مجلة (المُدبّـر) المجلد 09، العدد 01، 2022
ISSN 2392-5345/E-ISSN 2661-7129
ص ص: 128-153

التنقيب عن البيانات باستخدام أداة WEKA

Data mining with WEKA

تاريخ الاستلام: 2022-02-02 تاريخ قبول النشر: 2022-06-27

فتيحة, بوهرين[*]، جامعة قسنطينة(عبد الحميد مهري)،الجزائر
fatiha.bouhrine@univ-constantine2.dz
حسيبة, هدوقة، جامعة قسنطينة(عبد الحميد مهري)،الجزائر
hadouga.hassiba@yahoo.fr

فريال منال, عزي، جامعة ميلة(عبد الحميد بوصوف)،الجزائر
f.azzi@centre-univ-mila.dz

Abstract

Data mining is a technique for identifying patterns in large amounts of data and information, databases, data centers, data storage formats, or data that is dynamically streamed to a network and which are examples of data sources.Several programs are used for data mining, including the WEKA program, and accordingly this paper came to provide an overview of the data mining process, as well as its advantages and disadvantages, and data minin methodologies and tasks. This paper also discusses data mining techniques using the WEKA tool.

Keywords : data, data mining, data mining tools, WEKA tool

JEL Classification Codes: C00 , C45, C8

[*] - المؤلف المراسل

مجلة (المُدبّـر) المجلد 09، العدد 01، 2022
ISSN 2392-5345/E-ISSN 2661-7129
ص ص: 128-153

ملخص:

التنقيب عن البيانات هو أسلوب لتحديد الأنماط في كميات كبيرة من البيانات والمعلومات، قواعد البيانات ومراكز البيانات وتنسيقات تخزين البيانات أو البيانات التي يتم دفقها ديناميكيًا إلى الشبكة والتي تكون أمثلة لمصادر البيانات. ويتم استخدام عدة برامج للتنقيب عن البيانات منها برنامج WEKA، وعليه جاءت هذه الورقة لتقدم لمحة عامة عن عملية استخراج البيانات ، فضلاً عن مزاياها وعيوبها و عن منهجيات ومهام التنقيب عن البيانات، كما تناقش هذه الورقة أيضًا تقنيات التنقيب عن البيانات باستخدام أداة WEKA.

الكلمات المفتاحية: البيانات ، التنقيب عن البيانات، أدوات التنقيب عن البيانات، أداة WEKA

تصنيف JEL:. C00 , C45, C8

1. مقدمة:

في عصر يشار إليه غالبًا باسم عصر المعلومات ، و لأننا نعتقد أن المعلومات تؤدي إلى القوة والنجاح ، وبفضل التقنيات المتطورة مثل أجهزة الكمبيوتر والأقمار الصناعية وما إلى ذلك ، وبفضل هده التقنيات تم تجميع كميات هائلة من المعلومات، ففي البداية ، مع ظهور أجهزة الكمبيوتر ووسائل التخزين الرقمي الشامل ، تم البدأ في جمع وتخزين جميع أنواع البيانات ، معتمدين على قوة أجهزة الكمبيوتر للمساعدة في فرز هذا المزيج من المعلومات لكن لسوء الحظ ، أصبحت هذه المجموعات الضخمة من البيانات المخزنة على هياكل متباينة بسرعة كبيرة تشكل فوضى أولية في إنشاء قواعد بيانات منظمة وأنظمة إدارة قواعد البيانات (DBMS)، وعليه تم انشاء أنظمة إدارة قواعد البيانات الفعالة والتي كانت مهمة للغاية لإدارة مجموعة كبيرة من البيانات وخاصة لاسترجاع معلومات معينة بشكل فعال ، كما ساهم انتشار أنظمة إدارة قواعد البيانات أيضًا في التجميع الضخم لجميع أنواع المعلومات، فاليوم ، لدينا معلومات أكثر بكثير مما يمكننا التعامل معه، من المعاملات التجارية والبيانات العلمية ، إلى صور الأقمار الصناعية والتقارير النصية والاستخبارات العسكرية، واسترجاع المعلومات ببساطة لم يعد كافياً لاتخاذ القرار وهدا في مواجهة مجموعات ضخمة من البيانات ، وعليه تم انشاء احتياجات جديدة للمساعدة على اتخاذ خيارات إدارية أفضل تمثلت في موضوع التنقيب عن البيانات ، التي شكلت التلخيص التلقائي للبيانات ، واستخراج "جوهر" المعلومات المخزنة ، واكتشاف الأنماط في البيانات الأولية.

كما يسمح التنقيب عن البيانات بالبحث عن معلومات قيمة بكميات كبيرة من البيانات ، وتوليد النمو الهائل في قواعد البيانات التي هي بحاجة إلى تطوير تقنيات تستخدم المعلومات والمعرفة بذكاء، لذلك ، أصبح DMT مجال بحث مهم بشكل متزايد[1] .

كما يعد التنقيب عن البيانات تقنية جديدة قوية ذات إمكانات كبيرة لمساعدة الشركات على التركيز على المعلومات الأكثر أهمية في مستودعات البيانات الخاصة بهم، حيث

تم تعريفه على أنه، التحليل الآلي لمجموعات البيانات الكبيرة أو المعقدة من أجل اكتشاف الأنماط أو الاتجاهات المهمة التي قد لا يتم التعرف عليها لولا ذلك.

يستخدم التنقيب عن البيانات عدة برامج و تقنياتDMT والتي تعد فرعًا من الذكاء الاصطناعي التطبيقي (AI) ، منذ الستينيات.

حيث يوجد الكثير من البرامج التي يتم استخدامها للتنقيب عن البيانات، ندكر منها KNIME و ORANGE و IBM SPSS و WEKA كأمثلة على أدوات استخراج البيانات، ولكن كلما كان هناك شرط لنموذج تصنيف ما ، فإن WEKA هو الأنسب.

حيث تستخدم أداة WEKA لاستخراج البيانات بعض خوارزميات التصنيف في سياق مجموعة من البيانات، وعليه يتم طرح الاشكالية التالية: كيف يتم استخدام برنامج WEKA كأداة للتنقيب عن البيانات؟

للاحابة عن الاشكالية المطروحة، يتم التطرق إلى ما يلي:

أولا : الاطار النظري

ثانيا: الاطار التطبيقي

أهمية الدراسة: تكتسب الدراسة أهميتها من خلال تناولها لمتغيرين مهمين تنقيب البيانات DM وأداة WEKA لاستخراج البيانات ، حيث يمكن أن يجعل البحث مساهمة متواضعة في هذا المجال خاصة بعد ما لوحظ من نقص كبير جدا في المصادر العربية الخاصة بالمتغير الثاني والذي يعد من الموضوعات الحديثة جدا.

كما تأتي أهمية الدراسة من أهمية البيانات كونها المحرك الأساسي لعمل المنظمات والذي على أساسه تحدد قواعد البيانات وبموجبه تعمل تقنية تنقيب البيانات.

أهداف الدراسة: تهدف الدراسة إلى تحقيق جملة من الأهداف أهمها:

.التعرف على طبيعة البيانات ومقدارها وكيفية التعامل مع الحجوم الكبيرة لها تمهيدا لمعالجتها باستخدام أداة WEKA .

.العمل على تحسين مستوى استخدام قواعد البيانات من قبل الطلبة ، الأساتذة والمؤسسات.

. محاولة إجراء بحوث ودراسات مستقبلية في مجلات جديدة من خلال استخدام أداة التنقيب عن البيانات WEKA ومختلف الأدوات الأخرى.

2. الاطار النظري

2.1مفاهيم عامة حول التنقيب عن البيانات:

حدثت إحدى أولى حالات التنقيب عن البيانات في عام 1936 ، عندما قدم آلان تورينج فكرة آلة عالمية يمكنها إجراء عمليات حسابية مماثلة لتلك الموجودة في أجهزة الكمبيوتر الحديثة.[2]

لقد تطور التنقيب عن البيانات بشكل كبير عما كان عليه في البداية، يمكن تتبع جذور التنقيب في البيانات على ثلاثة مسارات:

–أقدم هذه المسارات هي الإحصاءات الكلاسيكية، فالإحصاء هو أساس التنقيب في البيانات، لن تكون هناك أي طريقة لقياس البيانات بدونها، وهو الجزء الأقدم والأكثر أهمية في التنقيب عن البيانات.[3]

– هناك مسار آخر للتنقيب عن البيانات وهو الذكاء الاصطناعي، حيث يركز الذكاء الاصطناعي على ما يسمى بالتقنيات القائمة على الخبرة لاكتشاف المعرفة، حيث يحاول تطبيق عمليات التفكير البشري على المشاكل الإحصائية.

–المسار الثالث للتنقيب في البيانات هو أكثر من مزيج من المسارين السابقين، بما يعرف بتعلم الآلة، لقد جمعت التقنيات القائمة على الخبرة جنبًا إلى جنب مع التحليل الإحصائي المتقدم [4].

وتمت صياغة مصطلح التنقيب عن البيانات في الستينيات، تم استخدام التنقيب عن البيانات للعثور على المعلومات الأساسية من المجموعات مثل إجمالي الإيرادات على مدى السنوات الثلاث السابقة، وكانت التقنيات التي جعلت هذا ممكناً تتكون من أشرطة وأقراص وأجهزة كمبيوتر.

ثم جلبت الثمانينيات قواعد البيانات الحقيقية للاستخدام على نطاق أوسع، فسمح هذا للوصول إلى البيانات بشكل أسهل عبر SQL، ثم جاءت التسعينيات وجلبت استخدام مستودعات البيانات ودعم القرار، كانت هذه هي الفترة الزمنية التي تم فيها تطوير الكثير من التنقيب عن البيانات التي نراها اليوم.[5]

ولمصطلح التنقيب عن البيانات تعريفات عديدة، حيث تم تقديم أحدهما بواسطة William J Frawley و Gregory Piatetsky Shapiro و Christopher J Matheus على النحو التالي: "التنقيب عن البيانات ، أو اكتشاف المعرفة في قواعد البيانات (KDD) هو الاستخراج غير التافه للكلمات الضمنية ، وغير المعروفة سابقًا ، و المعلومات المفيدة المحتملة من البيانات، و يشمل هذا عددًا من الأساليب الفنية المختلفة ، مثل التجميع ، ومقارنة مجموع البيانات ، وقواعد تصنيف التعلم ، وإيجاد شبكات التبعية ، وتحليل التغييرات ، واكتشاف الحالات الشاذة.[6]

تتمثل فكرة التنقيب عن البيانات في بناء نماذج من مجموعات البيانات هذه التي تساعد في استرداد المعلومات القيمة من المجموعة، يمكن تقسيم طرق الوصول إلى مثل هذه النماذج إلى مجموعات مختلفة من المهام الأساسية:

ˆ التنبؤ: قد يكون من الممكن التنبؤ بقيمة تلك السمة لسجل جديد لا يحتوي على هذه السمة ، على سبيل المثال ، يمكن التنبؤ بمتوسط رصيد العميل الجديد إذا كانت بعض السمات الأساسية معروفة عنه فقط.[7]

ˆ التصنيف: يشبه التنبؤ تقريبًا ، ولكن مع هذا الاختلاف تكون القيمة التي نريد توقعها اسمية مع التصنيف ، قد يكون من الممكن (على سبيل المثال) تحديد ما إذا كان العميل الجديد سيرد الأموال إذا تم إقراضها له.

ˆ التجميع: تجميع السجلات في مجموعات من السجلات "المتشابهة" ("متشابهة" في قيم سماتها)

ˆ الاقتران: تجميع السمات التي يبدو أن لها قيمًا متشابهة لقيم السجلات الخاصة بها [8]

2.2مفاهيم عامة حول برنامج WEKA

WEKA هي أداة لاستخراج البيانات تتيح معالجة البيانات مسبقًا، كما أن أداة Weka (بيئة Waikato لتحليل المعرفة) هي مجموعة شائعة من برامج التعلم الآلي المكتوبة بلغة Java ، تم تطويرها في جامعة Waikato ، نيوزيلندا، وهي برنامج مجاني متاح بموجب رخصة GNU العامة، حيث تحتوي طاولة عمل Weka على مجموعة من أدوات التصور والخوارزميات لتحليل البيانات والنمذجة التنبؤية ، جنبًا إلى جنب مع واجهات المستخدم الرسومية لسهولة الوصول إلى هذه الوظيفة.

ففي عام 1993 ، بدأت جامعة وايكاتو في نيوزيلندا تطوير النسخة الأصلية من Weka ، والتي أصبحت مزيجًا من Tcl / Tk و C و makefiles.[9]

في عام 1997 ، تم اتخاذ قرار بإعادة تطوير Weka من الصفر في Java ، بما في ذلك تطبيقات خوارزميات النمذجة.

في عام 2005 ، حصلت Weka على جائزة SIGKDD لخدمات التنقيب عن البيانات واكتشاف المعرفة.[10]

في عام 2006 ، حصلت شركة Pentaho Corporation على ترخيص حصري لاستخدام Weka في معلومات الأعمال، حيث يشكل عنصر التنقيب عن البيانات والتحليلات التنبؤية لمجموعة Pentaho Business Intelligence. استحوذت شركة هيتاشي فانتارا على Pentaho منذ ذلك الحين ، وتدعم Weka الآن مكون PMI (البرنامج المساعد لذكاء الآلة) مفتوح المصدر[11].

كان الإصدار الأصلي غير Java من Weka عبارة عن واجهة أمامية لـ TCL / TK لخوارزميات نمذجة (معظمها طرف ثالث) تم تنفيذها بلغات برمجة أخرى ، بالإضافة إلى أدوات معالجة البيانات المسبقة في لغة C ، ونظام قائم على Makefile لتشغيل تجارب التعلم الآلي، تم تصميم هذا الإصدار الأصلي بشكل أساسي كأداة لتحليل البيانات من المجالات الزراعية ، ولكن الإصدار الأحدث المستند إلى Java بالكامل

(Weka 3) ، والذي بدأ تطويره في عام 1997 ، يُستخدم الآن في العديد من مجالات التطبيق المختلفة ، ولا سيما للأغراض التعليمية و البحث، وتشمل مزايا Weka ما يلي:

– التوفر المجاني [12]

–قابلية النقل ، نظرًا لأنها مطبقة بالكامل في لغة برمجة Java وبالتالي فهي تعمل على أي منصة حوسبة حديثة تقريبًا

–تحتوي على مجموعة شاملة من معالجة البيانات وتقنيات النمذجة.

– سهولة في الاستخدام بفضل واجهات المستخدم الرسومية.

–لديها مجموعة متنوعة من الخيارات والخوارزميات المذكورة أدناه:[13]

ˆ المعالجة المسبقة

ˆ المصنفات

ˆ التجميع

ˆ المساعدون

سمات المقيمين

ˆ التصورات[10]

لا تقبل أداة WEKA الملفات بتنسيق xlsor xlsx العادي، ويعد ARFF هو "تنسيق ملف علاقة السمة" للملف الافتراضي لـ WEKA ، على الرغم من أنه يقبل أيضًا ملفات بتنسيق قيم مفصولة بفواصل (CSV) ، تنسيق C4.5 ، إلخ.

حيث يعرف ملف ARFF على أنه ملف نصي ASCII ؛ يعطي قائمة بالمثيلات التي تشترك في مجموعة من السمات. تحتوي ملفات ARFF على قسمين متميزين – قسم الرأس وقسم البيانات، ترد تفاصيل العلاقة والسمات وأنواعها في قارئ ملف ARFF ويحتوي قسم البيانات على البيانات ذات الصلة. يمكن إضافة التعليقات باستخدام علامة

النسبة المئوية (٪) واسم مجموعة البيانات باستخدام علامةrelation ومعلومات السمة باستخدامattributetag. السطر الأول في ملف ARFF هو اسم العلاقة. التنسيق هو

@relation <relation-name>

حيث <relation-name> عبارة عن سلسلة. إذا كان الاسم يحتوي على مسافات ، فيجب أن يتم اقتباس السلسلة.

تنسيق بيانattribute هو :

@attribute <attribute-name><datatype>

يدعم Weka أربعة أنواع من البيانات:

- رقمى - اسمي - سلسلة - التاريخ

قد تكون السمات الرقمية أرقامًا حقيقية أو أرقامًا صحيحة ، ويتم استخدام السمات الاسمية عند اختيار قيمة سمة من قائمة محددة مسبقًا. يتم تعريف السمة الاسمية من خلال توفير "مواصفات اسمية" تسرد القيم المحتملة على النحو التالي: <الاسم الاسمي 1> ، <الاسمي 2> ، <الاسم 3> ، ...[14]

3. الاطار التطبيقي

3.1تجميع البيانات في شركة Avon :

نرغب في تجميع عملاء **شركة Avon** بناءً على الميزات المشتركة ودلك عبر الإنترنت، لا تملك إدارة الشركة أي تسميات محددة مسبقًا لهذه المجموعات، بناءً على نتيجة التجميع ، سوف يستهدف الحملات التسويقية والإعلانية للمجموعات المختلفة. تتضمن المعلومات التي لديهم عن العملاء معرّف العميل ، واسم العميل ، وعدد العملاء ، و ProductSold ، وقناة المبيعات ، والوحدات المباعة ، وتاريخ البيع.

تحتوي WEKA على "مجموعات" للعثور على مجموعات من الحالات المتشابهة في مجموعة بيانات. مخططات المجموعات المتوفرة في WEKA هي k-Means و EM

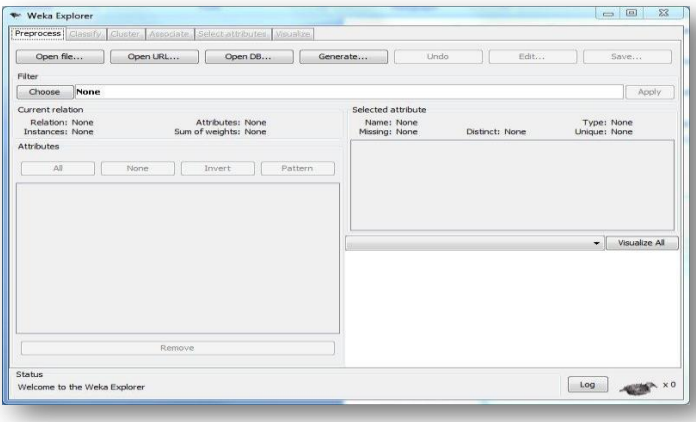
و Cobweb و X–means و Farthest First. يمكن تصور المجموعات ومقارنتها بالمجموعات "الحقيقية" (إذا تم تقديمها)، يعتمد التقييم على احتمالية السجل إذا كان مخطط التجميع ينتج توزيعًا احتماليًا.

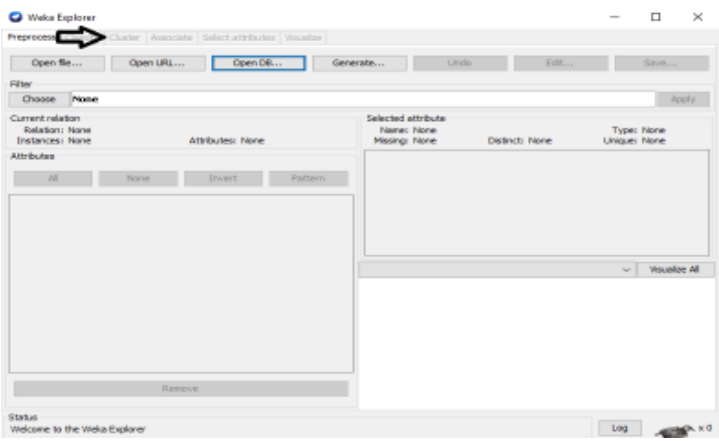
في مثالنا ، سنستخدم جزءًا من قاعدة البيانات للعملاء ،و اعتمادًا على نوع المنتجات المباعة ، ليست كل السمات مهمة.

في نافذة "Preprocess" ، يتم النقر فوق الزر "فتح ملف ..." وتحدد ملف "customers.csv". بالنقر فوق علامة التبويب "Cluster" أعلى نافذة WEKA Explorer يتم الحصول على نافدة بيانات برنامج WEKA

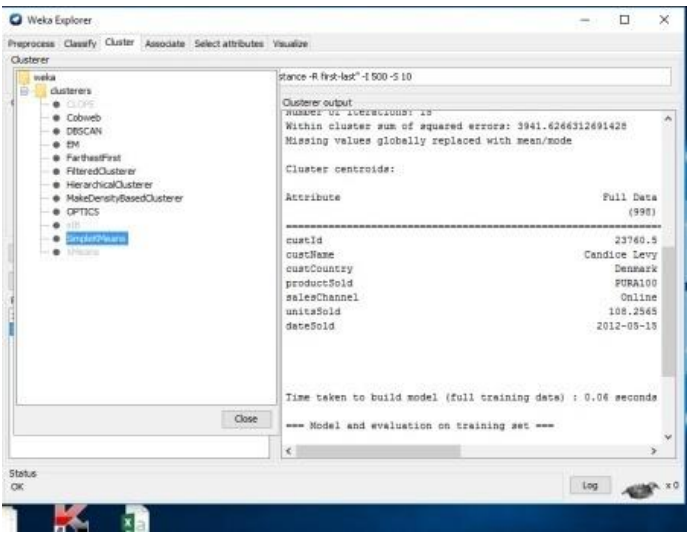
لكن يجب التحويل من تنسيق xlsv / xls إلى تنسيق csv، يمكن بسهولة تحويل البيانات المخزنة في ورقة عمل Excel إلى تنسيق ARFF / csv. يجب اتباع الخطوات التالية لتحويل جدول بيانات إلى تنسيق csv:

1. فتح جدول البيانات في MS Excel.
2. تحديد ملف–> حفظ باسم
3. في علامة التبويب Sava as type ، حدد CSV (محدد بفواصل) (*. csv)
4. تم حفظ الملف بتنسيق CSV. يمكن تحميله في WEKA. سيتم أخذ القيم الموجودة في الصف الأول كأسماء سمات وسيتم التعامل مع الصفوف اللاحقة على أنها مثيلات بيانات.

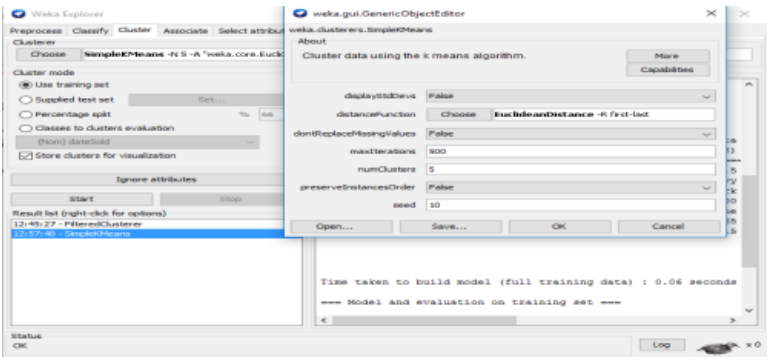




–اختيار نظام التجميع:في مربع "Clusterer" يتم بالنقر فوق الزر "اختيار". في القائمة المنسدلة ، تحدد WEKA Æ Clusterers ، وتحدد مخطط المجموعة "SimpleKMeans، فبعض تطبيقات K– تعني السماح فقط بالقيم العددية للسمات ؛ لذلك ، لا نحتاج إلى استخدام عامل تصفية وفق الشكل الموضح في الأسفل:



بمجرد اختيار خوارزمية التجميع ، بالنقر بزر الماوس الأيمن على الخوارزمية ، ويظهر "ضعيفة.gui.GenericObjectEditor" على الشاشة. تعيّن القيمة في مربع "numClusters" إلى 5 (بدلاً من الافتراضي 2) لأن لدينا خمس مجموعات في ملف .arff الخاص بنا. يتم ترك قيمة "البذرة" كما هي. يتم استخدام القيمة الأولية في إنشاء رقم عشوائي ، والذي يتم استخدامه لإجراء التخصيص الأولي لمثيلات المجموعات. يلاحظ أنه بشكل عام ، فإن K–mean حساسة جدًا لكيفية تعيين المجموعات في البداية. وبالتالي ، غالبًا ما يكون من الضروري تجربة قيم مختلفة وتقييم النتائج.



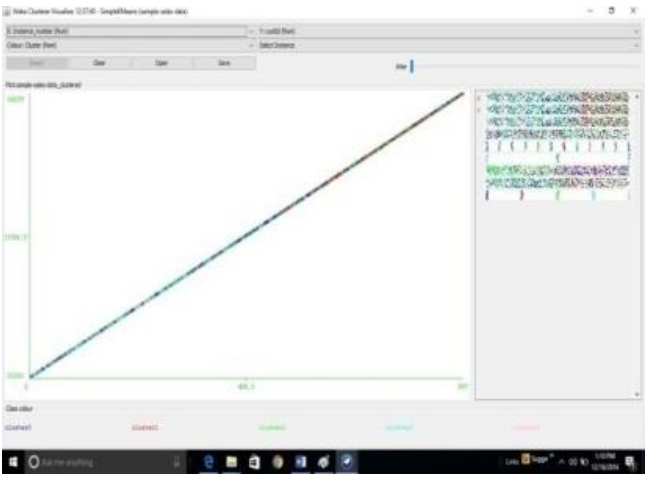
–ضبط خيارات الاختبار: قبل تشغيل خوارزمية التجميع ، تحتاج إلى اختيار "وضع المجموعةCluster mode " . بالنقر على زر الاختيار "فصول للتقييم العنقودي" في مربع "وضع المجموعةClasses to cluster evaluation " تحدد المربع المنسدل أدناه.

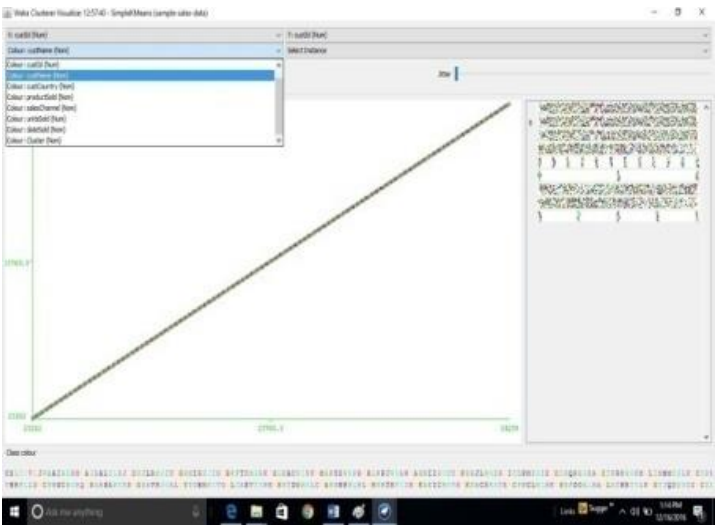
بمجرد تحديد الخيارات ، يمكن تشغيل خوارزمية التجميع. بالنقر فوق الزر "ابدأ" لتنفيذ الخوارزمية.

مجلة (المُدبّـر) المجلد 09، العدد 01، 2022
ISSN 2392-5345/E-ISSN 2661-7129
ص ص: 128-153

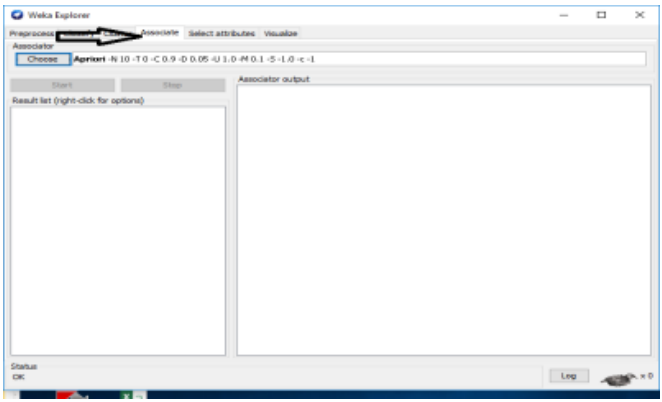
–تصور النتائج: طريقة أخرى لتمثيل نتائج التجميع من خلال التصور. بالنقر على زر الماوس الأيمن على الإدخال في "قائمة النتائجVisualize cluster assignments " تحدد "تصور تعيينات المجموعة" في النافذة المنسدلة.

يُظهر هذا نافذة "Weka Clusterer Visualize".

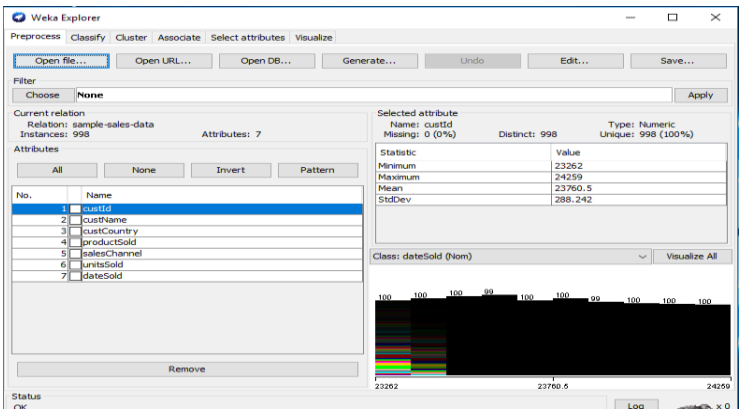




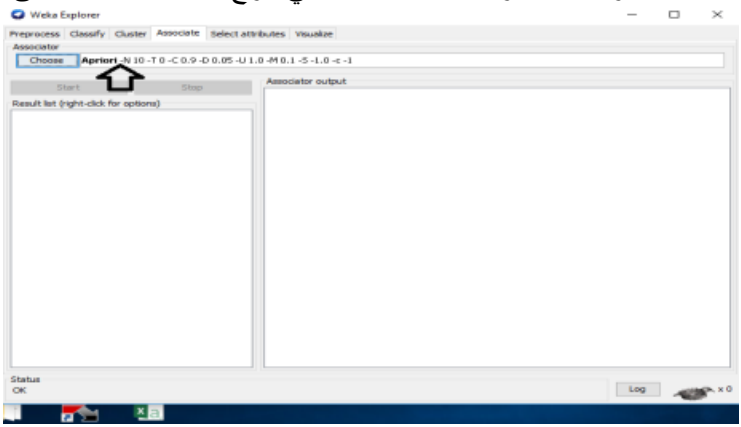
–تحديد ASSOCIATION: يحتوي WEKA على تطبيق خوارزمية Apriori لقواعد جمعية التعلم. هذا هو المخطط الوحيد المتاح حاليًا لجمعيات التعلم في WEKA. إنه يعمل فقط مع البيانات المنفصلة وسيحدد التبعيات الإحصائية بين مجموعات السمات. يمكن أن يحسب Apriori جميع القواعد التي لها حد أدنى معين من الدعم وتتجاوز ثقة معينة.



في هذا المثال ، ستستخدم بيانات المبيعات من ملف "sales-sample-data.csv".

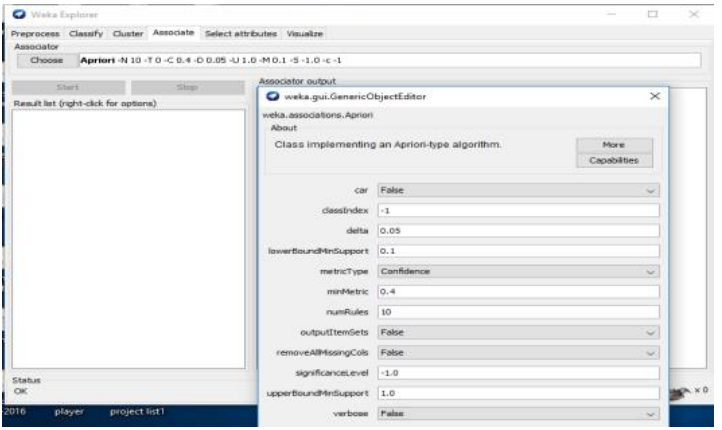


-ضبط خيارات الاختبار :يحدد حقل النص في مربع "Associator" أعلى النافذة.

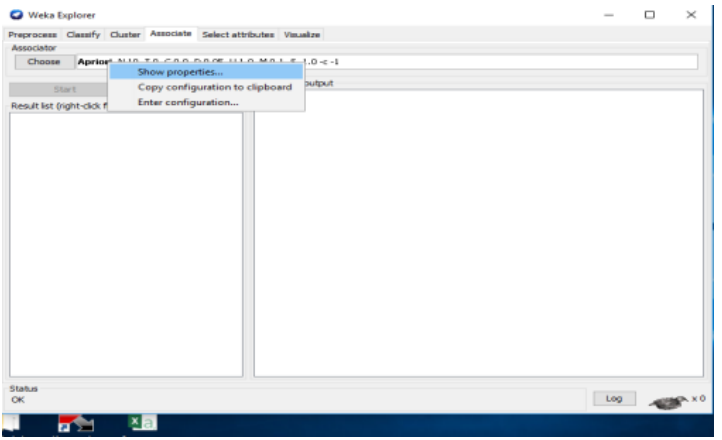


بالنقر على زر الماوس الأيمن فوق مربع "Associator" ، ثم النقر فوق إظهار الخصائص ، يظهر "GenericObjectEditor" على الشاشة. في مربع الحوار ، يتم تغيير القيمة في "minMetric" إلى 0.4 للثقة = 40٪. ثم التأكد من تعيين القيمة الافتراضية للقواعد على 100. يجب تعيين الحد الأعلى للدعم الأدنى "upperBoundMinSupport" على 1.0 (100٪) و "LowerBoundMinSupport"

إلى 0.1. يبدأ Apriori في WEKA بالدعم الأعلى ويقلل الدعم بشكل تدريجي (بزيادات دلتا ، والتي يتم تعيينها افتراضيًا على 0.05 أو 5٪). تتوقف الخوارزمية عند إنشاء العدد المحدد من القواعد ، أو الوصول إلى الحد الأدنى للدعم الأدنى. لا ينطبق خيار اختبار "مدى الأهمية" إلا في حالة الثقة ويكون (−1.0) افتراضيًا (غير مستخدم).

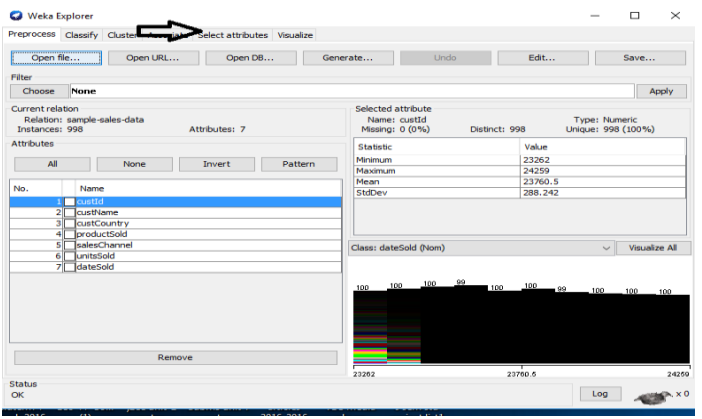


بمجرد تحديد الخيارات ، يمكن تشغيل خوارزمية Apriori. بالنقر فوق الزر "ابدأStart " لتنفيذ الخوارزمية.



في نافذة "Weka Clusterer Visualize" ، أسفل محدد المحور X ، توجد قائمة منسدلة ، "اللون" ، لاختيار نظام الألوان. يتيح لك هذا اختيار لون النقاط بناءً على السمة المحددة. أسفل منطقة الرسم ، توجد وسيلة إيضاح تصف القيم التي تتوافق معها الألوان.

يبحث تحديد السمات في جميع مجموعات السمات الممكنة في البيانات ويجد أي مجموعة فرعية من السمات تعمل بشكل أفضل للتنبؤ. تحتوي طرق اختيار السمة على جزأين: طريقة بحث مثل الأفضل أولاً ، التحديد الأمامي ، العشوائية ، الشاملة ، الخوارزمية الجينية ، الترتيب ، وطريقة التقييم مثل القائمة على الارتباط ، الغلاف ، اكتساب المعلومات ، مربع كاي. آلية اختيار السمات مرنة للغاية – تسمح WEKA (تقريبًا) بتوليفات عشوائية من الطريقتين. لبدء تحديد سمة ، انقر فوق علامة التبويب "تحديد السماتSelect attributes " .

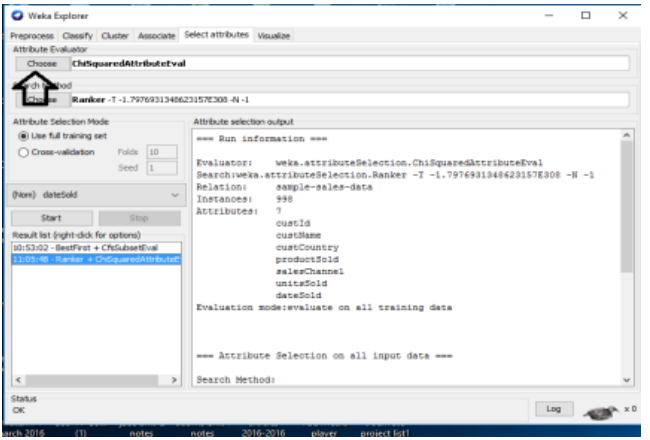


يظهر لنا الشكل أن أعلى السمات المشتركة بين فئات زبائن الشركة من حيث سمة مبيعات الشركة هي 24259 ، وفي المتوسط 23760,5 من زبائن الشركة قادري على دفع مبيعات الشركة .

للبحث في جميع مجموعات السمات الممكنة في البيانات والعثور على أي مجموعة فرعية من السمات تعمل بشكل أفضل للتنبؤ ، يتم التأكد من إعداد مقيِّم السمات إلى "CfsSubsetEval" وطريقة بحث إلى "BestFirst". سيحدد المقيم الطريقة التي يجب استخدامها لتعيين قيمة لكل مجموعة فرعية من السمات. ستحدد طريقة البحث أسلوب البحث المطلوب إجراؤه. الخيارات التي يمكنك تعيينها للاختيار في مربع "وضع تحديد السمةAttribute Selection Mode " هي:

1. استخدم مجموعة التدريب الكاملة**Use full training set** . يتم تحديد قيمة مجموعة السمة الفرعية باستخدام المجموعة الكاملة لبيانات التدريب.

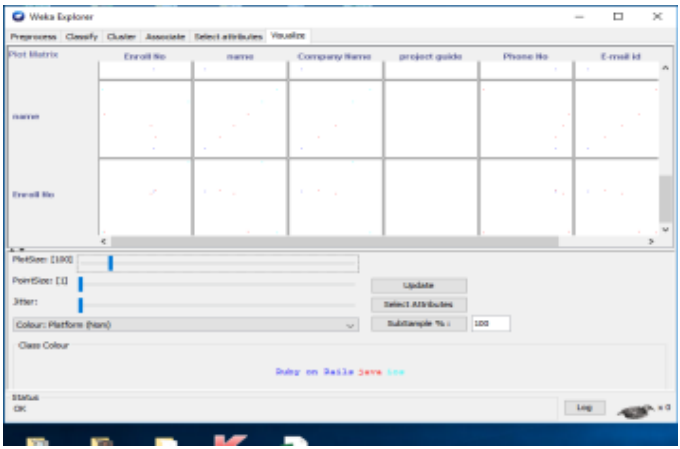
2. عبر التحقق من الصحة**Cross-validation** . يتم تحديد قيمة مجموعة السمة الفرعية من خلال عملية التحقق من الصحة. يحدد حقلا "الطي" و "المتغير" عدد



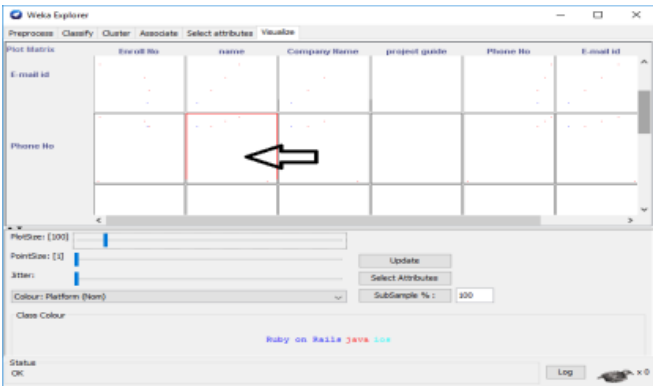
الطيات المراد استخدامها والمتغيرات العشوائية المستخدمة عند خلط البيانات.

تحدد السمة التي يجب معاملتها على أنها فئة في المربع المنسدل أسفل خيارات الاختبار. بمجرد تعيين جميع خيارات الاختبار ، يمكن بدء عملية اختيار السمة بالنقر فوق الزر "ابدأ".

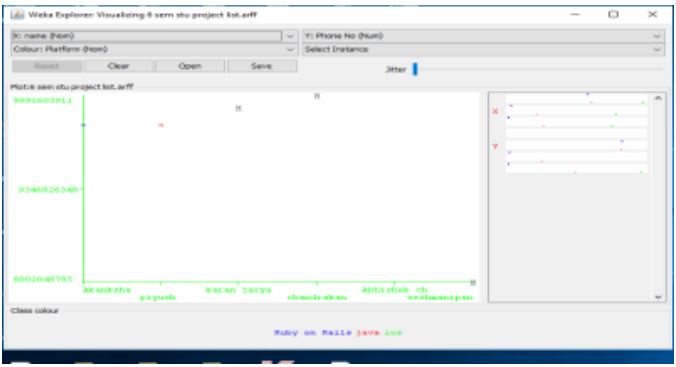
–عرض مرئي للمعلومات: يسمح لنا تصور WEKA بتصور مخطط ثنائي الأبعاد لعلاقة العمل الحالية. التخيل مفيد جدًا في الممارسة ، فهو يساعد على تحديد صعوبة مشكلة التعلم. يمكن لـ WEKA تصور سمات مفردة (1–د) وأزواج من السمات (2–د) ، وتدوير تصورات ثلاثية الأبعاد (نمط Xgobi). لدى WEKA خيار "Jitter" للتعامل مع السمات الاسمية واكتشاف نقاط البيانات "المخفية".



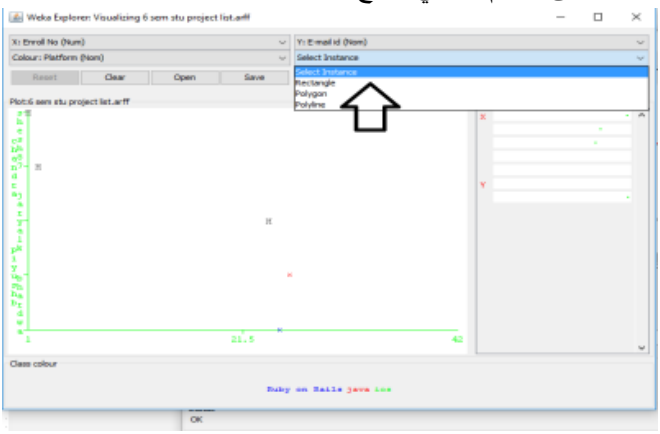
بتحديد مربعًا يتوافق مع السمات التي ترغب في تصورها. على سبيل المثال ، نختار "outlook" للمحور X و "play" للمحور Y –. ننقر في أي مكان داخل المربع الذي يتوافق " النظرة المستقبلية "في الأعلى.



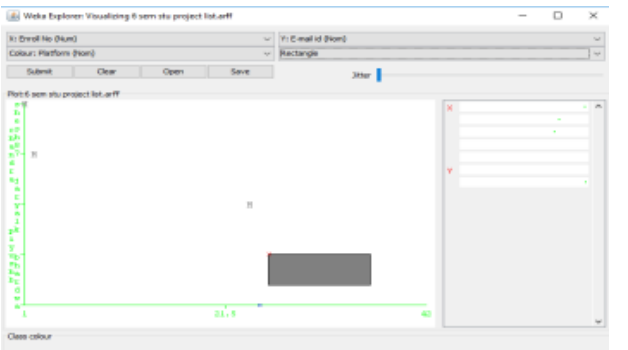
تظهر نافذة "تصور r Visualizing " على الشاشة

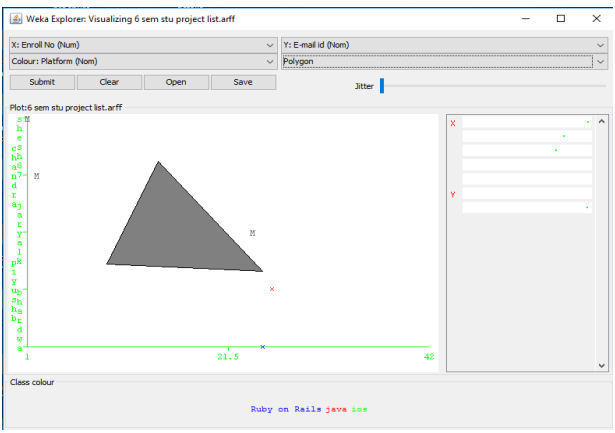


2.3اختيار المثيلات: في بعض الأحيان يكون من المفيد تحديد مجموعة فرعية من البيانات باستخدام أداة التصور. حالة خاصة هي "مصنف المستخدم" ، والذي يتيح لنا إنشاء المصنف الخاص بك عن طريق تحديد الحالات بشكل تفاعلي. يوجد أسفل المحور ص قائمة منسدلة تسمح لك باختيار طريقة تحديد. يمكن تحديد مجموعة من النقاط على الرسم البياني بأربع طرق.



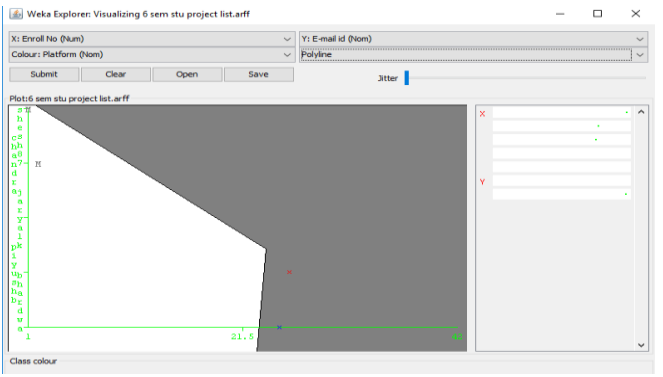
1. تحديد المثيل: بالنقر على نقطة البيانات الفردية، تظهر نافذة تسرد سمات النقطة. إذا ظهرت أكثر من نقطة واحدة في نفس الموقع ، فسيتم عرض أكثر من مجموعة سمات واحدة.
2. المستطيل: يمكن إنشاء مستطيل عن طريق سحبه حول النقاط.
3. مضلع: يمكن تحديد عدة نقاط من خلال بناء مضلع حر الشكل. بالنقر على زر





الماوس الأيسر على الرسم البياني لإضافة رؤوس إلى المضلع والنقر على زر الماوس الأيمن لإكماله.

4. متعدد الخطوط: للتمييز بين النقاط الموجودة على جانب واحد والنقاط الموجودة في الجانب الآخر ، يمكن بناء خط متعدد الخطوط. بالنقر على زر الماوس الأيسر على الرسم البياني لإضافة رؤوس إلى الشكل المتعدد الخطوط وبالنقر على زر الماوس



الأيمن للإنهاء.

4. خاتمة:

عمدت هده الورقة على اعطاء مفاهيم عامة حول أداة التنقيب عن البيانات – WEKA، حيث تمت مناقشة تنسيقات الملفات المختلفة التي يمكن استخدامها مع أداة WEKA ، فتنسيق الملف الافتراضي لـ WEKA هو ARFF، والدي عن طريقه تم توضيح كيفية تخزين البيانات بتنسيق ARFF وكيفية تحويل البيانات من تنسيقات الملفات الأخرى إلى تنسيق ARFF،كما تتكون مهمة التصنيف من فحص ميزات بيانات يتم تقديمها حديثًا وتعيين فئة محددة مسبقًا له.

كما توضح هده الورقة كيفية إنشاء وتقييم البيانات باستخدام أداة WEKA عن طريق استخدام مجموعة تجميع البيانات المتوفرة على أداة WEKA، فعند فحص النموذج باستخدام نفس مجموعة البيانات المستخدمة لبناء النموذج ، يتم تقسيم مجموعة البيانات إلى مجموعة بيانات التدريب ومجموعة بيانات الاختبار ؛ فيتم استخدام مجموعة البيانات الأولى لإنشاء النموذج ، ويتم التحقق من دقة النموذج باستخدام مجموعة البيانات الثانية، كما تم اظهار أيضًا كيف يمكن استخدام النموذج للتنبؤ بفئة تلك المجموعات ، والتي لا يُعرف تصنيفها مسبقا.

5. المراجع:

[1]ناهد عبدالعزيز العوضي ،(2010)، استخدام تقنية التنقيب عن البيانات لتطوير العملية التعليمية في نظم التعليم عن بعد ، جامعة الملك عبدالعزيز .

[2]Agrawal, R., Imielinski, T., Swami, A. (1993), "Database Mining: A Performance Perspective", IEEE Transactions on Knowledge and Data Engineering, p. 914

[3]Bengio Y., Buhmann J. M., Embrechts M., and Zurada J.M (2012). Introduction to the special issue on neural networks for data mining and knowledge discovery. IEEE Trans. Neural Networks.

[4]Berry, J. A., Lindoff, G. (1997), Data Mining Techniques, Wiley Computer Publishing (ISBN 0-471-17980-9)

[5]Craven M. W. and Shavlik J. W. (1997). Using neural networks for data mining. Future Generation Computer Systems, 13:211,p229

[6]W. M. Dlamini, (2011),A Data Mining Approach to Predictive Vegetation Mapping Using Probabilistic Graphical Models, Ecological Informatics 6, p111

[7]H. A. Edelstein, (2005),Introduction to Data Mining and Knowledge Discovery. Potomac, MD, USA: Two Crows Corporation, p43.

[8]R. R. L. Rokach and O. Maimon, (2010),Supervised Learning, In L. Rokach and O. Maimon, Data Mining and Knowledge Discovery Handbook ,p133.

[9]Llobodanin, I. A. Castro and R. Barbosa, (2019),Using Support Vector Machines and Neural Networks to Classify Merlot Wines from South America, Information Processing in Agriculture , p265.

[10]Prof. Dr. K. Coolsaet. (2004),De nieuwe api voor invoer en uitvoer in java (dutch), p89.

[11]Bill Venners. (2000),Inside the Java Virtual Machine. McGraw-Hill, 2nd edition, p45.

[12]D. L. Olson and L. G. (2019),Descriptive Data Mining, In Descriptive Data Mining ,p 129.

[13]L. Rokach, (2010),A Survey of Clustering Algorithms, In L. Rokach and O. Maimon, Data Mining and Knowledge Discovery Handbook ,p 269.

[14]L. Rokach and O. Maimon, (2010),Classification Trees, In L. Rokach and O. Maimon, Data Mining and Knowledge Discovery Handbook ,p148 .