التنقيب في بيانات القبول الجامعي في سورية

رند شعبان* د. زين جنيدي**

(الإيداع: 3 حزيران 2020 ، القبول: 20 تموز 2020)

الملخص:

لا يخفى على أحد أن عملية القبول الجامعي هي عملية استراتيجية على مستوى القطر، ومهمة جداً للمواطنين بمختلف فئاتهم، فهذه العملية تلقى الاهتمام من شرائح المجتمع كافة، الطلاب وأهاليهم والقطاع التعليمي بشكل عام. لذلك كان لا بد من الاهتمام بأدق تفاصيلها، ومحاولة تحسينها وتبسيطها. تعتمد مديرية تقانة المعلوماتية والاتصالات الإعلانات الصادرة عن وزارة التعليم العالي والبحث العلمي في سورية لتنفيذ المفاضلات الكترونياً، وتعتبر هذه الإعلانات بمثابة الناظم لعمل البرامج الالكترونية، هدفت هذه الدراسة إلى توظيف بيانات القبول الجامعي الخاصة بالفرع العلمي والمتراكمة خلال العشر سنوات السابقة، وتسخيرها لإنشاء نموذج تنبؤي يطبق على الطلاب المقبلين على القبول الجامعي، وكان من أهم نتائج الدراسة:

تصميم نموذج تنبؤي باستخدام منصة WEKA للتنقيب في البيانات باستخدام بيانات القبول الجامعي المتراكمة في السنوات العشر السابقة، عن طريق تطبيق ثلاثة خوارزميات هي (أشجار القرار – نايف بايز – المجاور الأقرب)، واختيار الخوارزمية ذات دقة التنبؤ الأعلى، وتصميم النموذج النهائي لتطبيقه على طلاب الفرع العلمي المقبلين على القبول الجامعي، وإرشادهم للجامعة والفرع المناسب لهم حسب بياناتهم، واستخدام النموذج لدعم القرار في وزارة التعليم العالي والبحث العلمي، بما يسهم في توزيع الطاقة الاستيعابية في الكليات والمعاهد في الجامعات الحكومية السورية كافة قبل البدء بأعمال القبول الجامعي. بلغت دقة التنبؤ للنموذج النهائي 66.4982%.

الكلمات المفتاحية: التنقيب في البيانات، منصة WEKA، القبول الجامعي، أشجار القرار، نايف بايز، الجار الأقرب، المتغير الهدف، التنبؤ، التصنيف، اختيار المتغيرات.

*ماجستير علوم الوب، الجامعة الافتراضية السورية

**دكتور مهندس في هندسة البرمجيات، نائب رئيس الجامعة الافتراضية السورية

Mining College Admission Data in Syria

Rand Chaaban* **Prof. Zein Juneidi****

(Received: 3 June 2020 , Accepted: 20 July 2020)

Abstract:

It is clear to everyone that college admission is a strategic process at the country level, and is very important for citizens of all categories. This process receives attention from all divisions of society, students, families, and the educational sector in general. It is necessary to pay attention to the smallest details, and try to improve and simplify it. The Directorate of Information and Communications Technology in the Ministry of Higher Education and Scientific Research in Syria adopts the announcements issued by the ministry to carry the admission process out programmatically, these announcements are considered as the regulator for the work of software programs. This study aims to use college admission data for the scientific branch gathered during the earlier ten years, to create a predictive model that applies to new students coming to college admission. The most important results of the study were:

This research aims to design a predictive model using the WEKA workbench. We utilized the college admission data gathered during the previous ten years. We applied three algorithms for data mining which are (Naïve Bayes – Decision Trees – Nearest Neighbor), then selected the best algorithm with the highest prediction accuracy, and designed the final model and applied it to students of the scientific branch coming to college admission. Students can use this model to get guidance to the proper university and college, according to their data. Furthermore, Ministry of Higher Education and Scientific Research can use the model for decision support to assign the proper seats for colleges and institutes in all Syrian Governmental universities, before starting college admission process. The prediction accuracy of the final model reached 66.4982%.

Keywords: Data Mining, Weka Workbench, College Admission, Decision Trees, Naïve Bayes, Nearest Neighbor, Class Attribute, Prediction, Classification, Attribute Selection.

* Master Of Web Science MWS, Syrian Virtual University

** Doctor engineer in software engineering, Syrian Virtual University Vice-President

1. مقدمة:

يعتبر التعليم العالي حاجة إنسانية وضرورة اقتصادية هامة لبناء شخصية الانسان وتنمية قدراته، لإسهامه في تربية الأجيال والنهوض بالإنسان وتنمية العقل البشري، باعتبار التعليم الجامعي يشغل قمة الهرم التعليمي الذي يوفر الأطر الضرورية التي يحتاجها المجتمع، إضافةً لذلك، لا يمكن إنكار أن قطاع التكنولوجيا والمعلومات أصبح مدخلاً أساسياً للعمل الحكومي بكل مكوناته، وتماشياً مع متطلبات المرحلة الحالية والمستقبلية لا بد من التوجه نحو مفهوم الحكومة الالكترونية بهدف تقديم الخدمات عن بعد وتعزيز وتسهيل التواصل مع المواطنين وتبسيط الإجراءات.

مما سبق كان لابد لوزارة التعليم العالي والبحث العلمي من مواكبة هذا التوجه ومشاركة باقي الجهات لتحويل تقديم الخدمات من الطريقة التقليدية إلى الطريقة الالكترونية، كما أن التعليم العالي يواجه تحدياتٍ صعبةً لم يعد بالإمكان مواجهتها بالطرق التقليدية، وهو مطلب مشروع وحيوي لكل من الطلبة والمجتمع، لرفع الكفاءة العلمية، وتزويد المجتمع بما يحتاجه من خريجين [1].

يتم القبول في المرحلة الجامعية بناءً على ركيزتين أساسيتين هما: السعة المكانية، وعلامات الطلبة في الثانوية العامة، لكن التخطيط للقبول الجامعي متفاوت من بلد إلى آخر من حيث المعايير المحدّدة للقبول، وكذلك من حيث الآلية المعمول بها، فهناك أنموذج القبول التقليدي الذي تتبعه فرنسا، وهو يعتمد على معدل الطلبة في الشهادة الثانوية، واجتياز اختبار في بعض التخصصات، وهناك أيضاً الأنموذج الأمريكي، إذ يكون لكل جامعة الحق في وضع الشروط والضوابط والمتطلبات الخاصة بها لانتقاء الطلبة، وهناك الأنموذج السويدي الذي يتم القبول فيه على مبدأ التكافؤ الاجتماعي، أي هو قبول مفتوح لمن تنطبق عليه شروط القبول [2]. تشترك جميع النماذج في اجتياز الامتحان النهائي للثانوية مع سجل الطلبة الدراسي في المدرسة الثانوية واجتياز اختبارات التحصيل التي يقرّها مجلس امتحان القبول في الكليات.

أما في البلدان العربية، فهناك بعض البلدان كالأردن الذي يتولى عملية قبول الطلبة فيه مكتبُ تنسيق موحد للقبول، يتلقى الطلبات، ويعلن أسماء المقبولين في قائمة واحدة، ثم يوزّع جميع المتقدمين على جميع الجامعات الرسمية وفق تسلسل رغبات الطلاب في طلب القبول على أن يراعى في القبول تسلسل المعدلات في شهادة الثانوية العامة والتخصص الذي يمكن للطالب قبوله فيه في ضوء المعدل الحاصل عليه. أما القبول في سورية فيقتصر على مجموع درجات الطالب في الثانوية العامة كمعيار للمنافسة مع الطلبة الآخرين وبناء عليه يتم قبوله للدراسة، ومما لا شك فيه أن مشكلة الالتحاق والقبول بالجامعات عموماً والسورية خصوصاً من المواضيع الأساسية في تطور البلد، إذ أن سياسة القبول في الجامعة تنبثق من السياسة العامة في توفير احتياجات البلد من القوى العاملة وربط سياسة القبول باحتياجات التنمية الاقتصادية والاجتماعية [3]، فالحاجة إلى التخطيط تتطلب إدراك التغيير في التعليم الجامعي والعالي على المستوى العالمي، من خلال ما سبق كانت فكرة هذه الدراسة لاستخدام التقنيات المعاصرة في دعم القرار وتحديد سياسات القبول الجامعي في سورية.

2. مشكلة الدراسة

لا يخفى على أحد أن عملية القبول الجامعي هي عملية استراتيجية على مستوى القطر، ومهمة جداً للمواطنين بمختلف فئاتهم، فهذه العملية تلقى الاهتمام من شرائح المجتمع كافة، الطلاب وأهاليهم والقطاع التعليمي بشكل عام. لذلك كان لا بد من الاهتمام بأدق تفاصيلها، ومحاولة تحسينها وتبسيطها.

تتراكم بيانات القبول والتسجيل في الجامعات والمعاهد السورية عاماً بعد عام، ويتم أرشفتها دون الاستفادة منها أو محاولة استخدامها لمواكبة التقنيات الحديثة في علوم البيانات بما يسهم في تحسين عملية القبول الجامعي وتطويرها.

3. أهداف الدراسة

تهدف هذه الدراسة إلى تبسيط عملية تسجيل الطلاب وإرشادهم لما يناسبهم بناءً على بيانات القبول من السنوات السابقة، من خلال استكشاف الأنماط في قاعدة البيانات التي تضم بيانات القبول الجامعي للأعوام السابقة، واستخدام منصة العمل WEKA في تصميم نماذج تنبؤية قادرة على توقع قبول الطلاب المقبلين على التسجيل في الجامعات والمعاهد السورية اعتماداً على بيانات الأعوام السابقة.

4. أهمية الدراسة

تنبع أهمية الدراسة من أن أنظمة الإرشاد (التوصية) هي أنظمة أساسية في الشركات المعتمدة على الوب والتي تعرض مجموعة ضخمة من الخيارات. ويهدف استخدام هذه الأنظمة لمساعدة الزبائن على إيجاد الاختيار الأمثل من بين المجموعة الضخمة من المعروضات. تقوم هذه الأنظمة على فهم طبيعة وسلوك المستخدمين ومن ثم إرشادهم إلى السلع التي تناسبهم، كما أن استخدام تطبيقات أنظمة الإرشاد الهجينة تمكِّن من تجاوز السلبيات في طرق الإرشاد الأساسية [4]. يمكن تطبيق هذه المبدأ على القبول الجامعي من خلال تحديد خصائص كل طالب واستنتاج ما يناسبه من بين مجموعة الرغبات المتاحة.

5. الدراسات السابقة

5-1 نظام دعم القرارات المتعلقة بالقبول الجامعي (المفاضلة) في الجمهورية العربية السورية

حاول الباحث إيجاد حل لمشكلة القبول التي تصطدم بها الجامعات السورية سنوياً من خلال رؤية جديدة باستخدام تقنيات الذكاء الصنعي والتفكير بصياغة منهج عمل مناسب لهذه المعطيات، ومحاولة تقديم اقتراح تحسينات من شأنها تطوير سياسة القبول الجامعي في ضوء تجارب بعض الدول العربية والعالمية المتقدمة في هذا المجال، من خلال الإمكانات التي يقدمها التنقيب في المعطيات الأكاديمية في استنباط معلومات من بيانات الطلبة الخاصة بالبطاقة الالكترونية. خلص الباحث على تقديم نموذج مقترح لحساب التوقعات الخاصة بالكليات الجامعية ونموذج آخر لتوقع أعداد المقبولين في الجامعات [5].

5-2 تصميم وتشغيل نظام إرشاد هجين للتنبؤ بالقبول الجامعي

يتضمن البحث نظام قبول جامعي جديد باستخدام نظام إرشاد هجين يعتمد على التنقيب في البيانات وقواعد استكشاف المعارف، لمعالجة مشكلة التنبؤ بالقبولات الجامعية. وذلك بسبب العدد الضخم من الطلاب الراغبين بالتسجيل في الجامعات كل سنة. يتألف النظام الجديد من نظامين يعملان معاً إلى جانب نظام التنبؤ. يقوم أول نظام بتحديد الطلاب المقبولين في السنة التحضيرية. بينما يقوم الآخر بتحديد الطلاب في الاختصاصات التي تلي السنة التحضيرية. بينما يقوم نظام التنبؤ باستخدام المعدل المئوي للطالب للتنبؤ بالكلية الأنسب. يقوم النظام بتحليل المزايا الأكاديمية للطالب وخلفيته وسجلاته مع معايير القبول الجامعي. ثم يقوم بالتنبؤ بالجامعة والكلية الأقرب لرغبة الطالب. تم تصميم نظام أولي واختباره على بيانات حقيقية لجامعة الملك عبد العزيز. بالإضافة إلى التنبؤ الدقيق يتميز النظام بالمرونة والتكيف [6].

5-3 نظام قبول معتمد على الوب للمرحلة المتقدمة في المدارس الخاصة

أجريت هذه الدراسة أجريت في تنزانيا، حيث تم إلقاء الضوء على مختلف التحديات التي تواجه عملية القبول. وتم توزيع استبيانات لجمع المعلومات من المستخدمين المحتملين للنظام. لمعرفة مدى رضاهم عن آلية القبول التقليدية المتبعة. بعد ذلك تم تحليل نتائج الاستبيان وتبين أن القبول الجامعي يتم بشكل ورقي. وهذا النظام اليدوي له مشاكله العديدة التي تتضمن صعوبة إيجاد الرغبة المناسبة للطالب عدا عن هدر الوقت والجهد. يقوم هذا البحث على تصميم نظام معالجة القبول المركزي في تنزانيا [7].

4-5: نظام إرشاد فعال باستخدام خوارزميات التصفية وخوارزميات استكشاف الأنماط

بينت الدراسة أهمية أنظمة الإرشاد ودعم القرار في تسهيل البحث عن معلومة ضمن الحجوم الهائلة للمعلومات على الانترنت والنمو المطّرد للبيانات ما يجعل عملية البحث عن معلومة عملية معقدة. تساعد أنظمة الإرشاد المستخدمين على الوصول إلى المعلومة بدقة. فنظام الإرشاد هو واحد من التقنيات التي تسمح بتخصيص الوب، حيث يوصي بصفحات وب محددة للمستخدم بناءً على سجل التصفح السابق. في هذا البحث يعتبر التنقيب في سجل الوب وتصفية المعلومات هي المصادر الرئيسية التي يعتمد عليها نظام الإرشاد [8].

5-5 القبول المركزي للكليات الهندسية في الهند

اعتمد الباحثون على الخوارزمية المعروفة DA (Differed Acceptance) في تخصيص المقاعد لأكثر من 500 برنامج دراسي موزعة على 80 جامعة ومعهد تقني في الهند [9]. يتقدم سنوياً إلى هذه المعاهد حوالي 1.2 مليون طالب ويقبل منهم ما نسبته 1% فقط، ومع ذلك تتبقى نسبة 6% من المقاعد شاغرة، أحد أسباب ذلك هو أن القبول في هذه المعاهد كان يتم بشكل منفصل بين المعاهد التكنولوجية (IITs) والمعاهد غير التكنولوجية (non-IITs) بين العامين 1960-2014، حيث يمكن للطالب الحصول على قبول في هذه المعاهد وبنفس الوقت قد يحصل على قبول في مجموعة المعاهد غير التكنولوجية وهذا يعني أن الطالب سيختار أحد هذين القبولين ويترك المقعد الآخر شاغراً. قام الباحثون بعدة تعديلات على عمل الخوارزمية من بينها نظام إرشاد مصمم بدقة ليقوم بدمج المعايير المشتركة بين عدة برامج دراسية، بالإضافة إلى منهجية لتخصيص المقاعد الشاغرة بما يضمن مبدأ تكافؤ الفرص والعدالة في التوزيع والشفافية. تم استخدام هذا النظام منذ عام 2015 وحتى الآن، مع إضافة بعض التحسينات بشكل مستمر وقد أثبت نجاحاً في تخفيض عدد المقاعد الشاغرة. حيث يتم تخصيص كل متقدم بمقعد واحد فقط حسب ترتيب الرغبات التي اختارها المتقدم ومستوى المتقدم في امتحانات القبول الوطنية المشتركة (JEE)[1].

6. خلاصة الأبحاث السابقة وموقع الدراسة الحالية منها

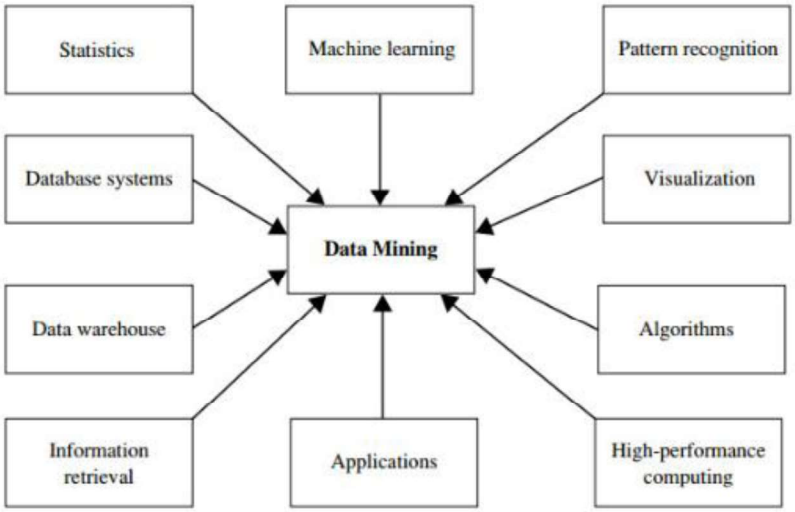
ركزت الأبحاث المنشورة على إنشاء نظام إرشاد للطلاب يقدم المشورة من خلال الاعتماد على البيانات من السنوات السابقة. بالإضافة إلى توظيف تقنيات الذكاء الصنعي والتنقيب في البيانات لإعطاء نتائج تحليلية دقيقة. وبينت مختلف الدراسات تنوع أنظمة القبول الجامعي واختلاف معاييرها من بلد لآخر، ولا شك أن كل بلد يشكل حالة خاصة بمفرده تندرج تحت سياق عام مشترك بين البلدان ولكن تختلف التفاصيل والمتغيرات الجزئية والوقائع الفريدة لكل بلد، وهذا ينطبق على حالة نظام القبول في سورية، لذلك وجب دراسته بشكل متعمق وإفراده عن بقية أنظمة القبول. لا شك أن السياق العام وهو نظام الإرشاد تم تطبيقه في مختلف الأنظمة إلا أنه لا يوجد تطبيق لنظام الإرشاد في سورية حتى الآن.

7. التنقيب في البيانات: دوره ومراحله

يشير مصطلح التنقيب في البيانات إلى استخراج المعرفة من كميات كبيرة من البيانات، وهي عملية البحث الآلي ضمن حجوم كبيرة من البيانات عن الأنماط، وهي تعتمد على تقنيات حسابية من علوم الإحصاء، استرجاع البيانات، التعلم الآلي وتمييز الأنماط، تبين الخطوات التالية مراحل التنقيب في البيانات:

[1] JEE: Joint Entrance Examination امتحان القبول المشترك

1. تنظيف البيانات: وهي الخطوة الأولى، ويتم فيها إزالة البيانات التي تتضمن سجلات فارغة أو تالفة.
2. تكامل البيانات: تجميع البيانات وتنسيقها ضمن بنية موحدة. عادةً لا تقدم مصادر المعلومات المتخلفة بنى موحدة أو تفسيرات للبيانات، ومن هنا تأتي أهمية هذه الخطوة.
3. اختيار البيانات: بالرغم من ذلك، ليست كل البيانات التي تم تجميعها بنفس الأهمية، تسمح هذه الخطوة باختيار البيانات ذات الصلة فقط.
4. تحويل البيانات: لا تزال البيانات التي اجتازت مرحلة التنظيف غير مهيأة للبدء بالتنقيب، لذلك يجب تحويلها إلى تنسيق مناسب لخوارزمية التنقيب.
5. التنقيب في البيانات: في هذه الخطوة يمكن تطبيق خوارزميات متنوعة على البيانات لاستكشاف المعرفة المحتملة المخفية في البيانات.
6. تقييم الأنماط: يجب تقييم أهمية نتائج التنقيب في البيانات، فقد لا تكون جميع النتائج ذات أهمية.
7. عرض المعارف: يتم اختيار النتائج التي تم تقييمها على أنها الأكثر أهمية من أجل تقديمها وتصويرها بأفضل شكل يمكن فهمه.



الشكل رقم (1): التقنيات المتنوعة للتنقيب في البيانات

8. منصة WEKA للتنقيب في البيانات

أظهرت التجارب العديدة في هذا المجال أنه لا يوجد نظام واحد للتنقيب في البيانات مناسب لجميع حالات العمل. حيث أن التنقيب في البيانات هو علم تجريبي [10]. منصة WEKA هي مجموعة من خوارزميات التعلم الآلي وأدوات معالجة البيانات، تتضمن تقريباً جميع خوارزميات التنقيب في البيانات، وهي مصممة بحيث يمكن للمستخدمين تجربة تطبيق الخوارزميات الموجودة على مجموعات البيانات بشكل مرن، وهي تقدم دعم كامل للتنقيب في البيانات، بما في ذلك تحضير بيانات الدخل واستعراضها بشكل رسومي والتقييم الإحصائي لنتائج التطبيق بشكل مفصل ومرئي.

تم تطوير منصة WEKA في جامعة ويكاتو في نيوزيلندا، وهي اختصار للعبارة التالية (Waikato Environment for Knowledge Analysis) وتعني بيئة ويكاتو لتحليل المعارف. تمت برمجة المنصة باستخدام لغة جافا وهو نظام مفتوح المصدر ومرخص للاستخدام العمومي. وهو قابل للعمل على أي نظام تشغيل.

بالإضافة إلى خوارزميات التنقيب الشهيرة، تتضمن منصة WEKA مجموعة واسعة من أدوات المعالجة المسبقة، يتم الوصول إليها من خلال واجهة مشتركة بحيث يمكن للمستخدم مقارنة الطرق المختلفة وتحديد الأساليب الأكثر ملائمةً للمشكلة المطروحة.

تقدم منصة WEKA خوارزميات التعلم العديدة التي من الممكن تطبيقها بسهولة على أي قاعدة بيانات. وهي تتضمن أيضاً مجموعة من الأدوات للمعالجة المسبقة للبيانات، وتحليل نتائج التنقيب ومقاييس الأداء دون كتابة أي سطر برمجي. ومن بين خوارزميات التنقيب في البيانات التي تقدمها منصة WEKA: تحليل الانحدار، التصنيف، العنقدة، قواعد الارتباط واختيار المتغيرات ذات الصلة.

تمتلك منصة WEKA واجهة المستخدم الرسومية الرئيسية وتسمى المتصفح (Explorer) وهي القسم الأكثر استخداماً في المنصة، يؤمن المتصفح وصولاً لكل الميزات بطريقة الاختيار من القوائم وملء النماذج، ويحتوي على ست لوحات مختلفة على شكل تبويبات في أعلى النافذة، يختص كل منها بفئة معينة من مهام التنقيب في البيانات التي توفرها المنصة.

تستطيع منصة WEKA التعامل مع البيانات بعد تجميعها في جدول واحد، والذي يمكن استيراده من أي منصة إدارة قواعد البيانات بصيغ مختلفة منها ملفات (Comma separated values: csv) أو ملفات (JavaScript object notation: json) كما يمكن استيراد ملف من خلال مورد على الانترنت URL. عند استيراد البيانات يمكن تطبيق خوارزميات التعلم الآلي المختلفة، تمتلك معظم الخوارزميات بارامترات قابلة للمعايرة، وتمكّن منصة WEKA من معايرة هذه البارامترات لتحسين دقة النموذج بشكل تجريبي.

9. الجانب التطبيقي

9-1 تحضير البيانات

تهدف العملية إلى فحص البيانات وتنظيفها وتنسيقها تحضيراً للخطوة التالية، وتنقسم إلى قسمين:

9-1-1 تنظيف البيانات Data Cleaning

1. حذف الشهادات ذات المصدر الأجنبي: تم حذف جميع سجلات الطلاب الحاصلين على الشهادة الثانوية من دول أخرى وذلك بهدف توحيد المعيار المطبق على الطلاب. حيث أن الطلاب الحاصلين على شهادات ثانوية من دول أخرى يخضعون لمعايير مختلفة باختلاف الدول وأنواع الاختبارات التي يجرونها. لذلك تم اصطفاء الطلاب الحاصلين على الشهادات الثانوية من سورية حيث أنهم يخضعون لنفس الاختبارات ونفس المناهج الثانوية.
2. تم حذف بعض السجلات التي تحوي على قبول في الكليات التي تم إغلاقها مثل قسم الاقتصاد المنزلي في كلية التربية وكذلك الأمر بالنسبة للمعاهد التابعة لوزارة التربية، حيث أصبح قبول الطلاب في هذه المعاهد يصدر عن وزارة التربية.
3. تم دمج الطلاب المقبولين في كليات (الطب البشري – طب الأسنان – الصيدلة) قبل العام 2015 تحت مسمى واحد (الكليات الطبية) وذلك بسبب اعتماد السنة التحضيرية للكليات الطبية بدءاً من ذاك العام.
4. تم حذف السجلات الخاصة بالطلاب الذين رفضت جميع رغباتهم بنتيجة المفاضلة وذلك لعدم تأثير هذه البيانات على الدراسة.
5. بسبب وجود أخطاء في الإدخال تم الاحتفاظ بالسجلات التي تتراوح قيم العمر فيها بين 17 و45 وحذف بقية السجلات.

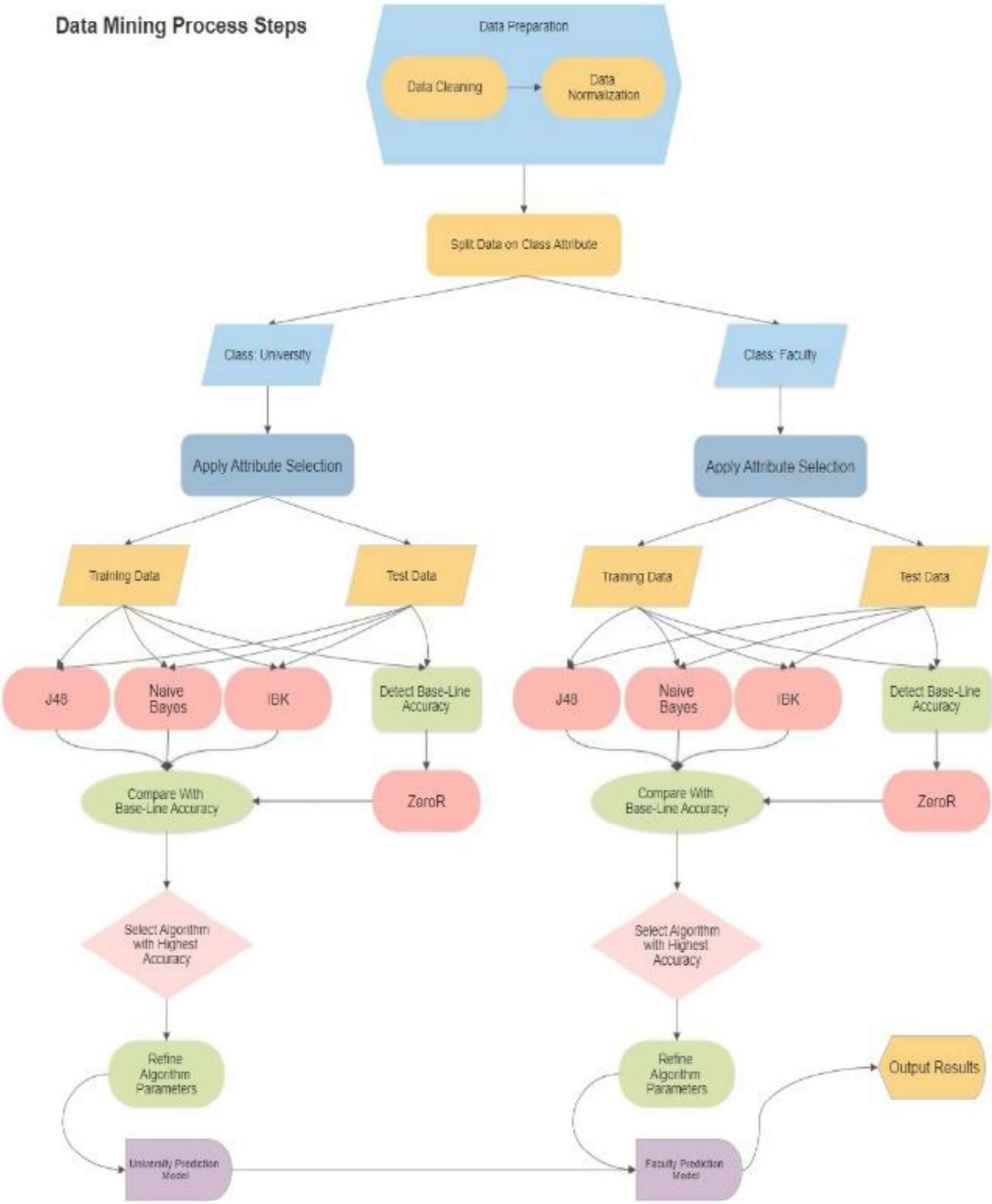
9-1-2 تنسيق البيانات Data Normalization

1. في العام 2013 قامت وزارة التربية بتغيير نظام الدرجات في الثانوية العامة، حيث تمت مضاعفة الدرجات في جميع المواد عشرة أضعاف مما كانت عليه سابقاً، فمثلاً أصبحت العلامة القصوى لمادة الرياضيات تحسب من 600 درجة بدلاً من 60، ولذلك وجب مضاعفة جميع الدرجات قبل العام 2013 بنفس القيمة بهدف توحيد التقييس على كافة السنوات التي يتم دراستها.
2. بسبب اختلاف الحد الأقصى لمجموع الدرجات لكل مادة، تم نسب علامات الطلاب في جميع المواد إلى 100، بحيث يصبح الحد الأقصى موحداً لجميع المواد.
3. بعد عملية النسب، تبين وجود عدة خانات بعد الفاصلة العشرية، لذلك تم تقريب العلامات إلى رقمين بعد الفاصلة، ومن ثم ضرب الناتج ب 100 للتخلص من الفاصلة العشرية. وبذلك يصبح الحد الأقصى لمجموع الدرجات هو 10000، كذلك الأمر بالنسبة للمجموع العام.
4. تضم البيانات قبول الطلاب في الفرع العلمي – المفاضلة العامة، وعند تدقيق البيانات تبين وجود اختلافات في ترميز قبول الطلاب مع اختلاف تسمية نفس القبول في بعض الحالات من سنة إلى أخرى، لذلك وجب توحيد التسميات والرموز.

تم تجميع الاختصاصات ضمن ستة مجموعات هي: العلوم الطبية – العلوم الهندسية – العلوم التطبيقية – العلوم النظرية – المعاهد التطبيقية – المعاهد التقانية.

9-2 تقسيم البيانات إلى مجموعتين حسب المتغير الهدف

تفترض معظم خوارزميات التنقيب في البيانات وجود متغير هدف واحد، وهو المتغير الذي سيتم التنبؤ به للحالات الجديدة أو ما يمثل الصنف. ويكون عادة آخر متغير في مجموعة البيانات كما تفترض منصة WEKA آلياً. تتمثل المشكلة هنا بوجود متغيرين يجب التنبؤ بقيمهما، وهما الجامعة والكلية. تم إنشاء نسخة ثانية من مجموعة البيانات بحيث يصبح لدينا نسختان تضم النسخة الأولى – إضافةً للمتغيرات الأساسية – المتغير الهدف الأول وهو الجامعة وتهمل المتغير الثاني وهو الكلية/المعهد، وتضم المجموعة الثانية المتغير الهدف الثاني وهو الكلية/المعهد وتهمل الأول أي الجامعة بعكس المجموعة الأولى.



الشكل رقم (2): خطوات العمل

يتوجب في هذه الحالة إنشاء نموذج تنبؤي مستقل لمجموعة البيانات الأولى للتنبؤ بالجامعة، وإنشاء نموذج تنبؤي لكل جامعة على حدة للتنبؤ بالكلية/المعهد. تم إهمال المتغير DiffYear الذي يمثل عام القبول، وذلك لاستخدامه في تقسيم البيانات في الخطوة التالية. تتألف المجموعة الأولى من البيانات من جميع المتغيرات المتاحة كمتغيرات الدخل والمتغير الهدف هو الجامعة، كما هو موضح بالجدول رقم/1/:

الجدول رقم (1): مجموعات البيانات والمتغير الهدف

Dataset 1		Dataset 2	
Attribute	**Type**	**Attribute**	**Type**
CityName	**Input**	CityName	**Input**
Gender	**Input**	Gender	**Input**
Age	**Input**	Age	**Input**
TotalScore	**Input**	TotalScore	**Input**
Math	**Input**	Math	**Input**
Physics	**Input**	Physics	**Input**
Chemistry	**Input**	Chemistry	**Input**
Biology	**Input**	Biology	**Input**
English	**Input**	English	**Input**
French	**Input**	French	**Input**
Arabic	**Input**	Arabic	**Input**
Nationalism	**Input**	Nationalism	**Input**
Religion	**Input**	Religion	**Input**
University	**Predict**	Acceptance	**Predict**

أما المجموعة الثانية فهي نسخة طبق الأصل من المجموعة الأولى باستثناء المتغير الهدف وهو الكلية/المعهد.

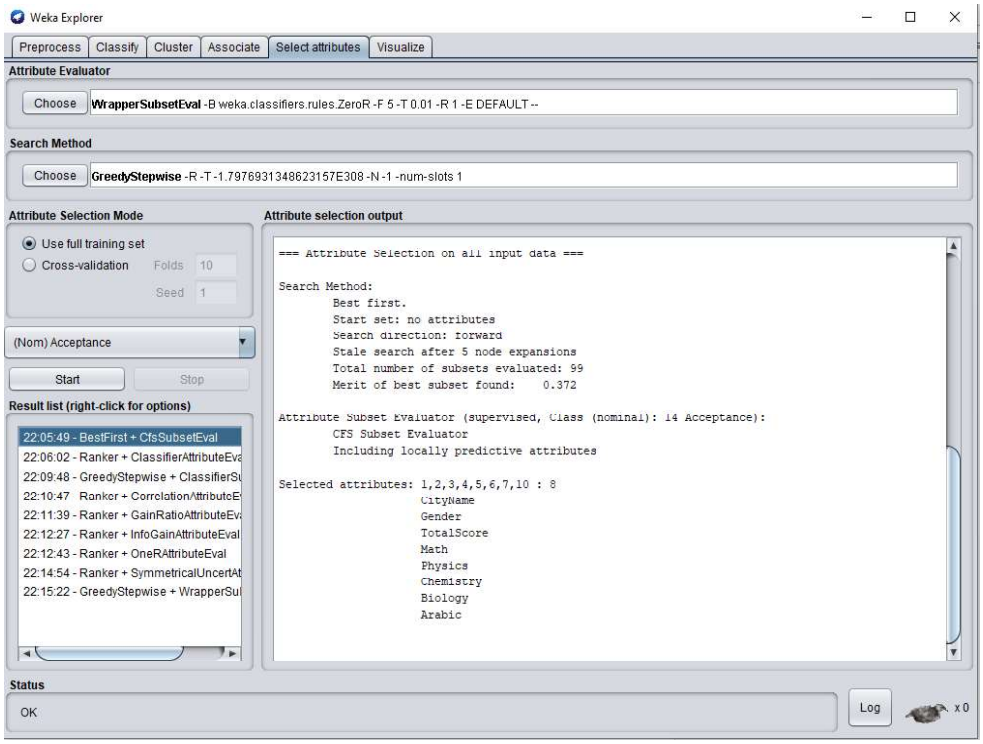
3-9 تطبيق خوارزميات اختيار المتغيرات:

تم تصميم معظم خوارزميات التعلم الآلي لتحديد المتغيرات الأنسب لاستخدامها في اتخاذ القرارات. على سبيل المثال، تختار طرق أشجار القرار المتغير الأكثر تأثيراً لتقسيم فروع الشجرة بناءً على قيم هذا المتغير في كل عقدة، ولا تقوم هذه الطرق – نظرياً – باختيار المتغيرات غير ذات الصلة أو التي ليس لها تأثير في النتائج، إن تأثير هذه العملية في نتائج الخوارزمية يجب أن يكون واضحاً في زيادة دقة التنبؤ، لأن إدخال المتغيرات غير ذات الصلة في الخوارزمية قد يؤدي إلى ظهور تشويش في نتائج نظام التعلم الآلي [10]. بسبب التأثير السلبي للمتغيرات غير ذات الصلة في معظم خوارزميات التعلم الآلي، فإنه من الشائع تطبيق خوارزمية اختيار المتغيرات لتحديد المتغيرات ذات الصلة وتحييد بقية المتغيرات، كمرحلة أولية تسبق عملية التنقيب. وهذا بالتأكيد سيحسن من أداء خوارزميات التعلم ويسرع من وتيرة العمل.

تعمل خوارزميات اختيار المتغيرات عادةً عن طريق البحث في فضاء مجموعات المتغيرات الجزئية وتقييم كل منها، ويتم ذلك بجمع واحدة من خوارزميات تقييم المتغيرات مع واحدة من طرق البحث.

لاختيار المتغيرات الأنسب، توفر منصة WEKA تبويباً مستقلاً للخوارزميات الخاصة بذلك، يبين الشكل /3/ النافذة الخاصة باختيار المتغيرات، حيث يمكن اختيار خوارزمية تقييم المتغيرات وطريقة البحث.

تم تطبيق عدة خوارزميات لاختيار المتغيرات الأكثر تأثيراً على المتغير الهدف، تقوم منصة WEKA بالاختيار الآلي لطريقة البحث المرتبطة مع الخوارزمية، ولدى معاينة النتائج تم اختيار الخوارزمية cfsSubsetEval لاختيار المتغيرات مع طريقة البحث GreedyStepWise وأعطت النتائج الموضحة في الشكل /3/:



الشكل رقم (3): نتائج تطبيق خوارزمية اختيار المتغيرات

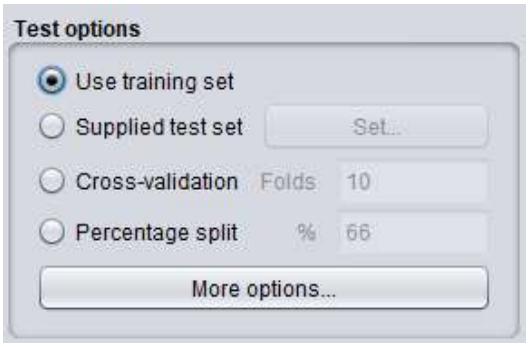
يتضح أن المتغيرات الأكثر تأثيراً على المتغير الهدف وعددها ثمانية متغيرات هي: مصدر الشهادة – الجنس – المجموع – الرياضيات – الفيزياء – الكيمياء – العلوم الطبيعية – اللغة العربية. لذلك تم اعتمادها لتشكل دخلاً لخوارزمية التنقيب وتم إهمال بقية المتغيرات وهي: العمر – اللغة الإنكليزية – اللغة الفرنسية – التربية القومية – التربية الدينية. تشكل هذه النتائج توافق منطقي مع المتوقَّع حيث يعتمد القبول في معظم الكليات والمعاهد على المجموع الكلي ويتم تثقيل علامات بعض المواد الاختصاصية المشمولة بنتائج الخوارزمية. أما المتغيرات التي تم إهمالها فهي متناقضة مع ما ذُكر إلى حدَّ ما، كما أن متغير العمر لا يشكل تأثيراً كبيراً خصوصاً أن مجال القيم التي يمكن ان يأخذها يعتبر ضيقاً، حيث تتركز نسبة 95% من المقبولين ضمن الفئة العمرية (17–18–19) عاماً.

4-9 تقسيم البيانات في كل مجموعة

إن هدف هذه الخطوة هو تقسيم البيانات قبل تطبيق خوارزميات التنقيب في البيانات، تقوم خوارزميات التنقيب بتقسيم بيانات الدخل إلى قسمين، يسمى القسم الأول بيانات التدريب Training Data والقسم الثاني بيانات الاختبار Test Data.

الشكل رقم (1): بيانات التدريب وبيانات الاختبار

لا يوجد فرق بين بيانات التدريب وبيانات الاختبار من حيث البنية، لكن يختلف استخدام كل منهما أثناء بناء النموذج، حيث يتم إنشاء النموذج باستخدام بيانات التدريب، بعد ذلك يتم اختبار النموذج باستخدام بيانات الاختبار، في حال كانت نتائج الاختبار قريبة من المتوقع، يتم اعتماد النموذج لأغراض التنبؤ وإلا يتم إعادة بناء النموذج بطريقة مختلفة.



الشكل رقم (2): خيارات تقسيم البيانات في منصة WEKA

توفر منصة WEKA عدة خيارات لإنجاز ذلك:

1. Use Training Set: في هذا الخيار يتم استخدام كامل مجموعة البيانات للتدريب ومن ثم استخدام نفس المجموعة للاختبار، وهذه في الحقيقة ليست فكرة جيدة لأن ذلك يعني اختبار النموذج على نفس البيانات التي استُخدمت في إنشاءه، قد تكون دقة النموذج مرتفعة لكن غير منطقية.
2. Supplied Test set: يتيح هذا الخيار تزويد الخوارزمية بمجموعة بيانات مستقلة لاستخدامها في الاختبار، قد تكون هذه البيانات جزءاً من البيانات المتوفرة كما في حالتنا هذه أو قد تكون بيانات من مصدر آخر.
3. Cross-validation: يقوم هذا الخيار على تقسيم مجموعة البيانات المتوفرة إلى عدد من الأقسام المتساوية، ومن ثم يتم تدريب النموذج على عدد هذه الأقسام، في كل دورة يتم تحييد أحد الأقسام للاختبار وإجراء التدريب باستخدام بقية الأقسام، بحيث يتم إعادة العملية لكل قسم مرة واحدة. من الواضح أن هذه الطريقة تستهلك وقتاً أكبر، لكنها تعطي نتائج أفضل. تختار منصة WEKA هذا الخيار افتراضياً ويتم تقسيم البيانات إلى عشرة أجزاء، ولكن يمكن تغيير عدد الأقسام وقياس النتائج تجريبياً.
4. Percentage split: يتيح هذا الخيار تقسيم بيانات الدخل إلى قسمين بنسبة مئوية لاستخدام القسم الأول للتدريب والقسم الآخر للاختبار، تعطي منصة WEKA نسبة 66% افتراضياً لبيانات التدريب ولكن يمكن تغيير هذه النسبة وقياس النتائج تجريبياً.

تم تقسيم بيانات الدخل المتوفرة باستخدام الخيار الثاني، حيث تم تخصيص بيانات الأعوام 2009 وحتى عام 2016 للتدريب وبيانات العام 2017 للاختبار. تم تحييد بيانات العام 2018 لتطبيق النموذج النهائي عليها.

الجدول رقم (1): تقسيم بيانات الدخل حسب متغير عام القبول

مجموعة البيانات وفق عام القبول	الاستخدام	النسبة من الإجمالي
2008 ← 2016	بيانات التدريب	%80
2017	بيانات الاختبار	%10
2018	تقييم التنبؤ	%10

9-5 تطبيق خوارزميات التنقيب في البيانات

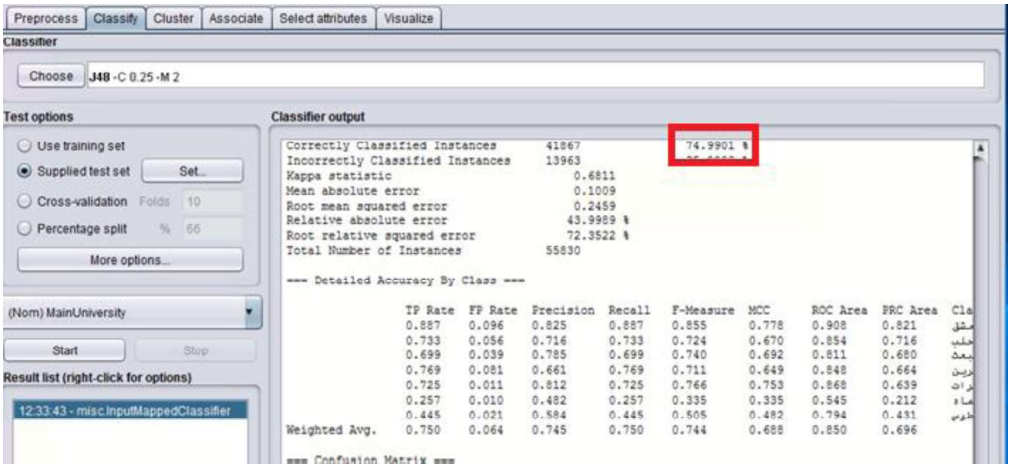
بعد تنظيف البيانات وإعدادها للعمل، تم تطبيق عدة خوارزميات من خلال منصة WEKA بما يتناسب مع المشكلة المطروحة وبتسلسل منطقي للحصول على نتائج فعالة وقابلة للاستخدام في نظام الإرشاد[1].

9-5-1 أشجار القرار Decision Trees

تستخدم أشجار القرار بشكل شائع في التنقيب في البيانات، تعتمد فكرة بناء أشجار القرار على اختيار أحد المتغيرات كجذر للشجرة وإنشاء فرع لكل قيمة ممكنة لهذا المتغير، ومن ثم تكرار هذه العملية بطريقة عودية لكل فرع، باستخدام القيم الممكنة في هذا الفرع فقط. يتوقف بناء الفروع من عقدة ما عندما يكون لكل القيم نفس التصنيف ضمن هذه العقدة. لكن الأهم هو تحديد المتغير الذي يجب البدء بالتقسيم بناءً على قيمه، يتم ذلك من خلال قياس كمية المعلومات الممكن الحصول عليها من كل متغير واختيار المتغير ذي كمية المعلومات الأكبر، ويقاس ذلك باستخدام مفهوم الإنتروبي.

تم تطوير وتحسين هذه الخوارزمية اعتماداً على مبدأ فرّق تسُد، باستراتيجية تنطلق من العقدة الجذر، والتي تكون في أعلى الشجرة، باتجاه الأسفل من قبل J. Ross Quinlan [12] من جامعة سيدني في استراليا. بالرغم من قيام بعض الباحثين بالعمل على نفس الفكرة، إلا أن بحث Quinlan كان في المقدمة. إن هذه الطريقة التي تعتمد على مفهوم ربح المعلومات هي في الحقيقة نفس الطريقة المعروفة باسم ID3. تشمل بعض التحسينات على طريقة ID3 استخدام مفهوم نسبة الربح، وخلال عدة سنوات تم إضافة عدة تحسينات على طريقة عمل الخوارزمية وصولاً إلى النسخة المعروفة ب C4.5 والتي قدمت حلولاً للتعامل مع المتغيرات الرقمية والقيم المفقودة وحل مشكلة الضجيج في البيانات.

توفر منصة WEKA خوارزمية تدعى J48[2] لتطبيق فكرة أشجار القرار، يمكن لهذه الخوارزمية تصميم شجرة سهلة الفهم وقابلة للتطبيق، وقد وصفها مصممو المنصة على أنها نقطة علام قد تكون الخوارزمية الأكثر استخداماً بين خوارزميات التعلم الآلي. يبين الشكل /6/ نتائج تطبيق هذه الخوارزمية على البيانات:



الشكل رقم (3): نتائج تطبيق الخوارزمية J48

[1] تم استخدام منصة WEKA النسخة 3.9.4

[2] تعرف أيضاً باسم C4.5

يتبين من خلال النتائج أن نسبة العينات التي تم تصنيفها بشكل صحيح هي 74.9901% وهذه النسبة تمثل دقة النموذج، تتيح منصة WEKA استخراج النص البرمجي الذي يحاكي شجرة القرار الناتجة لاستخدامه في أي تطبيق برمجي، يتم توليد هذا النص بلغة Java. كما تتيح المنصة إمكانية رسم الشجرة بشكل مرئي.

2-5-9 المجاور الأقرب K-NN

تسمى في بعض المراجع بطريقة التعلم القائم على العينة (Instance-based Learning)، وتقوم على مبدأ قياس المسافة بين العينات المعروفة مسبقاً والعينة الجديدة أو غير المعروفة بعد، يعطي مقياس المسافة المجاور الأقرب للعينة الجديدة، ويتم تصنيف العينة الجديدة بنفس الصنف الخاص بالعينة الأقرب لها. تتنوع طرق قياس المسافة لكنها غير معقدة بالمجمل وخصوصاً في حالة المتغيرات الرقمية. تستخدم معظم الطرق مقياس اقليدس (Euclidean distance) لقياس المسافة بين عينتين، وتحسب عن طريق إيجاد الجذر التربيعي لمجموع مربعات فروقات قيم المتغيرات المتقابلة من العينتين المراد قياس المسافة بينهما كما تبين المعادلة التالية:

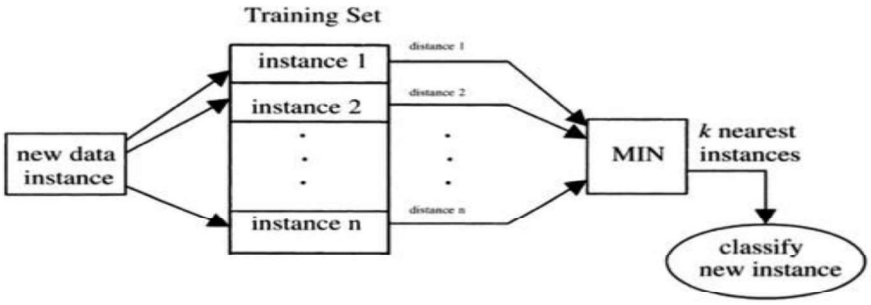
$$\sqrt{\left(a_1^{(1)}-a_1^{(2)}\right)^2+\left(a_2^{(1)}-a_2^{(2)}\right)^2+\ldots+\left(a_k^{(1)}-a_k^{(2)}\right)^2}.$$

حيث: $a_1^{(1)}$، $a_2^{(1)}$،.....، $a_k^{(1)}$ تمثل قيم المتغيرات العائدة للعينة الأولى.

$a_1^{(2)}$، $a_2^{(2)}$،.....، $a_k^{(2)}$ تمثل قيم المتغيرات العائدة للعينة الثانية.

k تمثل عدد المتغيرات في كل عينة.

ليس من الضروري تطبيق الجذر التربيعي عند مقارنة المسافات، يمكن مقارنة مجاميع المربعات مباشرةً والحصول على نفس النتيجة.



الشكل رقم (7): الفكرة الأساسية لخوارزمية المجاور الأقرب [11]

تستخدم بعض الطرق بديلاً عن المقياس السابق وهو مقياس مانهاتن (Manhattan metric)، وتحسب عن طريق جمع القيم المطلقة لفروقات قيم المتغيرات (بدون التربيع). تقوم بعض طرق قياس المسافة بعملية تسوية لقيم المسافات بين متغيرين عن طريق حساب القيمة العظمى والصغرى وأخذ النسبة لتقع بين 0 و 1، وهذه الطريقة مناسبة لحساب المسافة بين المتغيرات غير الرقمية، حيث تعتبر المسافة بين متغيرين غير رقميين 1 في حال الاختلاف و0 في حال التشابه.

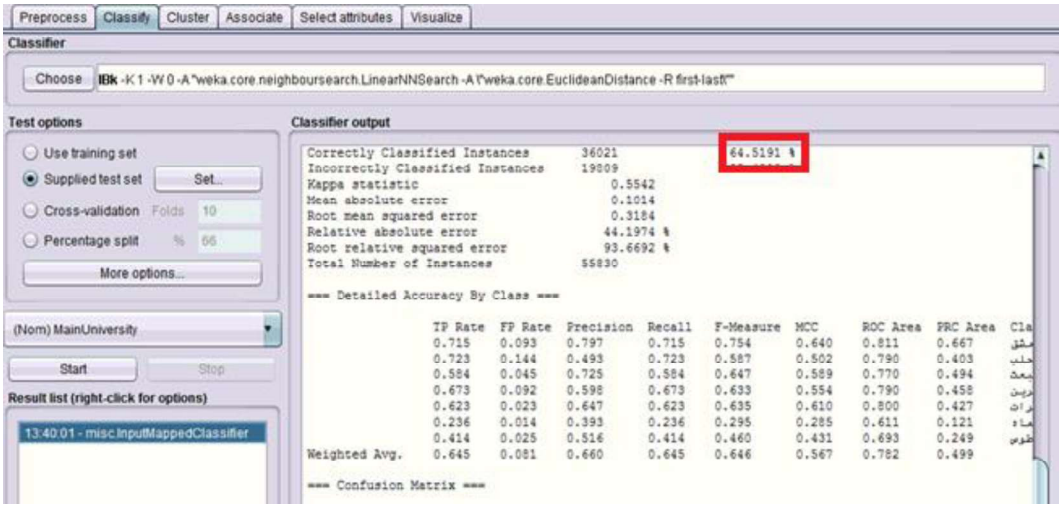
بالرغم من كون طريقة المجاور الأقرب بسيطة وفعالة، إلا أنها بطيئة في معظم الأحيان. ففي حالة التطبيق المباشر للخوارزمية يجب قياس المسافة بين العينة الجديدة من جهة وجميع العينات المعروفة من جهة أخرى واعتماد المسافة الأقصر. وهذا يعني عدد عمليات مساوٍ لعدد العينات المعروفة. وفي حال الحاجة إلى التنبؤ بعدد من العينات غير المعروفة يكون عدد العمليات التي يجب القيام بها مساوٍ لجداء عدد العينات المعروفة بعدد العينات غير المعروفة وهذا ما يؤدي إلى استهلاك

كبير للوقت والذاكرة. ولحل هذه المشكلة، تم تضمين الخوارزمية إمكانية تقسيم فضاء العينات إلى مناطق متقاربة لتقليل عدد عمليات حساب المسافات.

تتميز هذه الخوارزمية بإمكانية إضافة العينات التي تم تصنيفها إلى العينات المعروفة مسبقاً، وفي هذه الحالة، تقوم الخوارزمية بإعادة تقسيم فضاء العينات من حين إلى آخر لتحسين دقة التنبؤ. كما يمكن إيجاد عدد من الجيران (K) وتصنيف العينة الجديدة حسب صنف الأكثرية من هذه الجيران، ولكن يجب أخذ عدد العينات في الاعتبار عند اختيار قيمة K، ففي حال كون عدد العينات قليلاً لا ينصح باختيار قيمة كبيرة ل K والعكس صحيح.

كسبت طرق المجاور الأقرب شعبيتها في علوم البيانات من خلال البحث الذي قدمه D.W. Aha عام 1992، الذي أثبت أن دمج هذه الطريقة مع نماذج التشذيب (Pruning) وتثقيل المتغيرات يعطي نتائج جيدة مقارنةً مع بقية الطرق [10].

تقدم منصة WEKA نموذجاً لتطبيق خوارزمية المجاور الأقرب وتسمى IBK. وتصنفها المنصة ضمن تبويب Lazy Learning أو التعلم الكسول، يعود السبب في ذلك أن الخوارزمية لا تقوم فعلياً بأي عمليات على العينات قبل بدء التصنيف الفعلي بل تقوم باستخدام العينات المعروفة نفسها في عمليات المقارنة، بعكس الخوارزميات الأخرى "المتحمسة" التي تقوم بتوليد قواعد أو أشجار قرار أو غيرها انطلاقاً من العينات المعروفة.



الشكل رقم (4): نتائج تطبيق خوارزمية IBK

يتبين من خلال النتائج أن نسبة العينات التي تم تصنيفها بشكل صحيح هي 64.5191% وهذه النسبة تمثل دقة النموذج، تستخدم هذه الخوارزمية مقياس اقليدس لقياس المسافة بين العينات مع إمكانية التحكم بالبارامترات.

3-5-9 نايف بايز Naïve Bayes

تعد مشكلة تصنيف الوثائق من المشاكل المهمة في مجال التنقيب في البيانات، حيث تمثل الوثائق عينات البيانات والهدف هو تصنيف هذه الوثائق حسب مواضيعها. تمثَّل الوثائق بالكلمات التي تظهر فيها، تقوم إحدى طرق تصنيف الوثائق على معيار ظهور أو عدم ظهور بعض الكلمات فيها وينتج عن ذلك متغيرات بولياينة[1]. تعد طريقة نايف بايز طريقة شائعة لحل هذه المشكلة وهي تعطي نتائج سريعة ودقيقة.

[1] المتغير البولياني يأخذ القيم True أو False فقط.

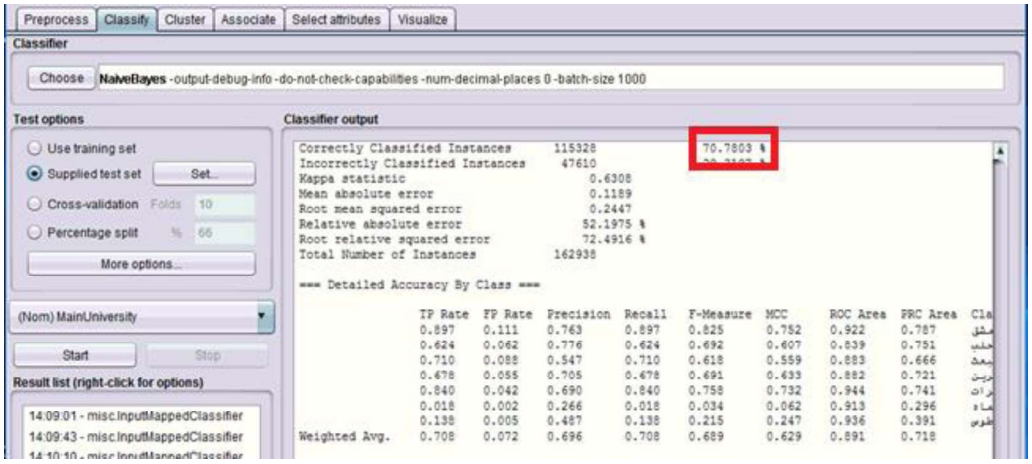
لكن المشكلة أن هذه الطريقة لا تأخذ بالحسبان معيار مهم وهو عدد مرات ظهور الكلمة في الوثيقة، وهذا ما أدى إلى ظهور نموذج معدل من خوارزمية نايف بايز تسمى multinominal Naïve Bayes، حيث يتم تصنيف الوثيقة اعتماداً على الكلمات التي تظهر فيها وعدد مرات ظهور كل كلمة في هذه الوثيقة.

تفترض هذه الطريقة استقلال المتغيرات عن بعضها، وتعتمد على جداء احتمالات ظهور هذه المتغيرات ببعضها البعض. لذلك يطلق عليها لقب "ساذجة" ولكنها تعمل بشكل جيد جداً على مجاميع البيانات الحقيقية، خصوصاً عند تطبيقها بالترافق مع إحدى خوارزميات اختيار المتغيرات والتي تحيّد المتغيرات غير ذات الصلة.

قد تواجه هذه الطريقة مشكلة أثناء التطبيق وهي حالة عدم وجود قيمة لأحد المتغيرات في عينة ما، وبما أن الطريقة تقوم على جداء احتمالات ظهور المتغيرات فإن ذلك سيؤدي إلى الحصول على قيمة صفرية، لحل هذه المشكلة اعتمدت بعض التطبيقات إضافة الرقم 1 إلى عدد مرات ظهور كل متغير لضمان عدم حدوث الحالة السابقة، وقد أعطت هذه الإضافة حلاً جيداً لهذه المشكلة دون التأثير على دقة النتائج [10].

أثبتت طريقة نايف بايز أنها منافسة للطرق الأخرى الأكثر تطوراً، وأنها قادرة على إعطاء نتائج مبهرة. كما أثبتت مقولة "قم بتجربة الطرق الأبسط أولاً".

الجدير بالذكر أن فكرة هذه الطريقة بدأت من مقالة حول حل المشكلات باستخدام علم الاحتمالات كتبها الفيلسوف البريطاني Bayes في القرن الثامن عشر. تنتج هذه الطريقة تقديرات احتمالية لانتماء العينة الجديدة إلى صنف ما، وقد تكون هذه النتائج ذات فائدة أكبر من مجرد توقع الصنف الذي تنتمي إليه العينة الجديدة، حيث يمكن ترتيب الأصناف وفقاً لاحتمالية انتماء العينة الجديدة لكل منها. تقدم منصة WEKA نموذجاً لتطبيق خوارزمية التصنيف هذه، مع عدة خيارات لتطبيق النسخ المحدثة من هذه الخوارزمية.



الشكل رقم (5): نتائج تطبيق خوارزمية Naive Bayes

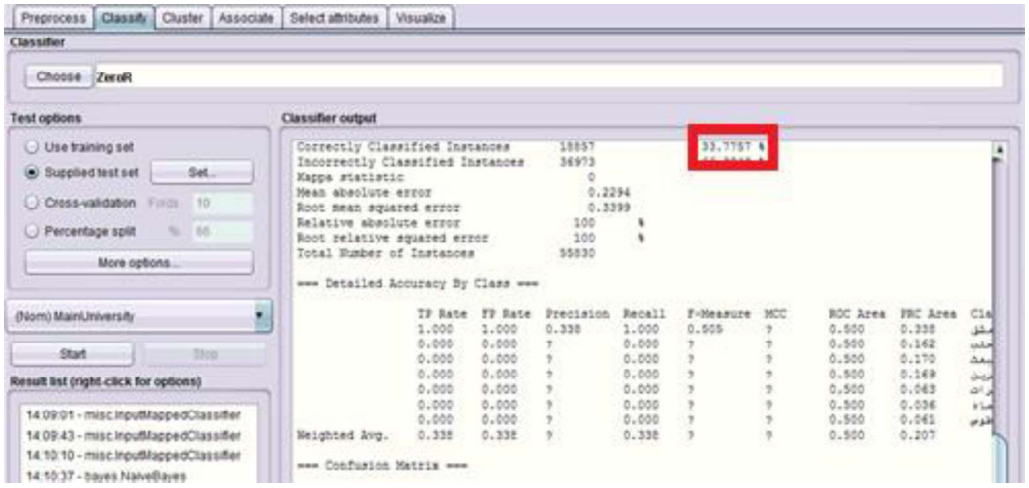
كما توفر منصة WEKA نافذة خاصة لضبط قيم بارامترات الخوارزمية لتحسين دقة النتائج، تم اعتماد القيم الافتراضية لجميع البارامترات، ويتبين من خلال النتائج أن نسبة العينات التي تم تصنيفها بشكل صحيح هي 70.7803% وهذه النسبة تمثل دقة النموذج.

9-5-4 خوارزمية ZeroR

تقوم كل من الخوارزميات سابقة الذكر بإنتاج نموذج تنبؤي ذي دقة معينة، ومن البديهي عند اعتماد الخوارزمية التي سيتم تطبيقها على البيانات اختيار الخوارزمية ذات الدقة الأعلى، لكن يوجد معيار مهم يجب أخذه بعين الاعتبار عند مقارنة دقة التنبؤ لكل من هذه الطرق، هذا المعيار هو الدقة الحدية (Base-Line Accuracy).

قبل مقارنة دقة التنبؤ للخوارزميات المطبقة، يجب أن تكون دقة التنبؤ لكل منها أعلى من الدقة الحدية، يمكن الحصول على الدقة الحدية من خلال تطبيق خوارزمية ZeroR والتي تقوم باختيار الصنف الأكثر شعبيةً لتصنيف العينات الجديدة في حالة كون المتغير الهدف نصي، أما في حال كون المتغير الهدف رقمي فتكون نتيجة التصنيف هي متوسط جميع قيم المتغير الهدف. من المفترض أن تكون دقة التنبؤ لكل الخوارزميات السابقة أعلى من دقة التنبؤ لخوارزمية ZeroR، وسيتم التحقق من ذلك من خلال نتائج التطبيق، ثم اختيار الخوارزمية ذات دقة التنبؤ الأعلى.

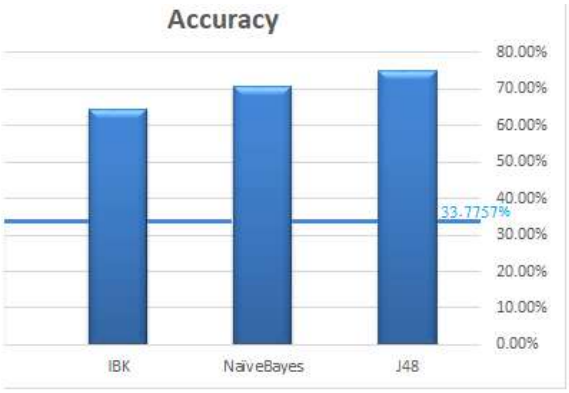
كغيرها من الخوارزميات، يمكن تطبيق خوارزمية ZeroR من خلال منصة WEKA ومعايرة البارامترات الخاصة بها من خلال نافذة مستقلة، يتبين من النتائج أن الدقة الحدية تبلغ 33.7757%.



الشكل رقم (6): نتائج تطبيق خوارزمية ZeroR

9-6 مقارنة نتائج تطبيق خوارزميات التنقيب في البيانات

بعد تطبيق الخوارزميات على مجموعة البيانات الأولى، يجب مقارنة النتائج واختيار الخوارزمية ذات الدقة الأعلى، ولكن قبل ذلك يجب اختبار مبدأ الدقة الحدية. كما تبين من الخطوة السابقة فإن الدقة الحدية الناتجة عن تطبيق خوارزمية ZeroR بلغت 33.7757% وبالمقارنة مع نتائج الخوارزميات الثلاث يتبين أن كل الخوارزميات أعطت دقة أعلى من الدقة الحدية، يبين الشكل /11/ دقة كل خوارزمية:



الشكل 7: مقارنة نتائج الخوارزميات

يتضح أن كلاً من الخوارزميات الثلاث اجتازت الدقة الحدية الناتجة عن تطبيق خوارزمية ZeroR، كما أن الخوارزمية J48 حققت أعلى دقة، لذلك تم اعتمادها لبناء النموذج التنبؤي.

9-7 معايرة البارامترات

بعد اختيار الخوارزمية ذات الدقة الأعلى، سيتم في هذه الخطوة معايرة البارامترات في محاولة لتحسين الدقة الناتجة عن التطبيق الأولي، تم معايرة البارامترات التالية:

الجدول رقم 2: معايرة بارامترات الخوارزمية J48

Parameter Name	Default Run	Run 1	Run 2	Run 3
Binary Split	False	True	True	**True**
Debug	False	True	True	**True**
minNumObj	2	160[1]	160	**56**
numDecimalPlaces	2	4	4	**4**
ReducedErrorPruning	False	True	True	**True**
useLaplace	False	True	True	**True**
useMDLCorrection	False	True	True	**True**
unpruned	False	True	False	**False**
Accuracy	%74.9901	77.8846%	78.0556%	**78.3468%**

بعد إعادة تطبيق الخوارزمية عدة مرات وتجربة عدة قيم للبارامترات، تحسنت دقة النموذج بمقدار 3.3567% لتصبح الدقة النهائية 78.3468%.

[1] متوسط الطاقة الاستيعابية من الطلاب في الكليات والمعاهد للفرع العلمي لعام 2018 (وزارة التعليم العالي والبحث العلمي، 2018)

من الواضح أن معايرة البارامترات هي عملية مهمة في تصميم النموذج النهائي، وهي عملية تجريبية بمعظمها، ولكن يمكن من خلال المعرفة المسبقة بطبيعة البيانات معايرة بعض البارامترات لإعطاء نتيجة أفضل، فمثلاً عند حساب متوسط الطاقة الاستيعابية للكليات والمعاهد من الطلاب للعام 2018 واعتبار هذه القيمة كعدد الحالات الأدنى في كل ورقة من أوراق الشجرة تحسنت الدقة بشكل ملحوظ ولكن عند التجربة تبين أن إنقاص هذه القيمة قد أدى إلى تحسن إضافي في النتائج.

تم بناء النموذج النهائي للتنبؤ بالجامعة وبطريقة مماثلة، تم بناء نموذج مماثل للتنبؤ بالكلية/المعهد لكل جامعة على حدة، بحيث يتم تقسيم البيانات الناتجة عن التنبؤ حسب الجامعة، واستخدامها كدخل للنماذج الخاصة بالتنبؤ بالكلية/المعهد.

7-9 تطبيق النموذج النهائي على بيانات العام 2018

تم تطبيق النموذج النهائي على مجموعة العينات الثالثة الخاصة بقبول الطلاب في العام 2018، تحتوي هذه البيانات على قبول الطلاب الحقيقي في الجامعة والكلية/المعهد، يبلغ عدد العينات (61937) طالباً وطالبة، تم التطبيق حسب الخطوات المذكورة سابقاً وفق التسلسل التالي:

1. إدخال البيانات إلى النموذج الأول الخاص بالجامعة، أعطى النموذج نتائج صحيحة لـ (45716) عينة، وبذلك تكون دقة التنبؤ 73.8104% وهي أقل من النسبة الناتجة عند بناء النموذج والبالغة 78.3468%.
2. تقسيم بيانات العام 2018 حسب الجامعة التي تم التنبؤ بها كنتيجة للخطوة السابقة. نتج عن التقسيم سبعة مجموعات.
3. إدخال كل مجموعة من المجموعات السبعة على النموذج التنبؤي الخاص بالجامعة المقابلة بهدف التنبؤ بالكلية/المعهد.

الجدول رقم 3: نتائج تطبيق نماذج التنبؤ الخاصة بكل جامعة

الجامعة	دقة النموذج	عدد العينات الكلي	العينات المصنفة بشكل خاطئ	العينات المصنفة بشكل صحيح	دقة التنبؤ
دمشق	%31.5145	19569	13189	6380	%32.60259
حلب	%29.6648	11977	8923	3054	%25.49887
تشرين	%30.2401	11090	8016	3074	%27.71867
البعث	%29.2844	9275	6562	2713	%29.25067
حماه	%52.4842	2624	1459	1165	%44.39787
الفرات	%26.8241	3591	2812	779	%21.69312
طرطوس	%47.3408	3811	2148	1663	%43.63684

نلاحظ انخفاض دقة نماذج التنبؤ الخاصة بالكلية/المعهد لكل الجامعات، حيث تتراوح دقة النماذج بين 26.8241% (جامعة الفرات) إلى 52.4842% (جامعة حماه)، ويعزى ذلك للأسباب التالية:

1. الخطأ التراكمي الحاصل في المرحلة السابقة أثناء تطبيق نموذج التنبؤ الأول الخاص بالجامعة، حيث كانت نسبة الخطأ 26.1896% من عدد العينات الكلي، وهذا أدى إلى تصنيف خاطئ لهذه العينات في المرحلة الثانية حكماً.
2. ازدياد عدد الأصناف في نماذج المرحلة الثانية، ففي المرحلة الأولى كان عدد الأصناف مساوٍ لعدد الجامعات أي (7) أما في المرحلة الثانية فبلغ عدد الأصناف (132) صنفاً مختلفاً.

3. تم تطبيق النماذج على بيانات العام 2018، وقد شهد هذا العام افتتاح عدة كليات جديدة مثل كلية العلوم الصحية في جامعة دمشق، وهذا سيؤدي أيضاً إلى أخطاء في التصنيف.
4. يُلاحَظ من خلال الجدول السابق انخفاض دقة التنبؤ في الجامعات التي لها فروع في محافظات أخرى، مثل جامعة دمشق والفرات، وارتفاعها بشكل نسبي في الجامعات الناشئة مثل جامعة طرطوس وحماه، وذلك بسبب تفوق الجامعات التي تمتلك فروع في المحافظات على بقية الجامعات من حيث عدد الكليات والمعاهد التابعة لها.

قد تكون نتائج التنبؤ في هذه المرحلة منخفضة نسبياً لكنها تعطي مؤشرات مفيدة جداً في دراسة توجهات الطلاب بناءً على معدلات نجاحهم في المرحلة الثانوية باعتبار متغيرات أخرى مثل الجنس ومصدر الشهادة، من غير الممكن التوصل إلى دقة أعلى من ذلك نظراً لوجود عوامل أخرى قد تؤثر في توجهات الطلاب، وهذه العوامل غير قابلة للقياس مثل تأثير الأهل والمجتمع في توجه الطالب، وإقبال الطلاب على دراسة الاختصاصات التي تزيد من فرصهم في سوق العمل، أو تفضيلهم لبعض التخصصات بسبب قربها من مكان سكنهم، كما يمكن أن يُتاح لبعض الطلاب اختصاصات جيدة لكنهم يفضلون اختيار اختصاصات ذات مدة دراسة قصيرة (معاهد) بهدف الدخول إلى سوق العمل بأسرع وقت ممكن، أو لأسباب أخرى مثل تأجيل الخدمة العسكرية الإلزامية للذكور، يتبين أيضاً من خلال نتائج التنبؤ أهمية متغير الجنس في توزيع المقاعد لكل من الجنسين في الكليات والمعاهد، كي لا يحصل تجمع لجنس واحد في كلية ما، أو تخصيص بعض الكليات لأحد الجنسين حسب أعداد المقبولين من الجنسين في السنوات السابقة. يوضح الشكل /12/ عينة من نتائج النموذج التنبؤي وما يقابلها من النتائج الحقيقية.

ID	CityName	Gender	TotalScore	Math	Physics	Chemistry	Biology	Arabic	actual	predicted	MainUniv	ActualGroup	PredictedGroup
152975	طرطوس	f	7904	8467	9050	7850	7100	8775	التربية - معلم صف-اللاذقية	الاقتصاد-اللاذقية	تشرين	العلوم النظرية	العلوم التطبيقية
152982	طرطوس	f	9552	9667	9475	9850	8533	9750	الهندسة المدنية-اللاذقية	الهندسة المعلوماتية-اللاذقية	تشرين	العلوم الهندسية	العلوم الهندسية
152985	طرطوس	f	8022	7467	7725	7800	8533	9675	الكيمياء-اللاذقية	الزراعة-اللاذقية	تشرين	العلوم التطبيقية	العلوم التطبيقية
152992	طرطوس	f	8122	8750	8450	8100	6367	9100	الرياضيات-اللاذقية	الاقتصاد-اللاذقية	تشرين	العلوم التطبيقية	العلوم التطبيقية
153005	طرطوس	f	6885	6750	7475	4550	6000	8825	الجغرافية-اللاذقية	ممرضة التمريض-اللاذقية	تشرين	العلوم النظرية	العلوم النظرية
153027	طرطوس	f	9226	9083	9875	9750	8933	9600	هندسة الحاسبات والتحكم الآلي-اللاذقية	الهندسة المدنية-اللاذقية	تشرين	العلوم الهندسية	العلوم الهندسية
153030	طرطوس	f	8304	9300	8225	8400	7400	8925	الرياضيات-اللاذقية	الزراعة-اللاذقية	تشرين	العلوم التطبيقية	العلوم التطبيقية
153052	طرطوس	f	8148	8033	8150	9050	7500	8900	الكيمياء-اللاذقية	الاقتصاد-اللاذقية	تشرين	العلوم التطبيقية	العلوم التطبيقية
153058	طرطوس	f	9459	9417	9500	9800	10000	9600	الهندسة المدنية-اللاذقية	الكليات الطبية-اللاذقية	تشرين	العلوم الهندسية	العلوم الطبية
153059	طرطوس	f	9026	8367	8425	9450	9233	9650	الزراعة-اللاذقية	الهندسة المعمارية-اللاذقية	تشرين	العلوم التطبيقية	العلوم الهندسية
153062	طرطوس	f	7589	6933	5950	8100	7933	9150	التقاني الطبي-اللاذقية	علم الحياة-اللاذقية	تشرين	المعاهد التطبيقية	العلوم التطبيقية
153073	طرطوس	f	7704	7883	7200	7800	7767	8475	الكيمياء-اللاذقية	علم الحياة-اللاذقية	تشرين	العلوم التطبيقية	العلوم التطبيقية
153076	طرطوس	f	6433	5400	7125	8150	6000	8275	التقانيم القانونية-اللاذقية	الجغرافية-اللاذقية	تشرين	المعاهد الثانية	العلوم النظرية
153091	طرطوس	f	9022	9800	8375	8950	8233	9575	هندسة التصميم والإنتاج-اللاذقية	الهندسة المعمارية-اللاذقية	تشرين	العلوم الهندسية	العلوم الهندسية
153093	طرطوس	f	8893	9400	9125	8700	8800	9350	هندسة الطاقة الكهربائية-اللاذقية	الهندسة المعمارية-اللاذقية	تشرين	العلوم الهندسية	العلوم الهندسية
153096	طرطوس	f	8841	8983	8800	8700	8933	9850	الزراعة-اللاذقية	الهندسة المعمارية-اللاذقية	تشرين	العلوم التطبيقية	العلوم الهندسية
153099	طرطوس	f	7015	5450	5975	7750	8567	7875	التقاني العلوم الحياتية والمخبرية-اللاذقية	الجغرافية-اللاذقية	تشرين	المعاهد الثانية	العلوم النظرية
153100	طرطوس	f	8267	7333	7350	9300	8900	8825	الحقوق-اللاذقية	الزراعة-اللاذقية	تشرين	العلوم النظرية	العلوم التطبيقية
153101	طرطوس	f	9111	9250	9700	9600	8900	9700	هندسة الاتصالات-اللاذقية	الهندسة المدنية-اللاذقية	تشرين	العلوم الهندسية	العلوم الهندسية
153103	طرطوس	f	8763	9167	9000	8400	9833	9825	هندسة القوى الميكانيكية-اللاذقية	الهندسة المعمارية-اللاذقية	تشرين	العلوم الهندسية	العلوم الهندسية
153125	طرطوس	f	8804	8967	8600	9500	8433	9000	الزراعة-اللاذقية	هندسة التصميم والإنتاج-ا...	تشرين	العلوم التطبيقية	العلوم الهندسية
153129	طرطوس	f	8904	9917	8500	9550	9300	9375	هندسة الاتصالات-اللاذقية	الهندسة المدنية-اللاذقية	تشرين	العلوم الهندسية	العلوم الهندسية
153130	طرطوس	f	9056	9717	9325	9700	7700	9500	هندسة الطاقة الكهربائية-اللاذقية	الهندسة المدنية-اللاذقية	تشرين	العلوم الهندسية	العلوم الهندسية

الشكل رقم (12): عينة من نتائج النموذج التنبؤي ومقارنتها مع النتائج الحقيقية

من ناحية أخرى، تم تجميع نتائج القبول الناتجة عن النماذج التنبؤية ضمن مجموعات القبول الستة المشار إليها سابقاً. ومقارنتها مع نتائج القبول الحقيقية، تبين من خلال مقارنة النتائج الملاحظات التالية:

1. عند مقارنة مجموعة القبول الحقيقية مع مجموعة القبول الناتجة عن التنبؤ، بلغت دقة التنبؤ 66.4982%.
2. عدد العينات المقبولة في مجموعة (المعاهد التقانية والمعاهد التطبيقية) والتي قُبلت بنتيجة تطبيق النموذج التنبؤي في مجموعات أخرى هو (7282) عينة، وتشكل نسبة 35.09% من العينات المصنفة بشكل خاطئ.
3. عدد العينات المقبولة في أي قبول عدا العلوم الطبية والتي قُبلت بنتيجة تطبيق النموذج التنبؤي في مجموعة الكليات الطبية هو (402)، وتشكل نسبة 1.937% من العينات المصنفة بشكل خاطئ.
4. عدد العينات المقبولة في مجموعتي العلوم التطبيقية أو العلوم النظرية والتي قُبلت بنتيجة تطبيق النموذج التنبؤي في مجموعة العلوم الهندسية هو (1084)، وتشكل نسبة 5.224% من العينات المصنفة بشكل خاطئ.
5. تم بنتيجة تطبيق النموذج التنبؤي اقتراح قبول بعض الطلاب في الكليات والمعاهد التي يتطلب القبول فيها اجتياز مقابلة أو اختبار، ولم يتم اعتبار هذا الشرط في بناء النموذج.

10. الخلاصة

نظام التنبؤ بالقبول الجامعي هو نظام وب يمكّن الطلاب من التسجيل وإدخال علاماتهم ومعلوماتهم الشخصية واستخدام هذه المعلومات في توقع قبولهم في الكليات. يمكن للمسؤول إضافة المعلومات التفصيلية المتعلقة بالكليات. باستخدام هذا النظام يصبح توزيع المقاعد أسهل وأكثر فاعلية. تظهر الفائدة الأساسية للنظام في أتمتة عملية توزيع المقاعد، وهذا ما يجعل عملية التوزيع أسرع ويختصر الوقت، وفي النتيجة يتم تقديم المساعدة للطلاب لاتخاذ القرار الصائب باختيار الكلية [13].

إن إنشاء نظام الإرشاد وتطبيقه سوف يساهم في زيادة فرص قبول الطالب الأنسب في الفرع الأنسب، عبر توجيه الطالب بالاعتماد على البيانات الحقيقية للسنوات السابقة، كما يسهم أيضاً بتوفير الطاقة الاستيعابية للكليات وتحسين توزيع المقاعد والطاقة الاستيعابية، وهذا ما يؤدي بالنهاية إلى رفع سوية المخرجات التعليمية.

يمكن تطبيق نظام الإرشاد على إحدى المفاضلات التي تجريها وزارة التعليم العالي والبحث العلمي وتخصيص مراكز دعم في الجامعات تستقبل الطلاب الراغبين بالتسجيل عبر المراكز، والطلب من الطلاب كتابة مراجعة توضح رأيهم بالنظام ومدى مقاربته للواقع الحقيقي. كما يمكن مقارنة نتائج هذه المفاضلة بنتائج مفاضلة العام السابق وقياس مدى التغير في القبول ومستوى الطلاب المقبولين في كل فرع.

11. التوصيات والمقترحات

خلصت هذه الدراسة إلى تصميم نموذج تنبؤي يمكن استخدامه في إرشاد الطلاب المقبلين على الجامعات إلى الفرع الأنسب لهم حسب بياناتهم في الثانوية العامة، توصي هذه الدراسة بتطبيق هذا النموذج على أرض الواقع ووضعه في الخدمة الفعلية بما يخدم الطالب في سورية، كما توصي بما يلي:

1. تصميم نموذج تنبؤي مشابه لطلاب الفرع الأدبي والمهني، وتصميم نظام إرشاد مستقل أو مدمج مع النظام السابق، مع الأخذ بالاعتبار المتغير الجديد هنا وهو فرع الشهادة.
2. تطبيق قواعد جديدة للقبول في الفرع العلمي بالاعتماد على التسجيل المباشر كما هو الحال في الفرع الأدبي، والاستغناء عن نظام المفاضلة المعمول به حالياً في الفرع العلمي، قد تكون نتائج تطبيق هذه الفكرة أفضل من المتوقع، ولا يمكن إثبات ذلك إلا بتجربته.
3. تطبيق الاختبارات والمسابقات والمقابلات في فروع أخرى إضافة إلى الاختبارات المعمول بها حالياً.
4. تطوير النموذج وإعادة تقييمه سنوياً من خلال البيانات التي تتجمع كل عام.

5. دراسة تغيرات الطاقة الاستيعابية في الجامعات السورية خلال الأعوام السابقة، والعوامل التي تؤدي إلى افتتاح أو إغلاق كليات وأقسام في الجامعات، وتصميم نظام تنبؤي قادر على تحديد الطاقة الاستيعابية بشكل تقريبي في كل كلية أو معهد في الجامعات السورية.

12. المراجع

1. هيفاء ابراهيم. (2013). **أنموذج مقترح لتطوير واقع سياسات قبول الطلبة في التعليم الجامعي في الجمهورية العربية السورية في ضوء تجارب بعض الدول المتقدمة.** جامعة دمشق – كلية التربية، سورية.
2. الشربيني الهلالي. (نيسان، 2008). **نظام مقترح للقبول بمؤسسات التعليم العالي في مصر.** اللجنة التحضيرية للمؤتمر القومي لتطوير الثانوية العامة وسياسات القبول في التعليم العالي في مصر.
3. نصرة رضا البناي، وفاء محمد بلحاضي، و محمد أحمد الخولي. (2004). **القيمة التنبؤية لمعايير القبول المستخدمة بجامعة قطر وعلاقتها بالمعدل التراكمي الجامعي.**
4. قصي عزام. (2015). **نظام دعم القرارات المتعلقة بالقبول الجامعي (المفاضلة) في الجمهورية العربية السورية.** دمشق: المعهد العالي للعلوم التطبيقية والتكنولوجيا.
5. Janusz Sobecki .(2006) .**Implementations of Web-based Recommender Systems Using Hybrid Methods** . *International Journal of Computer Science & Applications,* *3*(3)، 53-64.
6. Abdul Hamid M. Ragab, Abdul Fatah S. Mashat, and Ahmed M. Khedra .(2014) .**Design and Implementation of a Hybrid Recommender System for Predicting College Admission** .*International Journal of Computer Information Systems and Industrial Management Applications,*.44-35 ،(7988-2150)*6*
7. Mwapashua H. Fujo and Mussa Ally Dida .**(2018)** .**Web-based admission system for advanced level, private schools: case of Kilimanjaro region, Tanzania** . *International Journal of Advanced Technology and Engineering Exploration,*-407 ،(7454-2394)*(47)5* .418
8. R. Suguna and D. Sharmila .(2013) .**An Efficient Web Recommendation System using Collaborative Filtering and Pattern Discovery Algorithms** .*International Journal of Computer Applications,(0975-8887). 37-70.*
9. Baswana, S., Chakrabarti, P. P., Patanged, U., Kanoria, Y., & Chandran, S. (2019, September - October). **Centralized Admissions for Engineering Colleges in India**. *INFORMS JOURNAL ON APPLIED ANALYTICS*, pp. 338-354.

10. Ian H. Witten and Eibe Frank .(2005) . **Data Mining Practical Machine Learning Tools and Techniques** 2nd Edition. Elsevier.

11. Alex A. Freitas .(2002) .**Data Mining and Knowledge Discovery with Evolutionary Algorithms** .Berlin, Germany: *Springer*.

12. J. Ross Quinlan .(1994) .**C4.5: Programs for Machine Learning** .Morgan Kaufmann Publishers, Inc.

13. Annam Mallikharjuna Roa, Nagineni Dharani, A. Satya Raghava, J. Buvanambigai and K. Sathish .(2018) . **College Admission Predictor** . *Journal of Network Communications and Emerging Technologies (JNCET),*.147-142 ،(4)*8*