مجلة الحاسوب والتقانة العلمية
Scientific Journal of Computer and Technology

استخدام تقنية التصنيف للتنبؤ بالصادرات البستانية
باستخدام خوارزمية (J48)

Using the classification technique to forecast horticultural exports Using Algorithm (J48)

إعداد الباحث

أ.د / الطاهر محمد حسين

pro.altaher@oiu.ed.com

استاذ مشارك – جامعة أمدرمان الإسلامية

أعداد الباحثة

مناهل حسين حسن أحمد

mnahelhu2020@gmail.com

استاذ مساعدة – الكلية الأردنية

المستخلص

يقدم هذا البحث دراسة تحليلية وتنقيب البيانات لمجموعة من البيانات تقدر بـ 903 سجل للصادرات البستانية في وزارة الزراعة والموارد الطبيعية وذلك خلال الفترة من 2016–2021 ، حيث تكمن مشكلة البحث في وجود كمية كبيرة من البيانات التي لايتم الإستفادة منها وكذلك عدم الاهتمام بالمحاصيل البستانية يهدف البحث إلى إستخدام تقنيات التنقيب عن البيانات في تحليل بيانات الصادرات البستانية ، تم إستخدام المنهج العلمي التجريبي في هذه الدراسة وإستخدام خوارزمية التصنيف J48 لتوليد شجرة القرار والتنبؤ بالصادرات البستانية ، تم التوصل إلى عدة نتائج أهمها أن قطر والإمارات والسعودية ومصر من أكثر البلدان التي يتم تصدير المنتجات البستانية لها ، حيث كانت نسبة التصدير في عام 2018 هي الأعلى وبدأت تقل بشكل ملحوظ نسبة للعوامل الإقتصادية التي تمر بها السودان ، ومن أهم التوصيات التي نوصي بها هي السعي لبناء مستودع بيانات متكامل يساعد في تسجيل جميع الصادرات البستانية.

الكلمات المفتاحية : تنقيب البيانات ، خوارزمية شجرة القرار ، التنبؤ ، الصادرات البستانية

Abstract

This research presents an analytical study and data mining for a set of data estimated at 903 records of horticultural exports in the Ministry of Agriculture

and Natural Resources during the period from 2016-2021, where the research problem lies in the presence of a large amount of data that is not utilized as well as the lack of interest in horticultural crops. To the use of data mining techniques in analyzing the data of horticultural exports, the experimental scientific method was used in this study and the J48 classification algorithm was used to generate a decision tree and predict horticultural exports. For her, as the percentage of exports in 2018 was the highest and began to decrease significantly in proportion to the economic factors that Sudan is going through, and one of the most important recommendations that we recommend is to seek to build an integrated data warehouse that helps in recording all horticultural exports.

Keywords: data mining, decision tree algorithm, forecasting, horticultural exports

1. المقدمة

مع وجود كميات كبيرة من البيانات المخزنة في قواعد البيانات ومخازن البيانات إزدادت الحاجة إلى تطوير ادوات تمتاز بالقوة لتحليل البيانات وإستخراج المعلومات والمعارف منها ، من هنا ظهر ما يسمى بالتنقيب في البيانات كتقنية تهدف إلى إستخراج المعرفة من كميات هائلة من البيانات. وهي تقنية حديثة فرضت نفسها بقوة في عصر المعلوماتية ، وإستخدامها يوفر للشركات والمنظمات في جميع المجالات القدرة على إستكشاف والتركيز على أهم المعلومات في قواعد البيانات ، كما تركز تقنيات التنقيب على بناء التنبؤات المستقبلية وإستكشاف السلوك والإتجاهات مما يسمح بإتخاذ القرارات الصحيحة وإتخاذها في الوقت المناسب. والتى تعتبر بدورها مرحلة من مراحل عملية أكثر تعقيداً هى إستكشاف المعرفة في قواعد البيانات ، والمرتبطة إلى حد بعيد بعملية تطوير أخرى مهمة جداً هى مستودعات البيانات. حيث أن الكثير من الشركات والمنظمات الرائدة اليوم تستخدم عملية إستكشاف المعرفة في قواعد البيانات بشكل منهجي ومنظم بوصفها تشكل جوهر العمل الذي يعتمد عليه في تفعيل النشاط وتحقيق الميزة التنافسية.

إن التطور فى العلم والإقتصاد وتكنولوجيا المعلومات أدى إلى زيادة كمية البيانات الرقمية ، هذه الكميات الهائلة من البيانات لم تعد وسائل التحليل التقليدية قادرة على التعامل معها فظهر التنقيب عن البيانات منذ أواخر الثمانينات وأثبت وجوده كأحد الحلول الناجحة لتحليل كميات ضخمة من البيانات ، وذلك بتحويلها من مجرد معلومة متراكمة وغير مفهومة إلى معلومات قيمة يمكن إستغلالها والإستفادة منها بعد ذلك (معرفة) حيث ظهر مصطلح التنقيب فى البيانات في منتصف التسعينيات فى الولايات

المتحدة الامريكية. وهو ما يجمع ما بين الإحصاء وتكنولوجيا الإعلام (قواعد البيانات ، الذكاء الإصطناعي ، التعلم الآلي machine learning).

إن التنقيب في البيانات يهدف إلى إستخلاص المعرفة المخبأة فيها ، وهي تكنولوجيا حديثة فرضت نفسها بقوة في عصر المعلوماتية وإنتشار إستخدام قواعد البيانات (سيد،2016).

2. مشكلة البحث

تكمن مشكلة البحث في توفر الانتاج للصادر لبعض المنتجات البستانية في اوقات وجود فجوة في الاسواق العالمية وعدم الاستفادة من هذه الفرصة ، كما أن البيانات المتعلقة بالمحاصيل البستانية غير دقيقة ، وهنالك ضعف بالاهتمام بالبحث البستاني في كل حلقات سلسلة القيمة للمحاصيل البستانية في المؤسسات البحثية وعدم الاهتمام بالمحاصيل الواعدة التي تساهم في زيادة الدخل القومي .

3. أهداف البحث

يكمن الهدف الأساسي للبحث في استخدام تقنية التصنيف للتنبؤ بالصادرات البستانية
باستخدام خوارزمية (J48)

أما الأهداف الفرعية تتمثل في بناء نموذج يساعد على التنبؤ بالمنتجات البستانية التي يمكن تصديرها والاستفادة من البيانات المخفية في قاعدة البيانات والمرتبطة بالانتاج البستاني وكذلك الاستفادة من الميزة الانتاجية المتوفرة في السودان من اجل الحصول على العملات الصعبة وتوفير الانتاج للسوق المحلي وتحسين الجودة.

4. منهجية البحث

استخدام المنهج التحليلي والإحصائي وهي عبارة عن مناهج تعتمد على تحليل البيانات المعطاه وتمثيلها والإستفادة منها بشكل بياني وإحصائي.

5. أهمية البحث

1- الصادر البستاني يخلق مجال واسع للخدمات مثل صناعة مواد التعبئة ، الاتصال ، النقل ، مراكز الفرز والتعبئة والتزيين وطائرات النقل وتجارة المدخلات وخلق عمالة

2- توفر العائد المجزي والسريع في حالة توجة الزراعة للصادر

3- معظم المحاصيل البستانية يتوافق انتاجها في فترة الشتاء حيث يغطي الجليد البلدان الاوربية ولا تتوافر منتوجات طازجه لفترة 52 يوما من الدول المصدرة للاتحاد الاوربي

6. مساهمة البحث

مساهمة البحث تتمثل في الإستفادة من بيانات الصادارات البستانية التي تم الحصول عليها من الهيئة العامة للجمارك السودانية بغرض بناء نموذج تنبؤي يساعد في عملية إتخاذ القرارات الخاصة بالتصدير للمنتجات البستانية وكذلك في رفد المكتبات العربية بهذا البحث كونه فريد من نوعه ويتميز عن البحوث والدراسات السابقة في المنهجية المتبعة والأدوات المستخدمة حيث إهتمت كل الدراسات السابقة بذكر المحاصيل الزراعية وتحليل الفجوة بين الصادرات والواردات من ناحية إقتصادية بحته أو بإستخدام السلاسل الزمنية في عملية التحليل أما في هذه الدراسة فقد إعتمدت على تنقيب البيانات في عملية تحليل الصادرات البستانية والتنبؤ بها .

7. فرضيات البحث

1. استخدام تقنيات التنقيب عن البيانات يساعد في التنبؤ بالصادرات البستانية .
2. تعتبر خوارزمية التصنيف (J48) من أفضل الخوارزميات المستخدمة في التنبؤ .
3. هنالك فجوة بين الصادرات الخاصة بالمنتجات البستانية وبين العائد .
4. إستخدام تنقيب البيانات في التنبؤ بالصادرات ساعد في عملية إتخاذ القرارات المهمة الخاصة بالتصدير.

8. الدراسات السابقة

أ. دراسة ياسر عوض (2019)

تناولت الدراسة تصميم نظام تنبؤ بإستخدام تقنيات تنقيب البيانات ، إستخدمت فيه الطرق الإحصائية للتنبؤ بالصادرات الجمركية عبر النظام الإلكتروني للجمارك السودانية ، تكمن مشكلة الدراسة في عدم إستخدام هيئة الجمارك لأنظمة دعم القرار الذي يؤدي الى عدم الإستقلال الأمثل للبيانات في إتخاذ القرارات مما يستنزف الكثير من الجهود وتضيع العديد من الفرص علماً بأن هيئة الجمارك تتوفر لديها قاعدة بيانات ضخمة من الصادرات والواردات ، هدفت الدراسة إلى تقديم دراسة تطبيقية في مجال التنقيب في البيانات لدعم صانعي القرار بهيئة الجمارك السودانية مستخدماً أربع طرق إحصائية للتوصل إلى الطريقة الأفضل من بينها ، إعتمدت الدراسة على المنهج العلمي في التحليل والتصميم . وتم إستخدام كل من لغة النمذجة الموحدة (UML) في توثيق النظام ، لغة الجافا في بناء التطبيق ، أوركل (11g) في قواعد البيانات وبرنامج SPSS في التحليل الإحصائي ، أظهرت نتائج الدراسة أن الطرق الإحصائية

ناجحة جداً في دعم إتخاذ القرار وأن طريقة المربعات الصغرى هي صاحبة الأعلى تقدير من حيث القيمة وتليها طريقة الإنحدار الخطي البسيط ، ثم طريقة المتوسطات المتحركة وأخيراً طريقة الأنسبة المتصلة ، أوصت الدراسة بضرورة الإعتماد على التنبؤ في إعداد الخطط والبرامج التي تساعد في دعم الدخل القومي (عوض،2019).

ب. دراسة علي درب (2013)

تمثل مشكلة الفقد في بعض مشاهدات السلسلة الزمنية أهمية خاصة لاسيما في عملية تقدير النماذج الرياضية و الإحصائية ، الأمر الذي استوجب دراسة طرائق تقدير المشاهدات المفقودة و المتمثلة بطريقة متوسط السلسلة الزمنية و طريقة المتوسط الحسابي للقيم المتجاورة و الوسيط للقيم المتجاورة و طريقة الاستكمال الخطي و أخيرا طريقة التعويض بالانحدار و اختبارها باستخدام معايير الدقة التنبؤية و هي نسبة متوسط القيمة المطلقة للأخطاء (Mean Absolute Percentage Error) (MAPE)، و متوسط القيمة المطلقة للأخطاء معايير الدقة التنبؤية و هي نسبة متوسط القيمة المطلقة للأخطاء ((MAE Mean Absolute Error)) ومتوسط مربع الأخطاء , (Mean Square Error) (MSE، فضلا عن استخدام نماذج الانحدار البسيطة و مقارنة بعض الاختبارات الإحصائية المتحصل عليها من هذه النماذج كاختبار F و اختبار R2 لتدعم النتائج المتحصل عليها من اختبارات الدقة التنبؤية للحكم على أفضل الطرائق في تقدير المشاهدات المفقودة، و توصل البحث إلى أفضلية طريقة التعويض بالانحدار من خلال تفوقها باختبارات الدقة التنبؤية (درب،2013).

ج. دراسة : عبد الله سليمان (2013)

يهدف هذا البحث إلى تحليل السلسلة الزمنية لمؤشر أسعار الأسهم لسوق الخرطوم للأوراق المالية باستخدام منهجية بوكس – جنكنز ، حيث يتم التعرف على نمط المؤشر لبناء أفضل نموذج يساعد على التنبؤ بقيم المؤشر في الأحل القصير, و ذلك بالاعتماد على البيانات الشهرية للفترة من يناير 2005 إلى يونيو 2012 ، و قد أظهرت نتائج الدراسة أن السلسلة الزمنية للمشاهدات الأصلية غير الساكنة و تم استخدام التحويلة اللوغاريتمية و الفروق الأولى لتحويلها إلى سلسلة ساكنة.و عند استخدام المعايير الإحصائية لاختيار النموذج المناسب مثل اختبار عشوائية البواقي و تطبيق معايير (Akaike) و (Schwarze),توصلت الدراسة إلى أن أفضل نموذج يلائم بيانات مؤشر أسعار الأسهم هو نموذج الانحدار الذاتي المتكامل من الدرجة الأولى ARIMA(1,1,0) دون وجود تأثيرات موسمية. و بالاعتماد

على هذا النموذج تم التنبؤ بقيم مؤشر أسعار الأسهم للفترة من يوليو 2011 إلى يونيو 2012 و قد كانت القيم التنبؤية متناسقة مع القيم الأصلية للسلسة (سليمان،2013).

د. دراسة عبدالرزاق الشحادة (2012)

هدفت هذه الدراسة الى بيان مستوى الاهتمام بتطبيق وبالمجالات المكونة لبيئة مفاهيم وادوات التنقيب عن البيانات لأدارة العمليات المصرفية والمحاسبية في البنوك التجارية الاردنية. ولتحقيق تلك الاهداف استخدام الباحثان المنهج الوصفي التحليلي القائم على الاستبيان الموزع على افراد مجتمع الدراسة. وقد توصل الباحثان الى أن نسبة الاهتمام لدى أفراد مجتمع الدراسة عن تطبيق مفاهيم التنقيب عن البيانات لأدارة العمليات المحاسبية والمصرفية، كانت مرتفعة بشكل عام. وأن ترتيب المجالات التي قد ينظر إليها عند البحث في تطبيق مفاهيم التنقيب عن البيانات و قد أوصى الباحثان بضرورة البحث في اسباب انخفاض الاهتمام في بعض المجالات التي من شأنها التأثير بتطبيق مفاهيم التنقيب عن البيانات في أداء العمليات المصرفية والمحاسبية. وضرورة ايلاء التطوير المستمر واستثمار التكنولوجيا العناية اللازمة، وبما يساعد في تحقيق الاهداف المتوخاة منها لرفع كفاءة وفعالية العمليات المحاسبية والمصرفية (الشحادة ،2012).

هـ. دراسة : محمد علي محجوب (2017)

تعتبر نظم دعم القرارات من النظم الهامة التي تدعم قرار الإدارة العليا بصورة علمية وذلك لأنها تدمج مابين البيانات والنماذج والبرمجيات في نظام فعال لإتخاذ القرارات ، وبعد التطور الذي حدث مؤخرا في قطاع المعلومات والاتصالات أصبحت المعلومات متوفرة مما أدي الي ظهور برمجيات يسهل تطبيقها في معادلات إحصائية ليدعم عمل المؤسسات لذا ، جاءت هذه الدراسة لتدعم هذا التوجه من أجل التنبؤ بالايرادات المالية باستخدام سلاسل ماركوف الزمنية وتتناول هذه الدراسة المفاهيم الأساسية لجمع البيانات وإستخلاصها وتنفيذ المعادلات بنظام الاوراكل، ان الهدف من هذه الدراسة هو إيجاد نظام الي يقوم بالتنبؤ مما يؤدي الي زيادة كفاءة نظم دعم القرار في التنبؤ بالايرادات المستقبمية لوزارة المالية بولاية البحر الأحمر باستخدام سلاسل ماركوف الزمنية (محجوب،2017)

9. نشأة التنقيب في البيانات

إن التطور فى العلم والإقتصاد وتكنولوجيا المعلومات أدى إلى زيادة كمية البيانات الرقمية ، هذه الكميات الهائلة من البيانات لم تعد وسائل التحليل التقليدية قادرة على التعامل معها فظهر التنقيب عن البيانات منذ أواخر الثمانينات وأثبت وجوده كأحد الحلول الناجحة لتحليل كميات ضخمة من البيانات ،

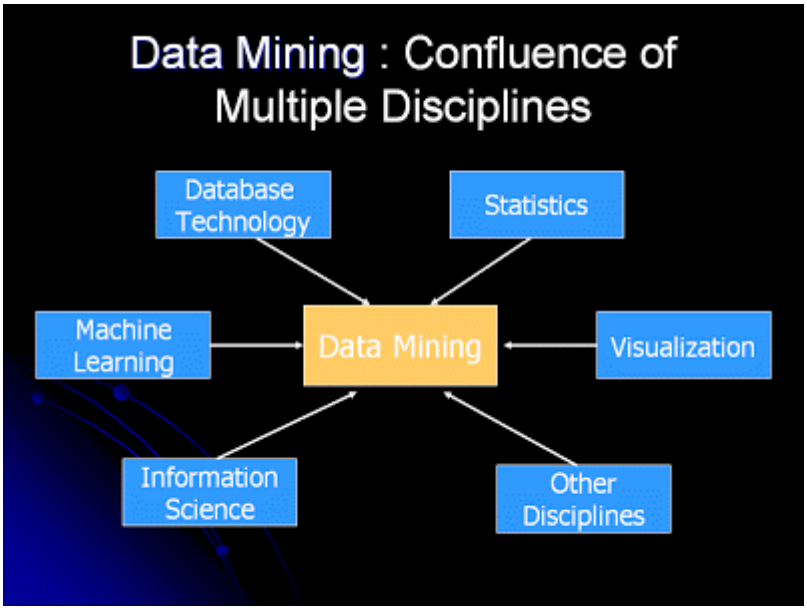
وذلك بتحويلها من مجرد معلومة متراكمة وغير مفهومة إلى معلومات قيمة يمكن إستغلالها والإستفادة منها بعد ذلك (معرفة) حيث ظهر مصطلح التنقيب فى البيانات في منتصف التسعينيات فى الولايات المتحدة الامريكية. وهو ما يجمع ما بين الإحصاء وتكنولوجيا الإعلام (قواعد البيانات ، الذكاء الإصطناعي ، التعلم الآلي machine learning).

فقد بدأ الإهتمام بالتنقيب عن البيانات عام 1989م اثناء إنعقاد ورشة عمل حول إكتشاف المعرفة في قواعد البيانات ، ومن ذلك الحين تم عقد هذه الورشة بصفة مستمرة سنويا فى 1994م ، اما في عام 1995م أصبح المؤتمر الدولي لإكتشاف المعرفة والتنقيب عن البيانات من أهم الأحداث السنوية (احسن،عمار،2018).

ومع وجود كميات كبيرة من البيانات المخزنة في قواعد البيانات ومخازن البيانات إزدادت الحاجة إلى تطوير ادوات تمتاز بالقوة لتحليل البيانات وإستخراج المعلومات والمعارف منها ، من هنا ظهر ما يسمى بالتنقيب في البيانات كتقنية تهدف إلى إستخراج المعرفة من كميات هائلة من البيانات. وهي تقنية حديثة فرضت نفسها بقوة في عصر المعلوماتية ، وإستخدامها يوفر للشركات والمنظمات في جميع المجالات القدرة على إستكشاف والتركيز على أهم المعلومات في قواعد البيانات ، كما تركز تقنيات التنقيب على بناء التنبؤات المستقبلية وإستكشاف السلوك والإتجاهات مما يسح بإتخاذ القرارات الصحيحة وإتخاذها في الوقت المناسب. والتى تعتبر بدورها مرحلة من مراحل عملية أكثر تعقيداً هى إستكشاف المعرفة في قواعد البيانات ، والمرتبطة إلى حد بعيد بعملية تطوير أخرى مهمة جداً هى مستودعات البيانات. حيث أن الكثير من الشركات والمنظمات الرائدة اليوم تستخدم عملية إستكشاف المعرفة في قواعد البيانات بشكل منجهي ومنظم بوصفها تشكل جوهر العمل الذي يعتمد عليه في تفعيل النشاط وتحقيق الميزة التنافسية(الحمامي،2008).

10. استعمالات تنقيب البيانات

بخلاف الطرق التقليدية للتحليل الاحصائي تستطيع هذه الاداة معالجة كميات كبيرة من البيانات حيث يتم اعتمادها بشكل رئيسي من قبل الشركات التي تركز على المستهلك في قطاعات التسويق ، التمويل ، الاتصالات ، أو تنقيب عن البيانات تستخدم تقنية تعدين البيانات في المجالات مختلفة ، كالرياضيات ، علم التحكم الالي ، وعلم الوراثة .



الشكل(1) يوضح تخصصات تعدين البيانات (الدوري،الحسين،2007).

11.أهداف التنقيب في البيانات

يشير Van der Aast إلى أن الهدف من عملية التنقيب هو إستخراج المعلومات في شكل نماذج عملية من السجلات ، حيث يتم إستخدام أى أحداث تم تسجيلها من نظم المعلومات لإستخراج معلومات حول الأنشطة والعلاقات المتداخلة بينها. أما (Ozgulbas and Koyuncugil) (صباح،2017) فيلخص هدف التنقيب في البيانات في خلق نماذج لصنع القرار التى تتنبأ بالسلوك في المستقبل إسناداً إلى تحليلات النشاط السابقة ، فهي عملية غربلة شاملة للبيانات التنظيمية (الداخلية والخارجية) لتحديد أنماط حاسمة لدعم إتخاذ القرار كما يؤكد جمعة (جمعة،2017) على أن إستخراج المعرفة من الكميات الكبيرة للبيانات المتوفرة بإستخدام الخوارزميات الإحصائية هو أحد الأهداف الأساسية للتنقيب في البيانات.

ويستعرض (Tiwari) (tiwari,2012) الهدف من عملية التنقيب بإستخراج والحصول على معلومات حول العمليات من سجلات المعاملة من أجل إكتشاف والتحكم في نموذج العملية. ومن أهم أهداف التنقيب في البيانات ما يلي :

1. من اجل تعليل بعض الظواهر المرئية.
2. من اجل التثبت من نظريه ما. مثال علي ذلك ان الاسر الكبيرة تهتم بالرعاية الصحية اكثر من الاسر الصغير.
3. من اجل تحليل البيانات للوصول والحصول علي علاقات جديده وغير متوقعه.

4. من خلالها يتم فهم مدي ترابط حقول هده البيانات وعلاقتها في بعضها البعض .وبالتالي فان استكشاف الانماط patterns او الشذوذ outliersوغيرها تصبح عملية سهله جداً.
5. إختزال كمية البيانات الكبيرة أو ضغطها بحيث تعبر بشكل بسيط عن كامل البيانات بدون تعميم ، من ثم إكتشاف المعلومات المخفية في تلك البيانات.

12. برنامج Weka

مفتوح المصدر يدمج عدد كبير من التقنيات، متكامل وربما يكون الأفضل بين البرامج مفتوحة المصدر، لكن يتطلب التدريب وقراءة متأنية للمستندات المرفقة. يقترح 4 طرق لمعالجة البيانات :واجهة بسطر الأوامر، ومستكشف لمعالجة البيانات، وتطبيق لتقنيات التنقيب في البيانات المتعددة على قواعد البيانات المتعددة، وبناءات رسومية لإجراءات معالجة البيانات. يوفر بناء المتغيرات، التعليم المراقب وغير المراقب، قواعد الارتباط.

ومن أهم مميزات هذا البرنامد ما يلي :

1. مفتوح المصدر.
2. يحتوي البرنامج علي مجموعة من ادوات تجهيز البيانات وتقنيات النمذجه.
3. يتميز البرنامج بواجهة مستخدم رسومية سهلة الاستخدام.
4. يدعم البرنامج عدة انواع مختلفة من تنسيقات الملفات مثل نوع (ARFF، CSV، C4.).
5. يمكنه التعامل مع برامج اخري مثل(MS.EXCEL2016) (Tiwari,2012)

13. التصنيف بإستخدام أشجار القرار

هي نموذج استكشافي يظهر علي شكل شجرة ويمثل كل فرع من فروعها سؤالا منطقيا وتمثل اوراقها أجزاء من قاعدة البيانات وتنتهي الي التصنيفات التي تم بناؤها ، وهي من إحدي الخوارزميات التي لا يتم فيها فقد اي بيانات من قاعدة البيانات لان البيانات التالية هي نفسها البيانات السابقة ولكن تم تقسيمها الي جزئيين علي حسب سؤال منطقي.

الجزء الاول من البيانات ينطبق عليهم السؤال والجزء الثاني لا ينطبق . وهكذا يتم من خلال سلسلة من الأسئلة بناء شجرة القرار بفروعها المتسلسلة ، لذلك نجد أن القاعدة الأساسية التي تحكم تنظيم البيانات والوصول إلي نتائج دقيقة هي السؤال ، مما يعني أنه يجب تحديد السؤال ، يجب وضع سؤال لكل فرع

بحيث تنقسم البيانات الي (ورقتين – جزئيين مقسمين) ، ولكن السؤال الذي فرض نفسة هو : متي يتم الوقوف عن وضع الأسئلة ؟ والجواب هو:

1- إذا كانت الشريحة تحتوي علي سجل واحد وهذا يعني أنه لايمكن تقسيمه.
2- إذا كانت الشريحة تحتوي علي مجموعه من السجلات ولكن هذه السجلات متطابقة.
3- إذا كان تقسيم البيانات لايؤدي الي فائدة او معلومة جديدة(محجوب،2017).

أ. من مميزات هذه الخوارزمية

أنها لا تحتاج أي خبير لأنها تعطيك عدد كبير جدا من الاستعلامات ، الهدف الاساسي من تقنية شجرة القرار هو تقسيم قاعدة البيانات علي حسب هدف معين سبق أن تم تعينة .

ب. استخدام شجرة القرار التنبؤ:

بالرغم من أن شجرة القرار تستخدم لإكتشاف وتحضير البيانات للعمليات الإحصائية إلا أنها أيضا تستخدم وبشكل أكثر للتنبؤ . ومن المهم جدا عند بناء خوارزمية شجرة القرار ان يؤخذ بعين الاعتبار ان تكون قابلة للتطبيق بقدر الإمكان وبشكل مثالي علي كل البيانات المتوفرة.

القاعدة الاساسية لبناء شجرة القرار هي إيجاد افضل سؤال عند كل فرع من فروع الشجرة بحيث يقسم هذ السؤال البيانات إلي قسمين ، القسم الأول منها ينطبق عليهم السؤال والقسم الثاني لا ينطبق ، وهكذا يتم من خلال سلسلة من الأسئلة بناء شرة القرار بفروعها المتسلسلة(Jiang,2011).

ج. محاسن التصنيف المعتمد علي أشجار القرار:

1- غير مكلفة وسهلة البناء.
2- دقة اكثر مقارنة مع باقي الخوارزميات.
3- السرعة تصنيف السجلات غير المعروفة.
4- سهلة القراءة.
5- قراءة القواعد بصورة سهلة جدا.

14. خوارزميه J48

تنتمي خوارزمية J48 لفرع classification وتعتبر امتداداً لخوارزمية ID3 إلا أنها تعالج النقاط التالية التي لا تعالجها خوارزمية ID3 وتستخدم لتوليد :Decision Trees (محجوب،2017)

1- تحديد كم سيكون عمق نمو شجرة القرارات.

2- معالجة القيم المستمرة مثل درجة الحرارة.

3- معالجة بيانات التدريب ذات الخصائص مفقودة القيم.

4- تحسين الفعالية الحسابية.

أ. طريقة عمل خوارزمية j48

1. تختار الخاصية ذات أعلي information gain والذي يحسب بالقانون:

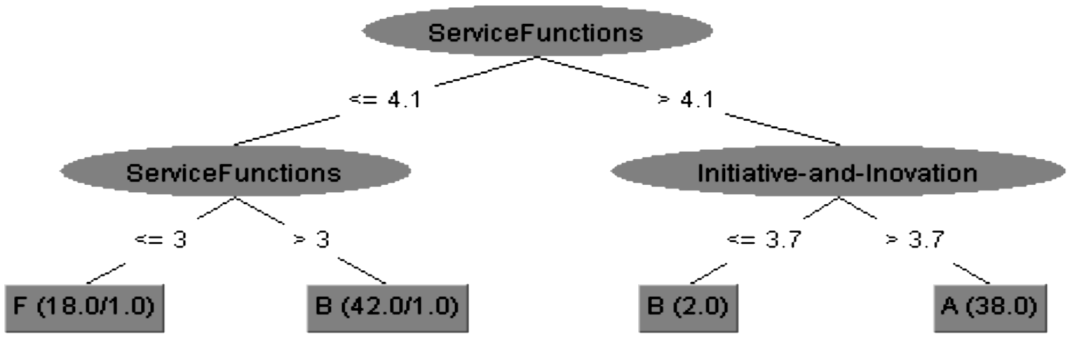
$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

2. بعد اختيار الخاصية ذات اعلي information gain ولتكن (A) نحسب لها:

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times I(D_j)$$

وهذا يعني ان البيانات بعد اختيار الخاصية (A) لتصنيف (D) انقسمت الي عدة اقسام أو تصنيفات تم بعد ذلك نقوم بحساب gain (A) والذي يحسب بالقانون:

$$Gain(A) = Info(D) - Info_A(D)$$



الشكل (2) شجرة القرارات الناتجة عن التصنيف J48 [5]

15. المحاصيل البستانية السودانية

يمتاز السودان بتباين بيئاته المناخية مما يساعد على انتاج انواع مختلفة من المنتجات البستانية فى اوقات مختلفة كما يساهم هذا التباين فى اتساع مواسم الانتاج لكثير من المحاصيل البستانية مثل المانجو الذى يبدأ انتاجه من ديسمبر وحتى اغسطس ، وعلى الرغم من ان المنتجات البستانية تشغل حوال 8.1 % فقط من المساحة الكليه المستثمره على مستوى السودان الا انها تساهم بنسبة 81 % من الناتج الاجمالى المحلى.

يعتبر السودان مثالياً لانتاج انواع عديدة من الخضر والفاكهه خاصة الخضر الشتوية والتى لها اسواق اوربيه رائجة تقل فيها المنافسة خاصة فى الفترة من يناير مارس والتى– تمثل فترة الشتاء الاوربى وذروة الانتاج البستاني للمحاصيل الشتوية مثل الشمام(القاليا) الفاصوليا الخضراء الشطة الشطة الخضراء الفلفلية.

عرف السودان زراعة الخضر لا سيما في المناطق المروية والمناطق عالية الأمطار، منذ زمن بعيد وأكتسبت أهمية أكبر في الآونة الأخيرة بسبب تزايد الوعي بقيمتها الغذائية والإقتصادية وبدأت تشكل جزءاً هاماً من الصادر ، تتركز زراعة الخضر في القطاع المروي، وتحتل أخصب الأراضي، ويأتي معظم الإنتاج من الحيازات الخاصة الصغيرة المنتشرة على ضفاف الأنهار والوديان وتروى رياً صناعياً. قدرت مساحة الخضر الرئيسية في السودان في عام 1999 بنحو 410 الف فدان وقدر مجمل إنتاجها بنحو 2.4 مليون طن(sudaneconomy,2021).

16. تجميع البيانات (Data Clustering)

مجموعة البيانات في هذا البحث مأخوذه من سجلات الصادرات البستانية من الهيئة العامة للجمارك (قسم الصادرات) لسنة 2016–2020 والتي بلغ عددها 903 سجل.

17. تهيئة البيانات(والمعالجة)

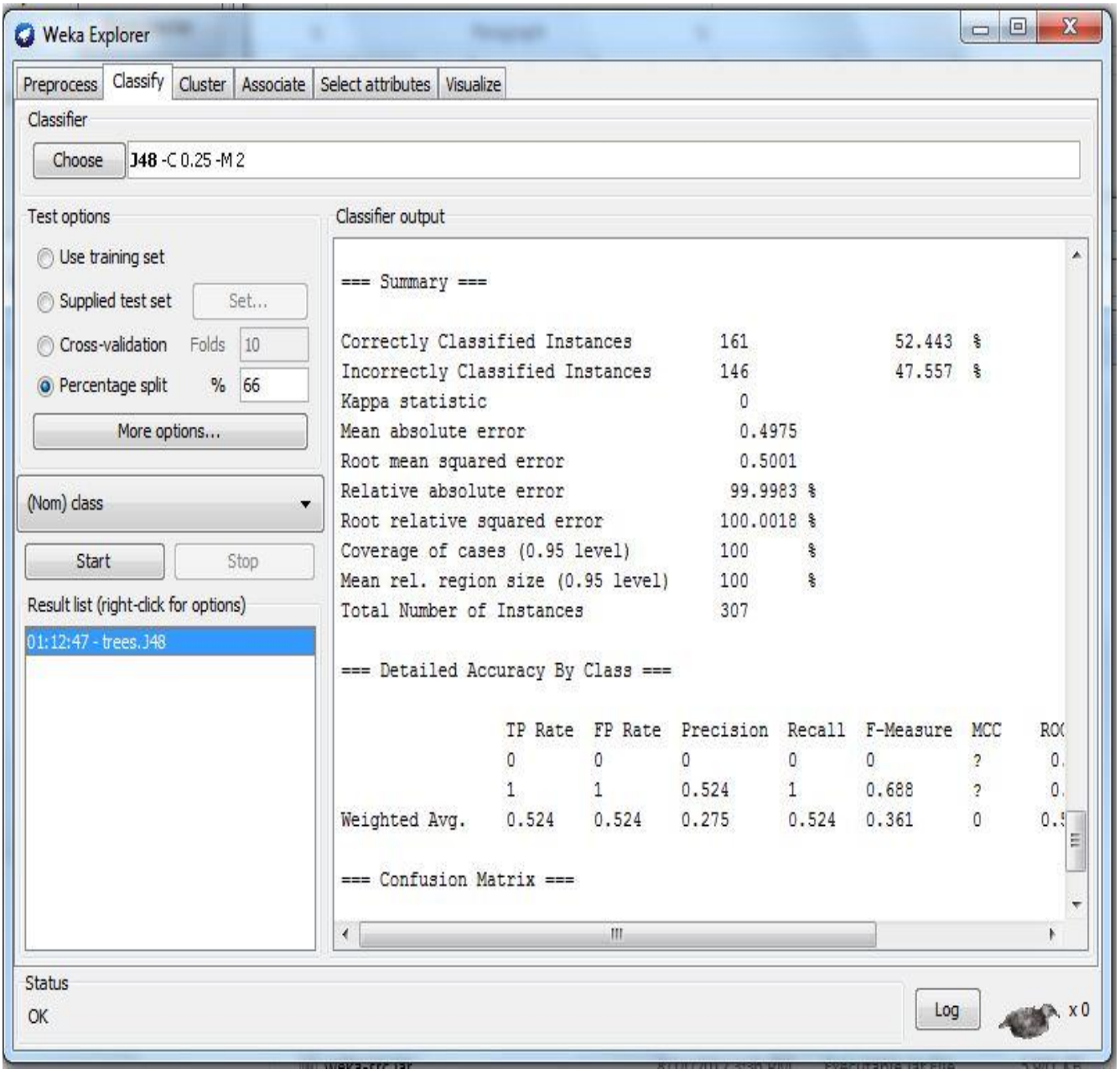
نحتاج لتهيئة البيانات لعده أسباب أهمها :

1. البيانات تكون عادة غيرنظيفة ونعني بها: غيرمكتملة(Incomplete)بمعني يمكن أن تكون بعض السجلات فارغة أوناقصة في الحقول المهمة،أو وجود شوائب(Noise) ونعني بالشوائب الاخطاء أو القيم الشاذة،أو غير متناسقة (Inconsistent) ونعني بها وجود تناقضات في بعض مسميات الحقول أو حتي في القيم .
2. عدم وجود البيانات بجودة عالية يعطي نتائج تنقيب غير جيدة .

تقنيات تهيئة البيانات تقوم بتحسين جودة البيانات وهذه تؤدي إلى نتائج جيدة في عملية التنقيب من حيث الكفاءة والدقة.

18. تنفيذ خوارزمية شجرة القرار

ويتم ذلك من خلال إختيار التصنيف (Class) وحجم البيانات المستخدمة (66%) والتي تمثل ثلث البيانات الخاصة بالتدريب ومن ثم الضغط على زر ابداء (Start) ليتم تنفيذ الخوارزمية كما يلي:



الشكل (3) يوضح تنفيذ خوارزمية شجرة القرار (J48)

حددت نسبة 66% =training set تم تدريب المودل على البيانات بنسبة 66% و ال34 =test set.

تم اختيار tree j48 و تم اختيار class الذي سيبني علي التعدين هو class او الفجوة حيث انه هو اكثر الحقول التي توضح هدف الدراسة والتي من أجلها تم بناء الفرضيات.

19. نتيجة شجرة القرار

Classifier output

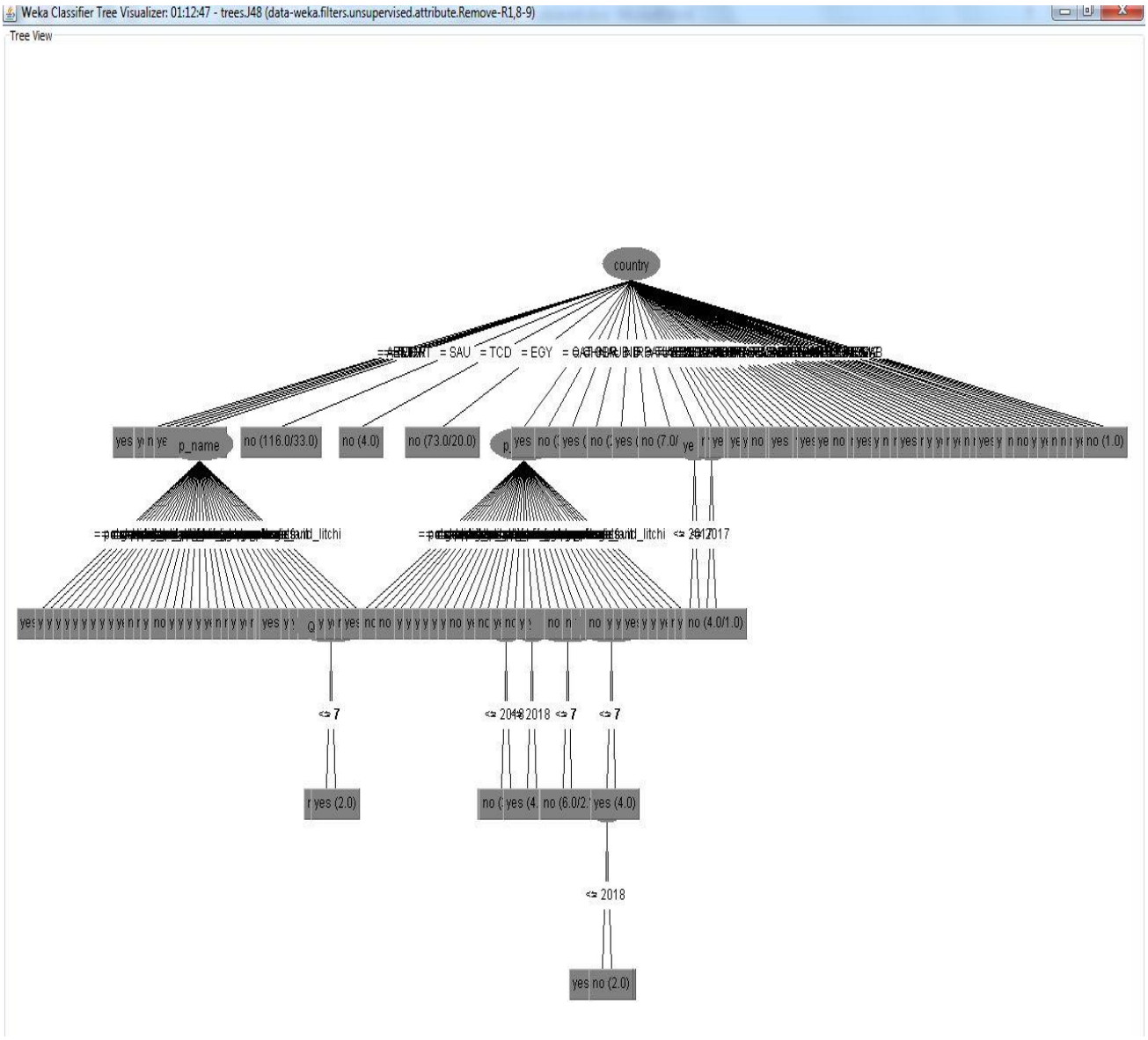
```
=== Classifier model (full training set) ===

J48 pruned tree
------------------

country = ARE: yes (111.0/45.0)
country = ERI: no (6.0/2.0)
country = TUR: yes (38.0/17.0)
country = ETH: no (26.0/10.0)
country = JOR: yes (22.0/7.0)
country = KWT
|   p_name = potatoes: no (2.0)
|   p_name = onions: yes (9.0/2.0)
|   p_name = garlic: yes (0.0)
|   p_name = cucumbers: yes (0.0)
|   p_name = peas: yes (0.0)
|   p_name = beans: yes (0.0)
|   p_name = asparagus: yes (0.0)
|   p_name = sweet_corn: yes (0.0)
|   p_name = olives: yes (0.0)
|   p_name = chickpeas: yes (1.0)
|   p_name = lentils: yes (0.0)
|   p_name = pigeon_peas: yes (0.0)
|   p_name = seed: no (2.0)
|   p_name = tomatoes: no (1.0)
|   p_name = carrots_and_turnips: yes (0.0)
|   p_name = aubergines: yes (0.0)
|   p_name = nuts: no (3.0/1.0)
|   p_name = spinach: yes (0.0)
|   p_name = pumpkins: yes (1.0)
|   p_name = prunes: yes (0.0)
|   p_name = durians: yes (0.0)
|   p_name = brazil_nuts: yes (0.0)
|   p_name = banana: no (1.0)
```

شكل (4) نتيجة شجرة القرار

20. تمثيل شجرة القرار



شكل (5) شجرة القرار

من الشكل اعلاه: كان التصنيف حسب الدول وعلاقتها بالصادرات البستانية حيث كانت توجد فجوة في الصادرات بشكل لعام لدولة الإمارات وتركيا والأردن وجنوب السودان و ألمانيا و فرنسا والترويج والصين وأستراليا و العراق وأيطاليا و السويد ، سوريا ، روسيا ، هنجاريا ، أفغانستان ، جيبوتي ، فنزويلا ، أما تونس وكينيا فقد ظهرت الفجوة بعد عام 2017 م أما الكويت وقطر فقد ظهرت الفجوة في صادرات بعض المنتجات البستانية مثل البصل والبطاطا والموز والبلح والبقوليات.

21. النتائج

1. استخدام تقنيات تنقيب البيانات يساعد في معرفة الفجوات الإقتصادية في تصدير المنتجات البستانية للدول والقارات.
2. كانت قطر والأمارات والسعودية ومصر من أكثر البلدان التي يتم تصدير المنتجات البستانية لها ، حيث كانت نسبة التصدير بالتوالي (127-111-116-74)
3. كانت نسبة التصدير في عام 2018 هي الأعلى وبدأت تقل بشكل ملحوظ نسبة للعوامل الإقتصادية التي تمر بها السودان ، حيث بلغت (213) وفي 2019 كانت (152)
4. يعتبر شهر ديسمبر ويناير من أكثر الشهور التي يتم فيها تصدير المنتجات وتنخفض هذه النسبة في منتصف السنة (مايو – يوليو) حيث كانت في ديسمبر ويناير (91-97) ومايو ويوليو (78-58)
5. من حيث القارات كانت قارة اسياء هي الأكثر إستيراداً للمنتجات البستانية تليها قارة افريقيا حيث بلغت نسبة الصادرات (611)

22. التوصيات

1. زيادة بيانات التدريب المستخدمة في الدراسة من خلال إعتماد البيانات الخاصة بالصادرات البستانية من وزارة الإستثمار والزراعة وإضافة مزيد من الخصائص مثل الجهة المصدرة ونوعها (خاصة – حكومية) والولاية والمحلية التي تم التصدير منها .
2. تصميم مستودع بيانات متكامل يساهم في جمع البيانات الخاصة بالصادرات البستانية وتوفير كل المعلومات والبيانات التي يحتاجها المحللون لمساعدتهم في تنقيب البيانات وإكتشاف المعرفة.
3. إمكانية إستخدام خوارزميات اخرى مثل التجميع Cluster أو الجار القريب ومقارنة النتائج بنتائج هذا البحث .
4. الإهتمام بعملية تخزين بيانات الصادرات البستانية بصورة جيدة لتكون فى متناول اليد مما يساعد فى إكمال مسيرة الابحاث فى هذا المجال.
5. تطبيق خوارزمية الترابط Association لتحديد العلاقات بين صدارات المنتجات والفجوات والشهور التي يكون فيها تزايد في نسبة الصادرات .

قائمة المصادر والمراجع

[1] أحمد فايز سيد ، أدوات التنقيب عن البيانات مفتوحة المصدر ، (دراسة تحليلية تقييمية) مجلة الآداب والعلوم الإنسانية ، جامعة طبة ، 2016، العدد 10

[2] ياسر عوض محمد علي : تصميم نظام تنبؤ بإستخدام تقنيات تنقيب البيانات حالة تطبيقية التنبؤ بالسلع الصادرة عبر النظام الإلكتروني للجمارك السودانية ، بحث مقدم لنيل درجة الدكتوراة في تقانة المعلومات ، جامعة النيلين ، 2019م

[3] علي درب كسار الحيالي ، إستخدام معايير الدقة التنبؤية في تحديد الطريقة المثلى في تقدير القيم المفقودة في المجال الزراعي ، مجلة العلوم الزراعية العراقية ، جامعة بغداد ، العدد 4 ، 2013م

[4] عبد الله سليمان محمود. 2013. تحليل السلسلة الزمنية لمؤشر أسعار الأسهم لشوق الخرطوم للأوراق المالية : استخدام منهجية (بوكس–جنكنز).مجلة الاقتصاد والعلوم السياسية والإحصائية،مج. 2013م

[5] محمد علي محجوب النيل ، زيادة كفاءة نظم دعم القرار في التنبؤ بالايرادات المستقبلية باستخدام السلاسل الزمنية، مجلة الدراسات العليا – جامعة النيلين – مجلد رقم (8) العدد (31) ، 2017م

[6] محمد علي محجوب النيل ، زيادة كفاءة نظم دعم القرار في التنبؤ بالايرادات المستقبلية باستخدام السلاسل الزمنية، مجلة الدراسات العليا – جامعة النيلين – مجلد رقم (8) العدد (31) ، 2017م

[7] طيار أحسن ، شلابي عمار ، التنقيب عن البيانات واتخاذ القرارات ، كلية العلوم الإقتصادية وعلوم التسيير ، جامع 20 اوت 1955 سكيكدة ، الجزائر ، 2018م.

[8] علاء حسين الحمامي ، تنقيب البيانات ، دار اثراء للنشر والتوزيع ، عمان ، الأردن ، 2008

[9]زكريا الدوري ، داليا الحسين ، دور تنقيب البيانات في زيادة أداء المنظمة "، دراسة تحليلية في المصرف الزراعي المجلد (13) جامعة بغداد ، 2007م .

[10] محمود مليك زروقي صباح ، "تنقيب البيانات كأداة مساعدة في إتخاذ القرارات "، جامعة المسيلة ، 2017م .

[11] عماد أحمد جمعة ، "تطبيق أساليب التنقيب عن البيانات لقياس مستوى تحصيل الطلاب"، رسالة ماجستير في تقانة المعلومات جامعة النيلين ، 2017م .

[12] Mahendra Tiwari, performance analysis of data mining algorithms in weka, university of Allahabad, India,2012

[13] Eltabakh.M (2010) . OLAP & Data mining. Available at : http://web.cs.wpi.edu/cs561/s12/lecturses/integrationOLAP/OLAPandMining.pdf.

[14] L Jiang, C. Li, “Scaling up the Accuracy of Decision -tree” Wuhan;China University of Geosciences, 2011.

[15] مقالة عن المحاصيل البستانية السودانية على الرابط

http://www.sudaneconomy.net/sects/agr/mhasil.htm