Journal of Education and Science (ISSN 1812-125X)

www.edusj.mosuljournals.com

Mining Students and Patients Data of Dentistry College in the University of Mosul

Marwa B. Mustafa[(1)] Ammar T. Y. Al Abd Alazeez [(2)]

[1,2] Department of Computer Science, College of Computer Science and Mathematics, University of Mosul, Mosul, Iraq

Article information

Article history:
Received October 22, 2023
Accepted December 26, 2023
Available online March 01, 2024

Keywords:
Database
Mining Database
Dentistry Applications

Correspondence:
Marwa Bassim Mustafa
Marwa.23csp@student.uomosul.edu.i

Abstract

This research paper includes the design and implementation of a system for mining student and patient data at the College of Dentistry at the University of Mosul using the Microsoft SQL Server database management system to design and implement the database system and WEKA program for database mining, and the Microsoft Visual C# .NET 2012 language was used to program system interfaces. The main steps of the database included analysis, design and implementation, and the mining process included seven steps; data collection, data preprocessing, data exploration, data transformation, data modeling, evaluation, and deployment. The database mining process was divided into two parts; the first part is a smart cluster process for students of the Faculty of Dentistry for the fourth and fifth stages on laboratories (i.e. the number of chairs available for each laboratory) using three famous algorithms (Canopy, K-Means, EM), the second part is the process of classifying patients into four classes according to the type of treatment that each patient needs using three also famous algorithms (SVM, Naïve Bayes, Random Forest).). After applying the system to the real data of the College of Dentistry at the University of Mosul, it was found that the best cluster algorithm is K-Means and the best classification algorithm is Naïve Bayes.

DOI: 10.33899/edusj.2023.143880.1398, ©Authors, 2024, College of Education for Pure Science, University of Mosul.

1. المقدمة

تعتبر صحة الفم مؤشراً رئيساً لصحة الانسان والراحة ونوعية الحياة. وقد يحتوي الفم على مجموعة من الحالات المرضية منها تسوس الاسنان، وأمراض اللثة، وفقدان الاسنان، وسرطان الفم، والمظاهر الفموية لعدوى فيروس نقص المناعة البشرية وحوادث الفم والاسنان والعيوب الخلقية مثل الشفة المشقوقة [1]. لذا ازداد اهتمام أطباء الأسنان في الآونة الأخيرة بتشخيص الآفات الفموية مما لذلك من أهمية في الكشف والتحري عن الآفات السرطانية والتشخيص المبكر لها، كما أن لبعض هذه الآفات الفموية دلالات للإصابة ببعض الأمراض في اجزاء الجسم الأخرى مما يساعد في الكشف عنها [2]. أدى التطور السريع في توليد وجمع البيانات الى وجود مجموعات من البيانات ذات الأحجام الهائلة في مجال الطب وكافة فروع المعرفة العلمية، إذ وجدت المؤسسات نفسها غير قادرة على ترجمة وفهم الكم الهائل من البيانات الموجودة، ولم تعد وسائل التحليل التقليدية الإحصائية قادرة على التعامل معها، فكانت تقنيات قواعد البيانات Database System(DB) والتنقيب في البيانات Data Mining(DM) واكتشاف المعرفة (knowledge Discovery Database)KDD احد الحلول الناجحة لحل هذه المشكلة. يعتبر البعض التنقيب في البيانات مصطلحا شائعا لاكتشاف المعرفة، في حين يضع البعض الاخر التنقيب في البيانات كخطوة أساسية في عملية اكتشاف المعرفة. فلقد ظهر التنقيب في البيانات في أواخر الثمانينات والذي دخل في العديد من التطبيقات منها التطبيقات الطبية[3]. تنقيب البيانات هو اكتشاف المعرفة وفي الواقع هي عملية تحليل مجموعات البيانات الضخمة من أجل استخراج معنى للبيانات. تستخدم عمليات وأدوات استخراج البيانات للتنبؤ بالاتجاهات التي تساعد في النهاية في اتخاذ قرارات استباقية قائمة على المعرفة [4].

تسببت الاعداد الكبيرة للطلبة المقبولين في كلية طب الاسنان في جامعة الموصل (والتي هي أكبر من الطاقة الاستيعابية للكلية والتي صُممت لاستقبال 50 طالباً فقط) بظهور مشكلة كثرة اعداد الطلبة في العيادات الطبية التابعة للكلية وكذلك صعوبة ايجاد الحالات

المرضية لسد النقص الحاصل في الحالات فضلا عن عدم وجود نظام الكتروني يحتفظ بمعلومات المرضى والطلاب. بمعنى اخر، تعاني كلية طب الاسنان في جامعة الموصل من عدم وجود نظام قاعدة بيانات للطلاب والمرضى المراجعين للعيادات الطبية التابعة للكلية وصعوبة توثيق الحالات المرضية وكذلك الزخم الحاصل على العيادات بسبب الاعداد الكبيرة للطلبة التي تعتبر خارج الطاقة الاستيعابية للكلية فضلا عن صعوبة ايجاد الحالات المرضية المناسبة للطلاب. ولحل هذه المشكلة تم في هذا البحث تصميم وتنفيذ نظام قاعدة بيانات للاحتفاظ بمعلومات الطلبة والمرضى فضلا عن استخدام تقنيات تنقيب البيانات لعنقدة بيانات الطلاب لتوزيعهم على العيادات وتصنيف الحالات المرضية على العيادات. بمعنى ادق، يهدف البحث إلى تصميم نظام الكتروني لتنقيب قاعدة بيانات الطلاب والمرضى في كلية طب الاسنان في جامعة الموصل والذي يتضمن تحقيق جملة من الأهداف الفرعية، أهمها:

A. تكوين قاعدة بيانات موحدة للطلاب والمرضى في كلية طب الاسنان في جامعة الموصل.
B. تقسيم الطلاب الى عدة مجاميع حسب المستوى العلمي، الجنس، والموقع الجغرافي للتخفيف من اعداد الطلاب المتواجدين في عيادة كل فرع وحسب عدد كراسي معالجة المرضى المتوفرة لكل فرع.
C. تصنيف المرضى الى اربعة أصناف وحسب أفرع الكلية.

1. اهمية البحث The Importance of the Research

في الوقت الحاضر، أدى التطور في تكنولوجيا المعلومات إلى تخزين كمية كبيرة من البيانات. ومع ذلك، لا تزال معظم استخدامات البيانات بسيطة، مثل استخرج البيانات من قاعدة البيانات. يمكن أن يكون تحليل البيانات ذو فائدة كبيرة في العمليات التنظيمية وصنع القرار للشركات والمؤسسات. بتعبير اخر، يمكن استخراج البيانات بأشكال عديدة اعتمادا على الهدف من استخراج البيانات. التصنيف هو أحد مهام التنقيب عن البيانات، وهو نمذجة للبيانات الفئوية من البيانات المصنفة مسبقا لاستخدام هذا النموذج لتصنيف البيانات الجديدة التي لم يتم تصنيفها من قبل. في حين أن العنقدة والتي هي أيضا أحد مهام تنقيب البيانات، تعمل على تجميع البيانات المتشابهة في مجموعة واحدة والبيانات المختلفة في مجاميع أخرى. فضلا عن ذلك، يعد تنقيب البيانات طريقة لاستخراج المعرفة من بيانات مختلفة للاستفادة من تلك المعرفة في صنع القرار. يمكن استخدام هذه المعرفة للتنبؤ أو إنشاء نماذج لتصنيف، او عنقدة البيانات أو عرض العلاقات بين الوحدات المختلفة، والتي يمكن تطبيقها في العديد من المجالات [5]. تكتسب هذه الدراسة أهمية من خلال تناولها لمتغيرين مهمين هما تكوين قاعدة بيانات لطلاب ومراجعي كلية طب الاسنان وتنقيب بيانات الطلاب والمرضى لاستخراج المعلومات والميزات التي قد تكون مخفية، حيث يمكن أن يقدم البحث مساهمة متواضعة في هذا المجال.

2. الدراسات السابقة Related Works

قدمت الدراسات السابقة مجموعة من الفوائد في تحديد الانظمة الالكترونية باستخدام قواعد البيانات وإدارة السجلات الطبية PHR(Personal Healthcare Records System). فضلا عن تحديد دور الابتكار الالكتروني في تثقيف المرضى والوقاية من الامراض وتحسين التشخيص والعلاج القائم على الابحاث وتقليل تكاليف الرعاية الصحية وتمكين المرضى من إدارة الحالات طويلة الاجل كما اوضحت اهمية ودور الانظمة الالكترونية في إدارة المستشفيات في التخلص من العمليات غير الضرورية والتوفير العام وتقليل التكاليف المالية.

في الورقة البحثية [3] تم استخدام تقنيات التنقيب عن البيانات لطلبة بعض المدارس الابتدائية في مدينة صبراتة في ليبيا، إذ تم تحليل البيانات وبناء نموذج شجرة القرار Decision Tree عن طريق برنامج WEKA3.8.5، للتنبؤ بمرض ألم الظهر والرقبة الناتج أثر حمل الطالب للحقيبة المدرسية. قدمت الدراسة [6] مراجعة عن بعض الموضوعات المتعلقة بخطوات تنقيب البيانات وتصف أيضا خطوات كيفية استخدام أداة WEKA لمختلف التقنيات والمرافق المختلفة لتصنيف البيانات من خلال خوارزميات مختلفة مثل خوارزميات التصنيف الشائعة المستخدمة في برنامج WEKA لتصنيف البيانات مثل شجرة القرار، وأقرب جار Neighbor K، وNaive Bayes، وSupport Vector Machine. تقترح الورقة البحثية [4] إطارا للتعامل مع السجلات الطبية للأسنان في باكستان. تركز الدراسة على تنقيب البيانات، وخاصة تقنيات التعلم غير الخاضعة للإشراف لاستخراج المعلومات من سجلات طبية مختلفة. تحوي هذه السجلات على جميع أنواع الأنظمة المستخدمة في مختبرات الأسنان. بشكل عام، تحتوي السجلات على معلومات حول خطوات بدء العملية وإنهائها والبيانات ذات الصلة. تم استخدام عملية التنقيب في اماكن متنوعة، مثل التصنيع عالي التقنية والعمليات السريرية في المستشفيات. يتضمن التنقيب الحصول على فهم لهذه العمليات (مثل معلومات الأداء). وبالتالي، أدت الى ان هناك اختلافاً كبيراً بين عملية استخلاص البيانات. في الورقة البحثية [7] تم تحليل أدوات استخراج البيانات WEKA وOrange على أساس تنفيذ المتغيرات (parameters). الهدف الرئيس من هذه المقارنة هو مساعدة الباحثين على اختيار الأداة المناسبة بينهما. من خلال استخدام الدراسة التجريبية، تم استنتاج أن أداة WEKA أفضل من Orange. يمكن القول أن WEKA لديها معظم الميزات المرغوبة لمنصة تعمل بكامل طاقتها وسهلة الاستخدام لمشكلات التصنيف والعنقدة. لذلك، يمكن التوصية بـ WEKA كبرنامج لحل مشكلات تنقيب البيانات.

3. المواد وطرائق العمل Materials and Approaches

قاعدة البيانات Database DB عبارة عن مجموعة من البيانات ذات العلاقة منتظمة ومخزونة الكترونياً. يمكن أن تحتوي قاعدة البيانات على انواع مختلفة من البيانات بما في ذلك الكلمات، الأرقام، الصور، مقاطع الفيديو، والصوت. يُستخدم برنامج نظام إدارة قواعد البيانات Database Management System DBMS لتخزين البيانات واسترجاعها وتحديثها. تنقيب البيانات (المعروف أيضا باسم اكتشاف المعرفة من قواعد البيانات) هو عملية استخراج المعلومات (المخفية وغير المعروفة سابقا والتي يحتمل أن تكون مفيدة) من قاعدة البيانات. في العالم اليوم، أصبح استخراج البيانات المفيدة مثيرا للاهتمام وشائعا في جميع التطبيقات. يتطلب استخراج البيانات المفيدة كمية ضخمة من مجموعات البيانات لاستخراج المعرفة منها. الهدف الرئيس من برامج استخراج البيانات هو السماح للمستخدم بفحص البيانات ومن ثم القرار بأهمية البيانات من عدمه [6] .

أصبح تنقيب البيانات في مجال الرعاية الطبية أكثر شيوعا في عالم اليوم لأنه يوفر قدرا كبيرا من المعلومات المعقدة التي تشمل خدمات المستشفيات والأدوية والمعدات الطبية والمرضى وتشخيص الأمراض وما إلى ذلك. يجب معالجة هذه البيانات المعقدة وتقييمها لاسترجاع المعلومات، وهو أمر فعال من حيث التكلفة ومفيد جدا في اتخاذ القرارات [8]. الأسنان هي عبارةٌ عن تركيبٍ صلبٍ يكون في جيوب الفكين، ويبلغ عددها اثنان وثلاثون سناً، وهي: القواطع، والأنياب، والطواحن، والضواحك أو النواجذ. ويتكون السن من: التاج وهو الجزء البارز، والجذر وهو الجزء المغمور أو الداخلي، وتُسمى الطبقة الخارجية للأسنان بالمينا، والطبقة الخارجية التي تُغطي الجذور تُسمى بالملاط السنيّ [9].

للأسنان أهمية كبيرة في حياة الإنسان، ومن أهمها أنها تستخدم لمضغ الطعام وتفتيته إلى قطع صغيرةٍ ليسهل وصوله إلى المعدة من أجل الاستفادة منه. هناك الكثير من الأشخاص وخاصة الأطفال يعانون من سوء التغذية لعدم قدرتهم على مضغ الطعام بسبب مشكلات في أسنانهم. كما تساعد الأسنان على النطق والكلام بالشكل السليم فالكثير ممن لديهم مشكلات في أسنانهم يعانون من التأتأة في الكلام أو عدم القدرة على إخراج الحرف بشكله الطبيعي. لذا تعتبر الأسنان وسيلة مساعدة للسان في النطق السليم فهي جزء من جهاز النطق في الجسم إذ توجد الكثير من الحروف التي من ضمن مخارجها الأسنان كحرف الثاء والدال وغيرها. ناهيك عن المظهر الجمالي لها إذ تعتبر الابتسامة الجميلة والأسنان المصفوفة والبيضاء من مقومات الجمال لدى الكثير من الأشخاص حول العالم فضلا عن أهميتها في تقوية الشخصية وزيادة الثقة بالنفس فغالبا ما تسبب الأسنان غير الجميلة كآبة وحالة مزاجية متقلبة للشخص [10] . من هنا أتت أهمية هذا البحث في تنقيب بيانات طلبة ومراجعي كلية طب الاسنان في جامعة الموصل. وتمت عملية تنقيب البيانات باستخدام اداة WEKA وبعض خوارزمياتها المشهورة ومنها:

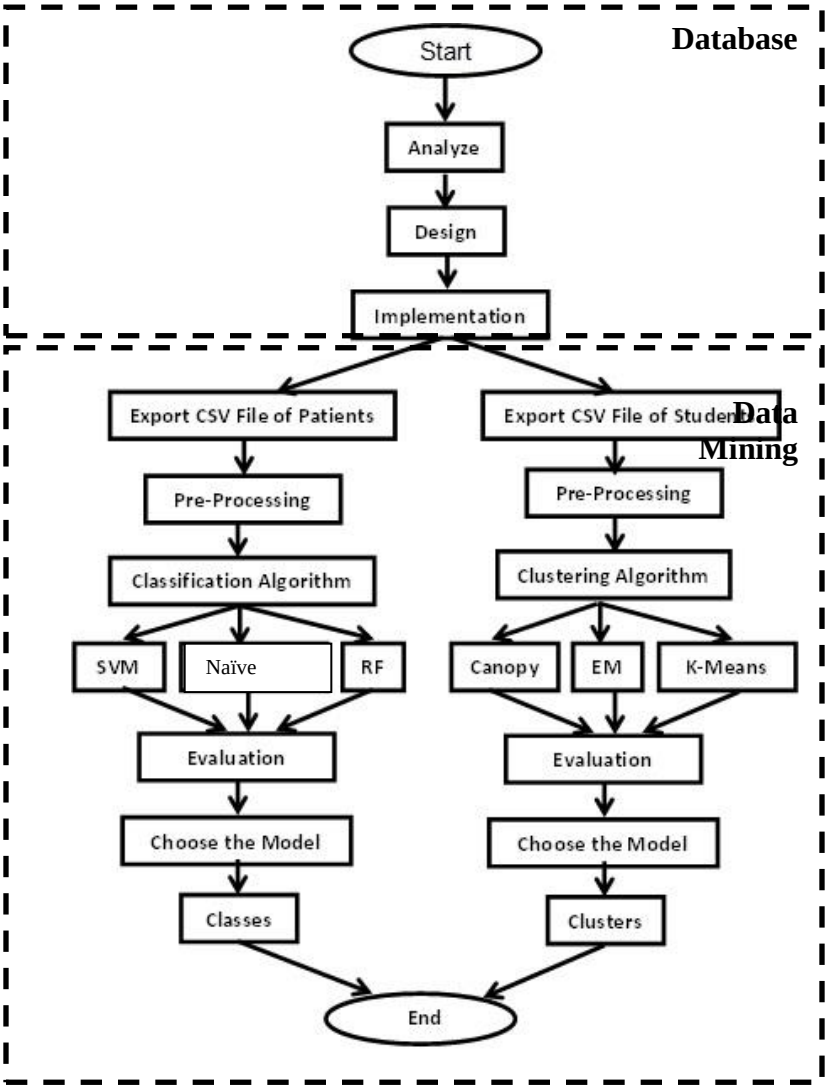
خوارزميات العنقدة، مع ذلك ركزنا على ثلاث منها فقط Canopy, K-Means, EM. خوارزمية التجميع (K-Means) من أكثر خوارزميات التجميع المعروفة بين علماء البيانات، وتقدّم في كثير من المحاضرات المتعلقة بعلوم البيانات؛ لسهولة فهمها وتطبيقها. ومن مميزاتها أنها سريعة جداً؛ لقلة الحسابات المستخدمة لتحديد المجموعات ومراكزها، إذ إن تعقيد الخوارزمية والوقت المستهلك لكامل العملية، فقط تزيد خطياً بزيادة مجموع عدد النقاط الموجودة في البيانات O(n) [11]. خوارزمية Canopy تعد أسرع ولكنها ذات دقة أقل. تختلف عن خوارزميات التجميع التقليدية (K-Means)، فإن الميزة الأكبر لتجميع البيانات هي أنها لا تحتاج إلى تحديد قيمة k (اي عدد المجموعات) مسبقًا ، لذلك لها قيمة تطبيق عملية كبيرة. خوارزمية Expectation Maximization (EM) تعمل على اختيار قيم عشوائية لنقاط البيانات وتستخدم التخمين لتقدير مجموعة جديدة من البيانات، ومن ثم تستخدم القيم الجديدة لإعادة التخمين مرة اخرة حتى الحصول على أفضل النتائج.

كذلك هناك مجموعة كبيرة من خوارزميات التصنيف، مع ذلك تم التركيز على ثلاث منها فقط SVM, Naïve Bayes, Random Forest. الة دعم المتجه Support Vector Machine SVM خوارزمية تستخدم للتصنيف، الانحدار، وتمييز الانماط. والهدف منها هو ايجاد أفضل دالة تصنيف وايضاً تهدف الى التمييز بين اعضاء فئتين من بيانات التدريب. فكرة الخوارزمية هي ايجاد مستوى مثالي Optimal Hyper Plane يفصل بين الفئتين والذي يستخدم للتصنيف وتحديد كل نمط. من ميزاتها الدقة العالية في التصنيف وتطبق في مجالات واسعة، منها تحديد فئات النص Text categorization، تصنيف الصور Image classification، وفي التطبيقات الطبية [12] . خوارزمية Naïve Bayes هي خوارزمية هامة لعدة أسباب، منها سهلة ولا تحتاج لمخططات التخمين Estimation schemes لاي متغيرات تكرارية معقدة، وقد تطبق بسهولة على مجموعة بيانات ضخمة والهدف من الخوارزمية بناء قاعدة تسمح بتخصيص هياكل مستقلة الى صنف معين وذلك بإعطاء متجهات للمتغيرات Vector of variable التي تصف ذلك الهيكل ويمكن بواسطتها اجراء المستخدم العديد من الاحصاءات لسهولتها [11] . خوارزمية Random Forest تصنف مجموعة البيانات الى مجموعات متفرعة ولكل مجموعة فرعية شجرة قرار خاصة بها تؤدي الى زيادة دقة النتائج.

4. الجانب العملي Practical Side

النظام المقترح يتضمن خطوتين رئيستين (انظر الشكل (1)):

الخطوة الاولى First Step: تشمل هذه الخطوة بناء نظام قاعدة البيانات والذي يتكون من ثلاث مراحل؛ التحليل والتصميم والتنفيذ.



الشكل 1. المخطط العام للنظام المقترح

1. تحليل النظام: بعدد من الخطوات المدروسة والمتسلسلة تم تحليل النظام بدأ من دراسة البنى التحتية الواقعية لنظام العيادات التخصصية لكلية طب الاسنان في جامعة الموصل وجمع المعلومات كافة عن كيفية دخول المريض وفحصه وتحويله الى الفرع المختص بمعالجة حالته ومن ثم وصف العلاج وإعطاء موعد المراجعة، فضلا عن جمع بيانات المرضى وطلبة الكلية للمرحلتين الرابعة والخامسة والمشمولين باستخدام العيادات.
2. التصميم: تشمل هذه الخطوة تحديد الكينونات Entities والصفات Attributes لكل كينونة Entity وتحديد العلاقات Relationship بين الكيانات ومن ثم رسم مخطط Entity Relationship (ER) Diagram . بعد ذلك يتم تحويل المخطط ER إلى Relation Model (RM) ومن ثم اجراء التطبيع Normalization على الجداول لتهيئتها للمرحلة التالية. يتم اختبار تسوية النظام لغرض التأكد من أن النظام يخضع لقوانين الـ Normalization "التطبيع او التسوية" وهو الاسلوب الاكثر ممارسة لتحليل قواعد البيانات العلائقية، الذي يهدف الى انشاء مجموعة من الجداول العلائقية مع الحد الادنى من تكرار البيانات والحفاظ على التناسق وتسهيل الادخال الصحيح، الحذف، والتعديل. تعتمد نظرية التطبيع بشكل كبير على نظرية التبعيات الوظيفية كما تمر هذه العملية بستة مراحل تسمى نماذج التطبيع (Normal Forms) والمراحل الاكثر استخداماً تصنف بالتوالي: انموذج التطبيع الاولي (1NF)، انموذج التطبيع الثاني (2NF)، انموذج التطبيع الثالث (3NF). وحسب الترتيب فإن النموذج الثاني لتطبيع الجدول أفضل من النموذج الاول، ويعد الثالث أفضل من الثاني، كلما قمنا بتسوية الجدول نكون قد حسَّنَّا من تصميمنا لقاعدة البيانات. هذا يعني ان أعلى (NF) يكون لديها عدد اقل من التكرار، ونتيجة لذلك، عدد اقل من مشكلات التحديث. ويتضمن كل مستوى مجموعة من

خصائص التبعية التي يجب ان يفي بها المخطط وكل منها يعطي ضمانات حول وجود او عدم وجود تغييرات غير طبيعية في التحديث كما وان لكل نموذج مجموعة من الشروط يجب تطبيقها على قاعدة البيانات من اجل الحصول على قاعدة بيانات خالية من مشكلات الحذف، التحديث، الاضافة والتكرار [13] .

فصلت العلاقات وقسمت الجداول الكبيرة الى جداول أصغر وتم الربط بينها بعلاقات لتكون جميعها ضمن صيغة (3NF) حيث اصبحت الجداول خالية من اي صفة متعددة واعتمادية جميع الصفات غير المفتاحية وظيفياً وكلياً على المفتاح الاساس ولا يوجد اي اعتماد وظيفي ما بين الصفات غير المفتاحية لتكون البيانات الاساسية للطلبة والمرضى.

3. التنفيذ: في هذه المرحلة يتم تحويل RM إلى قاعدة بيانات بواسطة ايعازات SQL وإدخال معلومات المرضى والطلاب وتهيئة البيانات للخطوة الثانية.

الخطوة الثانية Second Step: تشمل هذه الخطوة تنقيب قاعدة البيانات التي تم تكوينها في الخطوة الاولى. إذ يتم استدعاء البيانات من قاعدة البيانات بالاعتماد على مجموعة من الاستعلامات (مثلا، معلومات طلاب طب الاسنان – اسم الطالب، المستوى العلمي، الجنس، العنوان)، وتحويلها الى ملفات برنامج Excel بإتباع الخطوات التالية:

- بعد انشاء قاعدة البيانات ننقر على الملف المراد تحويلة (سواء اكان جدول Table ام منظور View).
- نختار من القائمة المنسدلة الخيار Task ثم بعد ذلك export database.
- نختار الأمر Next وبعد ذلك نختار Data Source وبعد ذلك SQL Server Native Client 11.0.
- نختار Next ثم Destination ثم Microsoft Excel ثم Browse ونختار الملف الذي سنضع فيه البيانات ثم Open.
- ننقر على زر Next ثم Finish.

ملاحظة: من الممكن استخدام أي صيغة ملفات اخرى (مثلا TXT) بدلاً من ملفات Excel عند سحب البيانات من قاعدة البيانات SQL SERVER لإجراء التنقيب عن البيانات.

نقوم بعد ذلك بمرحلة تنقيب البيانات والتي تتكون من سبع خطوات (انظر الشكل 2):

1- جمع البيانات: الخطوة الأولى هي جمع البيانات التي نرغب بتحليلها. وهنا تمثل ملف الـ Excel الذي تم الحصول عليه من قاعدة البيانات.
2- المعالجة المسبقة للبيانات: بمجرد الحصول على البيانات، نحتاج إلى معالجتها مسبقا لإزالة أي ضوضاء أو قيم مفقودة أو معلومات غير ذات صلة. هذه الخطوة حاسمة لأن جودة البيانات ستؤثر على دقة النتائج.
3- استكشاف البيانات: في هذه الخطوة، يتم استكشاف البيانات لفهم خصائصها وتحديد أي أنماط أو علاقات.
4- تحويل البيانات: قد نحتاج إلى تحويل البيانات لجعلها مناسبة للتحليل. على سبيل المثال، قد نحتاج إلى تطبيع Normalize البيانات أو تحويلها إلى تنسيق مختلف.
5- نمذجة البيانات: هذه هي المرحلة التي نقوم فيها بتطبيق خوارزميات التعلم الآلي على البيانات لإنشاء نموذج Model يمكنه عنقدتها او تصنيفها.
6- التقييم: بمجرد أن يكون لدينا نموذج، نحتاج إلى تقييم أدائه لمعرفة مدى قدرته على العنقدة او التصنيف.
7- النشر: المرحلة الأخيرة هي نشر النموذج في العالم الحقيقي، حيث يمكن استخدامه لعمل العنقدة او التصنيف على بيانات جديدة.

الشكل 2. خطوات التنقيب عن البيانات [14]

تم استخدام طريقة تعلم الآلأة Machine Learning للتنقيب عن البيانات والتي تستخدم خوارزميات عديدة مثل: EM, K-Mean, Canopy لعملية العنقدة وخوارزميات SVM, Random Forest, Naïve Bayes لعملية التصنيف.

عنقدة بيانات الطلاب Clustering Students Information

العنقدة(التجميع): هي آلية أساسية في معالجة البيانات وتطبيقات التعلم الآلي، وبالتالي فهي مجال بحث أساسي. إنها مهمة تجميع مجموعة كائنات بحيث تكون الكائنات في نفس المجموعة أكثر تشابها مع بعضها البعض من تلك الموجودة في المجموعات الأخرى، وبالتالي فهي تساعد في فهم واكتشاف التجميع الطبيعي في مجموعة البيانات [15] .

تمت عملية العنقدة على مرحلتين. الأولى هي تقسيم الطلاب أبجدياً الى أربع مجاميع حسب أيام الأسبوع A، B، C، D للأيام الاحد، الاثنين، الثلاثاء، والأربعاء على التوالي. الثانية هي بالاعتماد على عنقدة ذكية تم اقتراحها باتباع خطوات التنقيب المذكورة انفاً. حيث تمت عنقدة بيانات الطلاب حسب الخطوات الاتية:

- الدخول الى برنامج WEKA ومن الواجهة الرسومية GUI ننقر على الخيار Explorer (الشكل 6).
- من زر Open File نختار ملف Excel المعني من الموقع مع تحديد نوع البيانات ذات الامتداد CSV (الشكل 7). (الخطوة 1 للتنقيب).
- نقوم بالمعالجات المسبقة واستكشاف وتحويل البيانات. مثل تحويل صيغة الحقل S-Gender من صيغة نصية إلى صيغة عددية (Male إلى 1000 و Female إلى 2000). وكذلك حقل S_Region (Mosul إلى 5000 ، Erbil، Duhouk إلى 4000، Talkif، Hamdanyia، Namrood إلى 3000) (الخطوات 2-4 للتنقيب).
- من شريط القوائم نختار Cluster (شكل 8).
- نختار الخوارزمية المطلوبة ونجري بعض التعديلات على خصائصها (مثلاً Num Cluster و Seed ثم OK).
- من زر Ignore Attributes نحدد الصفات غير المهمة ثم Select .
- بالنقر على زر Start ستظهر لنا النتائج في نافذة Result List (الخطوة 5 للتنقيب).
- بالنقر على اخراج العنقدة بالزر الايمن للفأرة ستظهر قائمة منسدلة نختار منها Visualize Cluster Assignments ستظهر لنا اشكال العنقدة ونقوم بحفظها من خلال الامر Save للحصول على نموذج العنقدة المطلوب (الخطوات 6-7 للتنقيب).

تصنيف بيانات المرضى Classification Patients Information

عملية **التصنيف** تختص بمعلومات المرضى. بتعبير اخر، تحديد العلاج المطلوب لكل مريض وتحويله الى الفرع المعني. تصنيف بيانات المرضى تم كذلك بإتباع الخطوات الاتية:

- الدخول الى برنامج WEKA ومن الواجهة الرسومية GUI ننقر على الخيار Explorer (الشكل 6).
- من زر Open File نختار ملف Excel المعني من الموقع مع تحديد نوع البيانات ذات الامتداد CSV (الشكل 7). (الخطوة 1 للتنقيب).
- نقوم بالمعالجات المسبقة واستكشاف وتحويل البيانات. تم تحويل بعض البيانات الى الصيغة الثنائية، إذ إن لصفة Oral Maxillofaction الرقم 0 يعني ان المريض ليست لدية امراض في اللثة اما الرقم 1 فيعني ان المريض يعاني من امراض في اللثة. العمود الثاني Pediatrics الرقم 0 يعني ان عمر المريض أكبر من 12 سنة والقيمة 1 تعني ان عمر المريض اقل من 12 سنة. العمود الثالث Conservative القيمة 0 له تعني ان المريض لا يحتاج لحشو الاسنان اما 1 فيحتاج للحشو (الخطوات 2-4 للتنقيب).
- من شريط القوائم نختار Classify (شكل 8).
- نختار الخوارزمية المطلوبة.
- من نافذة Test Option نقوم باختيار نوع الاختبار وتحديد كمية البيانات لمرحلة التدريب Training او عدد Cross Validation Fold.
- بالنقر على زر Start ستظهر لنا النتائج في نافذة Result List . (الخطوة 5 للتنقيب).
- بالنقر على اخراج التصنيف بالزر الايمن للفأرة ستظهر قائمة منسدلة نختار منها Visualize Classification Error ستظهر لنا اشكال التصنيف ونقوم بحفظها من خلال الامر Save للحصول على نموذج التصنيف المطلوب (الخطوات 6-7 للتنقيب).

بعد تصنيف البيانات بالخوارزميات المحددة نختار أفضل النتائج بالاعتماد على نتائج المقياسين Recall وPrecision.

Recall: هو اظهار عدد الحالات الإيجابية الحقيقية التي يمكن للنموذج التنبؤ بها بشكل صحيح ويعد مقياسًا جيدًا لاستخدامه عندما تكون السلبية الكاذبة أكثر أهمية من الايجابية الكاذبة وحسب المعادلة التالية:

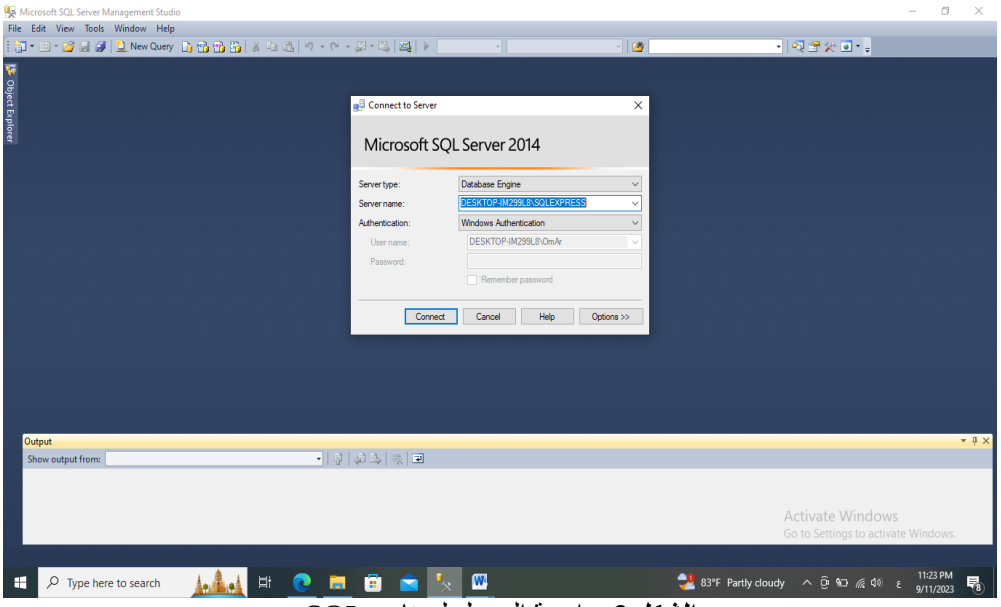
$$\text{Recall} = \frac{true\ positive}{true\ positive + false\ nagative} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 1$$

Precision: بالنسبة للفئة هو عدد الإيجابيات الحقيقية (أي عدد العناصر المصنفة بشكل صحيح على أنها تنتمي إلى الفئة الموجبة) مقسوما على إجمالي عدد العناصر المصنفة على أنها تنتمي إلى الفئة الموجبة وحسب المعادلة التالية [16]:

$$2\text{Precision} = \frac{true\ positive}{true\ positive + false\ positive} \quad \text{.....2}$$

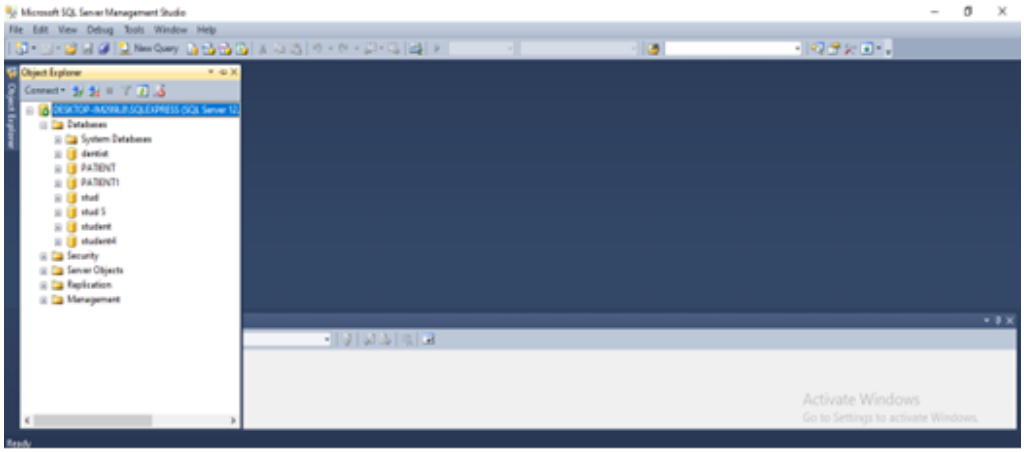
واجهات النظام المقترح Proposed System Interfaces

يحتوي النظام المقترح على مجموعة من الواجهات التي يمكن تلخيصها بالأشكال الاتية: الشكل (3) يمثل الواجهة الرئيسة لنظام قاعدة البيانات.



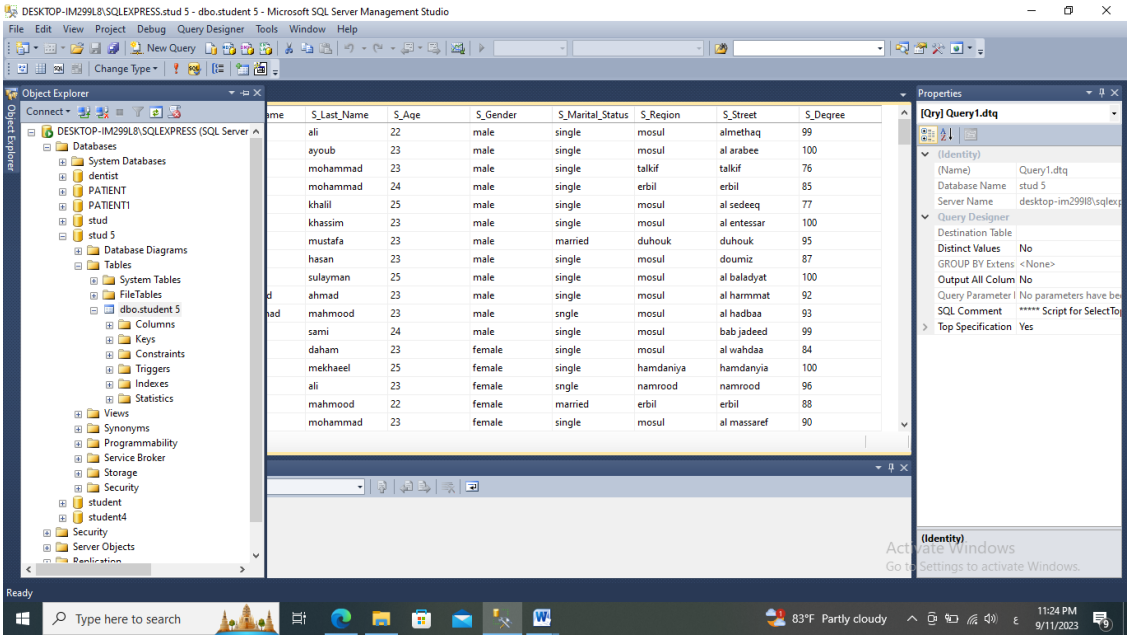
الشكل 3. واجهة الدخول لبرنامج SQL

الشكل (4) يمثل تكوين قاعدة البيانات.



الشكل 4. خطوات تكوين قاعدة البيانات

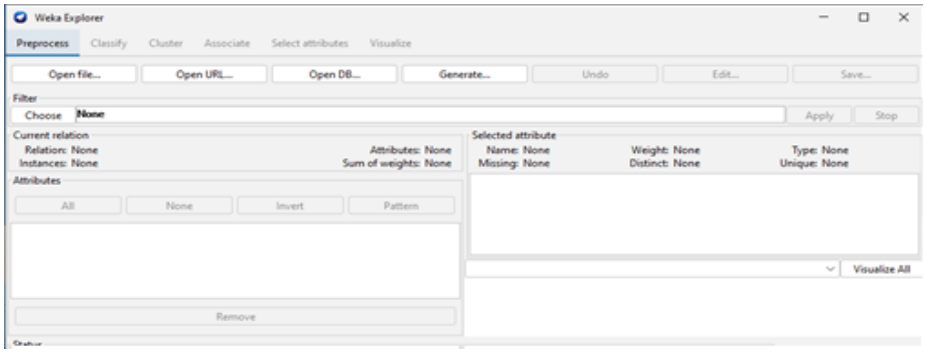
الشكل (5) يمثل ادخال البيانات الى قاعدة البيانات.



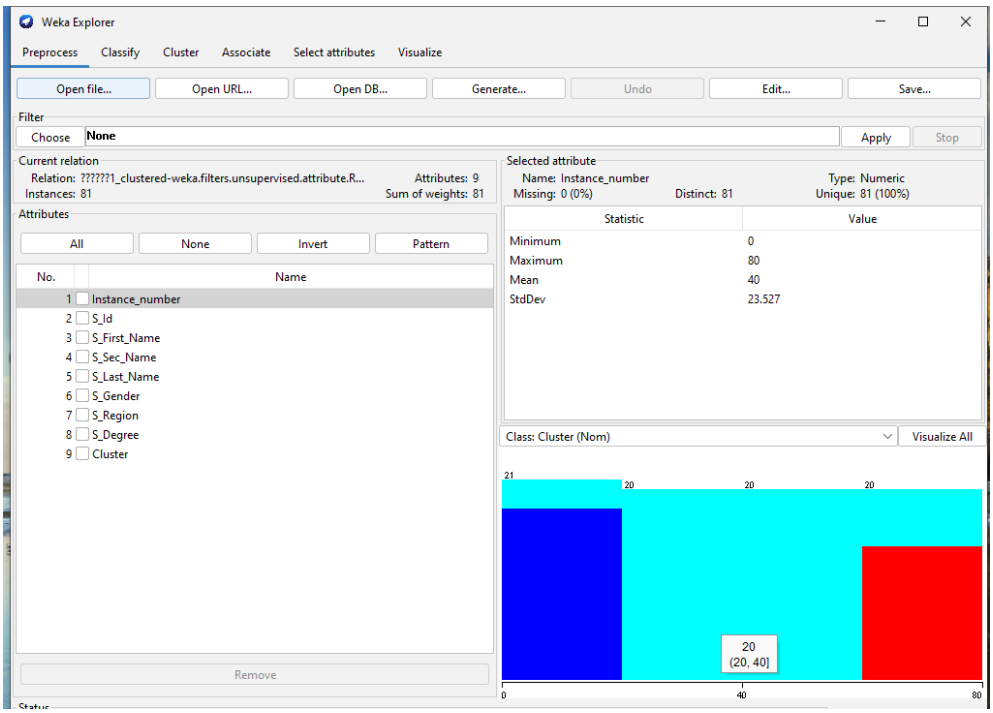
الشكل 5. خطوات ادخال البيانات الى قاعدة البيانات

برنامج التنقيب WEKA بدأ تطوير برنامج WEKA في عام 1997، وهو برنامج مفتوح المصدر يعمل على منصة تشغيلJava ، والذي تم تطويره من قبل جامعة Waikato في نيوزلندا [4] ، واستخدم في الكثير من المجالات التطبيقية المختلفة خاصة للأغراض البحثية والتعليمية. واجهة المستخدم في تطبيق WEKA تنقسم الى مجموعة من الأدوات والخوارزميات لتحليل البيانات وبناء نماذج التنبؤ معاً لتشكل واجهة رسومية لتسهيل استخدامه. كما أنه يدعم أنواع مختلفة من تنسيقات الملفات مثل[5] CSV, ARFF.

الشكل 6. الواجهة الرئيسة لبرنامج WEKA



الشكل 7. الواجهة الثانية لبرنامج WEKA



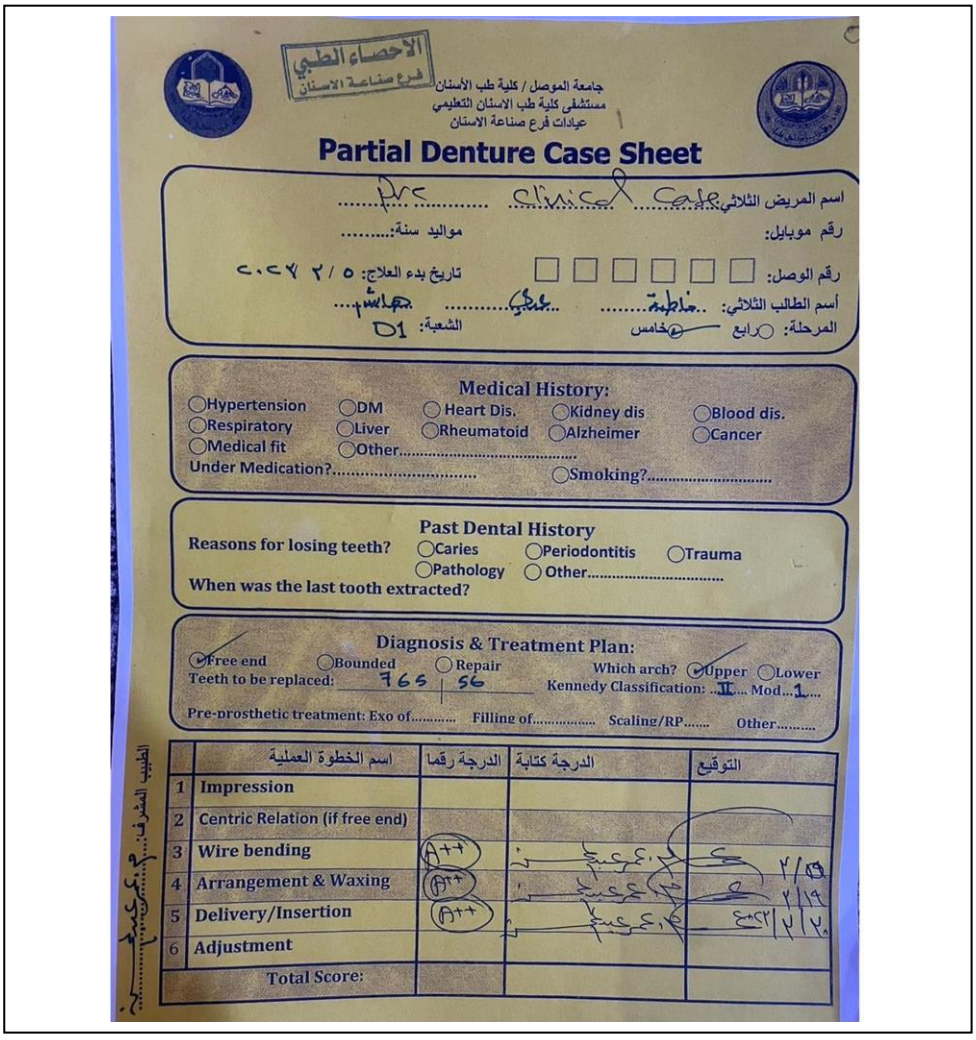
الشكل 8. واجهة المعالجة لبرنامج WEKA

6. المناقشة Discussion

في هذا البحث تم إنشاء نظام قاعدة بيانات لكلية طب الاسنان في جامعة الموصل يضم معلومات الطلبة والمراجعين وذلك من خلال جمع بيانات المرضى عن طريق Case Sheet الخاصة للمرضى (انظر الشكل 9) وتم الحصول على بيانات الطلاب (اسماء الطلاب فقط حفاظا على خصوصية معلوماتهم) من الكلية. يوفر النظام المقترح إمكانية الاستعلام عن معلومات المراجعين والطلبة بكل سهولة ويسر من خلال اوامر برنامج SQL. فضلا عن إمكانية تنقيب هذه البيانات باستخدام العديد من خوارزميات التعلم الالي وكذلك توزيع الطلاب على مختبرات العيادات التخصصية حسب الطاقة الاستيعابية لكل عيادة باستخدام طرائق العنقدة واختيار الخوارزمية التي تعطي نتائج بدقة عالية. كما تم توزيع الحالات المرضية على الطلاب باستخدام خوارزميات التصنيف واختيار الخوارزمية الافضل. مما سيوفر جهداً ووقتاً في تنظيم توزيع المرضى على الطلبة وكذلك تخفيف الزخم الحاصل داخل العيادات. من المعوقات التي واجهت الباحث هي صعوبة الحصول على بيانات المرضى (بسبب الخصوصية) وكذلك معلومات الطلاب (بسبب سرية معلومات الطلاب).

تتميز قواعد البيانات SQL-Server بكونها تعمل على بناء قواعد المعلومات المعرفية التي تمتاز بالسرعة والكفاءة في معالجة وتخزين واسترجاع المعلومات عند تنفيذ الانشطة والعمليات المطلوبة، كما وتساعد في تزويد كلية طب الاسنان بمختلف البيانات والمعلومات اللازمة للمرضى مع الاخذ بنظر الاعتبار اخفاء بعض معلومات المرضى بسبب اخلاقيات المهنة وخصوصية المريض وبناءاً

على طلبهم مثل رقم الهاتف وعنوان السكن والحالة الزوجية وعدد الاطفال والعمر والدخل الشهري للمريض. كما وتستعمل لغة البرمجة القياسية SQL للوصول الى قاعدة البيانات وتتبع خصائص ACID (الذرية، الاتساق، العزل، والمتانة) لقاعدة البيانات.



الشكل (9) Case sheet

بعد تطبيق ثلاث خوارزميات للعنقدة (EM، Canopy، K-Mean) على بيانات الطلاب ومقارنة النتائج تم اختيار خوارزمية K-Meansكأفضل خوارزمية وذلك لقيامها بتقسيم الطلاب حسب عدد العناقيد المطلوبة بدقة عالية وسهولة وسرعة عالية بسبب قلة العمليات الحسابية لها، اما خوارزميتا Canopy وEM فقدمتا نتائج صحيحة ايضا ولكن بدقة اقل. اما لتصنيف بيانات المرضى فتم اختيار ثلاث خوارزميات ايضا (SVM، Naïve Bayes، Random Forest). تم الحصول على افضل نتائج من خلال تطبيق خوارزمية Naïve Bayes وذلك بالاعتماد على مقياسي Recall وPrecision اللذين اعطيا افضل نتائج تصنيف للمرضى. تم مقارنة نتائج التصنيف واختيار اعلى نسبة وحسب الجدول التالي:

جدول 1. قيم مقاييس دقة التصنيف

Class	Naïve bayes		Random forest		SVM	
	Recall	Precision	Recall	Precision	Recall	Precision
Oral Maxillofaction	1.000	0.996	0.996	0.905	1.000	0.995
Pedicatric	0.988	0.993	0.798	0.815	0.982	0.989
Conservative	0.986	0.975	0.838	0.640	0.982	0.963
Prosthesis	0.940	0.971	0.798	0.830	0.928	0.971
Weighted Avg	0.990	0.990	0.865	0.851	0.987	0.987

7. الاستنتاجات Conclusion

تم في هذا البحث تصميم قاعدة بيانات لطلاب المرحلتين الرابعة والخامسة لطلاب كلية طب الاسنان في جامعة الموصل وتقسيم الطلاب الى اربع مجاميع وكل مجموعة قسمت بالاعتماد على عدد الكراسي المستخدمة في معالجة المرضى لكل فرع من خلال استخدام برنامج التنقيب WEKA وتطبيق خوارزميات العنقدة K-Mean، Canopy، EM وكذلك استخدام خوارزميات التصنيف SVM، Naïve Bayes، Random Forest إذ يسهم النظام المقترح بتسهيل توزيع الحالات المرضية وإحالتها للفرع المعني بالمعالجة وتوزيعها على الطلبة المعالجين للحالات وكذلك تخفيف الزخم الحاصل بالنسبة للطلبة من خلال تقسيمهم الى مجاميع بعدد مناسب بحسب عدد الكراسي المستخدمة لمعالجة المرضى. بمقارنة النتائج التي تم الحصول عليها مع الدراسة [4] تبين ان البحث يقدم تسهيلات أكثر من خلال توزيع الطلاب على مجاميع حسب الطاقة الاستيعابية لكل مختبر وكذلك تصنيف المرضى حسب الحالة المرضية على أفرع الكلية والتي يفتقر اليها البحث المذكور انفاً. من الاعمال المستقبلية الممكن اجراؤها على النظام المقترح هي تطبيق خوارزميات حديثة للتنقيب والتي من الممكن الحصول من خلالها على نتائج أفضل وبدقه اعلى، تطوير قاعدة البيانات لتشمل المرضى في المستشفيات الأخرى (مثل العيادات الاستشارية في كلية الطب) لغرض تقديم أفضل الخدمات المستقبلية لهم، اضافة امكانيات الأمنية للحفاظ على بيانات الطلبة والمرضى من السرقة.

8. الشكر والتقدير Acknowledgment

أقدم شكري وتقديري وامتناني لكلية علوم الحاسوب والرياضيات / جامعة الموصل / العراق على دعمها في إنجاز هذا البحث.

9. المصادر References

[1] World Health Organization, 2019, Oral Health, (who.int), https://www.who.int/news-room/fact-sheets/detail/oral-health

[2] T. Muhammad, A. Abeer, "A Clinical Study to Investigate the Prevalence of Tongue Lesions Among Patients Visiting the Clinics of The Faculty of Dentistry in Damascus University", *Journal of Hama University*, Vol 3, No. 2, 2020, https://hama-univ.edu.sy/ojs/index.php/huj/article/view/336

[3] A. Samira, W. Hamida, A. Al-Shawashi, E. Al-Attab, "Building A Decision Tree to Predict Back and Neck Pain Using Data Mining Techniques", Vol 4, No. 2, pp.: 108-114, 2022, sjst.scst.edu.ly

[4] B. Iqra and B. Dr, Shoab Ahmad Khan," A Framework for Clustering Dental Patients' Records Using Unsupervised Learning Techniques**",** *IEEE*, pp. 1-9, 2015,www.conference.thesai.org

[5] S. Saichon," Efficiency Comparison in Prediction of Normalization with Data Mining Classification", *ASTESJ*, pp. 130, 2021, www.astesj.com

[6] I. Fatma and A. Shiba, Omar " Data Mining: WEKA Software, " JOPAS, 2019, pp. 54, www.Suj.sebhau.edu.ly

[7] R. Ratra and P. Gulia," Experimental Evaluation of Open Source Data Mining Tools (WEKA and Orange)", IJETT *Journal*. Vol. 68, pp. 30-35, 2020, doi: 10.14445/22315381/IJETT-V68I8P206S

[8] R. Riad K. AL-Taie and B. Jumaa Saleh and A. Yousif Falih Saedi and L. Abdalhasan ySalman "Analysis Of Weka Data Mining Techniques For Heart Disease Prediction System", *IJECE*, Vol. 11, No. 6 pp. 5229 - 5239 , 2021 , DOI: 1op0.11591/ijece.v11i6.pp5229-5239

[9] https://se77ah.com/art

[10] L. Maryam, "A Scientific Article About the Importance of Oral and Dental Health", 2022, https://uomus.edu.iq/NewDep.aspx?depid=11&newid=12062#:~:tex
[11] The 5 Clustering Algorithms Data Scientists Need to Know - KDnuggets
[12] R.Zuhir Abdalgani and G.Abdalaziz Altaleb," A comparative study of algorithms for prospecting opinions and loading emotions and its applications", ", Raf. J. of Comp. & Math's., Vol. 12, No. 2, 2018,
[12] V. Archit," Evaluation of Classification Algorithms with Solutions to Class Imbalance Problem on Bank Marketing Dataset using WEKA", *IRJET*, Vol. 06, pp. 54-61, ,2019, www.irjet.net
[13] R. Tarik Elais Daoud," Designing and Implementation of an application for conducting auctions Electronically", University of Mosul, 2021.
[14] A. Thaher Yaseen," Data Mining Between Classical and Modern Applications ", Raf. J. of Comp. & Math's., Vol. 15, No. 2, 2021,
[15] Z. Uykan, "Fusion of Centroid-Based Clustering with Graph Clustering: An Expectation-Maximization-Based Hybrid Clustering," IEEE Transactions on Neural Networks and Learning Systems, vol. 34, no. 8, pp. 4068-4082, Aug. 2023, doi: 10.1109/TNNLS.2021.3121224.
[16] M.Ali Fadhe Abdullah."Arabic Alphabets Learning Application for Children Early Children based on Deep Learning",thesis not published, unvirsity of mosul, 2023.

تنقيب بيانات الطلاب والمرضى لكلية طب الأسنان في جامعة الموصل

مروة باسم مصطفى، عمار ظاهر ياسين ال عبدالعزيز

قسم علوم الحاسوب، كلية علوم الحاسوب والرياضيات، جامعة الموصل، الموصل، العراق

المستخلص

تتضمن هذه الورقة البحثية تصميم وتنفيذ نظام لتنقيب بيانات الطلاب والمرضى في كلية طب الاسنان في جامعة الموصل وذلك باستخدام نظام إدارة قواعد البيانات Microsoft SQL Server لتصميم وتنفيذ نظام قاعدة البيانات وبرنامج WEKA للتنقيب في قاعدة البيانات، واستخدمت لغة Microsoft Visual C#.NET 2012 لبرمجة واجهات النظام. الخطوات الرئيسة لقاعدة البيانات شملت التحليل والتصميم والتنفيذ كما تضمنت عملية التنقيب الخطوات السبع؛ جمع البيانات، والمعالجة المسبقة للبيانات، واستكشاف البيانات، وتحويل البيانات، ونمذجة البيانات، والتقييم، والنشر. عملية تنقيب قاعدة البيانات انقسمت الى جزئين؛ الجزء الأول هو عملية عنقدة ذكية لطلاب كلية طب الاسنان للمرحلتين الرابعة والخامسة على المختبرات (أي على عدد الكراسي المتوفرة لكل مختبر) باستخدام ثلاث خوارزميات مشهورة (Canopy، K-Means، EM)، الجزء الثاني هو عملية تصنيف المرضى الى أربعة أصناف حسب نوع المعالجة التي يحتاجها كل مريض باستخدام ثلاث خوارزميات مشهورة أيضا (SVM، Naïve Bayes، Random Forest). بعد تطبيق النظام على البيانات الحقيقية لكلية طب الاسنان في جامعة الموصل تبين ان أفضل خوارزمية عنقدة هي K-Means وأفضل خوارزمية تصنيف هي Naive Bayes.