تحسين أنظمة دعم القرار في نظم التعليم باستخدام تقنيات التنقيب عن البيانات و التعلم الآلي

د. كندة أبو قاسم *

لمى عدنان باشا**

(تاريخ الإيداع 19 / 5 / 2021. قُبِل للنشر في 10/ 10 / 2021)

□ ملخّص □

يهدف التنقيب عن البيانات التعليمية إلى دراسة البيانات المتوفرة في المجال التعليمي وإخراج المعرفة المخفية منه بغية الاستفادة منها في تعزيز عملية التعليم واتخاذ قرارات ناجحة من شأنها تحسين الأداء الأكاديمي للطالب.

تقترح هذه الدراسة استخدام تقنيات التنقيب عن البيانات لتحسين التنبؤ بأداء الطلاب، حيث تم تطبيق ثلاث خوارزميات تصنيف (Naïve Bayes, J48, Support Vector Machine) على قاعدة بيانات أداء الطلاب ، ثم تم تصميم مصنف جديد لدمج نتائج تلك المصنفات الفردية باستخدام تقنية الدمج Voting Method .

تم استخدام الأداة WEKA التي تدعم الكثير من خوارزميات و طرائق التنقيب في البيانات.

تظهر النتائج أن مصنف الدمج لديه أعلى دقة للتنبؤ بمستويات الطلاب مقارنة بالمصنفات الأخرى ، حيث حقق دقة تعرف وصلت إلى % 74.8084 .

و أفادت خوارزمية العنقدة simple k–means في تجميع الطلاب المتشابهين في مجموعات منفصلة بالتالي فهم مميزات كل مجموعة مما يساعد على قيادة وتوجيه كل مجموعة على حدى.

الكلمات المفتاحية: : الدمج ، Voting Method ، مستويات الطلاب ، WEKA ، العنقدة.

***أستاذ مساعد – قسم هندسة الحاسبات والتحكم الآلي – كلية الهندسة الميكانيكية والكهربائية – جامعة تشرين – اللاذقية –سورية.**

البريد الالكتروني:Ki.aboukassem@gmail.com

****طالب دراسات عليا(ماجستير)– قسم الحاسبات والتحم الآلي – كلية الهندسة الميكانيكية والكهربائية – جامعة تشرين –اللاذقية–سورية.البريد الالكتروني:lama.basha302@gmail.com**

Improving Decision Support Systems in Education Systems Using Data Mining and Machine Learning Techniques

Dr. Kinda Abu Qasim [*]
Lama Basha [* *]

(Received 19 / 5 / 2021. Accepted 10 / 10 / 2021)

□ ABSTRACT □

Educational data mining aims to study the available data in the educational field and extract the hidden knowledge from it in order to benefit from this knowledge in enhancing the education process and making successful decisions that will improve the student's academic performance.

This study proposes the use of data mining techniques to improve student performance prediction. Three classification algorithms (Naïve Bayes,J48, Support Vector Machine) were applied to the student performance database, and then a new classifier was designed to combine the results of those individual classifiers using Voting Method.

The WEKA tool was used, which supports a lot of data mining algorithms and methods.

The results show that the ensemble classifier has the highest accuracy for predicting students' levels compared to other classifiers, as it has achieved a recognition accuracy of 74.8084%.

The simple k-means clustering algorithm was useful in grouping similar students into separate groups, thus understanding the characteristics of each group, which helps to lead and direct each group separately.

Keywords: Ensemble, Voting Method, students' levels, WEKA, Clustering Algorithm.

[*]**Associate Professor, computer and automatic control engineering, faculty of Mechanical and electrical engineering, Tishreen university, Lattakia, Syria. E-mail:Ki.aboukassem@gmail.com.**

[**]**Postgraduate Student (Master degree) , Department of computer and automatic control Engineering, Faculty of Mechanical and Electrical Engineering, Tishreen University, Lattakia , Syria. E-mail:lama.basha302@gmail.com.**

مقدمة:

أصبح – في عالمنا اليوم عصر الانفجار المعرفي – الإنسان محاط بكم هائل من المعلومات المتنوعة في مختلف مناحي الحياة ، ويرجع السبب لهذا التدفق الهائل للمعلومات إلى تطور وسائل معالجة المعلومات فهي ثورة معلوماتية حيث باتت الطرق التقليدية (والتي هي مزيج من الطرق الإحصائية) لتحليل هذه البيانات تعاني من كثير من المشكلات في التعامل مع هذا النوع من البيانات اليوم. من بين العلوم التطبيقية الحديثة في هذا المجال يأتي علم اكتشاف المعرفة في قواعد البيانات (Knowledge Discovery in Database(KDD وعلم التنقيب في البيانات على رأس هذه العلوم .

قدمت الباحثة Dorina Kabakchieva عام 2013 بحثا علميا تطرقت فيه إلى الكشف عن الإمكانات المهمة التي يمكن أن تقدمها تطبيقات التنقيب عن البيانات لإدارة الجامعة(بيانات الطلاب) حيث أعطت نتائج جيدة ساهمت في تحسين أداء الطالب ، واستخدمت تقنيات التنقيب : J48, (NaiveBayes and Bayes Net), K-Nearest Neighbor algorithm OneR and JRip.[1]

أيضا في عام 2017 قدم الباحث Hilal Almarabeh بحثا لتحليل وتقييم أداء طلاب الجامعة من خلال تطبيق تقنيات استخراج المعرفة . اعتمدت هذه الدراسة على 5 مصنفات ,Naïve Bayes ,Bayesian Network, J48, Neural Network, ID3 . وأعطت نتائج جيدة حيث حسنت دقة التعرف بشكل كبير [2].

لتحليل هذا الموضوع يقوم البحث بتوظيف علوم التنقيب في المؤسسات التعليمية والأكاديمية لتحسين الأداء الأكاديمي باستنباط الأنماط السائدة في بيانات هذه المؤسسات حول الطلاب، وذلك بالاعتماد على تقنيات دمج المصنفات وبرنامج التنقيب WEKA لبناء النموذج التنبؤي المناسب.

أهمية البحث وأهدافه:

يهدف البحث إلى بناء نموذج تنبؤي للتعرف على الأنماط السائدة بمستوى التحصيل الدراسي للطلاب باستخدام تقنيات التنقيب عن البيانات والتعلم الآلي، ومن ثم الاستفادة من هذا النموذج لتحسين عملية اتخاذ القرار في نظم التعليم الجامعي باستخدام نظم التعلم الآلي .

تتبع أهمية هذا البحث من خلال استخدام النموذج المقترح لاستكشاف المعرفة الكامنة في المعطيات الأكاديمية للطلاب ومن ثم التنفيذ العملي لهذا النموذج ومقارنة تقنيات التنقيب التي تم استخدامها بما يساعد في تحسين عملية اتخاذ القرار.

مشكلة البحث:

تكمن في أن نظم معلومات التعليم الجامعي تضم البيانات المرتبطة بالفعاليات كجداول وفهارس تمهيدا لاستخدامها عبر الطرق الإحصائية التقليدية دون إدراك أهمية المعطيات المخزنة وما تحتويه من معرفة مضمنة يمكن استخدامها للتنبؤ المستقبلي بسلوك المعطيات ، أيضا انخفاض الدقة في تصنيف بيانات الطلاب المدروسة ولا سيما عند تحديد علاقة أوضاع الطالب بالانقطاع عن الدراسة.

طرائق البحث ومواده:

تقوم كل مؤسسة تعليمية بإنشاء الكثير من البيانات المتعلقة بالطالب المسجل وإذا لم يتم تحليل هذه البيانات بشكل صحيح فلن نستفيد منها بالشكل الذي يحسن الأداء الأكاديمي .

لحل تلك المشكلة تم اعتماد نظم دعم قرار ذكية وهي تطبيقات حاسوبية تقوم بجمع وتنظيم وتحليل البيانات بهدف مساعدة الإدارة على اتخاذ القرارات الصحيحة التي من شأنها أن تساهم في إيصال المنشأة لأهدافها الموضوعة.

تبعا لذلك فإن البيئة الأساسية المستخدمة في هذه الدراسة هي:

برنامج التنقيب عن البيانات Weka وهي أداة مفتوحة المصدر كتبت بلغة Java وتوفر العديد من مهام التنقيب وخوارزميات التعلم الآلي ، تم تطويرها في جامعة وايكاتو في نيوزيلندا.[3]

Waikato Environment for Knowledge Analysis

توفر بيئة Weka واجهة موحدة لكثير من خوارزميات التعلم بالإضافة لطرق ما قبل المعالجة وتحليل النتائج والتقييم.

تتضمن واجهته الرسومية العمليات الرئيسية التالية :

1- تحضير البيانات Pre_Proccessing.

2- التصنيف ، العنقدة ، قواعد الربط.

3- اختيار المعايير .

و يقبل ملفات بامتداد (Attribute Relation File Formate) arff : وهي ملفات تتكون من قائمة من السمات وقيمها بصيغة معينة حيث يكون في بداية الملف اسم العلاقة ثم السمات ونوعها وقيمها ، وبالتالي يجب تهيئة الملف(قاعدة البيانات) بحيث يناسب البرنامج Weka.

عينة البحث:

تم اختيار قاعدة البيانات student performance data set تتألف من 1044 سجل (طالب) و 24 سمة تشمل سمات اقتصادية واجتماعية و أكاديمية مختلفة للطالب مبينة في الجدول(1) .

الجدول رقم (1): ملخص عن مجموعة البيانات المدروسة

N	Attribute	Variable	Possible values
1	Sex-جنس الطالب	binary	M,F
2	Address-العنوان	binary	U,R
3	Famsize-حجم الأسرة	binary	LE3,GT3
4	Pstatus-الوضع العائلي للوالدين	binary	T,A
5	Medu-تعليم الأم	numeric	0,1,2,3,4
6	Fedu-تعليم الأب	numeric	0,1,2,3,4
7	Mjob-عمل الأم	nominal	Teacher,health,home ,service,other
8	Fjob-عمل الأب	nominal	Teacher,health,home ,service,other
9	Reason- سبب اختيار جامعة محددة	nominal	Home, reputation, Course,other
10	travel time- وقت السفر من المنزل إلى الجامعة	numeric	1,2,3,4

11	study time- وقت الدراسة الأسبوعي	numeric	1,2,3,4
12	Failures-عدد مرات الرسوب	numeric	n if 1<=n<3;else 4
13	universitySup- دعم الجامعة للطالب	binary	Yes, no
14	Famsup-دعم الأسرة التربوي	binary	Yes, no
15	Paid- دروس إضافية مدفوعة الأجر	binary	Yes, no
16	Higher- رغبة الطالب بمتابعة دراسات عليا	binary	Yes, no
17	Internet- الوصول إلى الانترنت في المنزل	binary	Yes, no
18	Romantic-العلاقات العاطفية	binary	Yes, no
19	Famrel-طبيعة العلاقة الأسرية	numeric	1,2,3,4,5
20	Health-الوضع الصحي للطالب	numeric	1,2,3,4,5
21	Absences- عدد الغيابات الجامعية	numeric	0 To 93
22	G1-درجة الفصل الأول	numeric	0 To 20
23	G2-درجة الفصل الثاني	numeric	0 To 20
24	G3-الدرجة النهائية	nominal	A,B,C,D,F

لتهيئة البيانات تم تحويل ال(G3) Class Attribute من شكلها الرقمي إلى الاسمي(A,B,C,D,F) لتكون متوافقة مع الخوارزميات المختلفة الجدول(2).

الجدول رقم (2): توصيف قيم (G3) Class Attribute

A	B	C	D	F
Excellent/ Very good	Good	Satisfactory	Sufficient	Fail
جيد جدا/ممتاز	جيد	مُرْضٍ	كافٍ	رسوب
16-20	14-15	12-13	10-11	0-9

الإطار النظري:

1- التنقيب عن البيانات:

هي عملية بحث عن المعرفة من البيانات دون فرضيات مسبقة عما يمكن أن تكون هذه المعرفة (KDD) . كما ويعرف التنقيب في البيانات على أنه عملية تحليل كمية بيانات عادةً ما تكون كمية كبيرة لإيجاد علاقة منطقية تلخص البيانات

بطريقة جديدة تكون مفهومة ومفيدة لصاحب البيانات. يطلق اسم (نموذج) على العلاقات والبيانات الملخصة التي يتم الحصول عليها من التنقيب في البيانات .[5] [4]

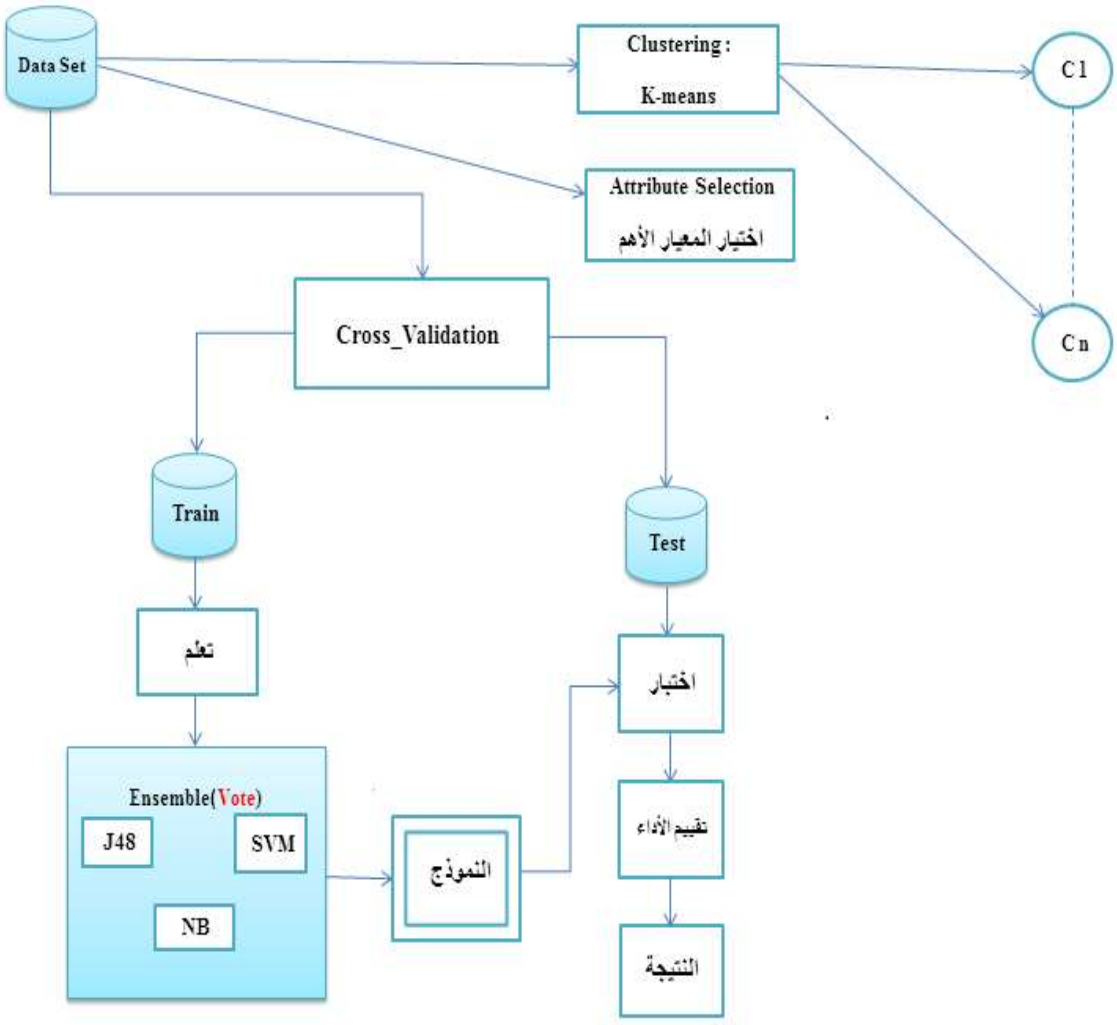
2-المخطط الصندوقي للنظام:

يبين الشكل(1) المخطط الصندوقي للنظام الذي تم تصميمه، وهو يهدف لتحسين الدقة التنبؤية الخاصة بأداء الطلاب ، و يبدأ من تحديد قاعدة البيانات وفقا لمجال الدراسة ليتم بعد ذلك العمل وفقا لثلاث اتجاهات:

أولا: إخضاع قاعدة البيانات لمجموعة من عمليات التقسيم بين عينات التدريب وعينات الاختبار(خوارزمية التدريب Cross-Validation) ، ومن ثم تمريرها إلى خوارزميات التصنيف المحددة والتي تقوم بمعالجة البيانات ، حيث تم تصميم مصنف دمج باستخدام تقنية Voting ، ومن ثم إعطاء النتائج وتقييمها.

ثانيا: انتقاء أهم المعايير والتي تمثل السمات الأكثر تأثيرا على نتيجة التنبؤ، استنادا لخوارزميات انتقاء المعايير.

ثالثا: تحليل بيانات الطلاب بالاستفادة من خوارزمية العنقدة K-means من خلال تقسيم مجموعة البيانات إلى مجموعات جزئية تدعى بالعناقيد (Clusters: $C_1, \dots, C_n$) بحيث تكون بيانات العنقود الواحد ذات خصائص مشتركة ومتشابهة فيما بينها أكثر من بيانات العناقيد الأخرى.



الشكل رقم (1):المخطط الصندوقي للنموذج التنبؤي الخاص ببيانات الطلاب.

3- تقنيات التنقيب عن البيانات:

يوجد العديد من تقنيات التنقيب عن البيانات منها:

3.1 العنقدة Clustering:

يمكن تعريف العنقدة بأنها عملية تقسيم مجموعة البيانات إلى مجموعات جزئية (Sub Sets) بحيث تكون بيانات المجموعة الواحدة ذات خصائص مشتركة ومتشابهة فيما بينها أكثر من بيانات المجموعات الأخرى.[6]

تعتبر عملية العنقدة أحد عمليات التنقيب الاستكشافي للمعطيات ويتم الاعتماد عليها في عدة مجالات منها التعلم الآلي (Machine Learning).

من أنواعها:

1-العنقدة القائمة على ارتباط المعطيات (Connectivity-based Clustering).

2- العنقدة القائمة على النقاط المركزية (Centroid-based Clustering).

ومن أهم خوارزميات العنقدة:

خوارزمية k-means:

وهي خوارزميه لجمع عدد من البيانات استنادا إلى خصائص وسمات هذه البيانات، وتتم عمليه التجميع من خلال تقليل المسافات بين البيانات ومركز التجمع (cluster centroid).

وتتم هذه الخوارزمية من خلال الخطوات التالية :

1. حساب إحداثيات مراكز التجميع.
2. حساب المسافة بين كل البيانات ومراكز التجميع.
3. تجميع البيانات وتنظيمها في مجموعات بناء على اقل المسافات بين المركز ونقاط البيانات.

إعادة تنفيذ الخطوات من 1 – 3 حتى الوصول إلى حاله الثبات.

3.2 التصنيفClassification:[7,8,9,10,11]

يستخدم التصنيف بشكل واسع في حل كثير من المشكلات من خلال تحليل مجموعة من البيانات و وضعها في شكل أصناف يمكن استخدامها فيما بعد لتصنيف البيانات المستقبلية.[7]

التحليل باستخدام خوارزميات التصنيف : هو عبارة عن عملية مكونة من خطوتين:

تسمى الأولى خطوة التعلم learn step حيث يتم فيها بناء نموذج التصنيف والخطوة الثانية هي خطوة التصنيف classification step حيث يتم فيها استخدام النموذج من أجل التنبؤ بالفئات أو السمات لبيانات محددة.

من أنواعها :

1- خوارزميات شجرة القرار.

2- خوارزميات الشبكات العصبية.

3- الخوارزميات المتعلقة بالنظريات الاحتمالية.

ومن أهم المصنفات :[8]

1- المصنف Naïve Bayes:

يستند هذا المصنف إلى نظرية بايز الاحتمالية (Bayes' theorem) القائمة على مبدأ الاحتمال الشرطي الذي يعمد لحساب احتمال وقوع أحد الأحداث الاحتمالية بناء على وقوع حدث آخر .

2- المصنف J48:

يندرج هذا المصنف ضمن خوارزميات أشجار القرار والتي على اختلاف أنواعها تشابه إلى حد ما خوارزمية تصنيف (Naive Bayes) من حيث اعتمادها على الاحتمالات الشرطية مع اختلاف رئيسي يكمن في أن هذه الخوارزمية تقوم بتوليد قواعد (Rules) لاستخدامها كجمل شرطية لتحديد السجلات والأحداث الاحتمالية بشكل عبارة شرطية (IF.....THEN).[9]

3- المصنف SVM:

أحد أقوى المصنفات التقليدية التي تعمل على إيجاد أفضل سطح فاصل بين بيانات التدريب وفق حالتين :

1- تصنيف خطي.

2- تصنيف غير خطي.

وهي خوارزمية قادرة على التعامل مع قواعد المعطيات ذات الأبعاد العالية بكفاءة مقارنة بعدد سجلات البيانات المتواجدة.[10]

ويتم تدريبها بالاعتماد على مبدأ تقليل الخطأ بين القيم المتوقعة والقيم الفعلية إلى الحد الأدنى لأن ذلك سيقلل الخطأ عند تطبيق الخوارزمية على الحالات الجديدة.

3.3 انتقاء المعايير Attributes Selection :

هي وسيلة تجهيز ضرورية لإزالة البيانات غير ذات صلة و زائدة عن الحاجة لتحقيق أفضل دقة تصنيف (اختيار أصغر مجموعة من المجموعة الأصلية التي تضم المعايير) .[11]

ربما تكون جميع المعايير مهمة في بعض الحالات ولكن من أجل بعض الأهداف يكتسب بعضها أهمية خاصة.

يساعد انتقاء المعايير في زيادة سرعة تنفيذ خوارزميات التصنيف ورفع أداء التنقيب.

ويوجد عدة خوارزميات للقيام بهذا العمل منها:

1- خوارزمية الترتيب(Ranker):

وهي أبسط خوارزميات انتقاء المعايير من خلال ترتيب المعايير بناء على التقييم الخاص بكل معيار على حدا ومن ثم توليد قائمة لجميع المعايير مرتبة حسب درجة تقييمها.

تستخدم هذه الخوارزمية مع عدد من مقيمات المعايير (Attributes Evaluators) نذكر منها :

- InfoGain : وهي طريقة لتقييم الميزات تعتمد على الانتروبيا، وتقيس أهمية السمة بواسطة قياس اكتساب المعلومة المحسوبة فيما يتعلق بالفئة المستهدفة class. يرمز للانتروبيا ب H في العلاقة التالية[11]:

$$InfoGain(Class,Attribute) = H(Class) - H(Class \mid Attribute) \quad (1)$$

- Correlation: هو من الأساليب الشائعة لاختيار السمات الأكثر صلة في مجموعة البيانات ، حيث يقيس قيم السمة عن طريق قياس الارتباط بينها وبين الأصناف classes، باستخدام معامل الارتباط بيرسون وهو معرف بالعلاقة[11]:

$$\text{Pearson's correlation coefficient} = \text{covariance}(X, Y) / (\text{stdv}(X) * \text{stdv}(Y)) \quad (2)$$

stdv():الانحراف المعياري. covariance() : التباين. X,Y:متغيرات.

4- دمج المصنفات Classifiers Ensembling:[12,13]

تعني الجمع بين مجموعة مصنفات للحصول على دقة للتصنيف أفضل من أي مصنف وحيد.

ولدينا طرق (طوبولوجيا) الدمج:[12]

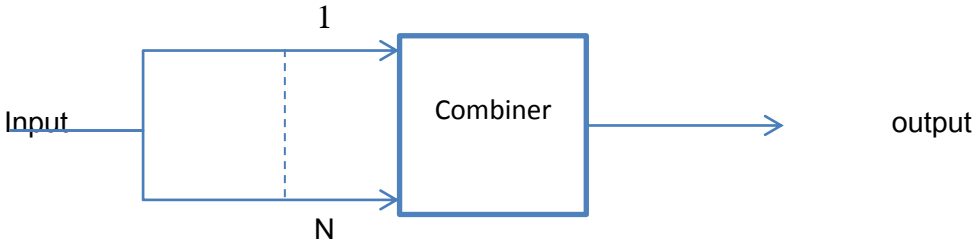
1- الطوبولوجيا التسلسليةCascading Topology :

حيث يتم استخدام خرج المصنف كدخل للمصنف التالي ، وبالتالي نحصل على التنبؤ من المصنف الأخير بالتسلسل.

الشكل رقم (2): طوبولوجيا المتتالية

2- طوبولوجيا الموازي Parallel Topology:

يتم جمع خرج جميع المصنفات بمكان واحد.



الشكل رقم (3): طوبولوجيا الموازي

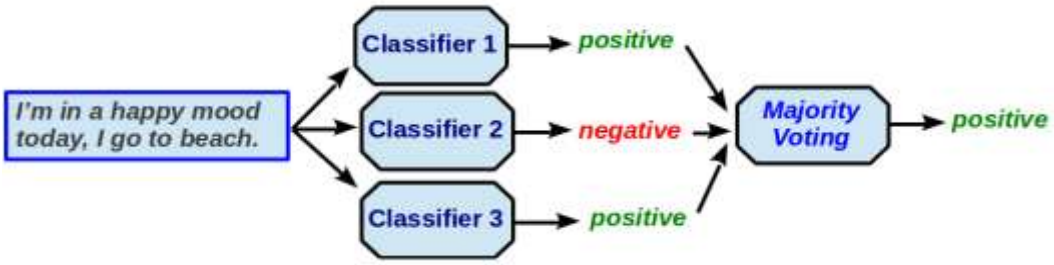
ونذكر مثال على ذلك تقنية Vote:

وهي أحد استراتيجيات اتخاذ القرار، تعمل من خلال الجمع بين التنبؤات من نماذج متعددة لتحسين أداء النموذج وتحقيق أداء أفضل من أي نموذج منفرد مستخدم بالمجموعة. حيث يتم تلخيص التنبؤات بكل تصنيف ويتم التوقع النهائي بأغلبية الأصوات (التوقعات هي تصويت غالبية النماذج المساهمة).

وهناك طرق لتوقع التصويت من أجل التصنيف:

1-Majority\hard Voting:

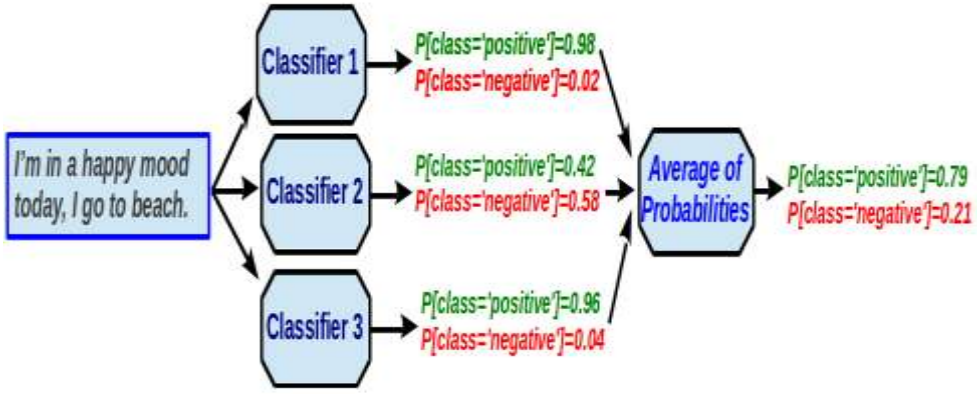
وتعني جمع التنبؤات لكل صنف من النماذج ، و يكون التوقع النهائي للصنف الحاصل على أكبر عدد من الأصوات.[13]



Majority/hard voting:(4) الشكل رقم

2-soft Voting:

بها يتم التنبؤ بناء على الاحتمالات المتوقعة للمصنف. وتتضمن جمع الاحتمالات المتوقعة لكل صنف (class label) من النماذج ، ويتم التنبؤ بالصنف النهائي عن طريق حساب متوسط الاحتمالات.[13]



الشكل رقم (5):Soft voting

5- معايير تقييم الأداء للمصنفات:

إن قياس الأداء يعتمد بشكل أساسي على مصفوفة الشك Confusion Matrix والتي تلخص عدد التسجيلات التي يتم التنبؤ بها بشكل صحيح أو خاطئ بواسطة نموذج التصنيف.[14]

الجدول رقم (3): مصفوفة الشك

Confusion Matrix		Predicted	
		Class1	Class2
Actual	Class1	TP	FN
	Class2	FP	TN

حيث نعرف:

- TN • عدد الحالات السلبية المتوقعة بشكل صحيح.
- FP • عدد الحالات السلبية المتوقعة بشكل خاطئ كحالات إيجابية.
- FN • عدد الحالات الإيجابية المتوقعة بشكل خاطئ.
- TP • عدد الحالات الإيجابية المتوقعة بشكل صحيح.

بارامترات الأداء التي تم اعتمادها في البحث :

1- دقة التصنيفAccuracy:

وتمثل عدد التسجيلات التي تم تصنيفها بشكل صحيح. تعطى علاقتها كالتالي[15]:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}} \qquad (3)$$

2- خطأ التصنيفClassification Error:

ويمثل عدد التسجيلات التي تم تصنيفها بشكل خاطئ. تعطى معادلته بالشكل التالي[15]:

$$\text{Classification Error} = 1 - \text{Accuracy} \qquad (4)$$

3- التحقيقPrecision :

نسبة التسجيلات التي تبين أنها موجبة صحيحة (TP) في المجموعة التي اعتبرها المصنف بأنها صنف موجب(TP+FP). تعطى علاقته كما يلي[15]:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad (5)$$

4-الاستدعاء Recall:

يقيس نسبة التسجيلات الموجبة التي تنبأ بها المصنف بشكل صحيح (TP) في المجموعة التي تبين أنها موجبة فعلا. يعطى بالمعادلة التالية[15]:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \qquad (6)$$

النتائج والمناقشة:

نتائج تطبيق الخوارزميات:

أولا: نتائج تطبيق دمج المصنفات مع تقنية vote :

تم تطبيق عدد من نماذج التنبؤ- مع 24 سمة من ضمنها class attribute (G3) و 1044 سجل - وذلك لدراسة خوارزميات التصنيف. في النموذج المدروس تم إدخال قاعدة البيانات للبرنامج ومن ثم إخضاعها لمجموعة من عمليات التقسيم بين عينات التدريب وعينات الاختبار ومن ثم تمريرها للخوارزميات المطلوبة.

تم دمج المصنفات(NB,J48 , SVM) وفق تقنية الدمج vote لتمثل مصنف جديد .

يبين تطبيق خوارزميات التصنيف أن مصنف الدمج قد حقق نتيجة أفضل من جميع الخوارزميات المستخدمة بالبحث، حيث يبين الجدول(4) تحليل نتائج الخوارزميات من أجل النموذج المدروس من قبل البرنامج weka، وتمت المقارنة بينها وفقا لمعايير أداء محددة (Recall ,Precision, Accuracy, Classification Error).

الجدول رقم (4): نتائج خوارزميات التصنيف

classifier	Class_label	Precision	Recall	Accuracy %	Classification Error %
Naïve Bayes	F	0.843	0.678	70.6897 %	29.3103%
	D	0.669	0.757		
	B	0.665	0.651		
	A	0.817	0.770		
	C	0.638	0.676		
	Weighted Avg	0.716	0.707		
J48	F	0.839	0.748	73.3716%	26.6284%
	D	0.696	0.829		
	B	0.661	0.657		
	A	0.850	0.746		
	C	0.693	0.639		
	Weighted Avg	0.739	0.734		
SVM	F	0.835	0.770		

	D	0.704	0.783	73.8506%	26.1494%
	B	0.704	0.651		
	A	0.885	0.754		
	C	0.658	0.704		
	Weighted Avg	0.745	0.739		
Ensemble (Vote: NB,J48, SVM)	F	0.860	0.748	74.8084%	25.1916%
	D	0.708	0.806		
	B	0.709	0.680		
	A	0.868	0.754		
	C	0.683	0.718		
	Weighted Avg	0.755	0.748		

من الجدول نجد أن مصنف الدمج قد تفوق بشكل عام على المصنفات الفردية ويظهر ذلك من خلال التقييم التالي:

أولا: التحقيق Precision الخاص بمصنف الدمج Ensemble(Vote) يملك 6 قيم ، خمسة منها خاصة بمستويات التصنيف (class labels: A,B,C,D,F) التابعة للواصفة الهدف G3، والقيمة الأخيرة تمثل المتوسط الموزون للقيم الخمس السابقة ، وتتم المقارنة بناء عليها ، وتبلغ (0.755) وهي أعلى من قيم Weighted Avg Precision الخاصة بالمصنفات الفردية الأخرى (NB:0.716 , J48:0.739, SVM:0.745) .

ثانيا: الاستدعاء Recall الخاص بمصنف محدد في الجدول يملك 6 قيم ، ويأخذ مصنف الدمج قيمة Weighted Avg Recall تبلغ (0.748) وتكون القيمة الأعلى بين المصنفات (NB:0.707 , J48:0.734, SVM:0.739).

ثالثا: دقة التنبؤ Accuracy لمصنف الدمج هي الأعلى في الجدول (74.8084 %) مقارنة ببقية المصنفات (NB: 70.6897%, J48: 73.3716%, SVM: 73.8506%).

رابعا: يملك مصنف الدمج أقل خطأ التصنيف Classification Error في الجدول (25.1916%) بالتالي فإن أداؤه هو الأفضل بين المصنفات (NB: 29.3103%,J48: 26.6284%, SVM: 26.1494%).

و تكون نسب التحسن التي قدمها مصنف الدمج بالنسبة لدقة التعرف Accuracy كما يلي:

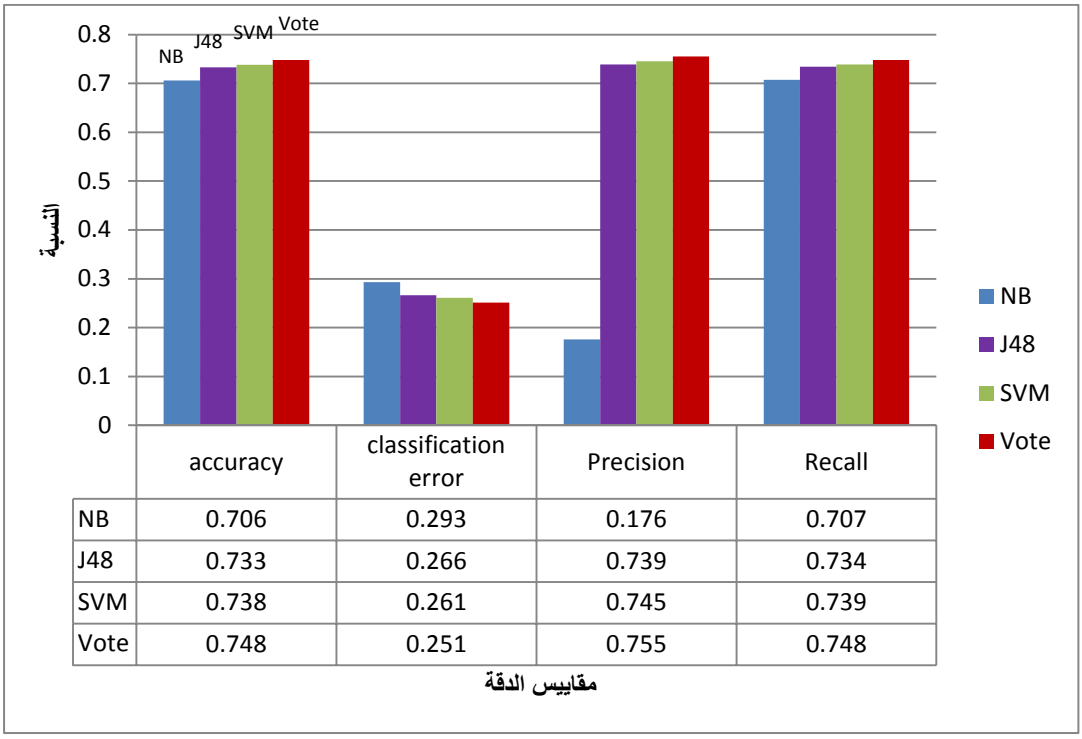
ازدادت دقة مصنف الدمج بمقدار:

1- (4.1187%) عن دقة المصنف NB كمصنف مفرد.

2- (1.4368%) عن دقة المصنف J48.

3- (0.9578%) عن دقة المصنف SVM.

ونرى ذلك أيضا في الشكل(6) :



	accuracy	classification error	Precision	Recall
NB	0.706	0.293	0.176	0.707
J48	0.733	0.266	0.739	0.734
SVM	0.738	0.261	0.745	0.739
Vote	0.748	0.251	0.755	0.748

مقاييس الدقة

الشكل رقم (6): المخطط البياني لمقاييس الدقة

من الشكل (6) نلاحظ أن الخط البياني الممثل لمصنف الدمج vote لديه أعلى accuracy عند النقطة 0.748 وأقل خطأ تصنيف عند النقطة 0.251 و أعلى precision عند 0.755 وأعلى recall عند 0.748 مقارنة بالمصنفات الأخرى المفردة.

ثانيا: نتائج تطبيق العنقدةClustering:

تم بناء نموذج للعنقدة باستخدام خوارزمية k-means بشكل مشابه لخوارزميات التصنيف مع مدخلات محددة وهي السمات التالية: (Sex,Medu,Fedu,Studytime,absences,G1,G2,G3).

وقمنا بتحديد قيمة البارامتر K=5 الخاص بالخوارزمية ويمثل عدد العناقيد. وتم دراسة توزع البيانات ضمن العناقيد.

وبذلك تكون المراكز النهائية final centroid :

```
Final cluster centroids:
                        Cluster#
Attribute    Full Data          0          1          2          3          4
               (1044.0)    (54.0)    (181.0)    (105.0)    (113.0)    (591.0)
=============================================================================
sex                  F          M          M          M          M          F
Medu            2.6034     3.3148     3.6519     1.4476     2.2212     2.4958
Fedu            2.3879     2.7778     3.3923     1.7714     1.5221     2.3198
studytime       1.9703     3.4444     1.6133     1.8286     1.0531     2.1455
absences        4.4349     2.6296     4.7238     4.1905     4.6903     4.5059
G1             11.2136    12.5741    11.9503    11.0571     9.1858    11.2792
G2             11.2462    12.7222         12    10.9524     9.2301    11.3181
```

الشكل(7):final clusters

يبين الشكل(7) ناتج تجزئة البيانات إلى خمس مجموعات(عناقيد)، كل عنقود ناتج يحتوي يحوي مجموعة معينة من الطلاب الذين تتشابه صفاتهم (من حيث المستوى والتقييمات).

تظهر محتويات العناقيد وفق قيم حقل معين ضمن هذه العناقيد ، مثلا نلاحظ ال cluster 0 تغلب عليه نسبة الأم والأب الحاصلين على التعليم الثانوي مع 3 ساعات دراسة تقريبا للطالب بالأسبوع ونسبة غياب أقل من العناقيد الأخرى مع درجات G1,G2 أعلى أيضا من العناقيد الأخرى لذلك تم تصنيف طلاب هذه المجموعة على أنهم الأفضل نتائجا.

من خلال إعادة التمعن بالنتائج يمكن لأصحاب القرار استخلاص :

أنه مع انخفاض نسبة الغياب بالدوام للطلاب مع مراعاة التحصيل العلمي المتقدم للأم والأب وزيادة عدد ساعات دراسة الطالب سنجد تحسنا بالأداء الأكاديمي للطالب. يمكن الاستفادة من هذه المعلومات لتوجيه كل مجموعة على حدى.

ثالثا :انتقاء المعايير Attribute selection:

بالاعتماد على الخوارزميات المتعددة التي تسمح باختيار المعايير ذات الأهمية في عمليات التصنيف ، تم إجراء عدد من عمليات اختيار أهم معايير التقييم وتم ترتيب النتائج بجدول.

تم اعتماد طريقتين لتقييم السمات هما:Correlation Attribute Evaluator, InfoGain Attribute Evaluator

باستخدام خوارزمية الترتيب Ranker التي تقوم بترتيب السمات بناء على التقييم الخاص بكل معيار على حدى.

الجدول رقم (5):

Attribute Evaluator	Search method	Attribute selection output
InfoGain Attribute Eval	Ranker	23,22,12,16,5,7,8,13,11,6,9,2,17 ,15,18,3,1,14,4,20,19,21,10
Correlation Attribute Eval	Ranker	23,22,12,16,5,6,11,10,21,13,2,17, 20,7,9,18,8,15,19,1,3,4,14

من تقاطع نتائج مقيمات المعايير يمكن التأكيد على أن G2,G1,Failures : 23,22,12 تعد أهم المعايير والتي يجب أخذها بعين الاعتبار في عدد من فعاليات التصنيف.

الاستنتاجات والتوصيات:

1- نجد أن مصنف الدمج الجديد أعطى أفضل دقة (تنبؤ) لنتائج الطلاب وفقا لعوامل مؤثرة مثل تعليم الأم وعمل الأب وسمات أخرى. وهذا يساعد متخذي القرار على تحسين أداء الطلاب من خلال اتخاذ خطوات استباقية تحد من المشاكل التي قد تواجههم، حيث إن إجراء العديد من التجارب في هذا المجال يمكّن المؤسسات التعليمية من تطوير الخطة المتبعة وزيادة الكفاءة .

2- من خلال خوارزمية ال k-means تم فرز الطلاب إلى خمس مجموعات ورؤية ما يميز كل مجموعة عن الأخرى . لكن لا يمكن القول بأنها أعطت أفضل توزع للبيانات ضمن العناقيد بسبب الاختيار العشوائي لمراكز العناقيد الابتدائية الذي تقوم به الخوارزمية ، وبسبب تكرار تنفيذها قبل الوصول للمراكز النهائية الذي قد يأخذ زمن تنفيذ كبير .

3- تمنح خوارزمية انتقاء المعايير Ranker القدرة على تحديد أهم معايير التقييم التي يجب أخذها بعين الاعتبار ضمن بيانات الطلاب الأكاديمية مثل العلامات وعدد مرات الرسوب.

يمكن التوسع في هذه الدراسة من خلال الاستفادة من أدوات وخوارزميات تنقيب أخرى لمعالجة هذا النوع من المعطيات وذلك للوصول لأفضل الخوارزميات التي يمكنها التعامل مع هذا النوع من المعطيات بأفضل النتائج.

References:

[1] KABAKCHIEVA ,D. *Predicting Student Performance by Using Data Mining Methods for Classification*. CYBERNETICS AND INFORMATION TECHNOLOGIES.Volume13,No 1,2013,61-72.

[2] ALMARABEH ,H . *Analysis of Students' Performance by Using Different Data Mining Classifiers*. International Journal of Modern Education and Computer Science(IJMECS). Vol.9, No.8,2017, pp.9-15.

[3] WITTEN,I. and FRANK,E. *Data Mining Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann (2000).

[4] ZAKI,M. and MEIRA JR,W . *Data mining and Analysis*, Cambridge University Press, 2014, Pp.606.

[5] WITTTEN,I., FRANK,E. and HALL,M. *Data Mining Practical Machine Learning Tools And Techniques*, Third Edition ,Elsevier, 2011, Pp.665.

[6] LESKOVEC,J. and RAJARAMAN,A . *clustering algorithms*, Stanford University, 2016, pp.46.

[7] HAN,J. and KAMBER ,M. *Data Mining: Concepts and Techniques* ,Morgan Kaufmann, 2006.nb

[8] Olson,D. and Delen,D. *Advanced data mining techniques*. Springer-Verlag, Berlin Heidelberg, 2008.svm

[9] SAWANT,T., POL,U. and PATANKAR,P. *EDUCATIONAL DATA MINING PREDICTION MODEL USING DECISION TREE ALGORITHM* , Journal of Emerging Technologies and Innovative Research (JETIR), Volume 6, Issue 5 ,May 2019.

[10] HSU,C., CHANG, C. and LIN,C. *A Practical Guide to Support Vector Classification*, Department of Computer Science National Taiwan University (2003).

[11] MEGHA , A .*Performance Analysis Of Different Feature Selection Methods In Intrusion Detection*, INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH, Vol 2,Issue 6 ,2013,225-231.

[12] KUNCHEVA,L. *Combining pattern classifiers methods and algorithms*, A JOHN WILEY & SONS, INC. PUBLICATION (2004).

[13] ZHOU, Z. *Ensemble Methods Foundations and Algorithms,* Chapman & Hall/CRC (2012).

[14] SINGH,R. and PAL,S. *Machine Learning Algorithms and Ensemble Technique to Improve Prediction of Students Performance,* International Journal of Advanced Trends in Computer Science and Engineering, Vol 9, No 3,2020,3970-3976.

[15] OLUKOYA,B. *Single Classifiers and Ensemble Approach for Predicting Student's Academic Performance,* International Journal of Research and Scientific Innovation (IJRSI) , Volume VII, Issue VI, June 2020 ,238–243.