مجلة جامعة طرطوس للبحوث والدراسات العلمية _ سلسلة العلوم الهندسية المجلد (6) العدد(5) 2022
Tartous University Journal for Research and Scientific Studies - engineering Sciences Series Vol. (6) No. (5) 2022

تحليل أداء خوارزميات التنقيب في البيانات في تصنيف مستوى المدارس

سالي محمد عيسى *

(تاريخ الإيداع 2/22 /2022 . قُبِل للنشر في 2022/6/14)

□ ملخّص□

إن التزايد الكبير والنمو الهائل في حجم البيانات وخاصة في ظل ثورة تكنولوجيا المعلومات جعل عملية استخلاص المعرفة أكثر صعوبة وتعقيد مع استخدام الأساليب والطرق التقليدية في التحليل والبحث، فأصبح هنالك حاجة ملحة للاعتماد على تقنيات ذكية وأدوات جديدة لتحليل مستودعات البيانات الضخمة وتحويل هذه البيانات إلى معلومات وأنماط معرفة مفيدة في التفسير واتخاذ القرارات الصحيحة ومنها تقنيات وخوارزميات التنقيب في البيانات Data Mining(DM) التي كان لها دور وأهمية كبيرة في صناعة المعلومات والاستخدام الأمثل للبيانات.

يسلط هذا البحث الضوء على خوارزميات التصنيف ودورها في تصنيف البيانات التربوية والتنبؤ بقيم الواصفات المجهولة وبالأخص خوارزميات Naïve Bayes وشجرة القرار J48 والغابة العشوائية Random Forest من خلال تقييم أداء هذه الخوارزميات بالاعتماد على مجموعة من البارامترات ومقارنة دقة مصنفاتها في التنبؤ بمستوى المدارس بمراحلها الثلاثة (الأولى – الثانية – الثالثة) وتصنيفها إلى مستوى جيد أو سيئ في محاولة لمراقبة أداء هذه المؤسسات التعليمية والعمل على تطويره لما لذلك من دور أساسي وفعال في تقدم المجتمع والارتقاء به إلى أعلى المستويات الثقافية والعلمية.

أظهرت النتائج النهائية لهذه الدراسة والتي تم تطبيقها على مجموعة بيانات تحوي 396 مدرسة لكل منها 21 سمة باستخدام برنامجWEKA أداء مميز للخوارزميات الثلاث من حيث الدقة العالية والتي حققت نسب أعلى من 92% لكل منها ومعدل الخطأ المنخفض جداً، مع تفوق لخوارزمية Naïve Bayes على خوارزميتيّ J48 و Random Forest في دقة المصنف التي وصلت إلى 94.94% والسرعة في التصنيف.

الكلمات المفتاحية: التنقيب في البيانات، خوارزميات التصنيف، التنبؤ، خوارزمية Naïve Bayes، خوارزمية شجرة القرار J48، خوارزمية Random Forest، Weka، الدقة، مستوى المدارس.

*مهندسة حاصلة على درجة الماجستير باختصاص هندسة تكنولوجيا المعلومات–قسم هندسة تكنولوجيا المعلومات –كليّة هندسة تكنولوجيا المعلومات والاتصالات–جامعة طرطوس–سوريا.

مجلة جامعة طرطوس للبحوث والدراسات العلمية _ سلسلة العلوم الهندسية المجلد (6) العدد(5) 2022
Tartous University Journal for Research and Scientific Studies - engineering Sciences Series Vol. (6) No. (5) 2022

Analyzing Performance of Data Mining Algorithms for school level Classification

Sally Mohammad Issa *

(Received 22/2/ 2022 . Accepted 14 /6/ 2022)

□ ABSTRACT □

The great increase and huge growth in the volume of data, especially in the revolution of technology, has made the process of knowledge extracting more difficult and complex, with the use of traditional methods and techniques in analysis and searches. There is an urgent need to rely on smart technologies and new tools to analyze huge data warehouses and transform this data into information and cognitive Patterns useful in interpretation and making the right decisions, including Data Mining techniques and algorithms that have had a great role and importance in the information industry and optimal use of data.

This research sheds light on classification algorithms and their role in classifying educational, data and predicting unknown descriptor values, especially the Naïve Bayes and the J48 decision tree and Random Forest algorithms. By evaluating the performance of these algorithms based on a set of parameters and comparing the accuracy of their classifiers in predicting the level of schools in its three stages (I. II - III) and categorizing them to a Good or Bad level. This is in in an attempt to monitor the performance of these educational institutions and work to develop it because of its essential and effective role in the advancement of society and its advancement to the highest cultural and scientific levels.

The results of this study, which was applied to a data set containing 396 schools, each with 21 traits, using the WEKA program, showed a great performance of the three algorithms in terms of high accuracy, which achieved rates higher than 92% for each of them and a very low error rate. With the superiority of the Naïve Bayes algorithm over the J48 and Random Forest algorithms. This is in classifier accuracy, which reached 94.94%, and classification speed.

KeyWords: Data Mining, Classification Algorithms, Prediction, Naïve Bayes algorithm, J48 Decision Tree algorithm, Random Forest algorithm, WEKA, Accuracy, School Level.

*An Engineer with a master's degree in Information Technology Engineering, Department of Information Technology Engineering, College of Information and communication Technology Engineering , Tartous University, Syria.

1. مقدمــة:

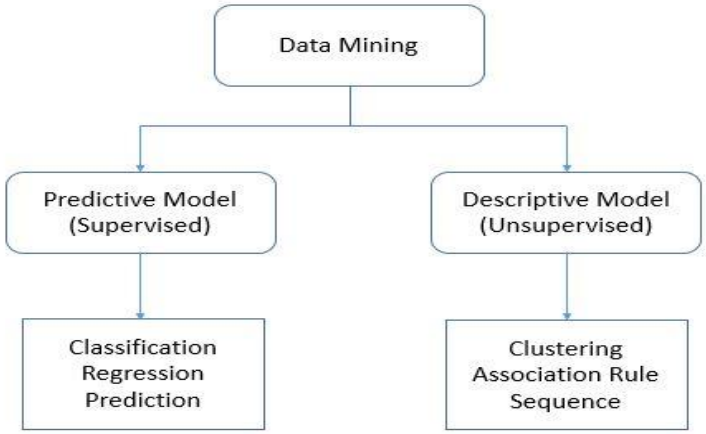
أدى تطور تقنيات المعلومات إلى توليد كميات كبير من مجموعات المعطيات والبيانات الضخمة في مختلف المجالات، والذي دفع بدوره إلى البحث في قواعد البيانات وتكنولوجيا المعلومات وظهور نُهج لتخزين هذه البيانات القيّمة ومعالجتها لاتخاذ العديد من القرارات الصحيحة والذكية بناءً عليها.

جذب التنقيب في البيانات الكثير من الاهتمام في الأوساط البحثية على مدار العقد الماضي في محاولة لتطوير خوارزميات ذكية قابلة للتوسع والتكيف مع كميات متزايدة من البيانات والبحث عن أنماط معرفية ذات معنى[1]، فهو يعتبر عملية شبه آلية تتمثل باستخدام ودمج التقنيات الإحصائية والرياضيّة بالإضافة لمفاهيم الذكاء الصناعي والتعلم الآلي لاستخراج وتحليل المعلومات المعرفيّة والمفيدة القابلة للاستنتاج والمخفيّة في مصادر البيانات المختلفة مثل أنظمة الملفات ومستودعات وقواعد البيانات.[2]

تهدف هذه التقنية إلى إيجاد واستخلاص أنماط ونماذج جديدة غير معروفة من قبل، واستخدامها في صنع قرارات تشكل ركيزة أساسية في عملية تطوير الأعمال بالإضافة إلى الكثير من الإحصائيات التي يمكن استخدامها لتلبية احتياجات عديدة.

يرتبط مصطلح Data Mining ارتباط وثيق بمفهوم اكتشاف المعرفة Knowledge Discovery (KD) والتي يشكل الأول جزء منها وخطوة فرعية من خطواتها للوصول إلى المعرفة النهائية ممثلة بالشكل الصحيح والمطلوب القابل للفهم والتطبيق.[1]

يدمج التنقيب في البيانات العديد من الخوارزميات المختلفة والتي تندرج تحت نوعين أساسيين من النماذج الموضحة في الشكل (1) وهي:



الشكل (1): نماذج التنقيب في البيانات

▪ **النموذج التنبؤي Predictive Model:** يعمل على التنبؤ المستقبلي بقيم السمات المجهولة وبآلية عمل البيانات بناء على نماذج مكتشفة من هذه البيانات أو من خلال نموذج تم تدريبه مسبقاً باستخدام بيانات سابقة، فهي تهدف لتحديد النتائج المستقبلية بدلاً من الاتجاهات الحالية، ويتضمن هذا النموذج أشهر تقنيات التنقيب في البيانات وهي: التصنيف Classification، العودية Regression، التنبؤ Prediction.[3]

▪ النموذج الوصفي Descriptive Model: يعمل هذا النموذج كطريقة لاستكشاف خصائص البيانات وصفاتها، ومن ثم تصنيفها إلى عدة فئات محددة مسبقاً تتقاطع معها في الصفات والخصائص، ويتضمن تقنيات التنقيب: العنقدة Clustering، قاعدة الارتباط Association Rule، التسلسل Sequence. [1]

2. هدف البحث:

يهدف هذا البحث إلى تقييم المدارس والتنبؤ بمستواها بناءً على قيم مجموعة من الواصفات باستخدام خوارزميات التصنيف Naïve Bayes وشجرة القرار J48 و Random Forest، مع دراسة أداء كل منها ومقارنة قيم دقة التصنيف Accuracy والحساسية Sensitivity والخصوصية Specificity التي تحققها بالإضافة إلى معدل الخطأ والزمن الذي تستغرقه في بناء نماذج التصنيف.

3. مواد وطرق البحث:

اعتمدنا في الدراسة العملية وتحليل النتائج المنبثقة عنها على برنامج WEKA 3.8وهو أداة مفتوحة المصدر ومكتوبة بلغة جافا تم تطويرها في جامعة وايكاتو Waikato في نيوزيلندا، تُستخدم على نطاق واسع في مجال التنقيب في البيانات.

يضم برنامج WEKA مجموعة من أدوات وخوارزميات التعلم الآلي المستخدمة لمعالجة مشكلات استخراج المعرفة في العالم الحقيقي والمساعدة في التنبؤ بقيم الواصفات المجهولة مع إظهار نتائج تطبيقها على مجموعات البيانات. [4]

4. آلية عمل خوارزميات التصنيف Classification:

- إنّ خوارزميات التصنيف هي شكل من أشكال تحليل البيانات والتي تستخلص نماذج تصف بشكل دقيق فئات وتصنيفات البيانات المهمة.
- يتم بناء نموذج التصنيف لتكون مهمته التنبؤ بالسمات المحددة التي تصف الفئة التي ينتمي لها العنصر المكتشف [5]، كما ويمكن لهذه الفئات والسمات المتنبئ بها أن تكون ذات قيم أسمية أو رقمية.
- التحليل باستخدام التصنيف هو عبارة عن عملية مكونة من خطوتين أساسيتين: [6]

1- الخطوة الأولى هي التعلم والتدريب Training and Learning: حيث يتم فيها بناء نموذج التصنيف بالاعتماد على مجموعة بيانات التدريب Training Dataset والتي تكون فيها السمة الهدف متاحة للخوارزمية ومرئية بالنسبة لها.

2- الخطوة الثانية هي التصنيف والاختبار Classification and Testing: حيث يتم فيها استخدام النموذج المبني في الخطوة السابقة واختبار صحته ودقته في التنبؤ بقيم الفئات أو سمات البيانات المحددة في مجموعة بيانات الاختبار.

- الهدف الرئيسي لخوارزميات التصنيف هو زيادة الدقة التنبؤيه التي حصل عليها نموذج التصنيف المبني عند تصنيف البيانات في مجموعة الاختبار والتي تكون غير مرئية أثناء التدريب .

ونستخدم في عملية التصنيف المدرجة ضمن هذا البحث خوارزميات التصنيف التالية:

4.1 خوارزمية شجرة القرار J48:

- شجرة القرار هي خوارزمية تعلم آلي خاضعة للإشراف supervised تُستخدم لحل مشاكل التصنيف المختلفة والعمل على التنبؤ بالفئة المستهدفة بالاعتماد على البيانات السابقة. [7]
- تعد أشجار القرار أدوات فعالة للغاية في العديد من المجالات مثل تحليل البيانات والنصوص واستخراج المعلومات والتعلم الآلي والتعرف على الأنماط من خلال تمثيل المعرفة بمخطط شجري Diagram Tree.

▪ تعمل خوارزمية الشجرة J48 على تحديد سمات البيانات المختلفة ومعالجتها بالاعتماد على تقسيم كل جانب من جوانب المعلومات إلى مجموعات فرعية ثانوية بناءً على قرار بالتالي تشكيل شجرة القرار المكونة من الأجزاء الرئيسية التالية: [8]

1. **عقدة الجذر root:** تصنف العقد الجذرية الحالات ضمن قاعدة البيانات بمختلف الميزات لدينا، ويمكن أن تحتوي العقد الجذرية على فرعين أو أكثر.
2. **العقد الداخلية enter nodes:** وتسمى أيضاً بـ "العقد غير الورقية" التي تشير بدورها إلى السمات وشروط الاختبار المطبقة عليها.
3. **العقد الطرفية(الأوراق) leaf:** والتي تمثل فئات التصنيف النهائية في أسفل الشجرة.

▪ تُحدد السمة عقدة جذر إذا كانت تحقق أعلى درجة ربح (IG) Information Gain بين جميع درجات الربح للسمات المتاحة والتي تعرف بـ التخفيض المتوقع في الانتروبيا Entropy الناتجة عن تقسيم البيانات وفقاً لسمة معينة وتُحسب وفق العلاقة(1): [9]

$$Gain(S,A)=Entropy(S) - \sum_{i=1}^{n} \frac{|Si|}{|S|} * Entropy(Si)$$ **العلاقة[1]**

حيث أنّ:

A: السمات

S: مجموعة حالات السمات.

N: عدد أجزاء السمة A.

|Si|: عدد الحالات في الجزء i

|S|: عدد الحالات في S

▪ ويمكن حساب درجة الانتروبيا Entropy والتي تمثل مقياس لدرجة الفوضى أو عشوائية مجموعة البيانات من العلاقة (2):

$$Entropy\ (A) = \sum_{i=1}^{n} -Pi\ log_2\ Pi$$ **العلاقة[2]**

حيث أنّ:

N: عدد أجزاء السمة S.

Pi: الاحتمال المتكرر للحالة i في مجموعة البيانات.

▪ تختار شجرة القرار العقدة في كل مرحلة من خلال تقييم أعلى نسبة ربح للمعلومات Information Gainبين جميع السمات حيث يرتبط الجذر والعقد الداخلية بالسمات، بينما ترتبط العقد الطرفية بالفئات.

▪ وبشكل أساسي، تحتوي كل عقدة غير ورقية على فرع صادر لكل قيمة محتملة للسمة المرتبطة بها، وبالتالي لتحديد فئة سجل جديد باستخدام شجرة القرار، بدءًا من الجذر، تتم زيارة العقد الداخلية المتعاقبة حتى يتم الوصول إلى عقدة طرفية.

▪ يتم إجراء اختبار في عقدة الجذر وفي كل عقدة داخلية، ومن ثم يتم تحديد نتيجة الاختبار للفرع الذي تم اجتيازه، والعقدة التالية التي تمت زيارتها، عندئذ تكون فئة المثيل هي فئة العقدة الطرفية النهائية في الشجرة (الورقة).

- من خصائص هذه الخوارزمية قدرتها على: [8]

1. التعامل مع مجموعة متنوعة من بيانات الإدخال: الاسمية والرقمية.
2. معالجة مجموعات البيانات الخاطئة أو القيم المفقودة.
3. سهلة الفهم من قبل المستخدم النهائي غير الاختصاصي.

4.2 خوارزمية Naïve Bayes:

- هي تقنية تصنيف تعتمد على نظرية الاحتمالات للعالم توماس بايز Thomas Bayes مع افتراض الاستقلالية بين المتنبئين، حيث يفترض مصنف Naive Bayes أنّ وجود سمة معينة في فئة ما لا علاقة لها بوجود أي سمة أخرى. [10]
- تعتبر نظرية Bayes أساس الخوارزمية حيث تقدم طريقة لحساب الاحتمال الشرطي، أي احتمال وقوع حدث بناءً على المعرفة السابقة المتوفرة عن الأحداث الأخرى وفق المعادلة التالية [11]:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

العلاقة[3]

حيث أنّ:

P (A | B): احتمال وقوع الحدث A بالنسبة للحدث B الذي وقع (الاحتمال الشرطي).

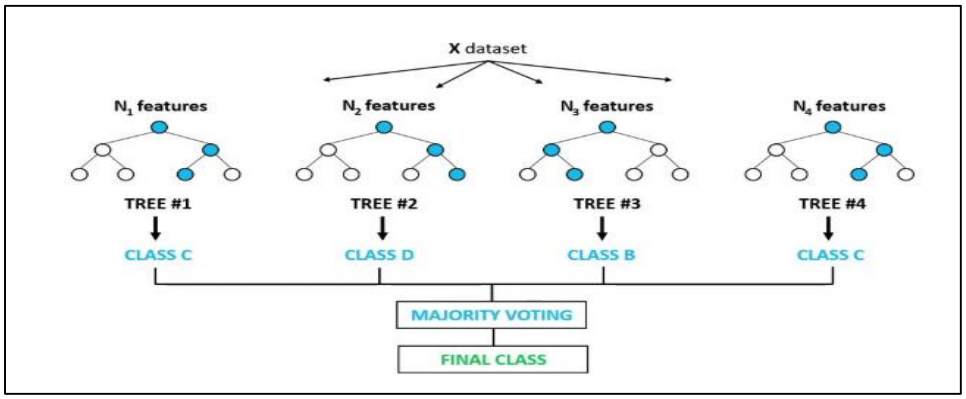
P (A), P (B): احتمالات وقوع الحدث A و B على التوالي.

P (B | A): احتمال وقوع الحدث B بالنسبة للحدث A الذي وقع.

- يعتبر نموذج Naive Bayes سهل البناء ومفيد بشكل خاص لمجموعات البيانات الكبيرة جدًا، إلى جانب البساطة فمن السهل بناء النماذج واجراء التنبؤات باستخدام هذه الخوارزمية، بالإضافة لتفوقها في الأداء على أساليب التصنيف المعقدة للغاية.
- خوارزمية Naïve Bayes هي الخوارزمية التي تتعلم احتمالية وجود عنصر بميزات معينة تنتمي إلى مجموعة أو فئة معينة وبالتالي تندرج تحت مسمى "مصنف احتمالي". [12]

4.3 خوارزمية الغابة العشوائية Random Forest:

- هي خوارزمية تصنيف مكونة من عدد كبير من أشجار القرار الفردية الناتجة عن الاختيار العشوائي لعينات من بيانات التدريب والتي تعمل كمجموعة.
- المفهوم الأساسي لـ Random Forestوالسبب الرئيسي وراء نجاح نموذج تصنيفها هو اعتمادها على مبدأ أنّه: "سيتفوق عدد كبير من النماذج (الأشجار) غير المترابطة نسبيًا التي تعمل كمجموعة على أي من النماذج الفردية". [13]
- تهدف هذه الخوارزمية إلى زيادة دقة التصنيف من خلال الاعتماد على تنبؤات مجموعة من الأشجار بدلاً من شجرة قرار واحدة وذلك وفق الخطوات التالية كما هو موضوح في الشكل (2):



الشكل (2): آلية عمل خوارزمية Random forest في التصنيف

1. يتم أخذ عدد من السجلات العشوائية من مجموعة البيانات التي تحتوي على عدد k من السجلات .
2. يتم إنشاء أشجار قرار فردية لكل عينة.
3. تقوم كل شجرة بتوليد قرار يمثل نتيجة تصنيفها الخاصة.
4. يتم تحديد نتيجة التصنيف النهائية على أساس تصويت الأغلبية.

4.4 معايير تقييم خوارزميات التصنيف:

- **دقة التصنيف classification accuracy:** وهي تمثل قدرة النموذج على توقع الفئة المستهدفة بشكل صحيح وتقاس بالنسبة المئوية وفق المعادلة (4): [7]

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN}$$

العلاقة[4]

حيث أنّ:

- **True Positive (TP)**: أي إنّ نتيجة التنبؤ هي positive والقيمة الفعلية هي أيضًاpositive وهذا يعتبر تصنيف صحيح.
- **False Positive (FP)**: أي إنّ نتيجة التنبؤ هي positive ولكن القيمة الفعلية negative، وهذا يعتبر تصنيف خاطئ.
- **True Negative (TN)**: أي إنّ نتيجة التنبؤ هي negative والقيمة الفعلية هي أيضًا negative وهذا يعتبر تصنيف صحيح.
- **False Negative (FN)**: أي إنّ نتيجة التنبؤ هي negative ولكن القيمة الفعلية positive، وهذا يعتبر تصنيف خاطئ.
- **Positive (P):** اجمالي عدد التصنيفات الإيجابية وهي مجموع كل من FP و TP .
- **Negative (N):** اجمالي عدد التصنيفات السلبية وهي مجموع كل من FN و TN .

- الحالات المصنفة بشكل صحيح Correctly classified records: مجموع كل من TP و TN .

- الحالات المصنفة بشكل خاطئ Incorrectly classified records: مجموع كل من FP و FN .
- **الحساسية Sensitivity:** وهي المقياس الذي يقيّم قدرة نموذج التصنيف على التنبؤ بالإيجابيات الحقيقية لكل فئة، وتحسب وفق العلاقة التالية:

$$Sensitivity = \frac{TP}{TP + FN}$$ **العلاقة[5]**

- **الخصوصية Specificity:** وهي المقياس الذي يقيّم قدرة نموذج التصنيف على

$$Specificity = \frac{TN}{TN + FP}$$ **العلاقة[6]**

التنبؤ بالسلبيات الحقيقية لكل فئة، وتحسب وفق العلاقة التالية:

- **الاسترداد Recall (R)** : يعبّر عن جزء التصنيفات الإيجابية المتوقعة بشكل صحيح True Positiveمن بين جميع التصنيفات الإيجابية Positive للخوارزمية وهو يمثل مقياس الجودة quality وفق العلاقة الرياضية التالية: [14]

$$R= TP/ (TP + FP) = TP / P$$ **العلاقة[7]**

- **Precision (P):** يعبّر عن جزء التصنيفات الإيجابية المتوقعة بشكل صحيح True Positiveمن بين

اجمالي البيانات الإيجابية.

- **معدل الخطأ Error-Rate:** يعبر عن نسبة التنبؤات الخاطئة لنموذج التصنيف (الإيجابية والسلبية) من اجمالي التوقعات ويعطى بالعلاقة:

$$Error\text{-}Rate = \frac{FP+FN}{TP+FP+FN+TN}$$ **العلاقة[8]**

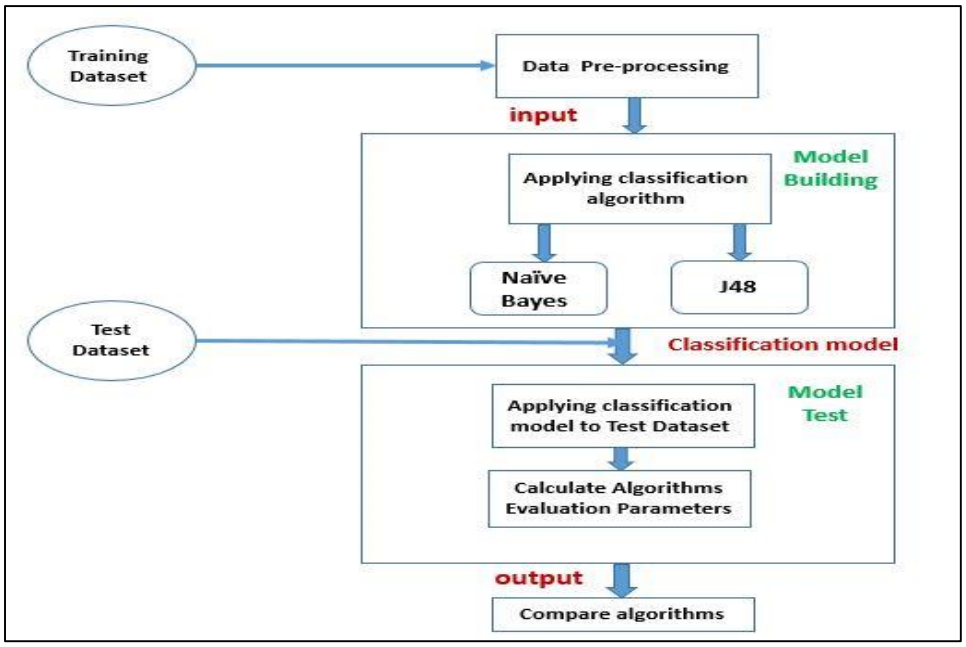
- **F-Measure** : تعبّر عن متوسط التوافق بين Precisionو Recall ويعطى بالعلاقة الآتية:

$$F= 2PR / (P+R)$$ **العلاقة[9]**

- **السرعة speed:** وتشير إلى الزمن الذي تستغرقه الخوارزمية في بناء النموذج.

5. الدراسة العملية:

يوضح الشكل (3) بنية مخطط البحث المتبع ومراحله الأساسية:



الشكل (3): المخطط العام ومراحل عملية التصنيف المقترحة

5.1 المعالجة المسبقة للبيانات Data Preprocessing:

تعد المعالجة المسبقة للبيانات مرحلة مهمة للتعامل مع البيانات قبل استخدامها في خوارزميات التنقيب وذلك لضمان الحصول على نتائج صحيحة ودقيقة والتي تتضمن:

- إزالة سجلات البيانات (الخاصة بالمدارس) المكررة.
- التأكد من عدم وجود حقول بيانات فارغة وحذفها في حال وجودها.
- التأكد من عدم وجود قيم مفقودة ضمن حقول البيانات.
- التخلص من القيم الشاذة: مثل معالجة حالة وجود القيمة 6 ضمن الحقل الخاص بعدد الدورات التدريبية NT–training والتي تعتبر قيمة شاذة كونها خارج المجال المحدد لقيم هذه الواصفة[5–0].
- تكامل البيانات: في بعض الحالات يتم استخدام قيم مختلفة للتعبير عن نفس الواصفة مثل قيم الواصفةModern–tools قد يكون هناك بعض السجلات التي تعبر عنها بـ 0 أو 1 بدلاً من No أو Yes وبالتالي قمنا بإجراء تكامل لهذه البيانات من خلال استبدال 0 بـ Noو 1 بـ Yes.

5.2 وصف مجموعات البيانات:

في هذه الدراسة، تم استخدام احدى وعشرون(21) واصفة تمثل مجموعة العوامل التي تلعب دور رئيسي في تحديد مستوى المدارس وتؤثر بشكل أساسي في تقييم المؤسسات التعليمية والموضحة قيمها وتفاصيلها في الجدول (1):

الجدول (1) مجموعة الواصفات المستخدمة في تقييم مستوى المدرسة وقيمها المحتملة.

Number	Attributes Name	Description	Possible Values
1	School-id	school identifier	Integer Number
2	stage	Educational stage	{first, second, third}
3	Location	school's geographical location	{ nearby, far, center}
4	NStudent	The number of school students	Integer Number
5	Attendance-rate	Student attendance rates	Percent[0%-100%]
6	Absenteeism-rate	Student absenteeism rates	Percent[0%-100%]
7	Ndisciplinary-cases	The number of disciplinary cases for school students	Integer number<student number
8	NRewards	The number of rewards for school students	Integer number<student number
9	Pass-rate	Certificate pass rates	Percent[0%-100%]
10	Fail-rate	Certificate fail rates	Percent[0%-100%]
11	NTeachers	number of teachers	Integer Number
12	NT-training	number of Teacher training courses	{0,1,2,3,4,5}
13	Curriculum-completion	School curriculum completion rate	Percent[0%-100%]
14	Compliance-laws	Compliance with regulations and laws	Percent[0%-100%]
15	Modern-methods	The use of modern methods in education	{yes , no}
16	Modern-tools	Availability of modern tools for education	{yes , no}
17	Attention-talent	Paying attention to students' talents	{yes , no}
18	Unclassed-activities	Participation in unclassed activities	{yes , no}
19	Parents-involvement	Involve parents in the educational process	{yes , no}
20	Harmony	Harmony between administrative and teaching staff	{yes , no}
21	School level	School level	(bad, good)

يُصنف مستوى المدارس إلى فئتين:

1- مستوى سيئ Bad level.

2- مستوى جيد Good level.

وكما يوضح الشكل (4) عينة من مجموعة البيانات المستخدمة في بحثنا:

school id	stage	Location	Nstudent	Attendance-rate	Absenteeis	Ndisciplina	NReward	Pass-rate	Fail-rate	NTeachers	NT-training	Curriculum-co	Compliance-l	Modern-methods	Modern-tools	Attention-talent	Unclassed-act	Parents-involvement	Harmony
1	first	nearby	100	90	10	3	0	95	5	13	2	76	80	yes	no	yes	no	yes	yes
2	second	center	85	80	20	6	1	83	17	15	1	85	95	yes	yes	yes	yes	yes	no
3	third	nearby	130	60	30	15	0	55	45	14	0	73	70	no	yes	no	yes	no	no
4	first	nearby	120	97	3	3	1	100	0	13	3	95	95	yes	yes	no	yes	yes	yes
5	second	far	95	88	10	2	0	70	30	14	0	87	70	no	no	no	no	yes	no
6	third	center	135	90	6	3	1	95	5	18	2	97	96	yes	yes	yes	yes	yes	yes
7	second	nearby	180	70	30	12	0	65	35	20	0	65	70	no	yes	yes	yes	yes	no
8	first	far	210	91	8	6	1	86	14	20	1	94	89	yes	no	no	no	yes	yes
9	third	center	300	85	10	12	1	92	8	28	2	94	90	yes	yes	no	yes	yes	yes
10	first	nearby	400	70	30	25	0	70	30	18	0	68	72	no	no	no	yes	no	no
11	third	far	279	75	18	14	0	80	20	27	1	81	81	yes	no	no	no	yes	yes
12	second	far	118	96	3	3	1	98	2	16	4	97	96	yes	yes	no	yes	yes	no
13	first	center	600	82	15	15	2	87	13	22	2	91	84	yes	no	no	yes	yes	no
14	third	far	200	93	5	4	3	97	3	23	2	97	95	yes	no	no	no	yes	yes
15	second	nearby	134	80	15	9	0	86	14	16	0	80	70	no	yes	no	yes	no	no
16	third	far	180	60	38	15	0	69	31	18	0	60	74	no	no	no	yes	yes	yes
17	first	center	800	89	10	5	3	96	4	36	3	97	99	yes	yes	yes	no	yes	yes
18	second	far	83	84	12	2	1	85	15	13	1	90	89	no	no	no	yes	yes	yes
19	third	nearby	318	70	28	22	1	69	31	27	2	82	64	no	yes	no	no	yes	yes
20	first	center	420	95	4	5	1	87	13	22	2	91	83	yes	no	no	yes	yes	no

الشكل (4): عينة من مجموعة البيانات الخاصة بالمدارس والمستخدمة ضمن البحث

5.3 بناء ملفات WEKA الخاصة بالبحث:

يعالج برنامج weka البيانات الموجودة في ملف ذو الامتداد (ARFF.) والذي قمنا بتحضيره من خلال استيراد مجموعة البيانات الموجودة في ملف excel إلى الملف School–Level.arff، والموضح في الشكل (5):

```
@RELATION School-Level
@ATTRIBUTE school-id NUMERIC
@ATTRIBUTE stage {first,second,third}
@ATTRIBUTE Location {nearby,center,far}
@ATTRIBUTE Nstudent NUMERIC
@ATTRIBUTE Attendance-rate NUMERIC
@ATTRIBUTE Absenteeism-rate NUMERIC
@ATTRIBUTE Ndisciplinary-cases NUMERIC
@ATTRIBUTE NRewards NUMERIC
@ATTRIBUTE Pass-rate NUMERIC
@ATTRIBUTE Fail-rate NUMERIC
@ATTRIBUTE NTeachers NUMERIC
@ATTRIBUTE NT-training {0,1,2,3,4,5}
@ATTRIBUTE Curriculum-completion NUMERIC
@ATTRIBUTE Compliance-laws NUMERIC
@ATTRIBUTE Modern-methods {yes,no}
@ATTRIBUTE Modern-tools {yes,no}
@ATTRIBUTE Attention-talent {yes,no}
@ATTRIBUTE Unclassed-activities {yes,no}
@ATTRIBUTE Parents-involvement {yes,no}
@ATTRIBUTE Harmony {yes,no}
@ATTRIBUTE School-level {good,bad}
@DATA
1,"first","nearby",100,90,10,3,0,95,5,13,2,76,80,"yes","no","yes","no","yes","yes","good"
2,"second","center",85,80,20,6,1,83,17,15,1,85,95,"yes","yes","yes","yes","yes","no","good"
3,"third","nearby",130,60,30,15,0,55,45,14,0,73,70,"no","yes","no","yes","no","no","bad"
4,"first","nearby",120,97,3,3,1,100,0,13,3,95,95,"yes","yes","no","yes","yes","yes","good"
5,"second","far",95,88,10,2,0,70,30,14,0,87,70,"no","no","no","no","yes","no","bad"
6,"third","center",135,90,6,3,1,95,5,18,2,97,96,"yes","yes","yes","yes","yes","yes","good"
7,"second","nearby",180,70,30,12,0,65,35,20,0,65,70,"no","yes","yes","yes","yes","no","bad"
```

الشكل (5): ملف البيانات بتنسيقARFF

5.4 تصنيف البيانات باستخدام خوارزميات التنقيب في البيانات:

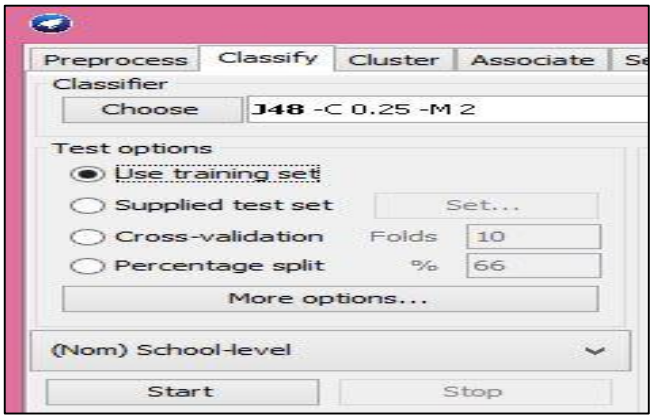
- بدايةً يجب اختيار مرحلة "**preprocess**" ضمن برنامج weka والتي يُراد من خلالها إعلام البرنامج عن موقع مجموعة البيانات التي سنستخدمها كبيانات تدريب.
- تتمثل الخطوة الثانية باختيار عملية المعالجة التي سيتم تطبيقها على مجموعة البيانات السابقة وهي التصنيف Classification لذلك نختار Classify".
- وأخيراً يترتب علينا تحديد ثلاث أمور أساسية:

1- نوع الخوارزمية التي نريد العمل عليها وهي إما J48 أو Naïve Bayes أو Random Forest.

2- خيارات الاختبار التي سيتم تطبيقها على البيانات وهي إما " **Use training set** " في مرحلة التدريب أو "**Supplied test set**" في مرحلة الاختبار.

3- الواصفة التي سيتم التصنيف وفقها والتي تمثل مستوى المدرسة **school-level** ضمن مجموعة البيانات.

كما هو موضح في الشكل (6):

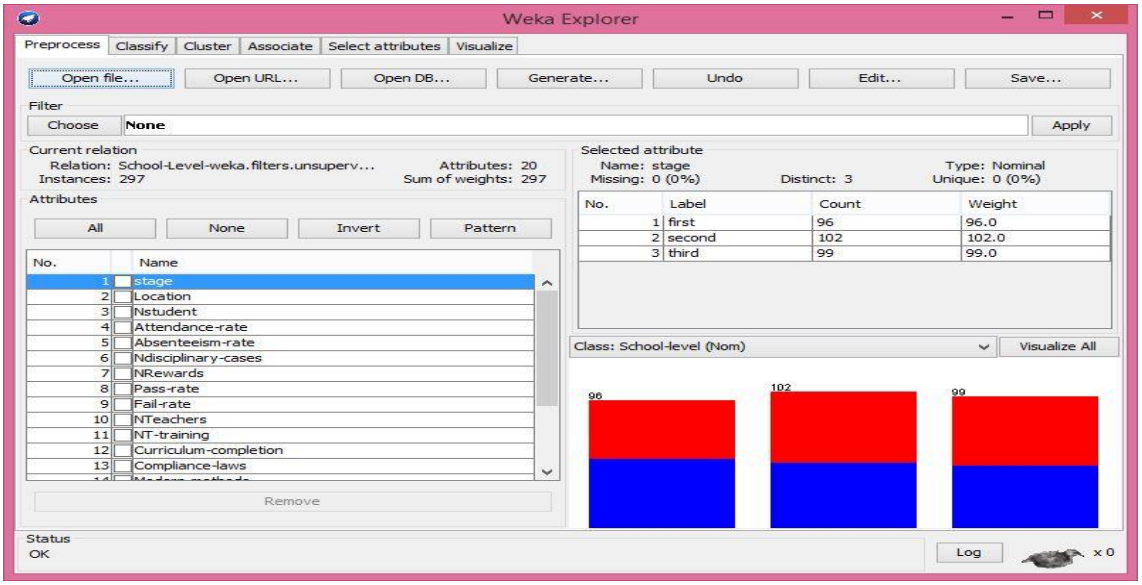


الشكل (6): تطبيق خوارزميات التصنيف على مجموعة بيانات التدريب في Weka

5.4.1 مرحلة التدريب Training:

5.4.1.1 تطبيق خوارزمية J48 على مجموعات بيانات التدريب:

قمنا بتطبيق خوارزمية شجرة القرار J48 على مجموعة بيانات التدريب المكونة من 297 سجل و20 واصفة (وذلك بعد حذف الواصفة School–Id والتي تمثل الرقم التسلسلي للمدرسة فمن غير المنطقي أنَ يؤثر على أداء الخوارزمية) والموضحة في الشكل (7)، حيث تتضمن واجهة البرنامج وصف لمحتويات الملف الذي تم فتحه (اسم قاعدة البيانات– مجموعة الواصفات وعددها –عدد السجلات) مع مجموعة معلومات إحصائية تتعلق بكل واصفة من الواصفات.



وبتطبيق الخوارزمية على مجموعة بيانات التدريب حصلنا على نموذج التصنيف الموضح في الشكل (8):

```
Classifier output

Time taken to build model: 0.06 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances         292               98.3165 %
Incorrectly Classified Instances         5                1.6835 %
Kappa statistic                          0.9663
Mean absolute error                      0.0315
Root mean squared error                  0.1255
Relative absolute error                  6.3047 %
Root relative squared error             25.1091 %
Coverage of cases (0.95 level)          98.6532 %
Mean rel. region size (0.95 level)      50.6734 %
Total Number of Instances              297
```

الشكل (8): نموذج التصنيف الناتج عن تطبيق خوارزمية j48 على بيانات التدريب

❖ **مصفوفة الارتباك confusion matrix:** تحتوي مصفوفة الارتباك على معلومات حول التصنيفات الفعلية والمتوقعة التي يقوم بها نظام التصنيف حيث يتم تقييم أداء هذه الأنظمة بشكل عام باستخدام البيانات الموجودة في

المصفوفة، ويوضح الشكل(9) قيم مصفوفة الارتباك لنموذج شجرة القرار J48 والتي تتضمن البيانات التالية:

```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.973     0.007      0.993     0.973      0.983      0.986    good
                 0.993     0.027      0.974     0.993      0.983      0.986    bad
Weighted Avg.    0.983     0.017      0.983     0.983      0.983      0.986

=== Confusion Matrix ===

   a   b   <-- classified as
 144   4 |   a = good
   1 148 |   b = bad
```

TP=144, FP=4, FN=1, TN=148

- TP rate for class good (a) = 144/(4+144) = 0.973
- FP rate for class bad (b) = 1/(1+148) = 0.007
- TP rate for class good = 148/(1+148) = 0.993
- FP rate for class bad = 4/(4+144) = 0.027
- Average TP rate = 0.983
- Average FP rate = 0.017
- Precision for class good = 144/(144+1) = 0.993
- Precision for class bad = 148/(4+148) = 0.974
- Recall for class good = 144/(144+4) = 0.973
- Recall for class bad = 148/(1+148) = 0.993
- F-measure for class good = 2*0.993*0.973/(0.993+0.973)= 0.983
- F-measure for class bad = 2*0.974*0.993/(0.974+0.993)= 0.983

ونستنتج من النموذج البيانات التالية:

- عدد البيانات التي تم تصنيفها بشكل صحيح: 292 بنسبة 98.31%
- عدد البيانات التي تم تصنيفها بشكل خاطئ:5 بنسبة 1.68%
- الزمن الذي استغرقه بناء النموذج: 0.06 seconds
- الدقة Accuracyوتحسب وفق العلاقة [4]:

$$\mathbf{Accuracy=144+148/144+4+1+148=0.9831=\ 98.31\%}$$

- الحساسية Sensitivity وتحسب وفق العلاقة [5]:

$$\mathbf{Sensitivity=144/144+1=0.9931=99.31\%}$$

- الخصوصية Specificity وتحسب وفق العلاقة [6]:

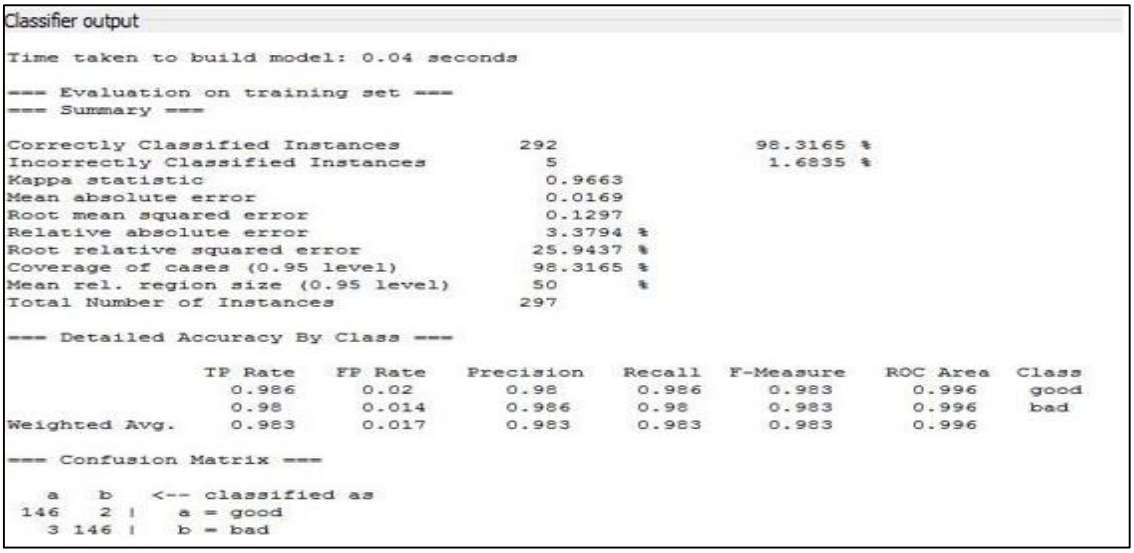
$$\mathbf{Specificity=148/148+4=0.9736=97.36\%}$$

- معدل الخطأ ويحسب وفق العلاقة [8]:

$$\mathbf{Error-Rate=4+1/144+4+1+148=0.01}$$

5.4.1.2 تطبيق خوارزمية Naïve Bayes على مجموعات بيانات التدريب:

بتطبيق خوارزمية التصنيف Naïve Bayes على مجموعة بيانات التدريب كما هو موضح في الشكل (10) نحصل على نموذج التصنيف ومصفوفة الارتباك التي تتضمن البيانات التالية:



```
Classifier output

Time taken to build model: 0.04 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances         292               98.3165 %
Incorrectly Classified Instances         5                1.6835 %
Kappa statistic                          0.9663
Mean absolute error                      0.0169
Root mean squared error                  0.1297
Relative absolute error                  3.3794 %
Root relative squared error             25.9437 %
Coverage of cases (0.95 level)          98.3165 %
Mean rel. region size (0.95 level)      50      %
Total Number of Instances              297

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.986     0.02       0.98      0.986     0.983      0.996    good
                 0.98      0.014      0.986     0.98      0.983      0.996    bad
Weighted Avg.    0.983     0.017      0.983     0.983     0.983      0.996

=== Confusion Matrix ===

   a   b   <-- classified as
 146   2 |   a = good
   3 146 |   b = bad
```

الشكل (10): نموذج التصنيف الناتج عن تطبيق خوارزمية Naïve Bayes على بيانات التدريب

- عدد البيانات التي تم تصنيفها بشكل صحيح:292 بنسبة 98.31%
- عدد البيانات التي تم تصنيفها بشكل خاطئ:5 بنسبة 1.68%
- الزمن الذي استغرقه بناء النموذج: 0.04 seconds
- قيم التصنيف ضمن مصفوفة الارتباك: TP=146, FP=2, TN=146, FN=3
- الدقة Accuracy: Accuracy=146+146/146+2+3+146=0.9831= 98.31%
- الحساسية Sensitivity: Sensitivity=146/146+3=0.9798=97.89%
- الخصوصية Specificity: Specificity=146/146+2=0.9864=98.64%
- معدل الخطأ: Error−Rate=2+3/144+4+1+148=0.01

5.4.1.3 تطبيق خوارزمية Random Forest على مجموعات بيانات التدريب:

بتطبيق خوارزمية التصنيف Random Forestعلى مجموعة بيانات التدريب نحصل على نموذج التصنيف ومصفوفة الارتباك كما هو موضح في الشكل (11):

```
Time taken to build model: 0.08 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances         297               100      %
Incorrectly Classified Instances         0                 0      %
Kappa statistic                          1
Mean absolute error                      0.0088
Root mean squared error                  0.0471
Relative absolute error                  1.7548 %
Root relative squared error              9.4283 %
Coverage of cases (0.95 level)         100      %
Mean rel. region size (0.95 level)      52.6936 %
Total Number of Instances              297

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         0          1         1         1          1        good
                 1         0          1         1         1          1        bad
Weighted Avg.    1         0          1         1         1          1

=== Confusion Matrix ===

   a   b   <-- classified as
 148   0 |   a = good
   0 149 |   b = bad
```

الشكل (11): نموذج التصنيف الناتج عن تطبيق خوارزمية Random Forestعلى بيانات التدريب

- عدد البيانات التي تم تصنيفها بشكل صحيح:297 بنسبة 100%
- عدد البيانات التي تم تصنيفها بشكل خاطئ:0 بنسبة 0%
- الزمن الذي استغرقه بناء النموذج: 0.08 seconds
- قيم التصنيف ضمن مصفوفة الارتباك: TP=148, FP=0, TN=149, FN=0
- الدقة Accuracy: Accuracy=148+149/148+0+0+149=1= 100%
- الحساسية Sensitivity: Sensitivity=148/148+0=1=100%
- الخصوصية Specificity: Specificity=149/149+0=1=100%
- معدل الخطأ: Error−Rate=0+0/148+0+0+149=0

5.4.2 مرحلة الاختبار Testing:

بعد تدريب الخوارزميات على تصنيف مستوى المدارس سنقوم بتقييم نموذج التصنيف المبني واختبار دقته في تحديد المستوى المجهول لمجموعة من المدارس ضمن مجموعة بيانات الاختبار Test Datasetالمكونة من 99 سجل.

5.4.2.1 تطبيق خوارزمية J48 على مجموعات بيانات الاختبار:

باختبار أداء خوارزمية شجرة القرار J48 في تصنيف البيانات ضمن مجموعة الاختبار نحصل على النتائج التالية:

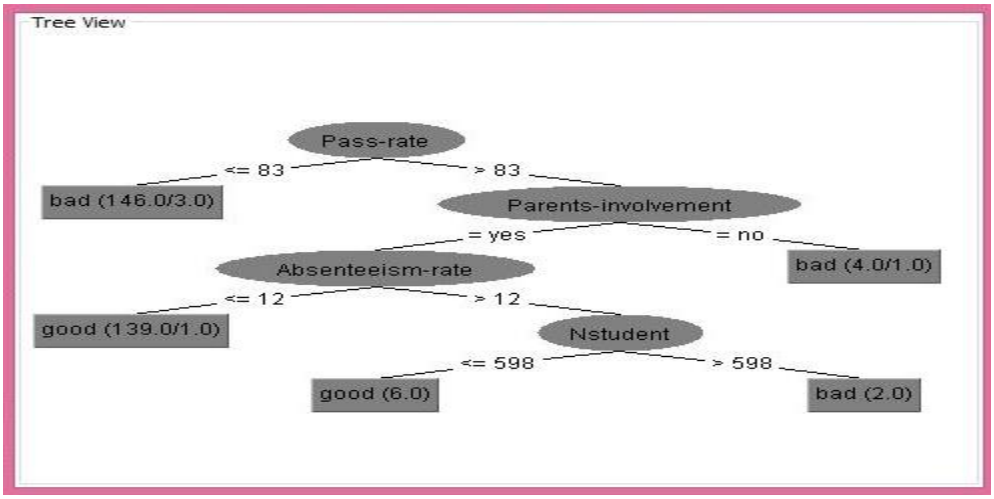
```
Classifier output

Time taken to build model: 0.02 seconds

=== Evaluation on test set ===
=== Summary ===

Correctly Classified Instances          92               92.9293 %
Incorrectly Classified Instances         7                7.0707 %
Kappa statistic                          0.8587
Mean absolute error                      0.0851
Root mean squared error                  0.2639
Relative absolute error                 17.0143 %
Root relative squared error             52.7711 %
Coverage of cases (0.95 level)          92.9293 %
Mean rel. region size (0.95 level)      50.5051 %
Total Number of Instances               99

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.9       0.041      0.957     0.9       0.928      0.919    good
                 0.959     0.1        0.904     0.959     0.931      0.919    bad
Weighted Avg.    0.929     0.07       0.931     0.929     0.929      0.919

=== Confusion Matrix ===

  a  b   <-- classified as
 45  5 |  a = good
  2 47 |  b = bad
```

الشكل (12): نموذج التصنيف الناتج عن تطبيق خوارزمية j48 على بيانات الاختبار

- عدد البيانات التي تم تصنيفها بشكل صحيح:92 بنسبة 92.92%
- عدد البيانات التي تم تصنيفها بشكل خاطئ:7 بنسبة 7.07%
- الزمن الذي استغرقه بناء النموذج: 0.02 seconds
- قيم التصنيف ضمن مصفوفة الارتباك: TP=45, FP=5, FN=2, TN=47
- الدقة Accuracy والموضحة تفاصيلها ضمن مصفوفة الارتباك في الشكل (10) والتي تحسب وفق العلاقة [4]: Accuracy=45+47/45+5+2+47=0.9292=92.92%

- الحساسية Sensitivity: Sensitivity=45/45+2=0.9574=95.74%
- الخصوصية Specificity: Specificity=47/47+5=0.9038=90.38%
- معدل الخطأ: Error−Rate=5+2/45+5+2+47=0.09

كما ويوضح الشكل (13) شجرة القرار المبنية من قبل الخوارزمية والمستخدمة في آلية التصنيف المتبعة من قبلها:



الشكل (13): شجرة القرار j48 الناتجة

4.5.2.2 تطبيق خوارزمية Naïve Bayes على مجموعات بيانات الاختبار:

بتطبيق خوارزمية naïve Bayes على مجموعة بيانات الاختبار ودراسة دقتها في تصنيف مستوى المدارس حصلنا على نموذج التصنيف التالي:

```
Classifier output

Time taken to build model: 0.01 seconds

=== Evaluation on test set ===
=== Summary ===

Correctly Classified Instances          94               94.9495 %
Incorrectly Classified Instances         5                5.0505 %
Kappa statistic                          0.899
Mean absolute error                      0.0506
Root mean squared error                  0.2247
Relative absolute error                 10.1214 %
Root relative squared error             44.9453 %
Coverage of cases (0.95 level)          94.9495 %
Mean rel. region size (0.95 level)      50      %
Total Number of Instances               99

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.94      0.041       0.959     0.94      0.949      0.957    good
                 0.959     0.06        0.94      0.959     0.949      0.968    bad
Weighted Avg.    0.949     0.05        0.95      0.949     0.949      0.962

=== Confusion Matrix ===

  a  b   <-- classified as
 47  3 |  a = good
  2 47 |  b = bad
```

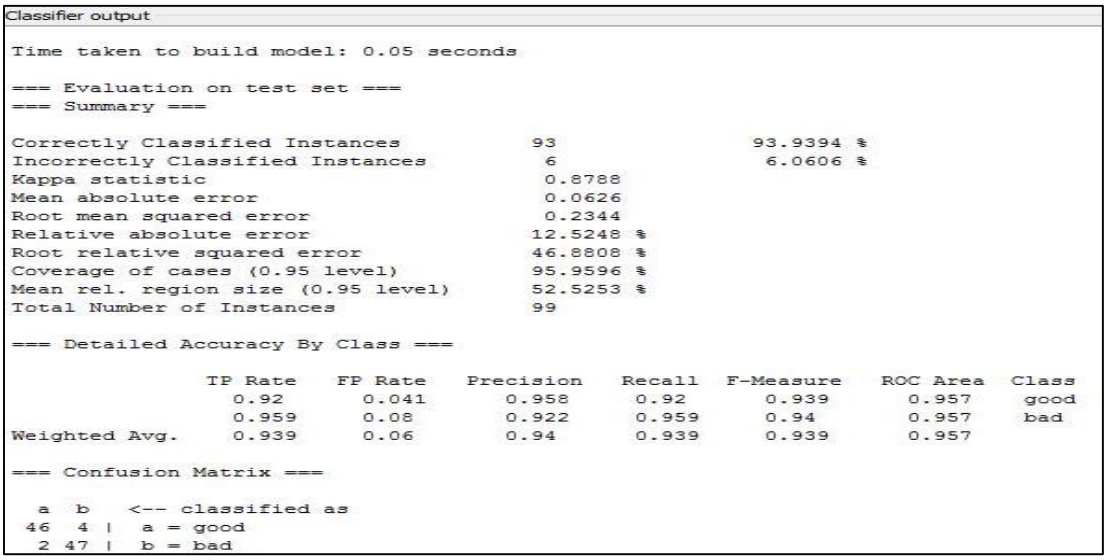
الشكل (14): نموذج التصنيف الناتج عن تطبيق خوارزمية Naïve Bayes على بيانات الاختبار

ونستنتج منه البيانات التصنيف التالية:

- عدد البيانات التي تم تصنيفها بشكل صحيح:94 بنسبة 94.94%
- عدد البيانات التي تم تصنيفها بشكل خاطئ:5 بنسبة 5.05%
- زمن الذي استغرقه بناء النموذج: 0.01 seconds
- قيم التصنيف ضمن مصفوفة الارتباك: TP=47, FP=3, FN=2, TN=47
- الدقة Accuracy: Accuracy=47+47/47+3+2+47=0.9494=94.94%
- الحساسية Sensitivity: Sensitivity=47/47+2=0.9591=95.91%
- الخصوصية Specificity: Specificity=47/47+3=0.94=94%
- معدل الخطأ: Error−Rate=3+2/47+3+2+47=0.05

4.5.2.2 تطبيق خوارزمية Random Forestعلى مجموعات بيانات الاختبار:

وأخيراً، قمنا بتطبيق خوارزمية الغابة العشوائية Random Forest على مجموعة بيانات الاختبار، وحصلنا على نموذج التصنيف النهائي ومصفوفة الارتباك الخاصة به كما هو موضح في الشكل (15):



```
Classifier output

Time taken to build model: 0.05 seconds

=== Evaluation on test set ===
=== Summary ===

Correctly Classified Instances          93               93.9394 %
Incorrectly Classified Instances         6                6.0606 %
Kappa statistic                          0.8788
Mean absolute error                      0.0626
Root mean squared error                  0.2344
Relative absolute error                 12.5248 %
Root relative squared error             46.8808 %
Coverage of cases (0.95 level)          95.9596 %
Mean rel. region size (0.95 level)      52.5253 %
Total Number of Instances               99

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.92      0.041      0.958      0.92      0.939      0.957    good
                 0.959     0.08       0.922      0.959     0.94       0.957    bad
Weighted Avg.    0.939     0.06       0.94       0.939     0.939      0.957

=== Confusion Matrix ===

  a  b   <-- classified as
 46  4 |  a = good
  2 47 |  b = bad
```

الشكل (15): نموذج التصنيف الناتج عن تطبيق خوارزمية Random Forest على بيانات الاختبار

ونستنتج منه البيانات التصنيف التالية:

- عدد البيانات التي تم تصنيفها بشكل صحيح:93 بنسبة 93.93%
- عدد البيانات التي تم تصنيفها بشكل خاطئ:6 بنسبة 6.06%
- زمن الذي استغرقه بناء النموذج: 0.05 seconds
- قيم التصنيف ضمن مصفوفة الارتباك: TP=46, FP=4, FN=2, TN=47
- الدقة Accuracy: Accuracy=46+47/46+4+2+47=0.9393=93.93%
- الحساسية Sensitivity: Sensitivity=46/46+2=0.9583=95.83%
- الخصوصية Specificity: Specificity=47/47+4=0.9215=92.15%
- معدل الخطأ: Error−Rate=4+2/46+4+2+47=0.06

6. النتائـــج والمناقشـــة:

وبحساب قيم مجموعة بارامترات التقييم ونتائج التصنيف لخوارزميات التصنيف Naïve Bayes وJ48 و Random Forest خلال مرحلتي بناء النموذج بالاعتماد على بيانات التدريب واختباره على بيانات الاختبار توصلنا إلى القيم الموضحة في الجدولين(2) و (3) :

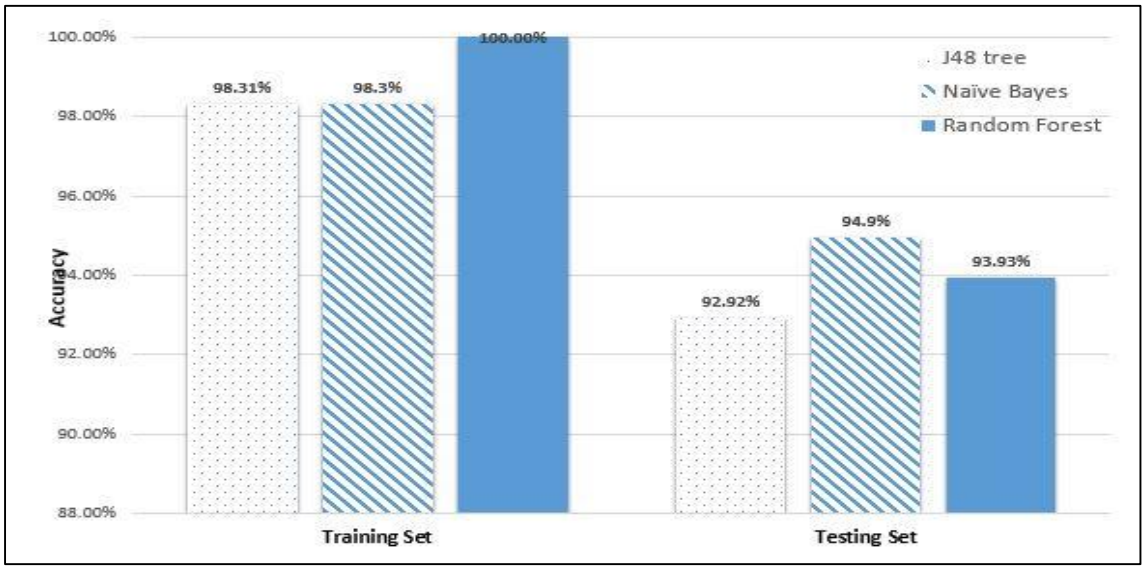
الجدول (2): البيانات المصنفة بشكل صحيح وخاطئ من قبل خوارزميات التصنيف الثلاث في مرحلتيّ التدريب والاختبار

Classification algorithm	Training set		Testing set	
	Correctly classified	Incorrectly classified	Correctly classified	Incorrectly classified
J48	98.31%(292)	1.68%(5)	92.92%(92)	7.07%(7)
Naïve Bayes	98.31%(292)	1.68%(5)	94.94%(94)	5.05%(5)
Random Forest	100%(297)	0%(0)	93.93%(93)	6.06%(6)

الجدول (3): بارامترات تقييم أداء خوارزميات التصنيف في مرحلتي التدريب والاختبار

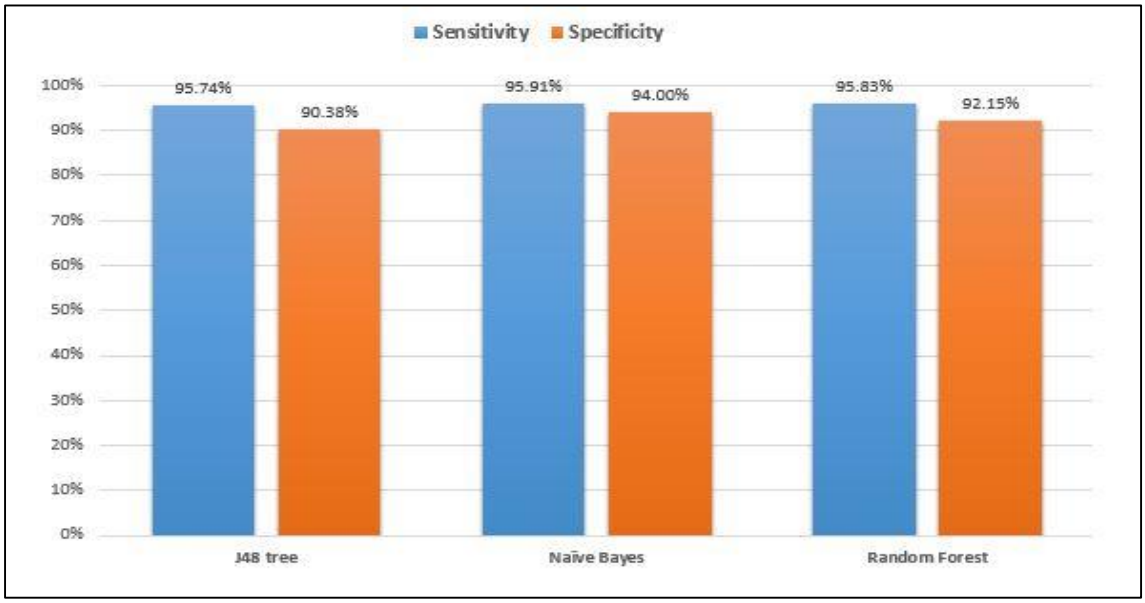
Classification algorithm	Training set					Testing set				
	accuracy	Sensitivity	Specificity	Error-rate	Modeling Time	accuracy	Sensitivity	Specificity	Error-rate	Modeling Time
J48	98.31%	99.31%	97.36%	0.01	0.06 sec	92.92%	95.74%	90.38%	0.09	0.02 sec
Naïve Bayes	98.31%	97.89%	98.64%	0.01	0.04 sec	94.94%	95.91%	94%	0.05	0.01 sec
Random Forest	100%	100%	100%	0	0.08 sec	93.93%	95.83%	92.15%	0.06	0.05 sec

توضح النتائج النهائية لتصنيف مستوى المدارس بالاعتماد على خوارزميات التنقيب المستخدمة في بحثنا أن خوارزميتي Naive Bayes و j48 متساويتين من حيث عدد البيانات المصنفة بشكل صحيح وخاطئ وتعملان بنفس الأداء في تصنيف مجموعة بيانات التدريب بدقة تصل إلى 98.31% في حين تتفوق خوارزمية Random Forest على الخوارزميتين السابقتين في مرحلة التدريب بدقة تصل إلى 100%، أما في مرحلة الاختبار نلاحظ تفوق خوارزمية Naïve Bayes على خوارزمية J48 وخوارزمية Random Forest في دقة تصنيف بيانات الاختبار حيث وصلت دقة خوارزمية Naïve Bayes إلى 94.94% وبفرق قدره 2% عن خوارزمية J48 و1% عن Random Forest كما هو موضح في الشكل (16):



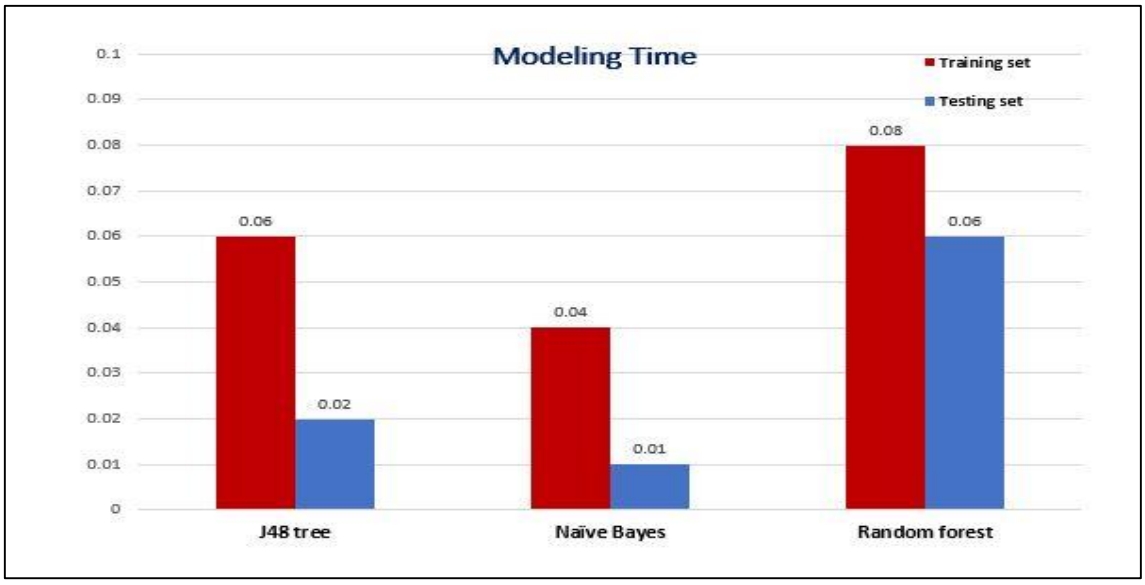
الشكل (16): دقة التصنيف التي تحققها الخوارزميات الثلاث في مرحلتي التدريب والاختبار

و بتتبع قيم كل من الحساسية والخصوصية للخوارزميات الثلاث والتي تلعب دور أساسي في تقييم جودة نموذج التصنيف الناتج (حيث ترتفع بارتفاع قيمة هذين البارمترين وتنخفض بانخفاضهما) نلاحظ تفوق خوارزمية Naïve Bayes أيضاً كما هو موضح في الشكل (17).



الشكل (17): قيم كل من الحساسية والخصوصية لنموذج الاختبار لدى الخوارزميات الثلاث

وبقياس الزمن الذي تستغرقه كل خوارزمية في بناء نموذج التصنيف الخاص بها نلاحظ من الشكل (18) أنّ خوارزمية Naïve Bayes أسرع من خوارزميتي Random forest وj48 في مرحلتي التدريب والاختبار بزمن قدره 0.04 seconds للتدريب و0.01sec للاختبار.



الشكل (18): الزمن الذي يستغرقه بناء نموذج التصنيف واختباره من قبل الخوارزميات الثلاث

7. الاســـتنتاجات والتوصيـــات:

يعتبر المجال التعليمي والتربوي ركيزة أساسية من ركائز التقدم والتطور في المجتمع لذلك من المهم جداً مراقبة أداء هذه المؤسسات التعليمية وسير العملية التدريسية فيها وتطويرها من خلال تحديد مستوى المدارس بمراحلها الثلاث في محاولة لحل مجموعة السلبيات والمشاكل التي تقف عقبة في ممارسة دورها بشكل فعال في انتاج جيل واعد علمياً وثقافياً وتربوياً بالإضافة للتركيز على إيجابيات المدارس الأخرى وتحفيز وتشجيع كوادرها لتقديم الأفضل دوماً في هذا المجال الحيوي الهام.

قدمت هذه الدراسة نظرة عامة عن تقنيات التنقيب في البيانات وخوارزمياتها التي توفر طرقًا واعدة وذكية للكشف عن الأنماط المجهولة والتنبؤ بها ضمن كميات كبيرة من البيانات، حيث تم التركيز على خوارزميات التصنيف Naïve Bayes وشجرة القرار J48 والغابة العشوائية Random Forestوآلية عمل كل منها.

وبدراسة أداء كل من هذه الخوارزميات ودقتها في تصنيف مستوى المدارس إلى صنفين أساسيين جيد good وسيئbad وفق مرحلتي التدريب والاختبار، أظهرت النتائج أداء عالي في التصنيف لكل الخوارزميات بحيث لم تحقق أي منها دقة أقل من 92%، وإمكانية التعامل مع أنواع مختلفة من البيانات (رقمية ونصية ومنطقية) مع تفوق لخوارزمية Naïve Bayes من حيث دقة التصنيف العالية والوقت المنخفض الذي تستهلكه عملية بناء النموذج وتصنيف البيانات وفقه مقارنة بخوارزمية J48 و Random Forest.

وبالنتيجة يقترح البحث التوصيات التالية:

- ❖ استخدام خوارزميات تصنيف أخرى في تحديد مستوى المدارس مثل k-Nearest Neighbors ومقارنة دقتها مع دقة الخوارزميات التي توصلنا إليها.
- ❖ دراسة أداء خوارزميات التصنيف المستخدمة ضمن بحثنا في تقييم مستوى المدارس مع حجم بيانات أضخم وعدد واصفات أكبر لمعرفة مدى تأثير حجم البيانات وعدد الواصفات على أداء كل منها.
- ❖ توسيع الدراسة السابقة لتشمل المدارس المهنية والتقنية التي تتضمن واصفات أخرى أساسية تؤثر في تقييم المستوى.

8. المراجع:

[1] HAN, J; PEI, J; KAMBER, M. 2011, *Data mining: concepts and techniques.* 3^{rd} Elsevier, NewYork,744.
[2] Nikam, S. S. (2015). A comparative study of classification techniques in data mining algorithms. Oriental Journal of Computer Science and Technology, 8(1), 13-19.
[3] Harkiran, K. (2017). A Study On Data Mining Techniques And Their Areas Of Application. International Journal of Recent Trends in Engineering and Research, 3, 93-95.
[4] Kamatkar, S. J., Tayade, A., Viloria, A., & Hernández-Chacín, A. (2018, June). Application of classification technique of data mining for employee management system. In International Conference on Data Mining and Big Data (pp. 434-444). Springer, Cham.

[5] Nagaparameshwara chary, S. D. (2017). A Survey on Comparative Analysis of Decision Tree Algorithms in Data Mining. International Conference on Innovative Applications in Engineering and Information Technology(ICIAEIT-2017) .

[6] Ba'abbad, I., Althubiti, T., Alharbi, A., Alfarsi, K., & Rasheed, S. (2021). A Short Review of Classification Algorithms Accuracy for Data Prediction in Data Mining Applications. Journal of Data Analysis and Information Processing, 9(3), 162-174.

[7] Chandrasekar, P., Qian, K., Shahriar, H., & Bhattacharya, P. (2017, July). Improving the prediction accuracy of decision tree mining with data preprocessing. In 2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC) (Vol. 2, pp. 481-484). IEEE.

[8] Venkatesan, E., & Velmurugan, T. (2015). Performance analysis of decision tree algorithms for breast cancer classification. Indian Journal of Science and Technology, 8(29), 1-8.
[9] Almarabeh, H. (2017). Analysis of students' performance by using different data mining classifiers. International Journal of Modern Education and Computer Science, 9(8), 9.

[10] Wickramasinghe, I., & Kalutarage, H. (2021). Naive Bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation. Soft Computing, 25(3), 2277-2293.

[11] Alpan, K., & İlgi, G. S. (2020, October). Classification of diabetes dataset with data mining techniques by using WEKA approach. In 2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT) (pp. 1-7). IEEE.

[12] CHEN, S; WEBB, G. I; LIU, L; MA, X. 2020, *A novel selective naïve Bayes algorithm. Knowledge-Based Systems*, 1^{st}, Elsevier, New York & London,102.
[13] Mohana, R. M., Reddy, C. K. K., Anisha, P. R., & Murthy, B. R. (2021). Random forest algorithms for the classification of tree-based ensemble. Materials Today: Proceedings.

[14] Hussain, S., Dahan, N. A., Ba-Alwib, F. M., & Ribata, N. (2018). Educational data mining and analysis of students' academic performance using WEKA. *Indonesian Journal of Electrical Engineering and Computer Science*, *9*(2), 447-459.