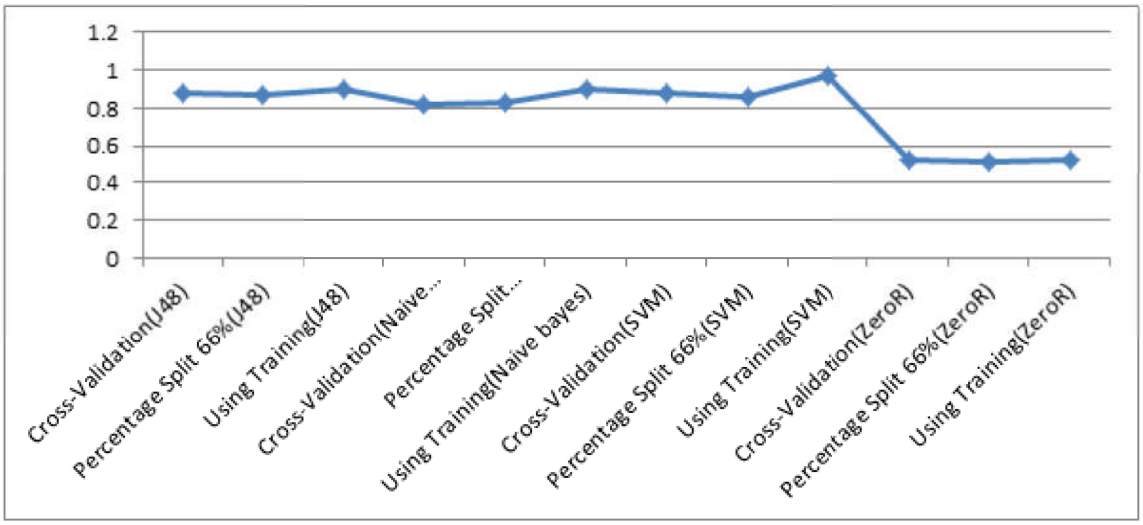
4.9.4 مقياس F-Meausure

هو مقياس لقياس دقة المصنف يعطى بالعلاقة التالية: $F-measures = \frac{2*precision*recall}{(presision+recall)}$

الخوارزميات	نوع الاختبار	Recall
Meta J48	Cross -validation (10) folds	0,882
	Percentage split 66%	0,867
	Use Training	0,903
Naïve Bayes	Cross-validation 10% folds	0,817
	Percentage split 66%	0,827
	Use Training	0,898
SVM	Cross-validation 10% folds	0,875
	Percentage split 66%	0,856
	Use Training	0,968
ZeroR	Cross-validation 10% folds	0,524
	Percentage split 66%	0,518
	Use Training	0,524

جدول 4.4: مقياس F-Measure

يظهر الجدول(4.4) نتائج مقياس F-Meausure اذ نجد ان خوارزميةJ48 حققت قيم عالية مقارنة بخوارزميات الأخرى مثلنا هذه النتائج على شكل منحنى (41.4).



شكل 41.4: مقياسF-Measurel

الشكل(4-41) عبارة عن منحنى لمقياس F-Meausure للخوارزميات إذ لاحظنا أن خوارزمية J48 حققت نسب عالية مقارنة بالخوارزميات الأخرى ، وهذا مما يدل على أن هذه الخوارزمية تمتاز بدقة عالية في تصنيف الفئات وتفوق الحالات المصنفة بشكل صحيح على المصنفة بشكل خاطئ.

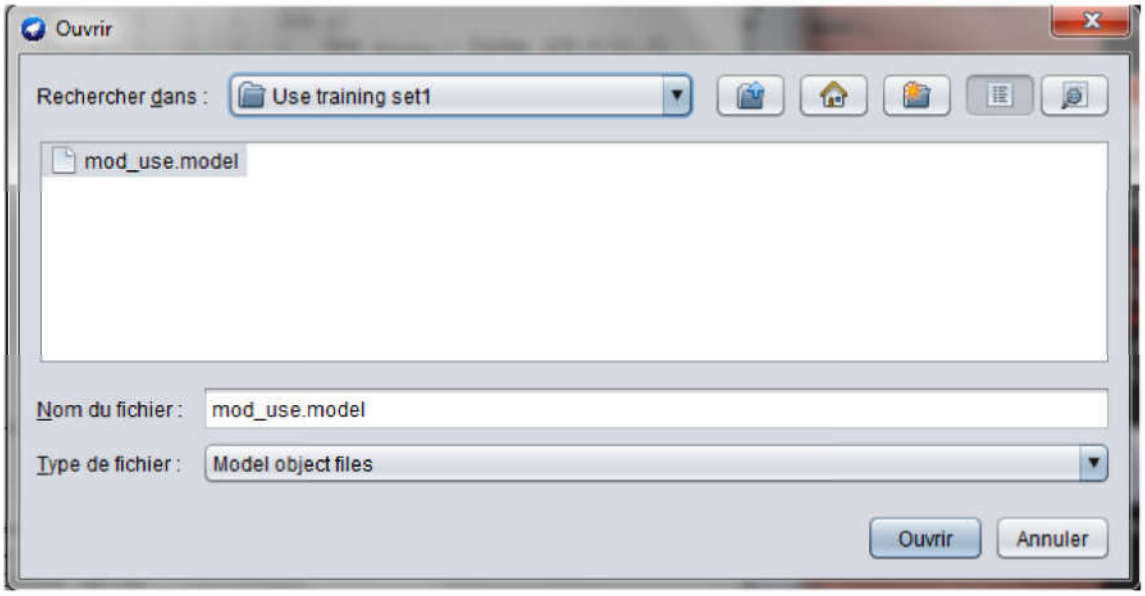
10.4 التنبؤ:

الغرض من استخدام آلية التنبؤ على حزمة بيانات التطبيق هو الكشف أو التنبؤ بفئات البيانات غير معروفة الفئة (التصنيف المتوقع حدوثه)، وذلك من خلال تطبيق النموذج الذي تم بناءه خلال مرحلة التصنيف ويكون ذلك على حزمة البيانات الجديدة ويتضح ذلك من خلال الشكل(4-9)، الشكل(4-42) والشكل(4-43) .

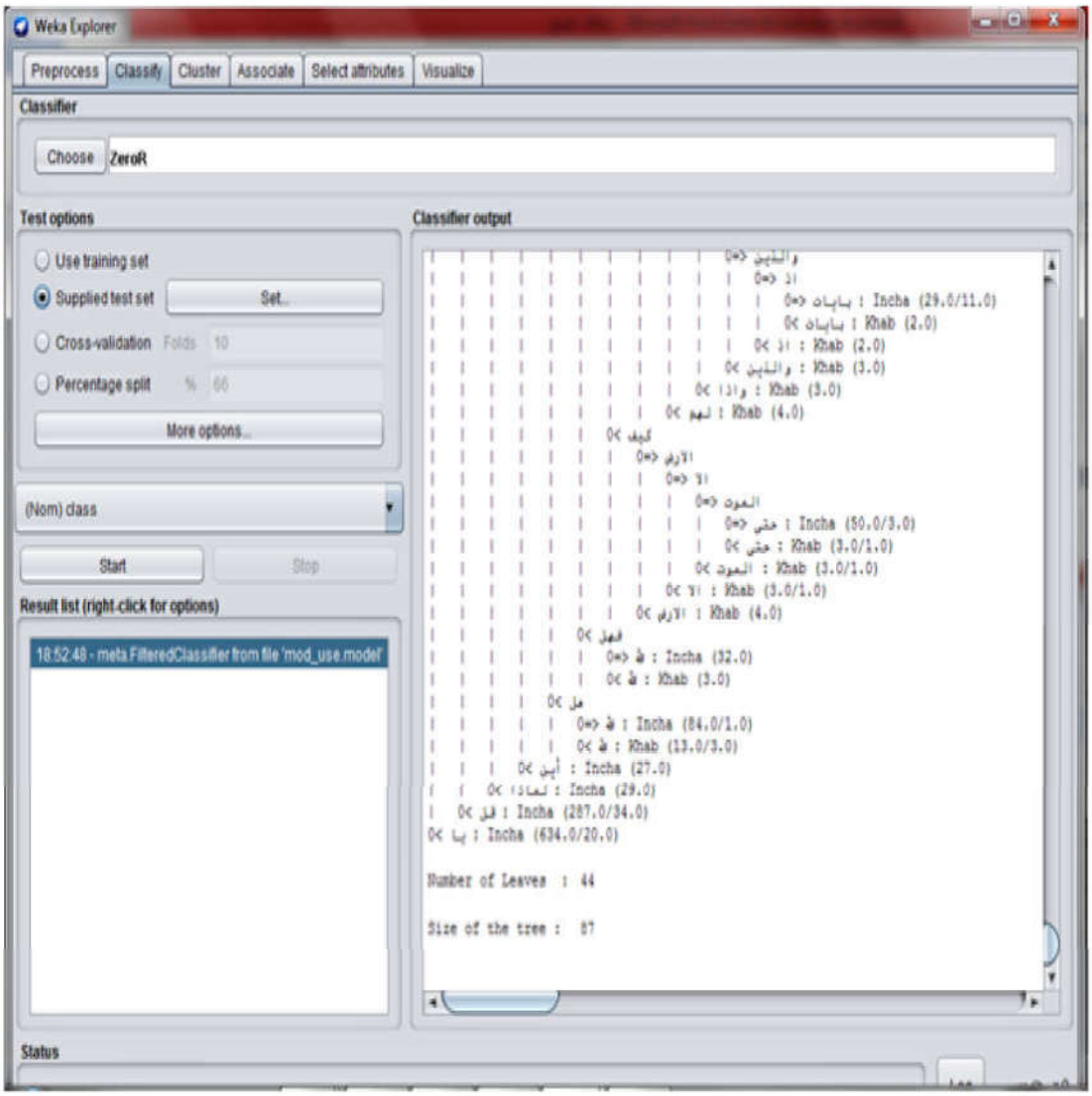
وللاختبار على الحزمة البيانية التي تم الاحتفاظ بها سابقا للاختبار والتي تقدر ب20% من نصوص المدونة من أجل إستخدامها للتنبؤ وجب ان تكون هيكليتها نفس هيكلية بيانات التدريب،

لذا تم استخدام طريقة Supplied Test Set لتنفيذ هذه المهمة الشكل(4-44).

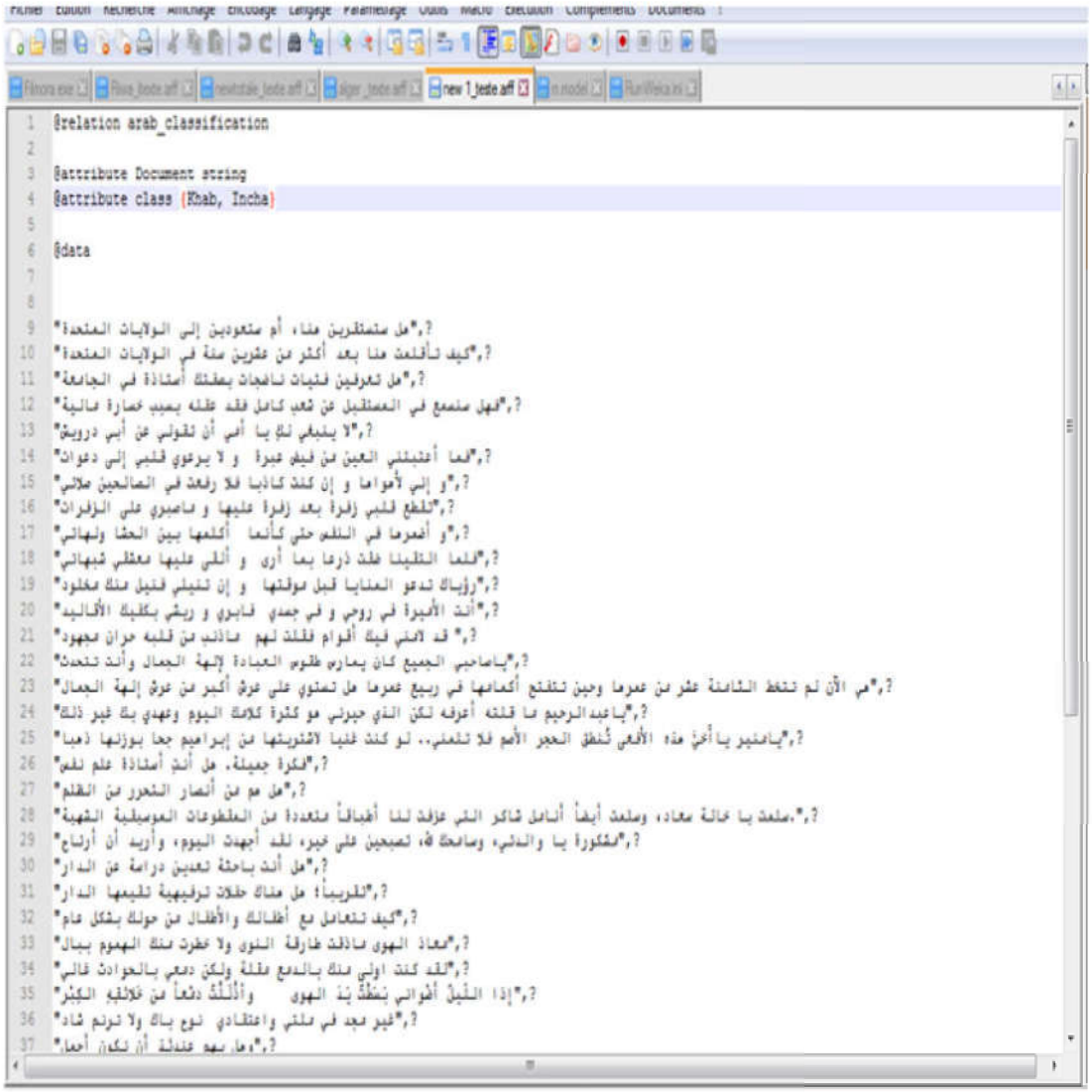
وبعد أن تم معرفة المصنف الاكثر فعالية كان لابد من تطبيقه على الحالات الجديدة لتحديد دقته التنبؤية، وبما أن مصنف J48 حقق دقة عالية ونسبة خطأ قليلة اكتفينا بعرض نتائجه التنبؤية اما بقية المصنفات الأخرى ضمنا نتائجها في الملحق.



شكل 42.4: مثال عن النموذج للخوارزمية J48



شكل 43.4: استخدام طريقة Supplied Test Set للتنبؤ

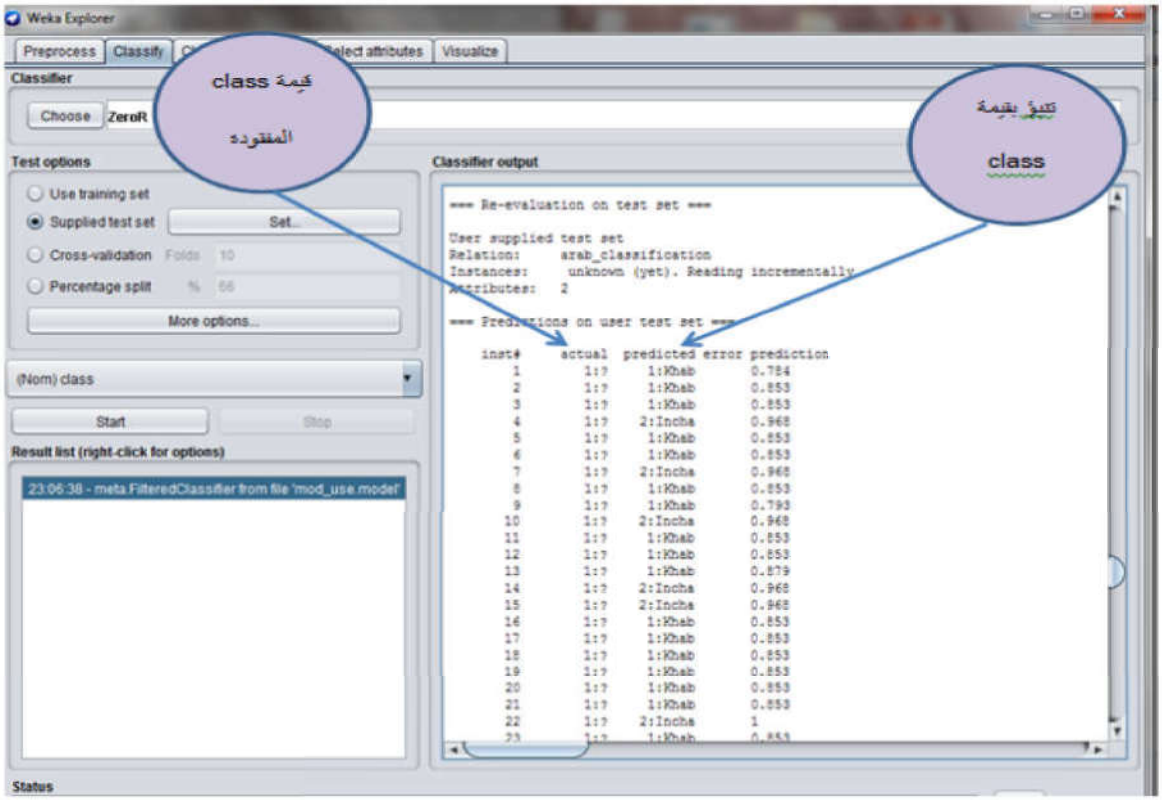


شكل 44.4: العينة التي تم تخصيصها للتنبؤ

1.10.4 نتائج الاختبار:

تظهر النتائج في شاشة « Classfiier output » تحت عنوان " Predictions on user test set " بمعنى التنبؤ بالبيانات المستخدمة للاختبار كما هو مبين في الشكل (45.4) و(4-46) نتيجة حزمة البيانات التي تم استخدامها للتنبؤ والمتعلقة بخوارزمية J48 و خوارزمية SVM حيث يوضح نتيجة تطبيق هذين المصنفين على البيانات الجديدة إذ يظهر الصف الحقيقي المحتمل (Actual class) وقيمته المجهولة والصف المتوقع (predicted class) من قبل الخوارزميتين.

2.10.4 التنبؤ بخوارزمية J48 :

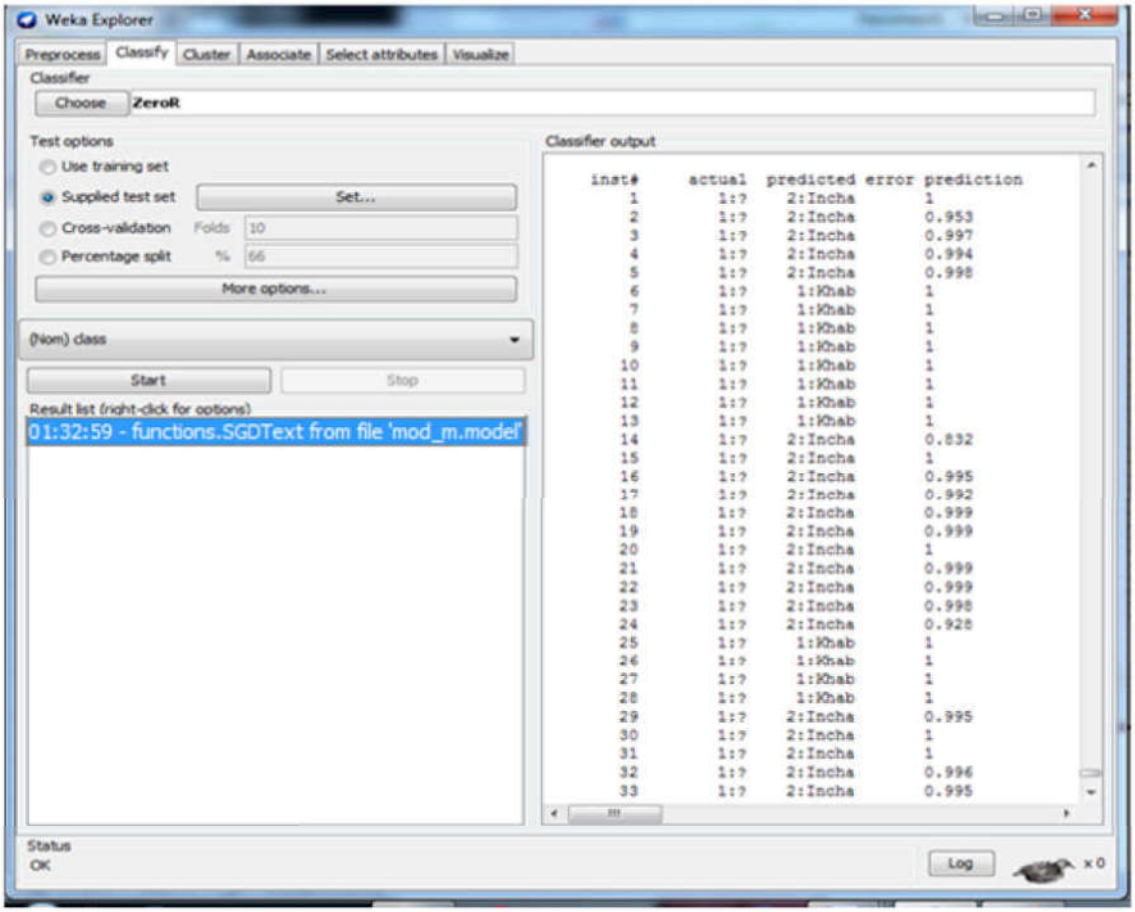


شكل 45.4: التنبؤ عل البيانات بخوارزمية J48

يوضح الشكل(4-45) ناتج عملية التنبؤ بخوارزمية (J48) تظهر نتائج التنبؤ من خلال القيمة الفعلية للصف المتوقع أن البيانات التي صنفت على أنها أساليب خبرية هي بالفعل خبرية أما الأساليب الإنشائية هي بالفعل إنشائية، يتضح ذلك من خلال المقارنة بين الشكل الذي يظهر عينة التطبيق والشكل الذي يظهر ناتج عملية التنبؤ، نأخذ المثال الأول من العينة وهو: «**هل ستستقرين هنا أم ستعودين إلى الولايات المتحدة**» هذا أسلوب إنشائي أتى بصيغة الاستفهام والمؤشر الدال على ذلك أداة الاستفهام "هل"، لو عدنا إلى شاشة النتائج التي يوضحها الشكل (4-45) نجد أن الخوارزمية صنفت المثال على انه أسلوب إنشائي وينطبق الأمر نفسه على جميع الأمثلة، وهذا يدل على أن تنبؤات خوارزمية (J48) كانت صحيحة ونتائجها كانت دقيقة.

3.10.4 التنبؤ بخوارزمية SVM

تم تطبيق خوارزمية SVM على نفس عينة التطبيق من اجل التنبؤ بالحالات الجديدة، ويظهر ذلك من خلال الشاشة التي يوضحها الشكل(4-46).



شكل 46.4: التنبؤ على البيانات بخوارزمية SVM

يوضح الشكل (46.4) نتائج عملية التنبؤ لخوارزمية SVM والتي أسفرت على أن تنبؤات هذه الخوارزمية صحيحة إلى حد ما.

من خلال المقارنة بين ناتج عملية التنبؤ لكل من الخوارزميتين لاحظنا أن خوارزمية J48 تتسم بدقة عالية، إذ نجد المثالين رقم (16-17) الظاهران في العينة صنفتهما خوارزمية SVM في فئة الإنشائي، أما خوارزمية J48 صنفتهما في فئة الخبري، نجد في المثالين مؤشرات إنشائية وهي أدوات النداء "يا" إلا أننا لو تمعنا الأمر لوجدنا أن هذين المثالين يحتويان على أكثر من جملة والمرجح فيها الأسلوب الخبري أكثر من الإنشائي لذا الجائز أن تصنف في خانة الخبري.

ما نخلص أن خوارزمية (J48) كانت تنبؤاتها كلها صحيحة ونتائجها كلها دقيقة مقارنة بخوارزمية SVM إلا أن هذا لا ينقص من كفاءة هذا المصنف الأخير.

خلاصة:

في هذه الدراسة استخدمنا أهم خوارزميات تنقيب البيانات خوارزمية j 48 ، Naive bayse ، SVM ، ZeroR والتي يتيحها برنامجWeka لتصنيف النصوص الأدبية وفق أساليبها الخبرية والإنشائية إذ أظهرت النتائج أن عملية التصنيف تختلف من خوارزمية إلى أخرى في الطريقة ودقة التنبؤ والنتائج المتوصل لها إذ وبعد تقييم دقة هذه المصنفات أظهرت النتائج المتحصل عليها تفوق خوارزمية J48 بشكل طفيف على SVM ، إذ الفارق بينهما هو دقة خوارزمية J48 في التعرف على النصوص ذات الأساليب الإنشائية والخبرية والقدرة على تصنيفها بشكل جيد مما يظهر كفاءة هذا المصنف في مجال التصنيف الآلي للنصوص، لذا يمكن القول انه أصبح بإمكان الباحثين في مجال اللغة العربية والأدب العربي الاستعانة بالبرامج الآلية مثل برنامج Weka أو برامج أخرى تعتمد على الخوارزميات في انجاز بحوثهم الأكاديمية، لما تتيحه هذه البرامج من تسهيلات وتوفير للجهد والوقت مع ضمان دقة النتائج المتوصل اليها.

الخاتمة

في هذا البحث تم تطبيق تقنية التصنيف الآلي على النصوص العربية لتحديد أساليبها الخبرية والإنشائية، بالاعتماد على برنامج حاسوبي متطور(Weka) يتيح لنا جملة من الخوارزميات المعروفة، اخترنا منها أربع خوارزميات قصد معرفة أدائها وكفاءتها في عملية التصنيف ومن خلال ذلك فقد تم التوصل إلى جملة من النتائج الهامة نعرضها كما يلي:

- يعد التصنيف الآلي للنصوص إحدى تقنيات التنقيب في البيانات، يعتمد على مجموعة كبيرة من البرامج الحاسوبية والخوارزميات المتطورة.
- يوفر برنامج Weka سهولة في تصنيف النصوص العربية من خلال مجموعة من الخوارزميات المعروفة التي يتيحها لنا.
- يوفر تطبيق برنامج Weka عامل السرعة والجهد حيث يمكن الوصول إلى فئات النصوص وأساليبها في غضون ثواني معدودة، مع السماح بتحديث هذه النماذج باستمرار هذا مما يساعد الدارس للغة العربية في الوصول إلى مبتغاه دون جهد أو ضياع للوقت.
- تبين من خلال اختبارات أداء واحتساب مقاييس دقة للخوارزميات المعتمدة في الدراسة تفوق خوارزمية J48 في عملية التصنيف والتنبؤ مما يؤكد كفاءتها العالية في هذا المجال.
- كفاءة كل من خوارزمية J48 و Naïve Bayes و SVM في التعرف على مختلف الأساليب الإنشائية والخبرية في مجموعات التدريب و التنبؤ وتراجع كبير بالنسبة لخوارزمية ZeroR .
- تبين انه يوجد فروق بين مختلف الخوارزميات والاختبارات وتمظهر ذلك في الأداء والكفاءة والسرعة والمرونة.

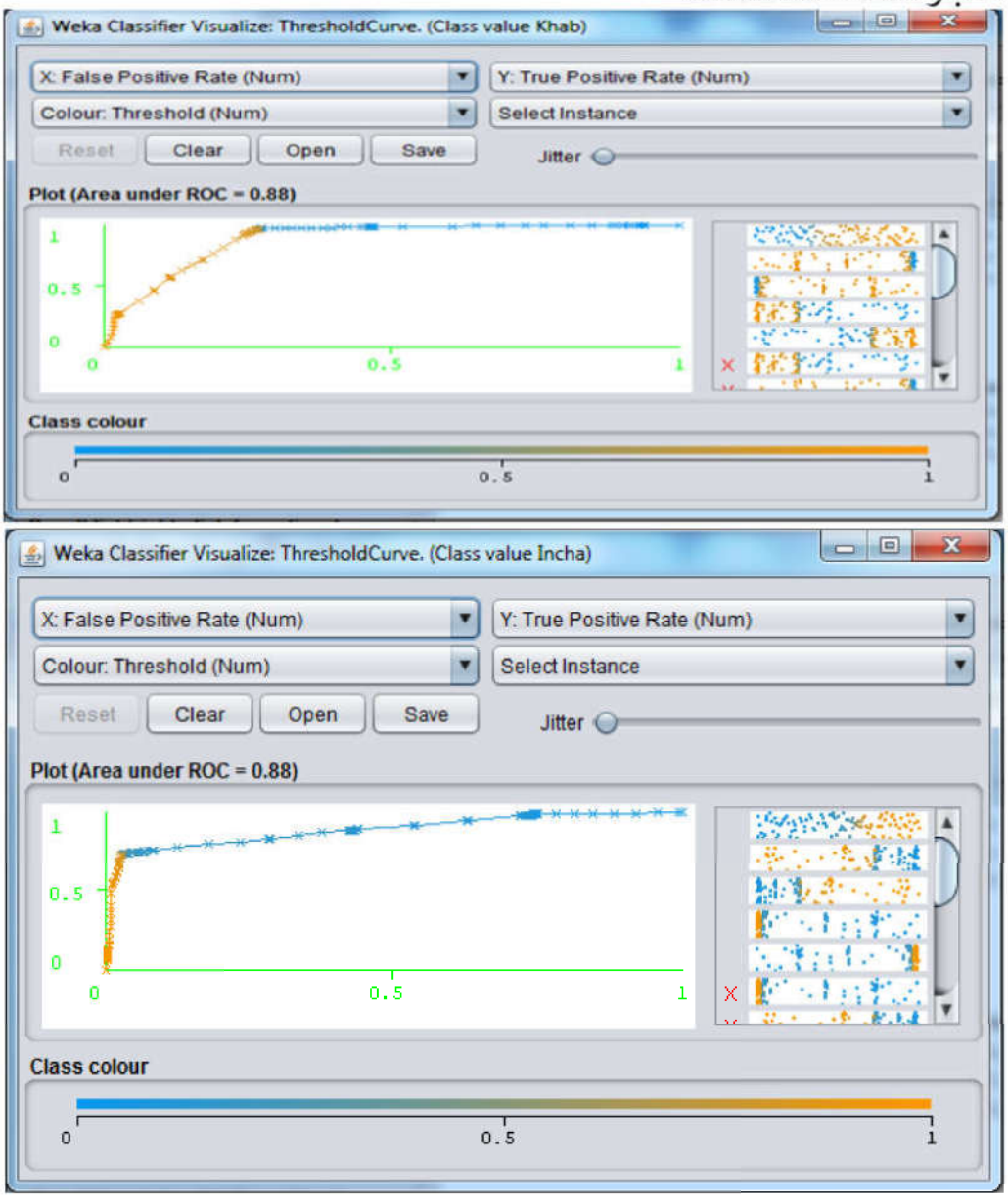
التوصيات: في ضوء ما توصلت إليه الدراسة من نتائج نوصي بما يلي:

- ضرورة توفير عناية أكبر بعملية تألية البيانات العربية باستخدام مختلف التقنيات والخوارزميات والأدوات التي يتيحها علم التنقيب في البيانات.
- العمل على زيادة نشر الوعي في الدراسات العربية لضرورة الاعتماد على مختلف البرامج الحاسوبية في التعامل مع البيانات النصية سواء أكان بغية التصنيف الآلي لها أو التعرف الموضوعي عليها أو تلخيصها وترجمتها و تشكيلها آليا.
- العمل على ضرورة إنشاء المدونات النصية العربية والتوسع في حجم ونوعية البيانات المطلوبة في أي دراسة قصد إتاحتها للباحثين وإجراء عليها مختلف الدراسات التقنية.
- إجراء الملتقيات والدراسات حول التقانات المتطورة التي تعمل على المعالجة الآلية للغة العربية، وتعريف الباحثين بأهم البرامج الحاسوبية وكيفية تطبيق مختلف الخوارزميات المعروفة.
- تقديم التسهيلات المادية لكل ما من شأنه أن يبقي اللغة العربية مواكبة للتطورات التقنية الحديثة من برامج وتقنيات وخوارزميات جديدة.
- تشجيع الباحثين والطلبة الأكاديميين على إجراء دراسات مماثلة وعلى مختلف النصوص العربية في حقول معرفية متعددة (الأدب العربي، الصحافة، العلوم الإنسانية والاجتماعية، العلوم السياسية...).

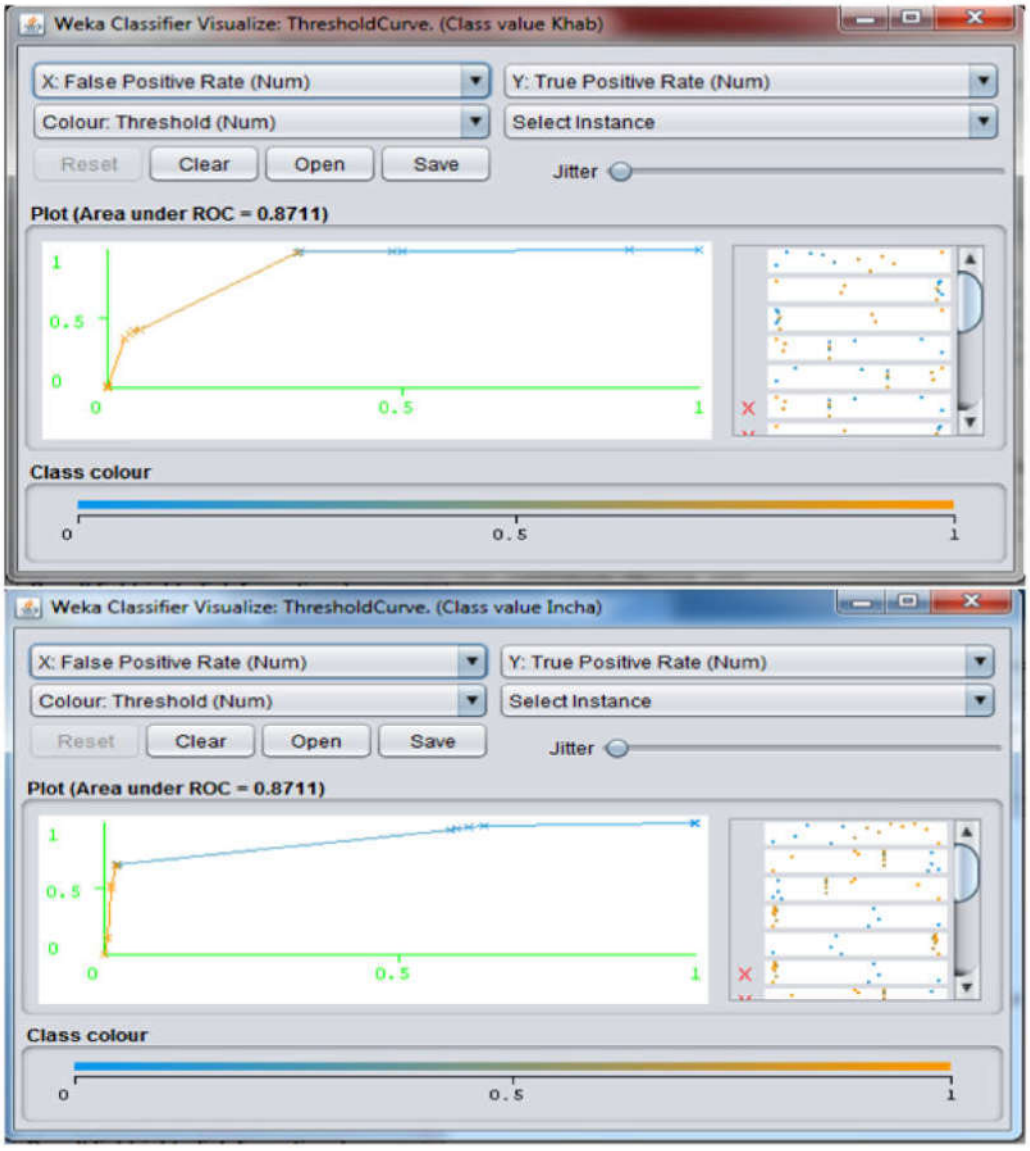
الملاحق

خورزمية Meta(J48)

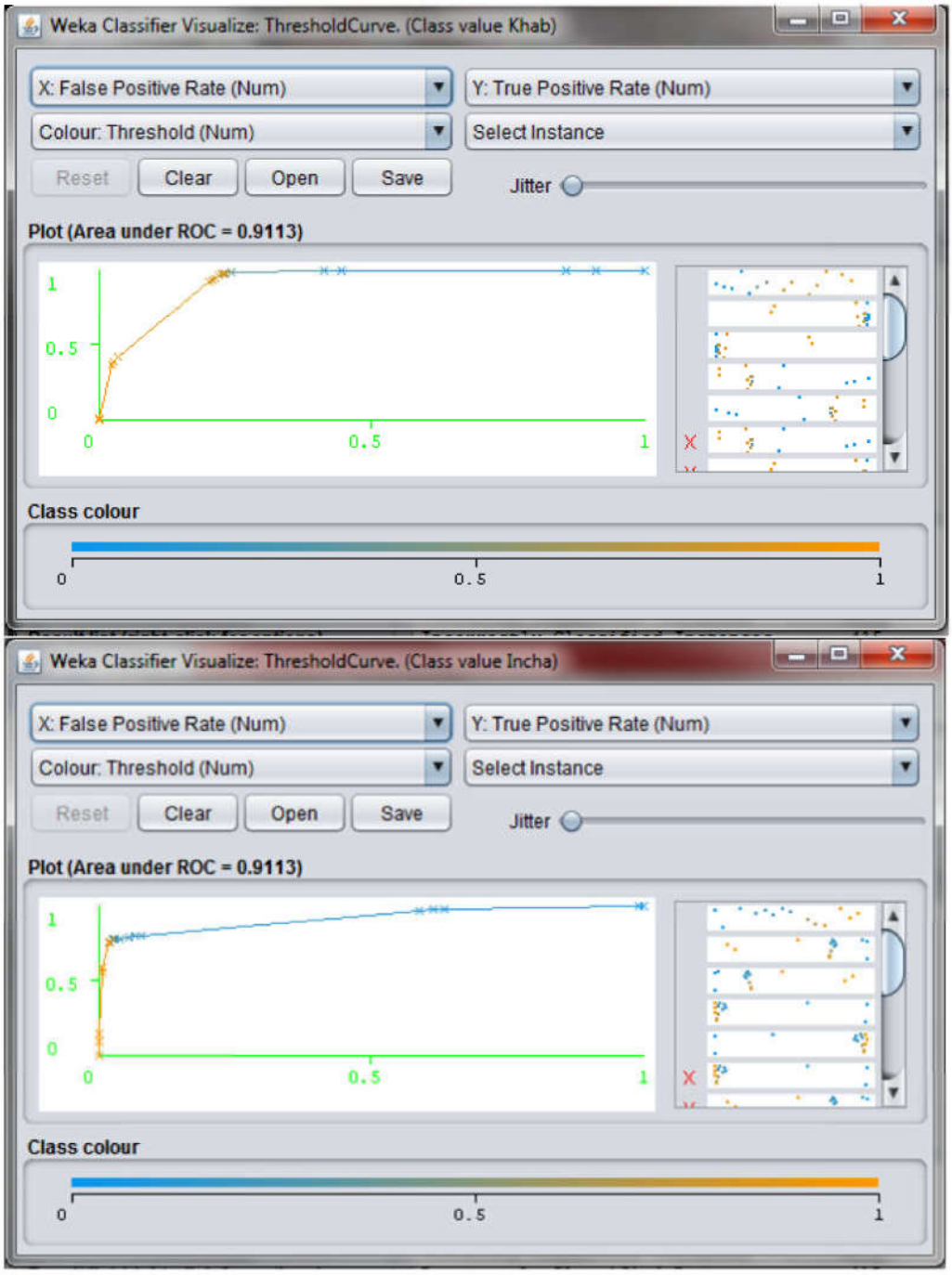
اختبار validation Cross



الشكل(1):نتائج ROC باختبار Cross validation

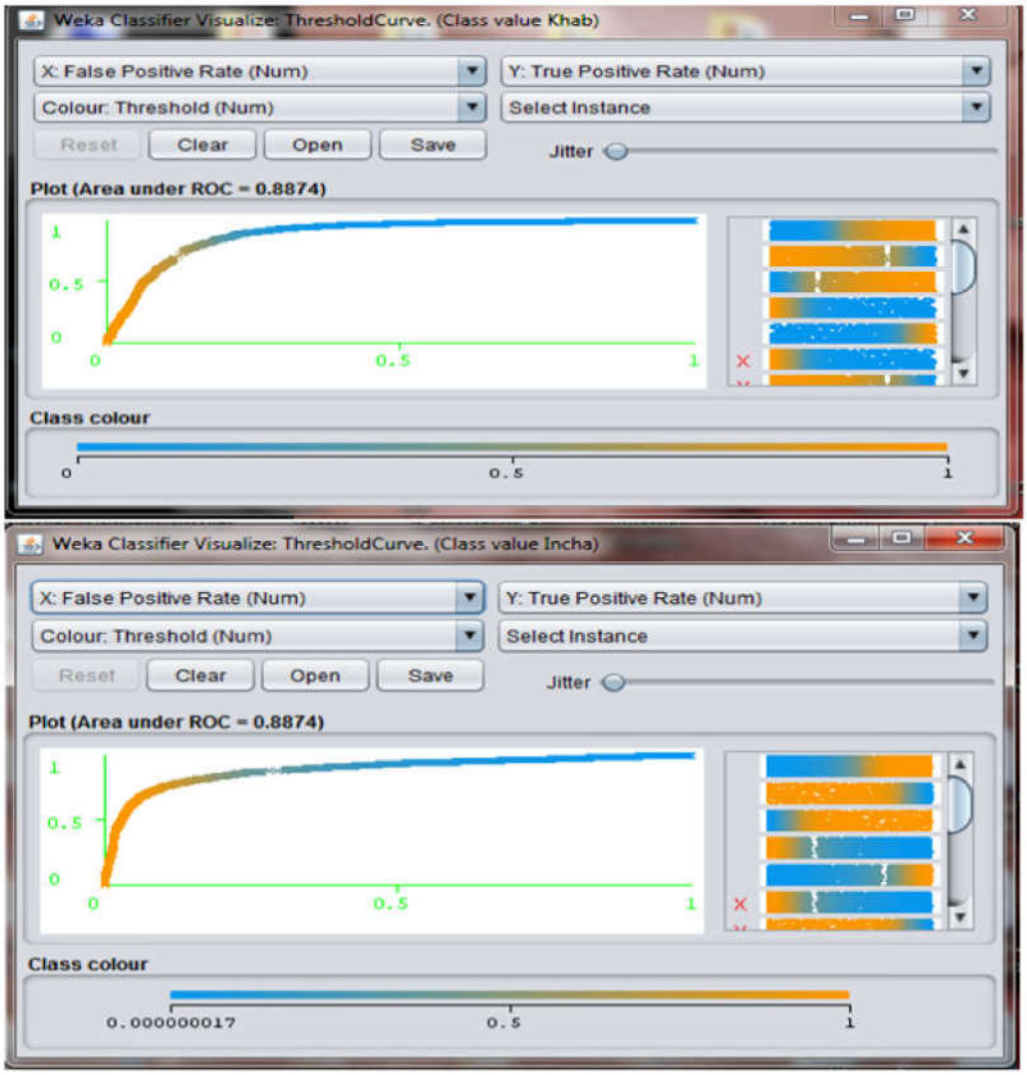


الشكل(2) :نتائج ROC باختبار Percetage split 66%

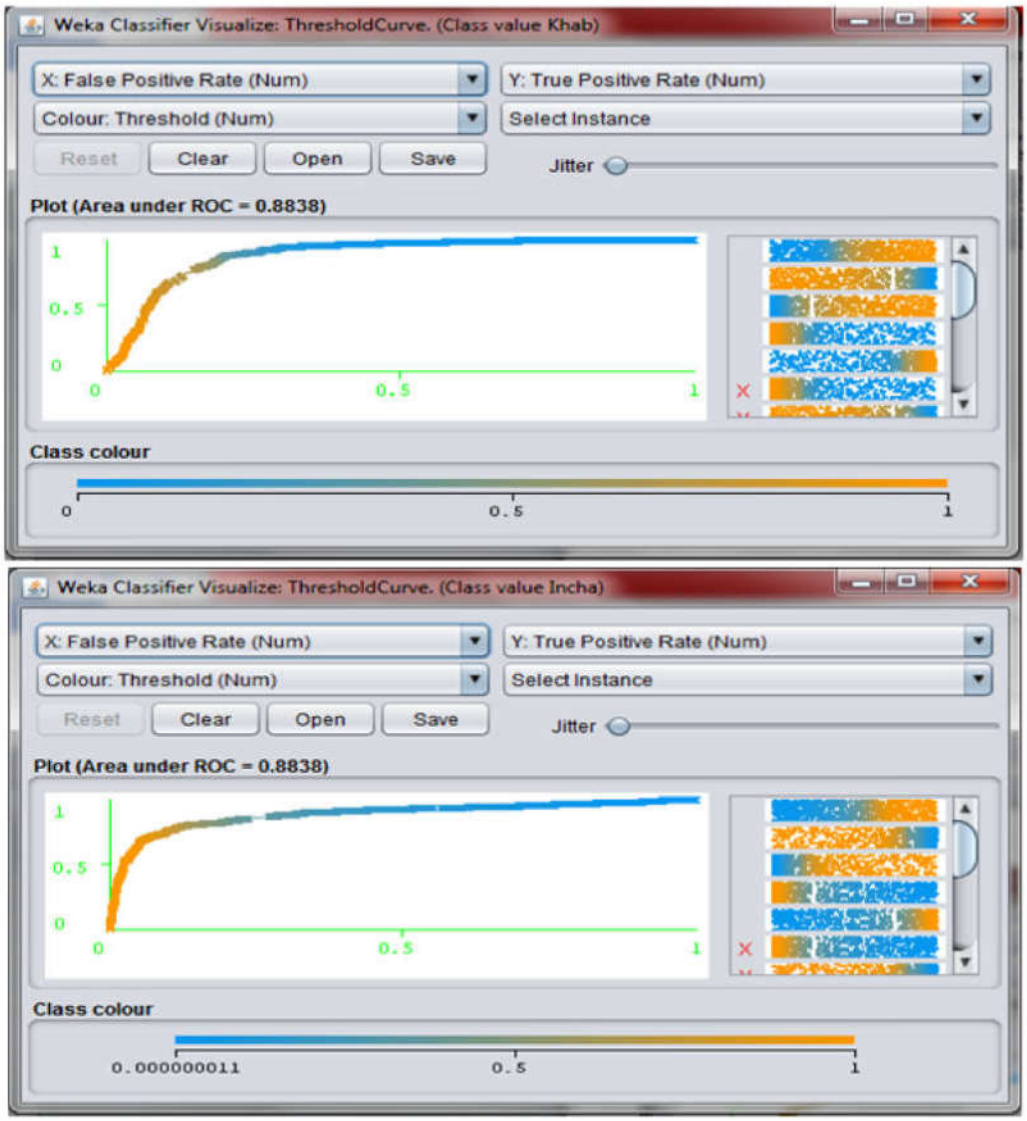


الشكل (3) :نتائج ROC باختبار Use training

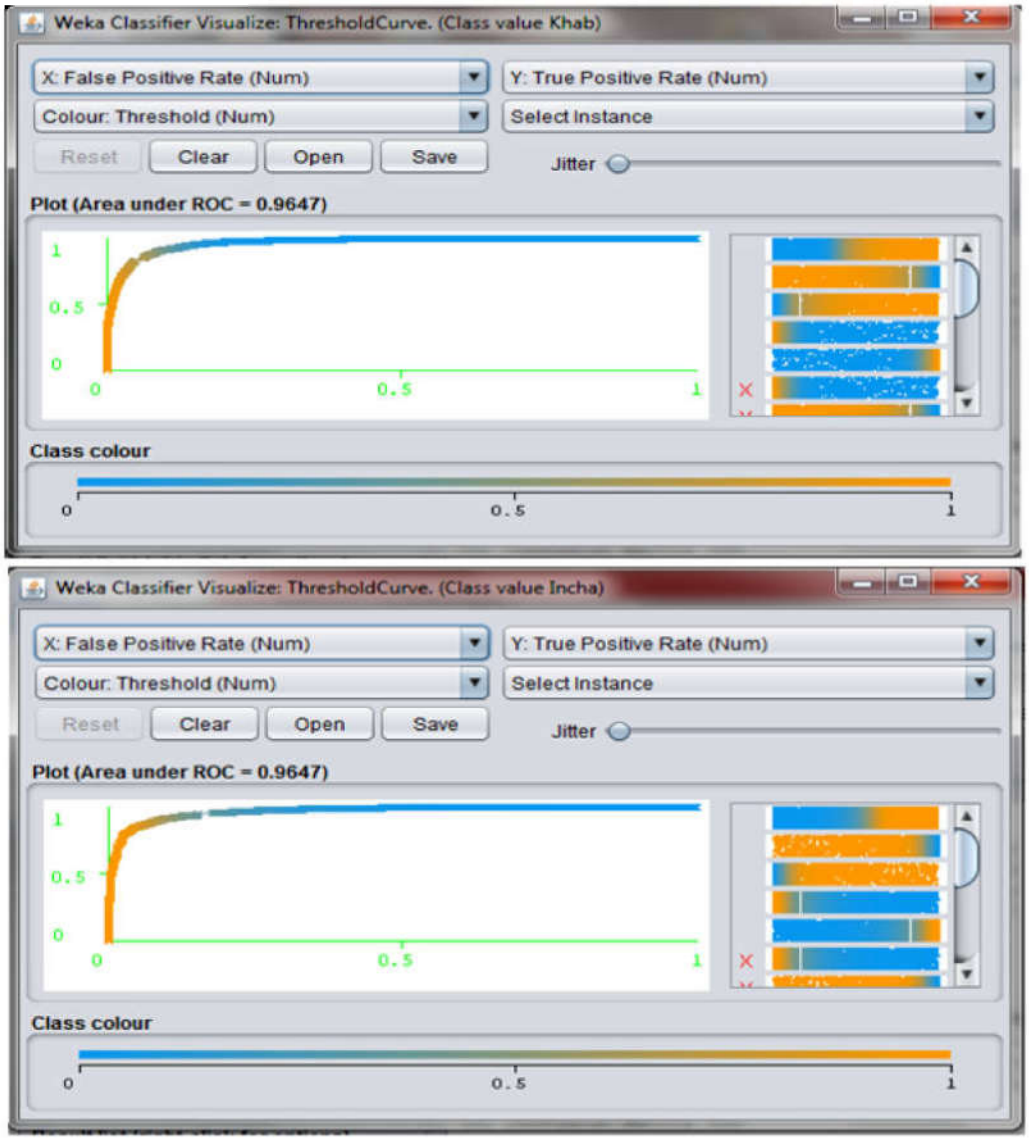
خوارزمية Bayes Naive



الشكل (4) :نتائج ROC باختبار Cross validation

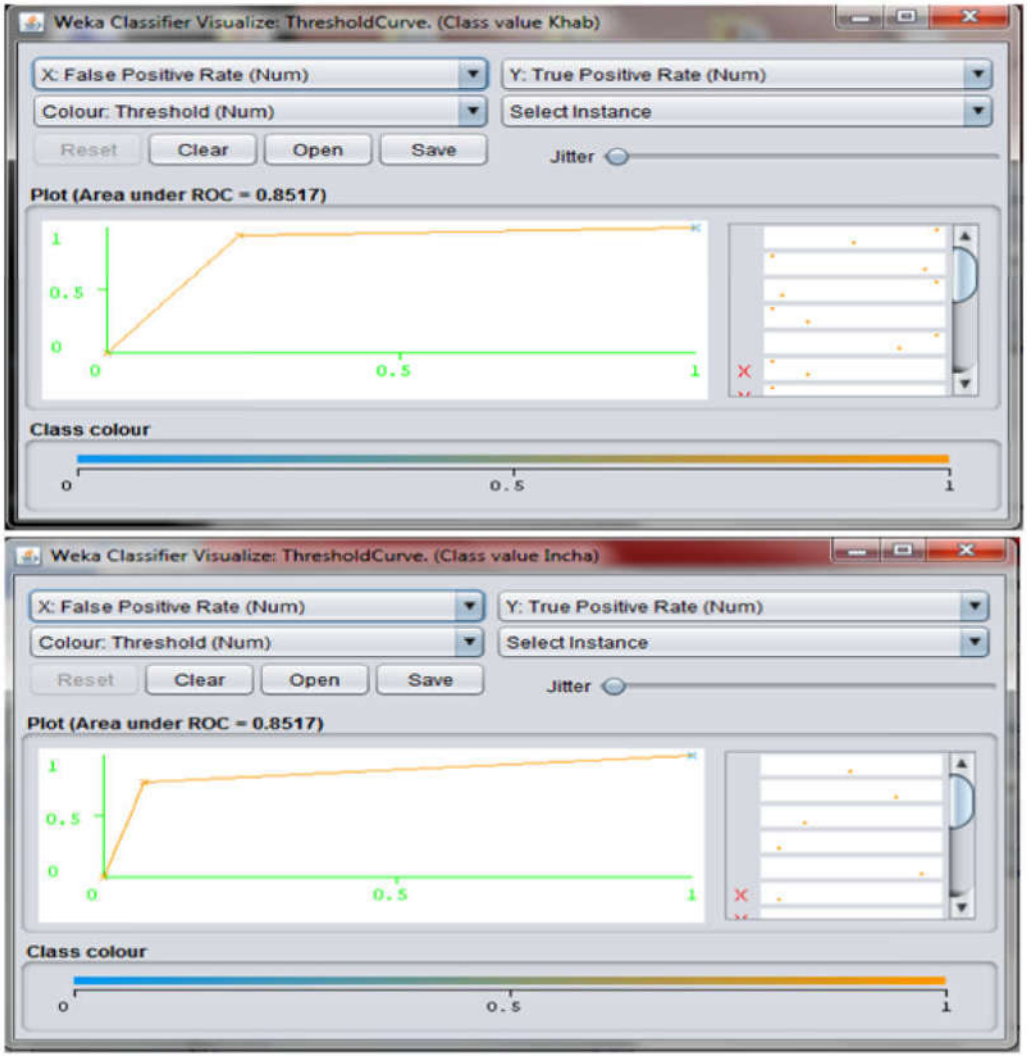


الشكل (5) : نتائج ROC باختبار Percetage split 66%

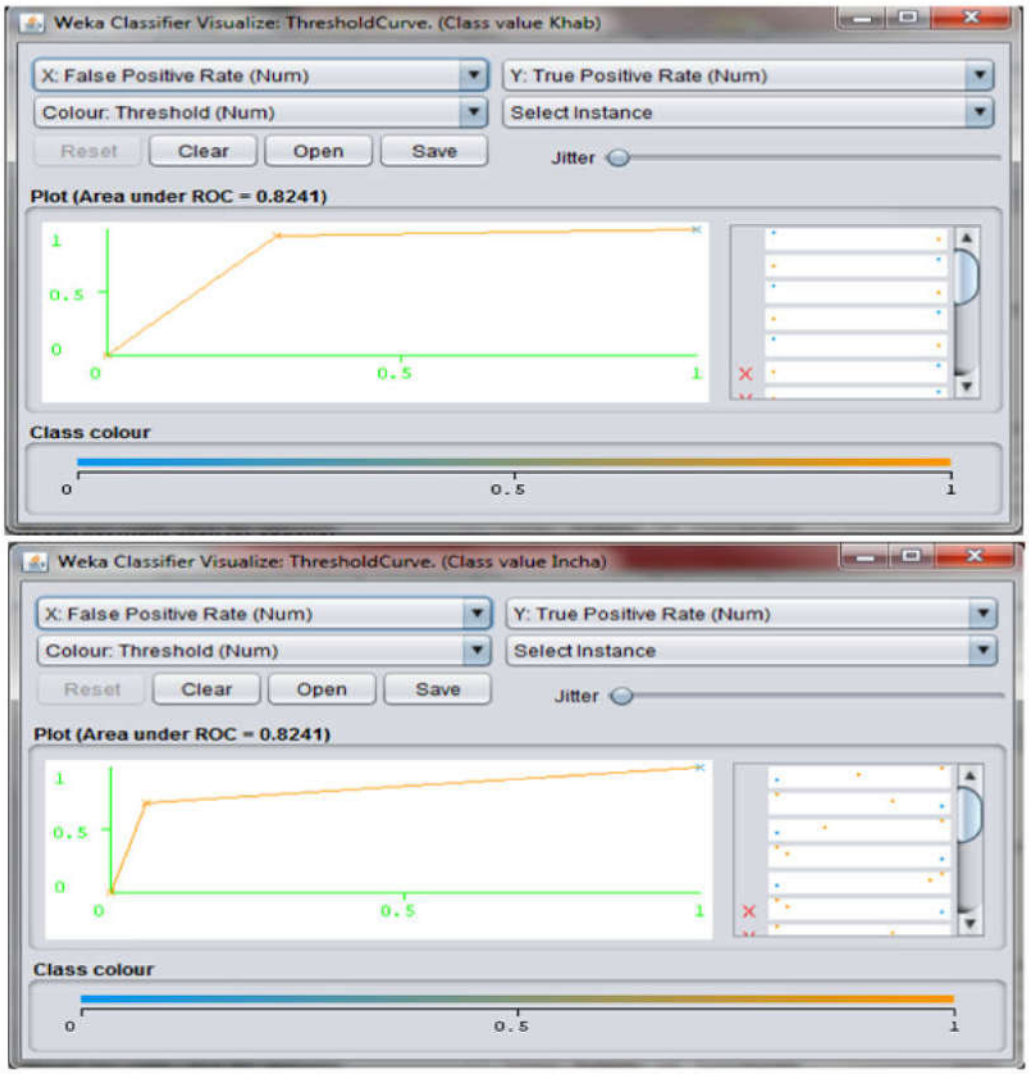


الشكل (6) : نتائج ROC باختبار Use training

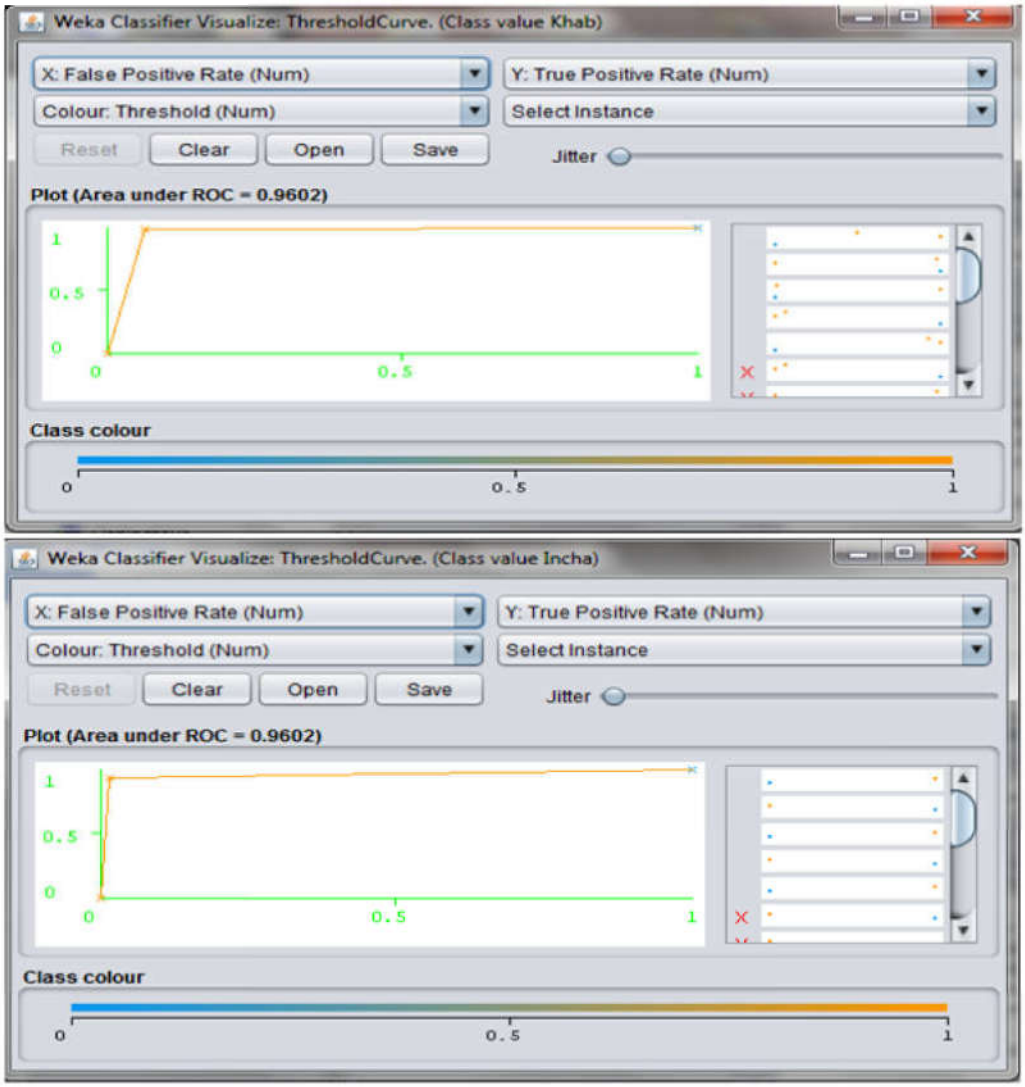
خوارزمية SVM



الشكل (7) :نتائج ROC باختبار Cross validation

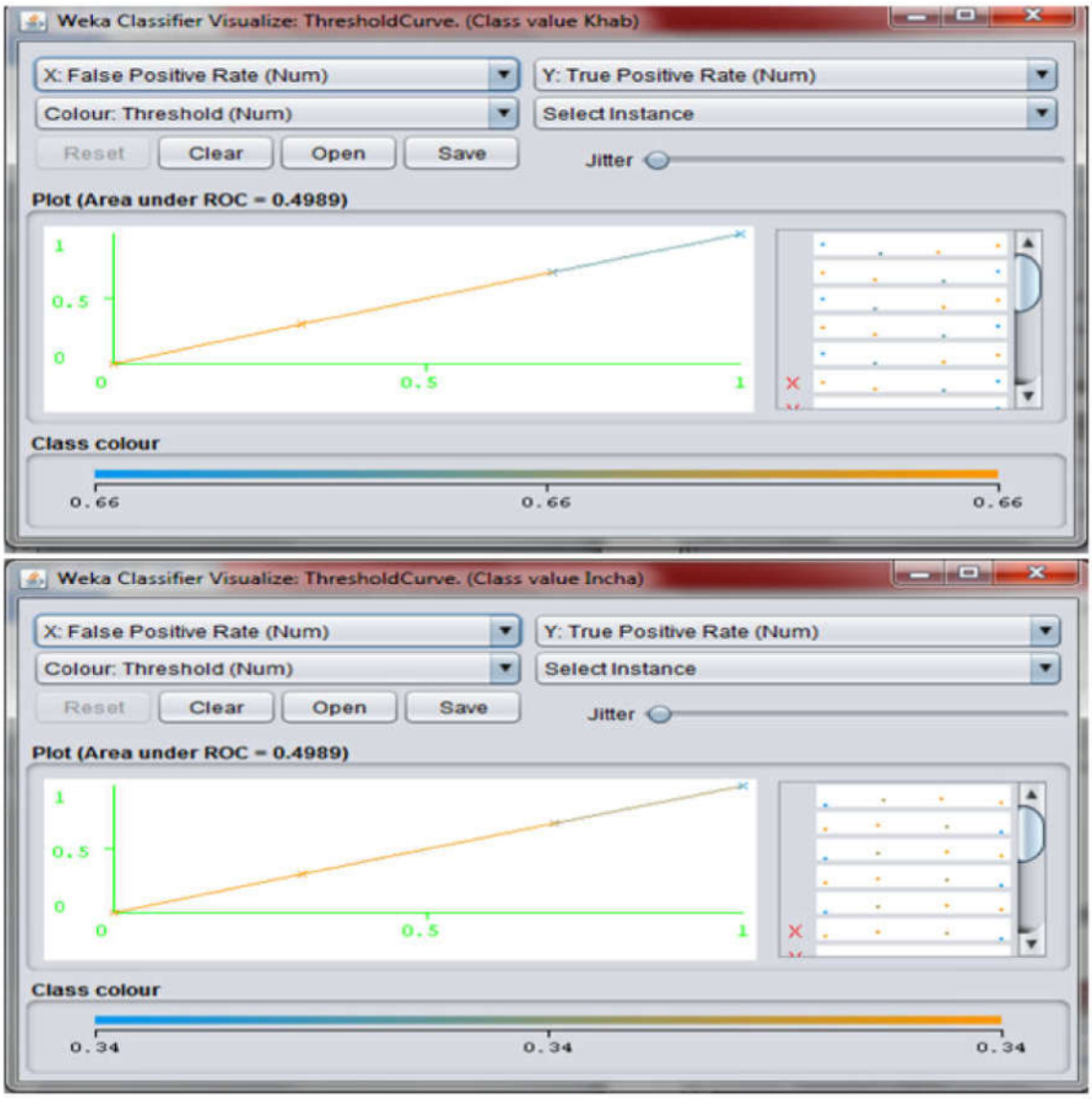


الشكل(8) : نتائج ROC باختبار Percetage split 66%

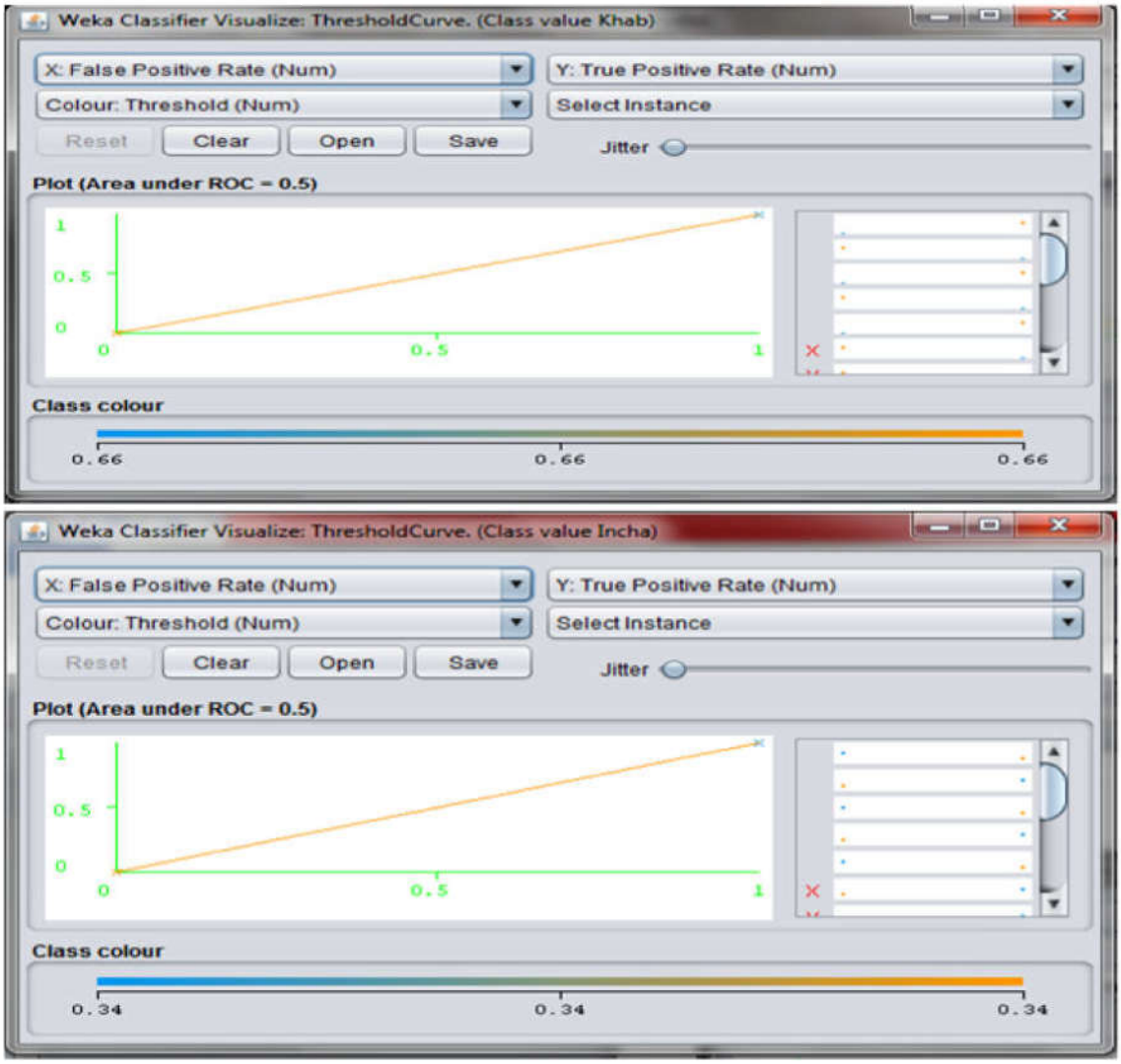


الشكل (9) : نتائج ROC باختبار Use training

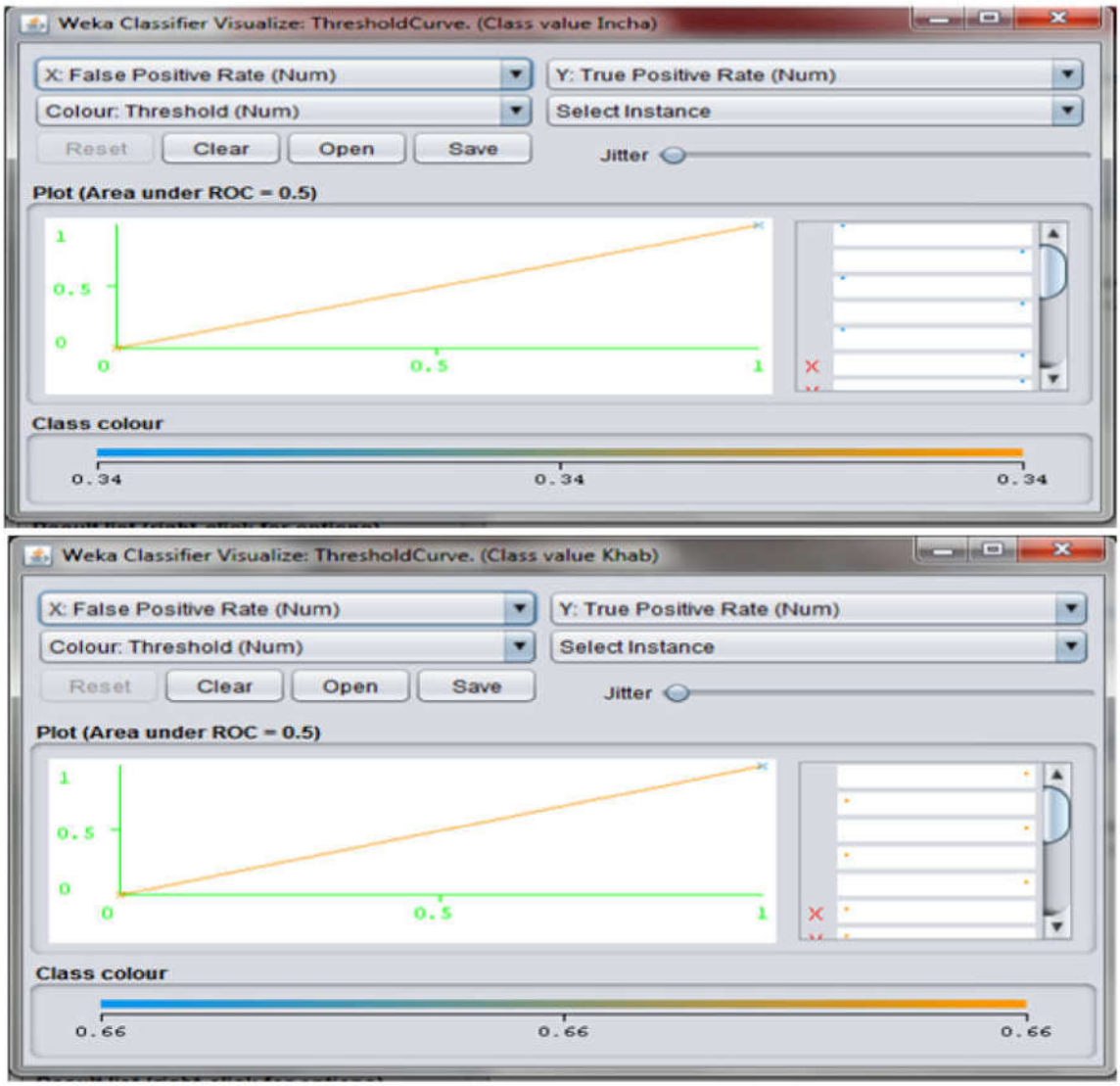
خوارزمية ZeroR



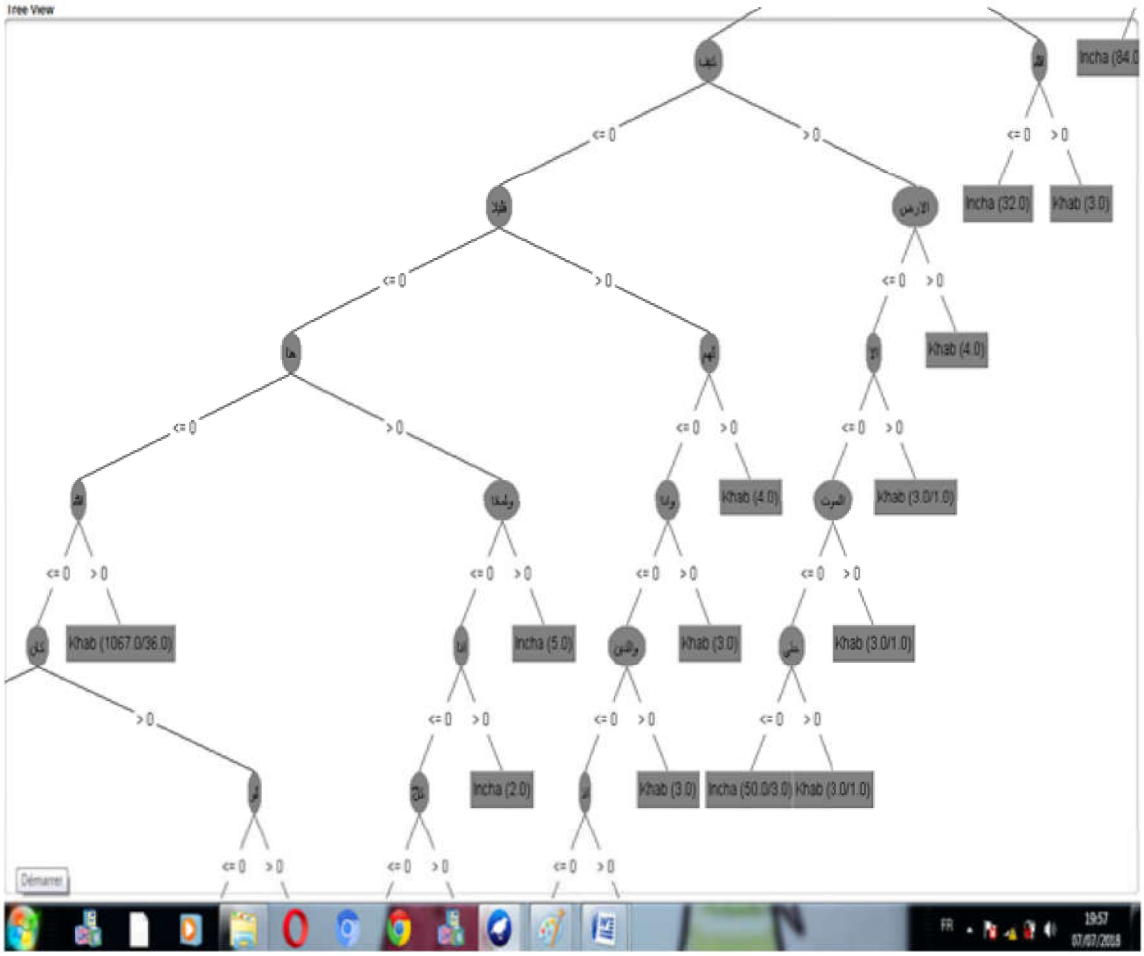
الشكل(10) :نتائج ROC باختبار Cross validation



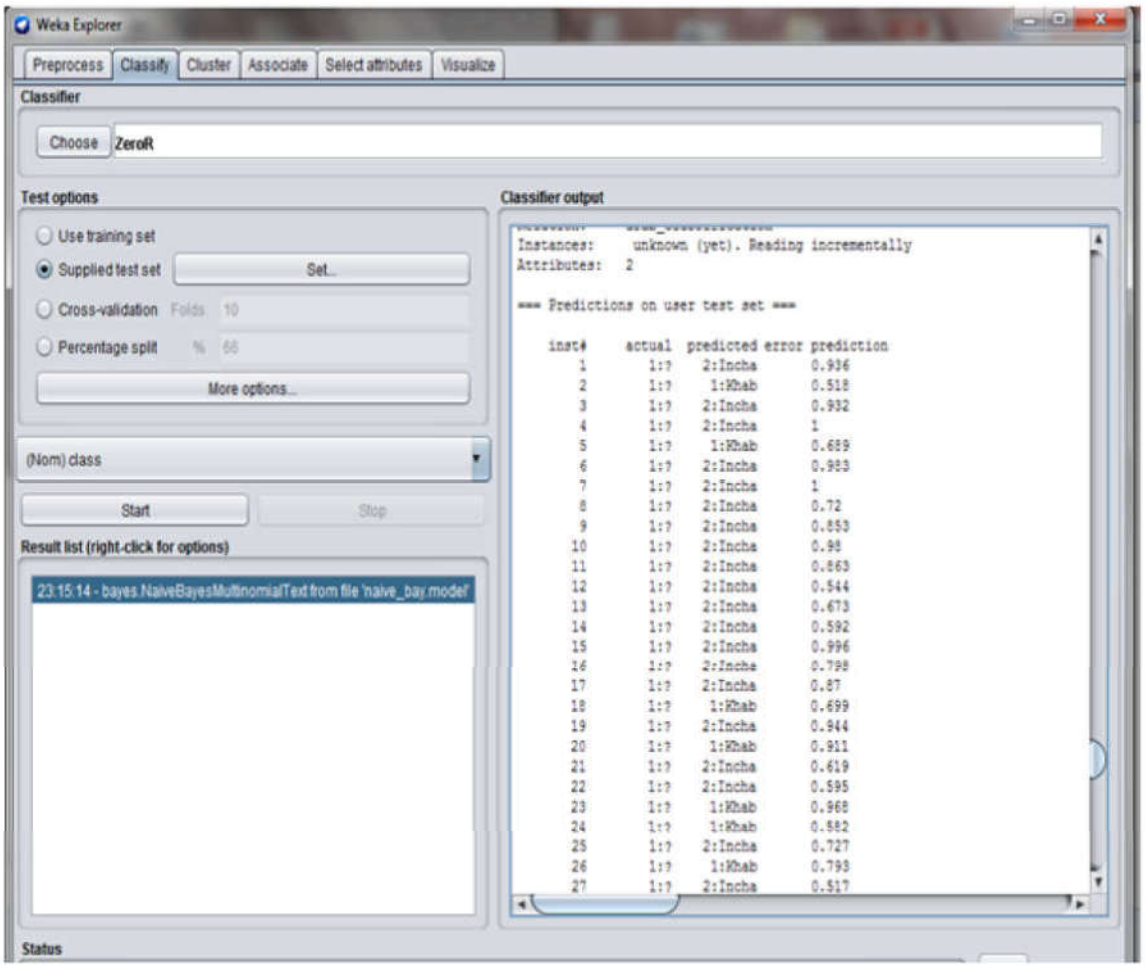
الشكل(11) : نتائج ROC باختبار Percetage split 66%



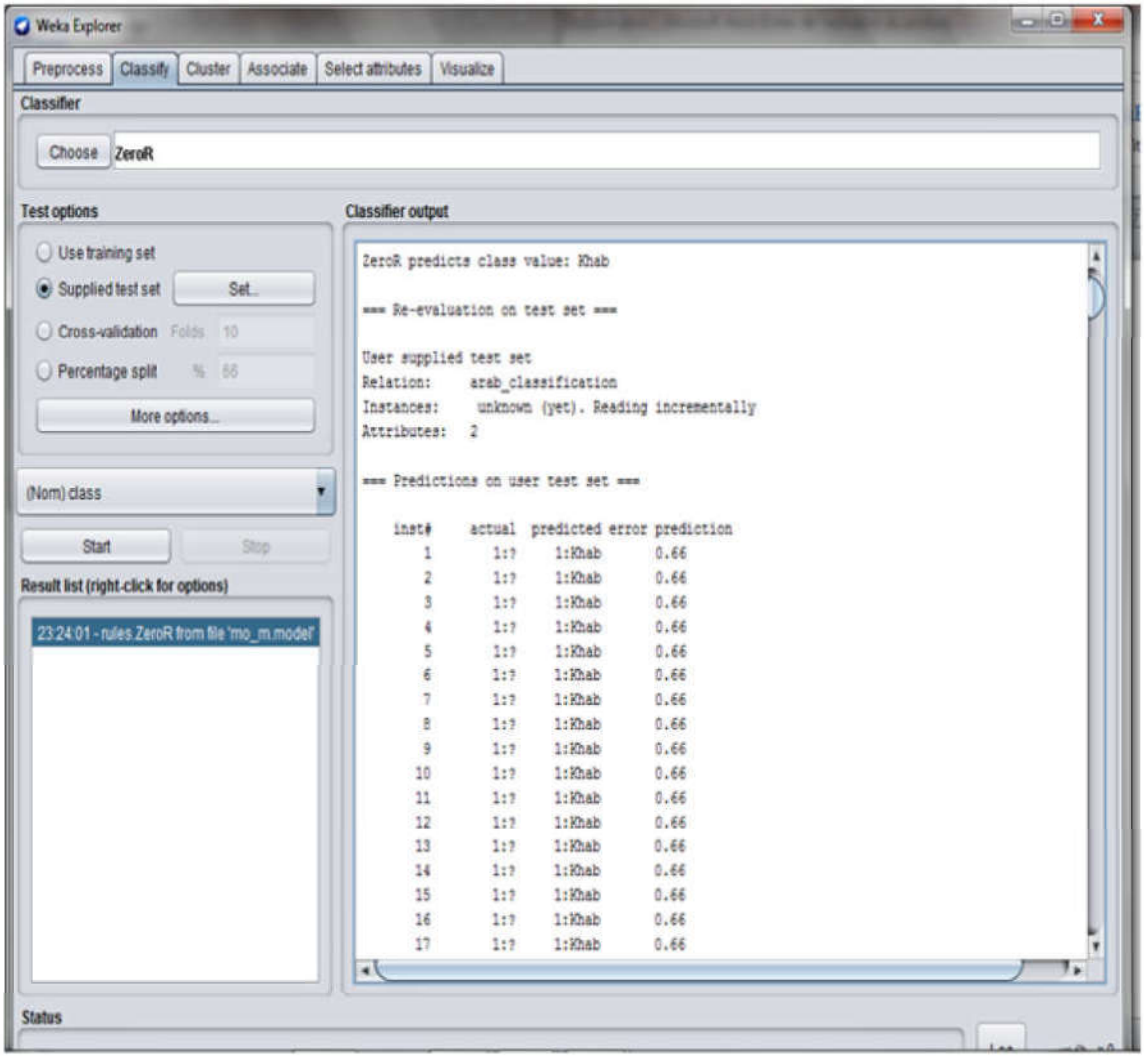
الشكل(12) : نتائج ROC باختبار Use training



الشكل(13) :ناتج شجرة القرار



الشكل(14) : التنبؤ على البيانات Naive bayes



الشكل(15) : التنبؤ على البيانات Zero R

فهرس الأشكال

قائمة الأشكال

قائمة الجداول

قائمة الجداول

المصادر والمراجع

قائمة المصادر والمراجع

- القرآن الكريم
- نصوص الشعر والروايات العربية.

قائمة المصادر باللغة العربية:

1. سميرة محمد علي القدم، تطبيق تقنيات التنقيب في البيانات لتقييم أداء طلاب قيم الحاسوب ،كلية العلوم،مذكرة بكالوريوس، إشراف محمود حفص الدين/ قسم الحاسوب، كلية العلوم ،جامعة سبها،2018.

2. إيهاب عثمان، عبد الرحمن عثمان، التنبؤات الانتخابية باستخدام تنقيب البيانات، سبتمبر 2016، https://www.researchgate.net/publication/307925413 .

3. بشير عباس، العلاق: الإدارة الرقمية المجالات والتطبيقات، مركز الإمارات للدراسات والبحوث الإستراتيجية، أبو ظبي، 2005.

4. رحاب فايز احمد، التنقيب عن بيانات مؤسسات العمل التطوعي على الويب، مجلة كلية الآداب، جامعة بني سويف، المجلد الأول، العدد 27، 2013.

5. عبد الحميد محمد العباسي، التنقيب في البيانات مجموعة محاضرات، فسم الإحصاء الحيوي والسكاني، معهد الدراسات والبحوث الإحصائية، جامعة القاهرة، مصر، 2013.

6. عبد المالك أمين، مقاربة لتحديد اللغات تلقائيا في مدينة نصوص متعددة اللغات، المجلة العربية الدولية للمعلوماتية، المجلد الثاني،العدد الرابع،2013.

7. فادي خلوف: تطوير أليات جديدة للتنقيب في المعطيات لإدارة علاقات الزبائن في بيئة مصرفية، مجلة جامعة دمشق للعلوم الهندسية ،المجلد 26، العدد الأول، 2010 .

8. كادي زين الدين، خديم خديجة: التنقيب المعلوماتي ودوره في تحليل احتياجات مستعملي المكتبات ومراكز المعلومات.

9. مراد عباس، تقييم طرق التعرف الموضوعي للنصوص العربية، مجلة الخليج العربي للبحوث العلمية، 29(3/4)، 2011.

10. مصعب شاهين، شادي صالح:مشروع تخرج بعنوان: تطوير نظام لتصنيف المستندات العربية،اشراف الدكتور ناصر ناصر، قسم البرمجيات ونظم المعلومات، جامعة تشرين سوريا، 2011/2012.

11. هالة حسن، تعدين بينات التمويل الاصغر باستخدام تقنيات التصنيف والعنقدة اشراف طارق عبد الكريم، مذكرو ماجستير، تخصص تقانة المعلومات، كلية علوم الحاسوب وتفانة المعلومات. جامعة النيلين، مارس 2013.

12. بسام الديب، تصنيف النصوص العربية باستخدام الخصائص الغرضية في قواعد البيانات، مجلة جامعة البعث، المجلد 38، العدد 15، 2016.

13. جلال الضاهر، تصميم نموذج نظام دعم القرار لإدارة الموارد البشرية بالاعتماد على تقنيات الذكاء الصنعي، مذكرة ماجستير، الجامعة الافتراضية السورية، 2013/ 2014.

14. سعد بن هادي قحطاني، تحليل اللغة العربية بواسطة الحاسوب، مركز اللغة الانجليزية، معهد الادارة ، الرياض.

15. سيف الدين عثمان،الشفيع جعفر: التنقيب في البيانات واتخاذ القرارات، مجلة النيل

الأبيض للدراسات والبحوث ، العدد الثالث، مارس 2014.

16. بسام محمد احمد السالمي، التصنيف الآلي للنصوص العربية باستخدام تعليم بايزين الاحتمالي، 2011.

17. سامي أدهم، الذكاء الاصطناعي، ثنائية الآلة والدماغ، مجلة كتابات معاصرة،ع 29-28، دجنبر 1996، يناير 1997.

18. سعد بن هادي قحطاني، تحليل اللغة العربية بواسطة الحاسوب، مركز اللغة الانجليزية، معهد الإدارة، الرياض.

19. قنديلجي وآخرون، المدخل إلى إدارة المعرفة، دار المسيرة، عمان، 2006.

20. نبيل محمد لطف مصلي، محاضرات في تنقيب البيانات، جامعة المستقبل لعلوم الإدارة وتكنولوجيا المعلومات.

21. محمد حسن عبد الله، تنقيب بيانات نتيجة التعليم الأساسي، مذكرة ماجستير في تقانة المعلومات، كلية الدراسات العليا، جامعة النيلين، 2016.

المصادر باللغة الأجنبية:

.22 (Brown & Chong, 1998) G.Brown, H.A.Chong « The Guru System in TREC-6 ».

.23 (Clech&Zighed, 2004)J.Clech, D.A.Zighed « Une technique de ré-étiquetage dans un contexte de catégorisation de textes ».

.24 (Fayet-Scribe, 1997)S.Fayet-Scribe « Chronologie des supports, des dispositifs et des outils de repérage de l'information ».

.25 (Lefévre, 2000)P. Lefévre « La recherche d'information - du texte intégral au thésaurus ».

.26 (Loupy& El-Bèze, 2000)C.de Loupy, M.El-Bèze « Using few cues can compensate the small.

.27 (Moulinier, 1996) I.Moulinier « Une approche de la catégorisation de textes par l'apprentissage symbolique ».

.28 (Sebastiani, 2002) F.Sebastiani « Machine learning in automated text categorization .1996«

.29 Advances in Knowledge Discovery and Data Mining, AAAi/MIT Press,amount of resource savailable for WSD »Discovery in Databases, AAAI/MIT Press, .1991

.30 Kareem Darwish. "Building Shallow Arabic Morphological Analyzer in One Day", Association for Computational Linguistics. 40th Anniversary Meeting.July 6-12, .2002 pp. .47-54 University of Pennsylvania.

.31 M. Elkourdi, A. Bensaid, and T. Rachidi, "Automatic Arabic Document Categorization Based on the Naïve Bayes Algorithm", in Proc. of COLING 20th Workshop on Computational Approaches to Arabic Script-based Languages, .2004

.32 Manning, P. Raghavan, and H. Schütze, An Introduction to Information Retrieval, Cambridge University Press, .2009

.33 P. F. Strawson (Introduction to Logical Theory) Methuen & Co. Ltd., London, UK, .1960

.34 R. Carnap (The Logical Syntax of Language) 5th edition, Routledge&Kegan Paul Ltd., London, UK, repérage.1959 de l'information ».

.35 S. Ghwanmeh, G. Kanaan, R. Al-Shalabi and A. Ababneh, "Enhanced Arabic Information Retrieval System based on Arabic Text Classification", 4th International Conference on Innovations in Information Technology, pp.461 - 465, .2007

.36 Fayyad,U., Piatesky-Shapiro,G., Smyth, P., and Uthurusamy, R, (1996).

.37 Piatetsky-Shapiro, G., and Frawley,W., (Eds) (1991). Knowledge.

المواقع الالكترونية:

1. موقع موسوعة الشعر العربي (الديوان): https://www.aldiwan.net .

2. https://hakaya.com موقع حكايا للرواية العربية.

فهرس المحتويات

المحتويات