Explore some patterns affecting the academic performance of students of the University of Science and Technology using data mining techniques/استكشاف بعض الأنماط المؤثرة في الأداء...

Article · May 2011

CITATIONS
0

READS
190

2 authors:

Awab Noori
Universiti Utara Malaysia
6 PUBLICATIONS **47** CITATIONS
SEE PROFILE

Ammar Thabit Zahary
Sana'a University
47 PUBLICATIONS **148** CITATIONS
SEE PROFILE

استكشاف بعض الأنماط المؤثرة في الأداء الأكاديمي لطلاب جامعة العلوم والتكنولوجيا باستخدام تقنيات التنقيب في البيانات

أواب الجناعي الحسين الحداد علي البار عمار الزهاري+

الملخص

يقدم هذا البحث دراسة تطبيقية في مجال اكتشاف المعرفة Knowledge Discovery باستخدام تقنيات التنقيب في البيانات Data Mining. الهدف الأساسي من الدراسة هو اكتشاف بعض الأنماط السائدة في البيانات الأكاديمية للطلاب في جامعة العلوم والتكنولوجيا اليمنية منذ العام 1994م وحتى العام 2005م ومن ثم الخروج بمؤشرات عامة حول الأداء الأكاديمي لدعم السياسات التعليمية لدى متخذي القرار في الجامعة، لا سيما وأن حجم البيانات وكذلك البعد الزمني الكبير نسبياً لهذه البيانات يدعّم من نتائج هذا البحث. في هذه الدراسة تم اكتشاف بعض الأنماط Patterns السائدة في هذه البيانات، وقد خلص البحث إلى وجود مجموعة من الأنماط التي يمكن أن تعطى مؤشرات ذات دلالة في الجانب التعليمي. من هذه الأنماط وجود ارتباط بين مستوى تحصيل الطالب لبعض المواد، وبين معدل الطالب في الثانوية واختيار التخصص في الجامعة، وكذلك علاقة المنح الدراسية بمستوى تحصيل الطالب أكاديمياً. قدم البحث محاولة لقراءة هذه النتائج وتفسيرها وعرضها والتحقق من مستواها ونوعيتها وذلك باطلاع متخذي القرار بالجامعة عليها. تم اختيار تقنيات التنقيب في البيانات كونها الأنسب للاستفادة من حجم هذه البيانات وكذلك لأنها تستخدم خوارزميات استنباطية ذكية تستخدم غالباً لدعم اتخاذ القرار. استخدمت طرائق مختلفة من تقنيات التنقيب في البيانات لدعم النتائج المكتشفة وهي قواعد الارتباط Association Rules والتصنيف بأشجار القرار Classification Using Decision Trees وذلك بعد عملية المعالجة الأولية Preprocessing لقاعدة البيانات وإعادة هيكلتها على شكل مستودع بيانات منطقي Data Warehouse Logical. وقد استخدمت خوارزميتي Apriori, Predictive Apriori في تقنية قواعد الارتباط، وخوارزميتي ID3, J48 في تقنية التصنيف بأشجار القرار. هذه الطرائق والخوارزميات تم تطبيقها من خلال الأداة WEKA التي تدعم الكثير من الخوارزميات والطرائق للتنقيب في البيانات. وفي الأخير تم استخلاص الاستنتاجات واقتراح بعض التوصيات التي تهم صانع القرار للاستفادة منها في تحسين الأداء الأكاديمي في الجامعة.

1. المقدمة

في ظل التطور المتسارع لتكنولوجيا الحاسوب والمعلوماتية فإن كمية البيانات التي يمكن للمؤسسات المختلفة تخزينها يتعاظم بشكل دراماتيكي، ولا ينتهي المطاف بالقدرة على تخزين هذه البيانات. حيث تأتي الخطوة الأهم وهي كيفية استثمار هذه الكميات الضخمة من البيانات. يمكن القول أن هذه البيانات تمثل ذاكرة المؤسسة وثروة حقيقية يمكن أن تهمل كما يمكن أن تستثمر بشكل ذكي في تحليل ما مضى من معاملات وأحداث وإجراءات داخل المؤسسة، ومن ثم يمكن أن يتم رسم السياسات والخطط الإستراتيجية على بصيرة ورؤية وموروث معلوماتي يصف مواضع نجاح وإخفاق المؤسسة ويتنبأ بفرص التحسين المستقبلي.

مع هذه الكميات الضخمة من البيانات فإن الطرائق التقليدية لتحليل البيانات والتي هي مزيج من الطرائق الإحصائية وبعض النظم الحاسوبية المصممة لإدارة قواعد البيانات باتت تعاني الكثير من المشكلات في التعامل مع هذا النوع من البيانات. بالإضافة إلى ذلك، فإن تلك الطرائق التقليدية تعتمد بشكل كلي على القدرات الذهنية والعملية والفنية ومن ثم على خبرة محلل البيانات في توجيه التحليل لاستخراج مؤشرات قيمة وعميقة لصناع القرار؛ حيث يفترض المحلل فرضيات سابقة للعلاقات الموجودة بين المفردات المختلفة للبيانات وجل ما يفعله بعد ذلك هو استخدام التحليل لبرهنة أو دحض هذه الفرضيات.

+قسم علوم الحاسوب وتقنية المعلومات، كلية العلوم والهندسة – جامعة العلوم والتكنولوجيا، ص.ب: 13064، صنعاء، اليمن

وبعيداً عن الطرائق التقليدية اتجه الباحثون لإيجاد طرق بديلة يمكن أن نسميها بالاستنباطية لتقوم بدور يحاكي دور محلل البيانات لاستكشاف العلاقات والأنماط السائدة في البيانات كما تصفها سجلات البيانات الفعلية لا كما يفترضها محلل البيانات. وكما هو متوقع تستخدم هذه الطرائق تقنيات وخوارزميات ذكية – أي أنها تحاكي نمط الاستنباط عند الانسان – فتلاحظ وتعمم وتستنبط وأخيراً تستنتج. من بين العلوم التطبيقية الحديثة في هذا المجال يأتي علم اكتشاف المعرفة في قواعد البيانات Knowledge Discovery in Databases (KDD) وعلم التنقيب في البيانات Data Mining على رأس هذه العلوم في توفير أطر عامة وطرائق وتقنيات وخوارزميات بل وأدوات مؤتمتة توجه وتسهل إجراء تحليلات ذكية وعميقة ومعقدة، واستكشاف أرقى أنواع المعلومات، والتي تسمى في هذا المجال بالـ "المعرفة" (Knowledge) ومن ثم توفيرها لصناع القرار بسرعة قياسية وبجودة عالية [7] [9].

تعد عملية إدارة المؤسسات التعليمية من الصعوبات التي تواجه القائمين عليها وذلك لكبر حجمها وتشعب هيكليتها وتعدد مصادر بياناتها ولذلك فإن المؤسسة التعليمية تواجه عدة مشاكل خلال إدارة العملية التعليمية منها مشاكل أكاديمية ومالية وإدارية. تحتاج هذه المشاكل إلى دراسة واستنتاج وتوصيات تساهم في دعم في عملية اتخاذ القرار الذي يسهل سير العملية التعليمية بناء على نظام معلومات مبني مسبقاً بطريقة علمية حديثة، وهذه من المشكلات الحقيقية التي ما زالت تقف في وجه أي نظام تنقيب في البيانات، في اليمن ومعظم الدول العربية، نظراً للأخطاء التي تصاحب بناء أنظمة المعلومات سواء من ناحية التحليل أو التصميم أو التنفيذ أو الصيانة أو بناء مخازن البيانات بشكل غير علمي وغير مدروس. في المؤسسات التعليمية والأكاديمية يمكن توظيف علمي استكشاف المعرفة والتنقيب في البيانات لتحسين الأداء الأكاديمي [6] [12] [14] باستنباط الأنماط السائدة في بيانات هذه المؤسسات حول الطلاب والخريجين وأعضاء هيئة التدريس، وارتباط كل ذلك بأهم مؤشرات الأداء كتحصيل الطلاب ومعدل بقائهم أو تسربهم ونوعية أداء أعضاء هيئة التدريس. كما يمكن للجامعات مثلاً أن تتنبأ بالطلاب الذين سيتسربون، والطلاب الذين سيكون تحصيلهم العلمي ضعيفاً، والطلاب الذين سيتخرجون، وغيرها من المعلومات الإستراتيجية، ومن ثم يمكنها بعد ذلك مراجعة وتطوير سياساتها التعليمية لمساعدة هؤلاء الطلاب في رفع مستوى تحصيلهم العلمي أو توجيههم لتخصصات تناسب استعداداتهم وميولهم وقدراتهم وما إلى ذلك من السياسات والتدابير التي تحسن من مستوى الأداء الأكاديمي في المؤسسة.

2. المشكلة التي يعالجها البحث

تتوفر لدى صانعي القرار بجامعة العلوم والتكنولوجيا اليمنية قاعدة بيانات ضخمة عن الطلاب والبرامج والمقررات والنتائج والمخرجات التعليمية وغيرها منذ تأسيس الجامعة عام 1994 وحتى اليوم. وهذا الكم الهائل من البيانات رغم ثرائه بالمعرفة لم يتم استغلاله حتى الآن بشكل فعال في معرفة عوامل النجاح وعوامل الفشل ولم يتم كذلك معرفة الأنماط السائدة المؤثرة على سلوك الطالب وهو من أهم عناصر التعليم في الجامعة، من قبيل معرفة أسباب التسرب أو التحويل وكذلك أسباب فشل الطالب أو تفوقه وماهية الأنماط التي تلعب دوراً في ذلك كله. أضف إلى ذلك أنه لم يتم حتى الآن تحويل قاعدة بيانات الجامعة إلى مستودع للبيانات لكي يسهل دراسة وتحليل البيانات بشكل دوري. وقد ظلت قاعدة البيانات تلك الثرية بالمعرفة حبيسة المزودات Servers الخاصة بالجامعة رغم تطور تقنيات التنقيب في البيانات Data Mining التي يمكن أن تستخدم لدعم صانعي القرار، هذه التقنيات يمكن أن تزودنا بنتائج تظهر الارتباطات بين الأنماط المختلفة داخل هيكل الجامعة. تلك النتائج تظهر لصانع القرار مواطن الضعف والقوة ليقوم بما هو مفيد في تقويم الأداء داخل الجامعة.

3. أهداف البحث

يقدم هذا البحث دراسة تطبيقية في مجال التنقيب في البيانات Data Mining لدعم صانعي القرار في جامعة العلوم والتكنولوجيا اليمنية، وذلك عن طريق استخلاص بعض الأنماط التي يمكن أن تساهم في تطوير العملية التعليمية في الجامعة من خلال تطبيق تقنيات التنقيب في البيانات مثل Decision Trees و Association Rules وذلك لكشف جوانب المعرفة في قاعدة بيانات الجامعة من خلال النتائج التي تظهر لصانع القرار ما يجب فعله من أجل تحسين الأداء الأكاديمي الجامعة. ويمكن تقسيم الهدف الرئيس لهذا البحث إلى الأهداف الفرعية التالية:

1. اكتشاف الأنماط Patterns السائدة في بيانات الطلاب الأكاديمية في جامعة العلوم والتكنولوجيا اليمنية.
2. الخروج بنتائج تساعد في اتخاذ القرار في الجانب الاكاديمي في جامعة العلوم والتكنولوجيا.
3. بناء مستودع بيانات منطقي لقاعدة بيانات الجامعة.

4. حدود البحث

تتمحور حدود البحث فيما يلي:

1. تم تطبيق البحث في جامعة العلوم والتكنولوجيا اليمنية – المركز الرئيسي في صنعاء فقط.
2. تم تطبيق البحث على بيانات الطلاب دون الطالبات.
3. تم تطبيق البحث على البيانات الأكاديمية دون المالية والشخصية.
4. تم تطبيق البحث على بيانات يتراوح عمرها بين عامي 1994 و2005.
5. تم تطبيق البحث باستخدام تقنيتي قواعد الارتباط وأشجار القرار فقط دون غيرهما من التقنيات.
6. لم يتم في هذا البحث دراسة جميع تخصصات الجامعة ولكن تم اختيار التخصصات ذات العدد الأكبر للطلاب.

5. منهجية البحث

في هذا البحث تم اعتماد منهجية تتكون من عدة مراحل تتلخص فيما يأتي:

1. جمع البيانات (الحصول على قاعدة بيانات الطلاب المنتظمين في جامعة العلوم والتكنولوجيا).
2. المعالجة الأولية (Preprocessing).
3. بناء مستودع بيانات منطقي (Logical Model).
4. اختيار التطبيق المناسب لعملية التنقيب في البيانات .
5. تصدير البيانات من مستودع البيانات على الشكل الذي يتطلبه التطبيق.
6. استكشاف الأنماط المؤثرة بتطبيق تقنيات التنقيب في البيانات على البيانات المُصدَّرة في مرحلة معالجة البيانات Processing Stage.
7. مناقشة وتقييم النتائج.
8. صياغة المقترحات والتوصيات للجهات المعنية وصانعي القرار.

6. الخلفية النظرية

يتضمن هذا المقطع خلفية نظرية للبحث تعرض فيها خلفية تاريخية إضافة إلى أبرز المفاهيم النظرية المرتبطة بموضوع البحث.

1.6 مدخل تاريخي

لقد أدى التطور المتسارع لتكنولوجيا التخزين إلى تضخم حجم البيانات في المؤسسات المختلفة بصورة كبيرة، حيث يستفاد من تلك البيانات على شكل معلومات معرفية مفيدة. هذا التضخم الدراماتيكي في حجم البيانات في المؤسسات لم يواكب بطرق فعالة لاستثمار هذا الحجم الهائل للبيانات، لذا ظهر في الآونة الأخيرة تحدٍ جديد وهو كيفية تجاوز فكرة قواعد البيانات التقليدية والتي تقوم بالتخزين والبحث عن المعلومة فقط عن طريق أسئلة يوجهها الباحث إلى تقنيات تستخدم في استنتاج المعرفة من خلال استكشاف الأنماط السائدة في البيانات بهدف اتخاذ القرار والتخطيط وتكوين رؤية مستقبلية للمؤسسات [2] [5] [9]. من هذه التقنيات تقنية التنقيب في البيانات Data Mining الذي يعد من أبرز العلوم الحديثة المتخصصة في التعامل مع البيانات والمعلومات. ويمكن القول أن عمر التنقيب في البيانات يزيد بقليل عن عشرة أعوام حيث بدأ فعلياً مع بداية الألفية الثالثة [8] [10]. وقد تبلور مفهوم التنقيب في البيانات عبر عدة أجيال:

الجيل الأول .. كان على شكل إحصائيات تقليدية مثل معامل الارتباط والانحراف المعياري والانحسار والتباين وغيرها.

الجيل الثاني .. ظهر مع ظهور علم الذكاء الاصطناعي Artificial Intelligence وبهذه المرحلة تطورت هذه الإحصاءات وأصبحت ذكية في فهم العلاقات وكانت النتائج مهمة ومفيدة في صنع القرار.

الجيل الثالث .. وهو ما يعرف الآن بـ "تعليم الآلة – Machine Learning"، حيث يصبح جهاز الكمبيوتر قادرا على إطلاق التحذيرات أو التوجيهات وقد يقترح أيضاً بعض القرارات لصانع القرار [11] [20].

ما يعرف الآن بـ "التنقيب في البيانات" يضم الأجيال الثلاثة السابقة [8]، وهو الآن يحتوي تقنيات وخوارزميات كثيرة نحاول استثمار البيانات بالشكل الأمثل لاستنتاج المعرفة ودعم اتخاذ القرار.

2.6 مستودعات البيانات Data Warehouses

لأن علم اكتشاف المعرفة والتنقيب في البيانات نما وتبلور بعد استخدام وبناء المؤسسات المختلفة لقواعد البيانات الخاصة بها فإن الصورة الحالية لقواعد البيانات في هذه المؤسسات لا تمكنها من تطبيق خوارزميات التنقيب في البيانات مباشرة حيث ينبغي أن نجمع البيانات في مستودعات موحدة بشكل متكامل ومتناغم وخال من العيوب وهو ما يعرف اليوم بإسم مستودعات البيانات. ومستودعات البيانات هي مجموعة من البيانات المتكاملة والكبيرة والموضوعية التي تساعد في اتخاذ القرارات عن طريق تحليلها واستخلاص النتائج منها، وتتضمن هذه المستودعات (المخازن) بيانات من عدة مصادر مختلفة هي قواعد البيانات المنتشرة في أنحاء المؤسسة والتي تجرى عليها العمليات اليومية.

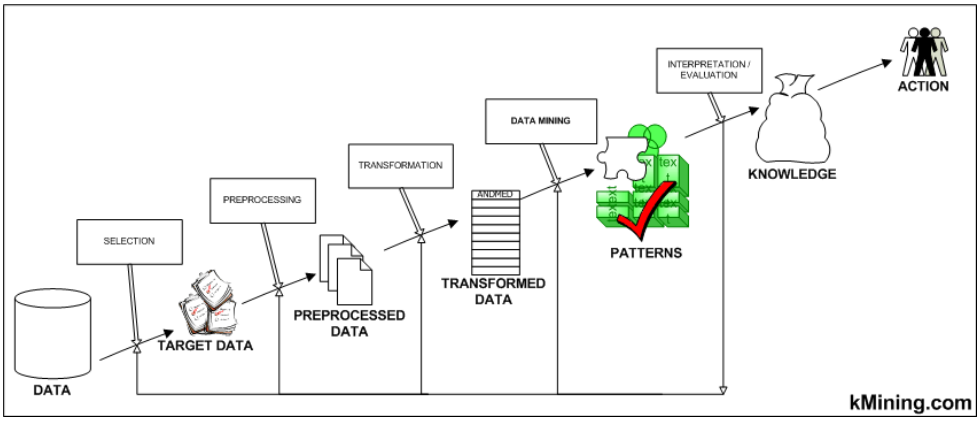
وقد برزت الحاجة مؤخراً إلى مخازن البيانات في المؤسسات الكبرى حيث أن كل قسم من هذه الشركات يدير قواعد بيانات مستقلة عليه Data Mart خاصة به (كأقسام التسويق– الحسابات– المشتريات – المبيعات...) وعند تجميع هذه الـ Data Marts (المستودعات المصغرة) يتكون مستودع البيانات الشامل (Data Warehouse) الذي يتضمن كثيراً من البيانات المشتركة التي تلبي متطلبات واحتياجات الشركة لإيجاد كامل البيانات المتعلقة بموضوع محدد كالزبائن مثلاً من خلال البحث في قاعدة بيانات واحدة متكاملة (Integrated DB) بدلاً من البحث في عدة قواعد بيانات منفردة، مع الاحتفاظ بقواعد البيانات الأصلية على أن يتم تغذية مستودعات البيانات دورياً في حالة حصول تعديل أو تبديل في قواعد البيانات الأصلية، وللعلم فأن أغلب الشركات العملاقة تستخدم تقنيات متطورة لتخزين مستودعات البيانات مثل تقنية Redundant Array Independent Disks (RAID) وذلك لاعتمادها على الخزن المتوازي بفعالية وأيضا لكونها أحد أساليب النسخ الاحتياطي المتطورة في مزودات الشبكات [1] [20].

3.6 التنقيب في البيانات Data Mining

1.3.6 نبذة عن التنقيب في البيانات

هي عملية البحث في قواعد البيانات (مستودعات البيانات) عن معرفة غير مكتشفة وغير متوقعة أي الحصول على معرفة جديدة غير موجودة في قواعد البيانات الأصلية وتكون هذه المعرفة مهمة بحيث تساعد في اتخاذ القرار [2] [13]. يعد التنقيب في البيانات مرحلة من مراحل استكشاف المعرفة في قواعد البيانات والتي تشير إلى استكشاف الأنماط الضمنية غير الاعتيادية والتي لم تكن معروفة سابقاً [19]، وعملية استكشاف المعرفة في قواعد تتضمن عدد من المراحل التي تبدأ من جمع البيانات الخام (Raw Data) وصولاً إلى مرحلة الحصول على المعرفة الجديدة، كما هو موضح بالشكل (1). وفيما يلي عرض لهذه المراحل [3]:

1. **اختيار البيانات Data Selection:** وهي مرحلة اختيار البيانات المرشحة للدراسة من مستودع البيانات الكلي بحسب الغرض من الدراسة.
2. **تنقية البيانات Data Preprocessing:** وهي مرحلة البيانات التي تحتوي على تشويش Noise أو شوائب من مجموعة البيانات بحيث يتم الحصول على قاعدة بيانات نقية.
3. **تحويل البيانات Data Transformation:** وهي عملية تحويل البيانات التي تم اختيارها إلى شكل ملائم للخوارزميات والتطبيقات التي ستستخدم في الدراسة. حيث قد تشترط بعض الخوارزميات وجود البيانات على هيئة معينة قبل تطبيقها.
4. **التنقيب في البيانات Data Mining:** في هذه المرحلة يتم تطبيق طرق وخوارزميات ذكية وبتسلسل مناسب لاستكشاف أنماط مفيدة.
5. **تقييم الأنماط Pattern Evaluation:** بعد استكشاف الأنماط المهمة والتي تمثل المعرفة يتم تقييمها بطرائق كمية ووصفية مختلفة وملائمة للتقنيات التي طبقت وباستخدام مقاييس محددة في بيئة المشكلة.



الشكل (1): خطوات استكشاف المعرفة

2.3.6 تقنيات التنقيب في البيانات

يوجد العديد من التقنيات التي يمكن استخدامها للتنقيب في البيانات، ومن أشهرها:

- **التقنية الأولى: قواعد الارتباط Association Rules**

قواعد الارتباط Associations Rules هي إحدى التقنيات الواعدة في الـ Data Mining كأداة من أدوات استكشاف المعرفة KDD ولديها القدرة على معالجة كميات هائلة من البيانات، وتسمح باستنتاج كل القوانين الممكنة التي تشرح بعض الصفات الموجودة اعتماداً على وجود الصفات الأخرى [16].

وبمعنى آخر هي قواعد ارتباط معينة بين عدة مجموعات من البيانات في قاعدة بيانات واحدة [3]، وتعتمد طريقة قواعد الارتباط على ايجاد Large Item Set من خلال المعادلة (1):

$$Support = \frac{Number\_of\_Transactions\_That\_Contains\_X\_and\_Y}{Total\_Number\_of\_Transactions} \quad \ldots\ldots\ldots (1)$$

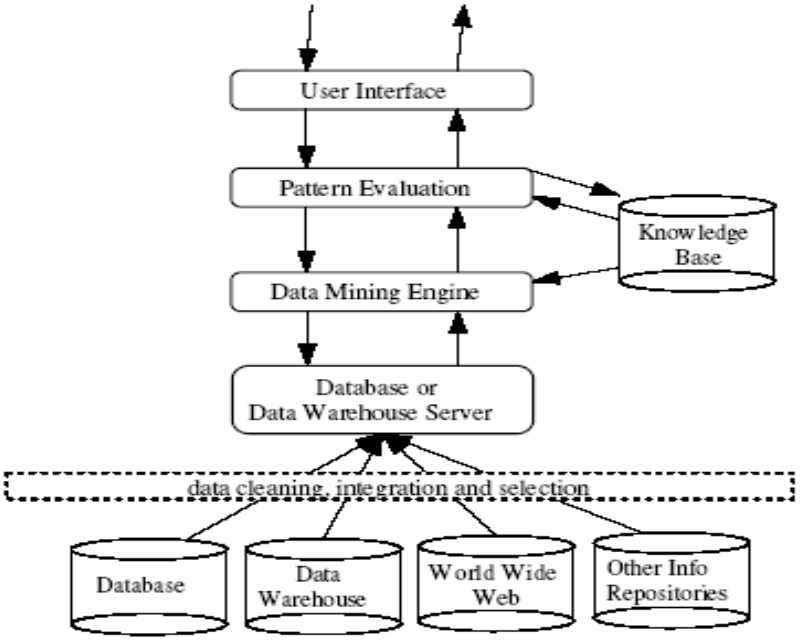
حيث X, Y تمثلان الصفتان المطلوب إيجاد الارتباط بينهما. لا تشترط هذه التقنية وجود حقل قرار معين لذا غالباً ما تستخدم في بداية الدراسة والتحليل للبيانات [13] [16].

- **التقنية الثانية: التصنيف Classification**

يستخدم التصنيف بشكل واسع في حل الكثير من المشكلات خاصة تلك التي تتعلق بالأعمال Business من خلال تحليل مجموعة من البيانات ووضعها على شكل أصناف أو أقسام يمكن استخدامها فيما بعد لتصنيف البيانات مستقبلاً [3], وهنا يكمن الفرق بين التصنيف وإنشاء العناقيد، فالتصنيف يقصد به تقسيم البيانات إلى مجاميع يتم تحديدها مسبقاً، أما إنشاء العناقيد أو ما يسمى بالـ Clustering فهو يعني تقسيم البيانات إلى مجاميع ليست معروفة مسبقا. هناك عدد من الطرق التي يمكن استخدامها في تصنيف البيانات باستخدام خوارزميات مختلفة مثل الخوارزميات الإحصائية Statistical Algorithms والشبكات العصبية Neural Network والخوارزميات الجينية Genetic Algorithms وطريقة الجار الأقرب Nearest Neighbor Method. إحدى الطرائق المستخدمة في التصنيف (وهي المستخدمة في هذا البحث) طريقة الـ Decision Trees وهي هيكل شجري يقدم مجموعة من القرارات التي تولد قواعد لمجموعة البيانات المصنفة (Classified Data) [15]. تشترط هذه التقنية وجود حقل قرار يتم تصنيف البيانات بناءً عليه.

3.3.6 معمارية نظام التنقيب في البيانات

الشكل (2) يوضح معمارية نظام عام للتنقيب في البيانات [4]، وسيتم فيما يلي توضيح أهم المفاهيم الواردة في الشكل.



الشكل (2): معمارية نظام عام للتنقيب في البيانات

- **قاعدة المعرفة Knowledge Base**

وتمثل مجال المعرفة الذي يمكن أن يستخدم لتوجيه عملية البحث أو تقييم الأنماط المفيدة واستكشافها. مثل هذه المعرفة يمكن أن تحتوي على المفهوم الهرمي والذي يستخدم لتنظيم الخواص أو قيم هذه الخواص في مستويات متعددة.

- **آلة التنقيب في البيانات Data Mining Engine**

وهو ضروري لنظام الـ Data Mining ويشمل مجموعة من الوحدات الوظيفية مثل Association - Characterization – Evolution – Outlier - Analysis.

- **تقييم الأنماط Pattern Evaluation**

هذا المكون يستخدم كمقياس ذو كفاءة عالية في استكشاف الأنماط المتواجدة في قاعدة البيانات.

- **واجهة المستخدم User Interface**

هذا المكون يعمل كحلقة وصل بين المستخدم ونظام التنقيب في البيانات ويسمح للمستخدم بما يلي:

1. التفاعل مع النظام بتحديد استعلام أو مهمة معينة.
2. التزود بمعلومات للمساعدة على تركيز البحث.
3. تصفح قاعدة البيانات وهيكلية مستودع البيانات.
4. بتمثيل الأنماط الناتجة صوريا في عدة أشكال.

7. الدراسات السابقة

يعرض هذا المقطع أربع دراسات سابقة مرتبطة بالموضوع سواءً بشكل مباشر أو غير مباشر حيث اهتمت كثير من الدراسات على مستوى العالم بتطبيق خوارزميات التنقيب في البيانات لاكتشاف المعرفة في الجامعات، من أهم هذه الدراسات تلك الموجودة في [12] و [17]. تهتم الدراسة الأولى [12] بتطبيقات التنقيب في البيانات في مجال التعليم العالي، وتركز على مدخلات العملية التعليمية ومخرجاتها وكيفية تأثير كل منها على الأخرى، استخدمت الدراسة الأولى طريقة الشبكات العصبية للتنقيب في البيانات، تظهر النتائج علاقات متنوعة بين المناهج والساعات المعتمدة وطبيعة الطلاب وبين الخريجين والوظائف التي يشغلونها، إضافة إلى استنتاجات أخرى مفيدة لصانع القرار في الجامعات. الدراسة الثانية [17] تركز على التنقيب في موقع جامعة سيتي، من خلال دراسة ملفات تسجيل دخول الزوار، وتكتشف الدراسة الكثير من العلاقات بين طبيعة الزوار وطبيعة التخصصات التي يدخلونها إضافة إلى أسباب تسرب العملاء من المؤسسة، وتخرج بتوصيات كثيرة لصناع القرار في الجامعة. أما الدراسة الثالثة [5]، والتي يعتبر هذا البحث امتداداً مباشراً لها، فقد قامت الباحثة فيه باستكشاف أنماط هامة في بيانات طلاب جامعة العلوم والتكنولوجيا اليمنية وذلك بتطبيق التقنيات الثلاث للتنقيب في البيانات والتي سبق ذكرها في المقطع السابق من هذا البحث. النتائج التي تم التوصل إليها في هذه الدراسة يمكن تلخيصها فيما يلي:

1. أن أكثر نسبة رسوب تكون في المراحل الأولى من الدراسة وأن هذا الرسوب مرتبط إلى انسحاب الطلاب.
2. أن فشل الطالب في المراحل الأولى مرتبط بضعف معدلاته في المرحلة الثانوية.
3. يرتبط الفشل لدى بعض الطلاب بتواضع القدرات الشخصية للطالب. إحدى الطرق التي اقترحتها الباحثة لعلاج هذه المشكلة هي بإضافة سنة تحضيرية للطلاب.
4. تسرب الطلاب مرتبط بصعوبة بعض المواد وخاصة في المراحل الأولى لهم.

تتلخص نواقص الدراسة الثالثة فيما يلي:

- عدم شمولها بيانات الطلاب المنتسبين وبيانات الكادر التعليمي وبيانات فروع الجامعة الأخرى.
- عدم شمولها بيانات المنح الدراسية والكثير من البيانات الشخصية للطلاب.
- عدم شمولها بيانات الحضور والغياب بالنسبة للطلاب.

في الدراسة الرابعة [1]، قام الباحث بمقارنة المستودعات الموزعة مع الأسلوب التقليدي وذلك اعتمادا على تحليل الأحمال في النظام التقليدي وفي النظام الموزع عبر شبكة محلية وباستخدام تقنيتي التنقيب في البيانات والـ OLAP معاً. النتائج التي تم التوصل إليها في هذه الدراسة يمكن تلخيصها فيما يلي:

- تخفيف الحمل على المستودع الرئيس حيث لم يعد هناك حاجة لعرقلة الاتصال بالسيرفر الرئيس لعدم حصول أي ضغط ناتج عن تزاحم المستخدمين.
- وضع أسلوب مستودعات البيانات الموزعة نوعاً من الاستقلالية للبيانات، وذلك لأنها توضع بناء على أساس فئات معينة تربطها خصائص متشابهة وليس حفظها في مستودع عام يضم أنواع متعددة من البيانات من جميع الفئات.
- تحسين النتائج المستخرجة والتي في السابق كانت تدل على مجموعات غير محددة من الفئات. بالمقابل كان الأسلوب الجديد عرض نتائج التحليل بشكل أكثر دقة. بحيث يتمكن صانع القرار من إصدار قرارات معتمداً فيها على قيم واضحة ولعدة مستويات وليس بشكل عام.
- تخفيف الحمل المترتب على المعالجة الكلية للبيانات، حيث أثبت الباحث أنه عند إجراء التحليل على جزء معين من المستودع "مستودع فرعي" فإن نسبة استخدام المعالج بكامل طاقته تقل وذلك بسبب معالجة البيانات ذات الصلة فقط وليس خليط من البيانات.

أما نواقص الدراسة فتتلخص في كون البحث اقتصر على مقترح لبناء مستودعات بيانات موزعة لكلية العلوم والهندسة ولم يتطرق إلى كيفية استكشاف المعرفة باستخدام طرائق التنقيب في البيانات.

في الدراسة الخامسة [6]، قام الباحث باستعراض تطبيقات التنقيب في البيانات في التعليم العالي، وقد قدم البحث دراسة نظرية على بيانات طلاب جامعة IUS للفترة (1989 - 1999) بحيث طبق إحدى تقنيات التنقيب في البيانات وهي الـ Association Rules وكون الدراسة نظرية فقد كان التركيز الأكبر على شرح تقنيات Data Mining وخوارزمياتها بالتفصيل وبيان مدى أهميتها في دعم اتخاذ القرار في مؤسسات التعليم العالي. تتلخص نواقص الدراسة في أن حجم البيانات التي تم تطبيق التقنيات عليها للمقارنة بينها صغير نسبياً.

في الدراسة السادسة [2]، تم تصميم نظام لدعم القرارات في مبيعات المنتجات الغذائية باستخدام طرق تقليدية سواء في البرمجة أو في قواعد البيانات كالاستعلامات. يساعد النظام المصمم صانع القرار في الحصول على معلومات ملخصة ومهمة تساعد في اتخاذ القرارات السليمة، والحصول على الاستعلامات العصية والمعقدة بسرعة ودقة، مع إمكانية تميز الأفضل وتقديمه للزبون، وإمكانية توقع مشاكل الزبون واستباق احتياجاته، وأخيراً معرفة العلاقة والارتباط بين المنتجات. تم تطبيق النظام على قاعدة بيانات North Wind المضمنة في MS SQL Server. وتتلخص نواقص الدراسة فيما يلي:

- اقتصرت البيانات على المبيعات الغذائية فقط دون المشتريات والأقسام الأخرى.
- لم يتم استخدام خوارزميات ذكية في الدراسة لتصنيف العملاء ودرجة ارتباطهم بالمنظمة كخوارزمية Decision tree مثلاً.
- تم اقتراح خوارزميات وطرائق معينة لتنظيف البيانات وتهيئتها قبل نقلها إلى المستودع مثل: وضع قيم متوسطة للحقول العددية، ولكن لم يتم تنفيذها في هذه الدراسة.

الدراسات السابقة في مجملها لم تركز كثيراً على بناء مستودع للبيانات، وتهمل كل منها عوامل مؤثرة في استخراج الأنماط السائدة في قاعدة البيانات، وفي بحثنا هذا نحاول التركيز على معظم العوامل المؤثرة مع التركيز على بناء مستودع البيانات في المراحل الأولى من خطوات تطبيق البحث.

8. تنفيذ الدراسة

تم تنفيذ الدراسة في هذا البحث اعتماداً على مراحل منهجية البحث المذكورة في المقطع الثالث من هذا البحث:

1.8 مرحلة الحصول على قاعدة البيانات

إن قاعدة البيانات التي طبقت عليها تقنيات التنقيب في البيانات هي لجامعة العلوم والتكنولوجيا، وقد شملت بيانات الطلاب دون الطالبات والمنتظمين دون المنتسبين ومن العام 1994 إلى العام 2005م بواقع 330000 سجل تقريباً، تعتبر هذه الجامعة من أكبر الجامعات الخاصة في الجمهورية اليمنية حيث تغطي قدراً كافياً من التخصصات التقنية والإنسانية الحديثة على نطاق واسع وحديث. تتكون قاعدة البيانات التي تم الحصول عليها من 76 جدولاً على هيئة قواعد بيانات MS-Access وذلك بعد تحويلها من قواعد بيانات Oracle واستثناء الكثير من الجداول المالية والخاصة بالهيئة التدريسية بالجامعة.

2.8 مرحلة المعالجة الأولية للبيانات (Preprocessing)

تم تطبيق هذه المرحلة على البيانات قبل نقلها إلى مخزن البيانات وهي المرحلة الأكثر جهداً والأطول وقتاً في البحث، حيث واجهت الباحثين الكثير من المشاكل والمعوقات في البيانات المخزنة، نسرد هنا بعضا منها:

- **مشكلة تواريخ شهادات الثانوية**

تتمثل هذه المشكلة في أن مدخلي البيانات أثناء تسجيل الطلاب في الجامعة كانوا يدخلون التاريخ بشكل غير ثابت، فأحياناً يدخلونها بالتاريخ الميلادي وأحياناً أخرى بالتاريخ الهجري (خصوصاً للطلاب القادمين من السعودية). فتم توحيد التواريخ جميعاً إلى التأريخ الميلادي.

- **مشكلة تاريخ الميلاد**

هذه المشكلة مماثلة للمشكلة السابقة، وقد تم حل المشكلتين بتنفيذ استعلام يقوم بفحص التأريخ إذا كانت السنة أقل من 1500 فيكون تأريخاً هجرياً، لذلك يتم إضافة 580 إلى الرقم فتتوحد التواريخ كما هو موضح في الشكل (3).

```
SELECT DD_URS_STUD_INFO.STUD_ID, Year(DD_URS_STUD_INFO!SI_JOIN_DATE) AS JoinYear,
IIf(DD_URS_STUD_INFO!SI_CER_YEAR>0,IIf(DD_URS_STUD_INFO!SI_CER_YEAR<1500,DD_URS_STUD_INFO!SI_CER_YEAR+580,DD_URS_STUD_INFO!SI_CER_YEAR),Year(DD_URS_STUD_INFO!SI_JOIN_DATE)-1) AS CERyear, [JoinYear]-[CERyear] AS After_Cer INTO afterCer
FROM DD_URS_STUD_INFO;
```

الشكل (3): الاستعلام المستخدم لحل مشكلة تأريخ الميلاد

- **وجود أكثر من مادة بنفس الاسم**

وهذا خطأ يقع فيه مدخلي البيانات أيضاً، فيتشتت الطلاب المسجلين بنفس المادة، وقد تمت معالجة هذا الخطأ باعتماد المادة التي تم تسجيل الطلاب فيها أكثر أو ضم طلاب المواد المتشابهة تحت مادة واحدة.

- **مكان الميلاد، مكان شهادة الثانوية**

كانت هذه البيانات مكتوبة بشكل نصي، فلم يكن بالإمكان تصنيفها بشكل رقمي (باستخدام مفتاح أساسي) واستخدام هذه البيانات التي يمكن من خلالها التوصل إلى نتائج جيدة ومفيدة. لذلك تم ترميزها لتسهيل التعامل معها.

- **ظهور بيانات شاذة أو غير منطقية**

مثل وجود تأريخ ميلاد بالأرقام السالبة أو قيم (1652م)، فتمت معالجتها بالتعديل أو الحذف من القاعدة.

- **تجهيز البيانات للتصدير**

بعد أن تم تركيب وإعادة هيكلة قاعدة البيانات الأصلية إلى قاعدة بيانات على شكل مستودع بيانات وذلك باختيار كل الجداول والحقول المطلوبة والاستغناء عن الجداول غير المهمة، تم تنفيذ استعلام لكل هذه الجداول حتى يتم تجميع البيانات في جدول واحد يمكن من خلاله التنقيب والاستكشاف بعد تحويله إلى شكل بيانات نصية، الاستعلام المستخدم لذلك موضح بالشكل (4).

```
SELECT DD_URS_STUD_INFO.STUD_ID, Year([SI_JOIN_DATE]) AS join_year, IIf([FINANCE]>0,"SUPPORTED","OWN") AS
SUPPORTED, DD_URS_COLLEAGE.COLG_L_NAME, DD_URS_COUNTRY.NAT_L_NAME,
DD_URS_SPEC_NAME.SPEC_L_NAME, DD_URS_COURSE.CRS_L_NAME,
IIf([SR_MARK]>=90,"Excellent",IIf([SR_MARK]>=80,"VeryGood",IIf([SR_MARK]>=70,"Good",IIf([SR_MARK]>=60,"Accepted",IIf([
SR_MARK]>=50,"Pass","Fail"))))) AS MARK,
IIf([SI_CER_PERC]>=90,"Excellent",IIf([SI_CER_PERC]>=80,"VeryGood",IIf([SI_CER_PERC]>=70,"Good",IIf([SI_CER_PERC]>=60,"
Accepted",IIf([SI_CER_PERC]>=50,"Pass","Fail"))))) AS CER, DD_URS_REG_STATUS.REG_STATUS_DESC_L,
DD_URS_STUD_STATUS.STUD_STATUS_DESC_L, afterCer.After_Cer,
IIf([AccumMarks]>=90,"Excellent",IIf([AccumMarks]>=80,"VeryGood",IIf([AccumMarks]>=70,"Good",IIf([AccumMarks]>=60,"Accepte
d",IIf([AccumMarks]>=50,"Pass","Fail"))))) AS AccAvrg
FROM DD_URS_REG_STATUS INNER JOIN (DD_URS_COURSE INNER JOIN (DD_URS_COUNTRY INNER JOIN
(DD_URS_COLLEAGE INNER JOIN (((((DD_URS_SPEC_NAME INNER JOIN (DD_URS_STUD_INFO INNER JOIN
DD_URS_STUD_SPECS ON DD_URS_STUD_INFO.STUD_ID = DD_URS_STUD_SPECS.STUD_ID) ON
DD_URS_SPEC_NAME.SPEC_NO = DD_URS_STUD_SPECS.SPEC_NO) INNER JOIN DD_URS_STUD_REG ON
DD_URS_STUD_INFO.STUD_ID = DD_URS_STUD_REG.STUD_ID) INNER JOIN DD_URS_STUD_STATUS ON
DD_URS_STUD_INFO.STUD_STATUS = DD_URS_STUD_STATUS.STUD_STATUS) INNER JOIN afterCer ON
DD_URS_STUD_INFO.STUD_ID = afterCer.STUD_ID) INNER JOIN StudCourses ON DD_URS_STUD_INFO.STUD_ID =
StudCourses.STUD_ID) ON DD_URS_COLLEAGE.COLG_NO = DD_URS_STUD_INFO.COLG_NO) ON
DD_URS_COUNTRY.CNTRY_NO = DD_URS_STUD_INFO.SI_NAT_CNTRY_NO) ON DD_URS_COURSE.CRS_NO =
DD_URS_STUD_REG.CRS_NO) ON DD_URS_REG_STATUS.REG_STATUS = DD_URS_STUD_REG.REG_STATUS;
```

الشكل (4): الاستعلام المستخدم لتجهيز البيانات للتصدير

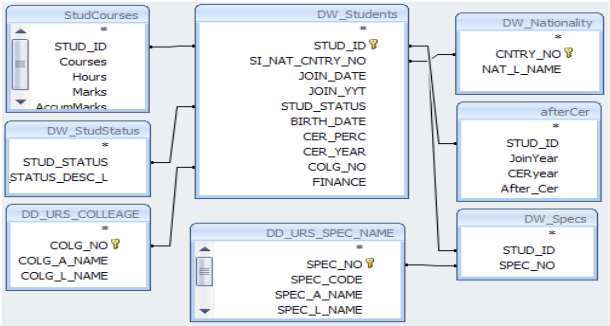
ينتج الاستعلام الموضح بالشكل (4) صفوفاً مرتبة وفقاً لمتطلبات التطبيق وبذلك تكون البيانات جاهزة للتصدير إما على شكل ملفات Excel أو ملفات نصية Text files، وسيظهر تفصيلها في مراحل لاحقة من هذا المقطع.

3.8 مرحلة بناء مستودع بيانات منطقي (Building Logical Data Warehouse)

تم في هذا البحث إعادة تصميم وهيكلة قواعد البيانات على هيئة مستودع بيانات نجمي Star Schema Data Warehouse أي جدول البيانات الرئيس مع جداول الترميز لمناسبة هذا الهيكل لطبيعة البيانات وبالاقتصار على البيانات المهمة في بناء هذا البحث. في هذا البحث تم إنشاء مستودعين للبيانات هما:

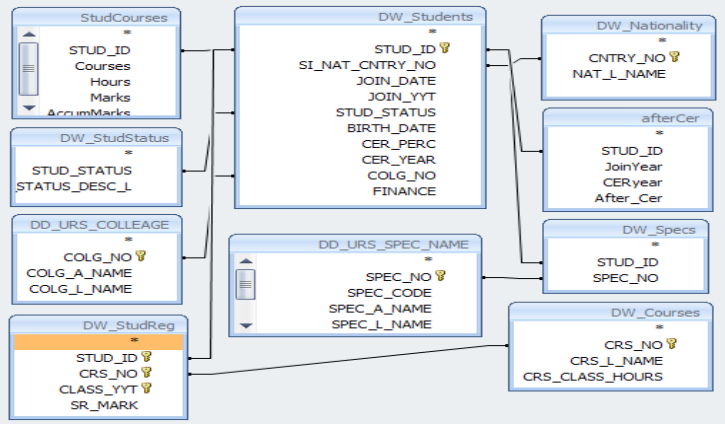
1. مستودع بيانات يحوي كل بيانات الطلاب ومعدلاتهم النهائية شكل (5).
2. مستودع بيانات يحوي كل بيانات الطلاب إضافة إلى بيانات كل المواد الملتحق بها الطالب مع درجة المادة كما هو موضح بالشكل (6).

مستودع البيانات الأول يؤدي الغرض في استنتاج العديد من الأنماط بشكل عام حول أداء الطلاب الأكاديمي بغض النظر عن ارتباط ذلك بالمواد، وبنفس الوقت بعدد سجلات أقل (10463 سجل) مقارنة بمستودع البيانات الثاني الذي يحوي على (330524 سجل).



الشكل (5): هيكلية مستودع البيانات الأول

في المستودع الأول تم إلغاء العديد من الجداول غير الهامة وجعل الحقول داخل الجداول محدودة جداً ومقتصرة على ما سوف يستخدم في الدراسة، ويتضح ذلك في الشكل (5)، أما في المستودع الثاني الموضح في الشكل (6)، فقد تم إضافة جدول تسجيل الطلاب للمواد وجدول المواد الذي يتضمن أسماءها وعدد ساعاتها.



الشكل (6): هيكلية مستودع البيانات مع المواد

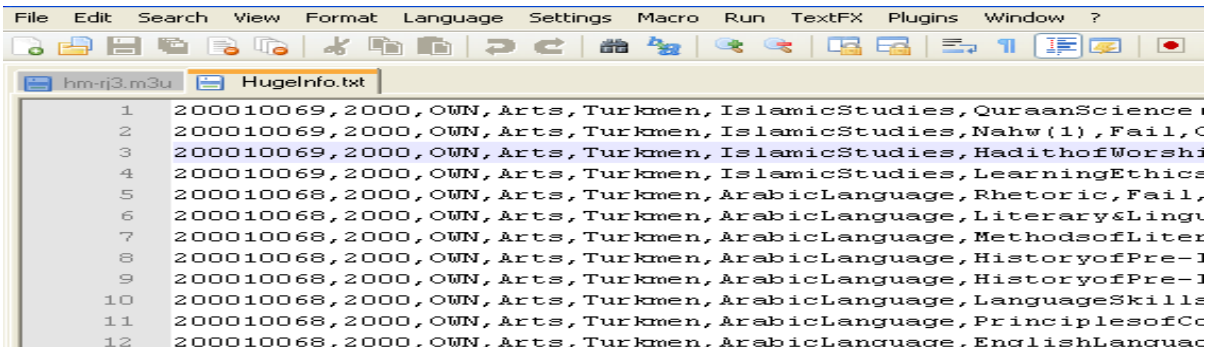
4.8 مرحلة اختيار التطبيق المناسب وتصدير البيانات إليه وتطبيق التقنيات المختلفة

تمثل هذه المرحلة معالجة البيانات بدءاً من تصدير البيانات وانتهاءً بتطبيق برنامج WEKA والذي تم اختياره حيث أنه متاح بالمجان ويحقق الغرض من البحث كما أنه سهل الاستخدام ويمكن تصدير البيانات إليه بسهولة. يتعامل البرنامج مع ثلاث صيغ من البيانات وهي إما بيانات من نوع اكسل أو بيانات على شكل ملفات نصية أو ODBC. عند تصدير البيانات من Access إلى Excel ظهرت مشكلة أن الأكسل لا يتقبل أكثر من 65536 سجلاً كما هو موضح في الشكل (7) بينما تحتوي قاعدة البيانات التي يدرسها هذا البحث على أكثر من 330 ألف سجل.

65529	2E+08	Science&E	Computer	Yemeni	Good	ABORTED	Accepted	ABS_FINA	Physics(1)	1996
65530	2E+08	Science&E	Computer	Yemeni	Good	ABORTED	Accepted	ABS_FINA	ArabicLan	1996
65531	2E+08	Science&E	Computer	Yemeni	Good	ABORTED	Pass	PASS	Differentia	1996
65532	2E+08	Science&E	Computer	Yemeni	Good	ABORTED	Good	PASS	Digitallma	1996
65533	2E+08	Science&E	Computer	Yemeni	Good	ABORTED	Accepted	ABS_FINA	ComputerI	1996
65534	2E+08	Science&E	Computer	Yemeni	Good	ABORTED	Accepted	ABS_FINA	CompilerD	1996
65535	2E+08	Science&E	Computer	Yemeni	Good	ABORTED	Accepted	ABS_FINA	EnglishLa	1996
65536	2E+08	Science&E	Computer	Yemeni	Good	ABORTED	Accepted	ABS_FINA	ArabicLan	1996

الشكل (7): توضيح مشكلة الإكسل في عدم استيعاب أكثر من 65536 سجل

ولحل هذه المشكلة تم تصدير البيانات على شكل ملفات نصية وفق شروط يفرضها برنامج WEKA على الملفات النصية التي يقبلها وهي أن لا تحتوي هذه البيانات على فواصل منقوطة أو علامات تنصيص أو فراغات كما في الشكل (8).



الشكل (8): صيغة النصوص المقبولة لدى WEKA

ولكي يتقبل WEKA هذه الملفات (النصوص) لابد من إضافة أوامر خاصة به لتعريف حقول البيانات Attributes في رأس الملف، وذلك لتعريف كل عمود من الأعمدة في هذه البيانات ماذا يمثل وما نوع البيانات بداخله، ومن ثم حفظه بامتداد arff وقد تكون هذه Attributes من نوع Number مثل (ID-Join Year-Birthday) أو تكون من نوع String مثل (Courses-Name) أو Nominal مثل (Own-Supported)، وتصبح صيغة الملفات النصية بعد إضافة هذه Attributes كما في الشكل (9).

```
1  @relation StudentData
2
3  @attribute StudID numeric
4  @attribute JoinYear {1993,1994,1995,1996,1997,1998,1999,2000,2001,2002,2003,2004,2005}
5  @attribute Finance {OWN, SUPPORTED}
6  @attribute College {MedicalSciences,Science&Engineering,Arts,AdministrativeSciences,Ed
7  @attribute Nationality {Turkmen,Russian,Bulgarian,Kamar,Kuwaiti,Omani,Bosnian,Yugoslav
8  @attribute SPEC_L_NAME {Accounting,ArabicLanguage,ArchitecturalEngineering,Banking,Civ
9  @attribute CRS_L_NAME {(Elective1)ElectricalMachines,Familylaw,(Elective2)NeuralNetwor
10 @attribute CRS_Mark {Excellent,VeryGood,Good,Accepted,Pass,Fail}
11 @attribute CER {Excellent,VeryGood,Good,Accepted,Pass,Fail}
12 @attribute REG_STATUS_DESC_L {ABS_Failed,ABS_FINAL,Abs_Oothr,Altrnated,Canceld,CANCELE
13 @attribute STUD_STATUS_DESC_L {GRADUATED,Repeted,WITHDRAWED,ESCAPE,BEHAVIOURDROPPED,AC
14 @attribute Avg {Excellent,VeryGood,Good,Accepted,Pass,Fail}
15 @attribute AfterCer {VeryOld,Old,Average,Normal}
16
17 @data
```

الشكل (9): الملفات النصية بعد إضافة أسماء الحقول Attributes

بعد ذلك تكون البيانات جاهزة لتطبيقها في برنامج WEKA.

1.4.8 تطبيق التقنيات في برنامج WEKA

بعد استقبال برنامج WEKA للبيانات بالصيغة المطلوبة كما تم شرحه أعلاه، تصبح البيانات جاهزة لاختيار الـ Attributes المراد تحليلها وتطبيق تقنيات التنقيب في البيانات عليها. وقبل البدء بتطبيق هذه التقنيات يجدر بالذكر هنا أن برنامج WEKA يوفر مرونة كافية في إظهار الرسومات البيانية على شكل Bar-Chart ثنائي الأبعاد بحيث يمكن تصنيف العمود الواحد إلى عدة ألوان حسب الـ Attribute المختار. وبعد أن يتم اختيار الحقول المناسبة يتم اختيار التقنية المناسبة من تطبيق WEKA حيث أن WEKA يقوم بتصنيف الخوارزميات على هيئة شجرة تحتوي على مجلدات حسب التقنيات والعمليات الأخرى بعد ذلك تم اختيار الخوارزمية المناسبة ضمن هذه التقنية لاستكشاف المعرفة الكامنة في قاعدة البيانات.

2.4.8 التقنيات المستخدمة في البحث

هناك تقنيات عدة يوفرها WEKA للتنقيب في البيانات، وسوف نسرد هنا تلك التقنيات التي تم استخدامها في هذا البحث، حيث تم استخدام ثلاث خوارزميات ضمن تقنية قواعد الارتباط Association Rules وهي Apriori، Predictive Apriori، وTertius. كما تم استخدام خوارزميتين ضمن تقنية أشجار القرار Decision Trees وهما ID3 و J48.

5.8 مرحلة استكشاف الأنماط المهمة والمؤثرة

وتتم هذه المرحلة بالحصول على نتائج تطبيق التقنيات المختارة على مستودع البيانات المنطقي، وسيتم مناقشة هذه النتائج بالتفصيل في المقطع التالي من هذا البحث.

9. مناقشة وتقييم النتائج

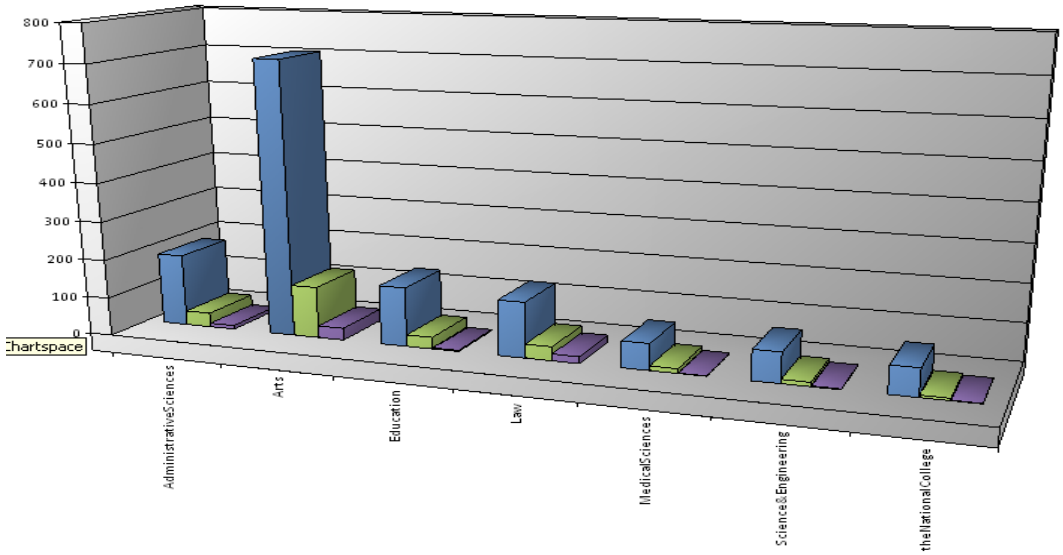
هذه المرحلة توضح خلاصة نتائج تطبيق تقنيات التنقيب في البيانات على مستودع البيانات المنطقي الذي تم إنشاؤه على قاعدة البيانات. وللتسهيل تم تجميع الأنماط المكتشفة في المجموعات والتي تم اختيارها دون غيرها بعد مقابلات ومناقشات تمت مع العديد من صانعي القرار في مستويات إدارية مختلفة بالجامعة ولاعتقادنا بأهميتها وتأثيرها أكثر من غيرها للمساعدة في اتخاذ القرار:

1. علاقة الفترة الفاصلة بين تخرج الطالب من الثانوية العامة والتحاقه بأي من كليات الجامعة.
2. علاقة المعدل باختيار الطالب للتخصص.
3. علاقة المنح الدراسية بمستوى الطالب الأكاديمي.
4. علاقة مختلف المواد بمستوى الطالب أو حالته الأكاديمية.

بعد ذلك تم التحقق من صحة النتائج وذلك بتنفيذ استعلامات في قاعدة بيانات الجامعة والتأكد من مدى تقارب نتائج خوارزميات التنقيب في البيانات مع نتائج التحقق وبيان أسباب أو تخمينات لوجود هذه الظاهرة أو تلك. وسيتم شرح النتائج بالتفصيل فيما يلي:

- **علاقة الفترة الفاصلة بين تخرج الطالب من الثانوية والتحاقه بأي من كليات الجامعة**

اتضح من خلال تطبيق تقنيات التنقيب في البيانات كما يوضح شكل (10) أن الطلاب الملتحقين بكلية العلوم الطبية هم من الذين لم يمضِ عليهم أكثر من سنتين منذ تخرجهم من الثانوية. عندما التحقت نفس الشريحة من هؤلاء الطلاب بكليات مختلفة كانت مستوياتهم متفاوتة خصوصاً في كليات الهندسة. الشريحة الأخرى، وهم الطلاب المتأخرين بعد تخرجهم من الثانوية بأكثر من سنتين، نجد أن أغلبيتهم يفضلون الالتحاق بتخصصات العلوم الإنسانية والدراسات الإسلامية المختلفة.



الشكل (10): اتجاهات الطلاب إلى مختلف الكليات حسب الفترة الفاصلة

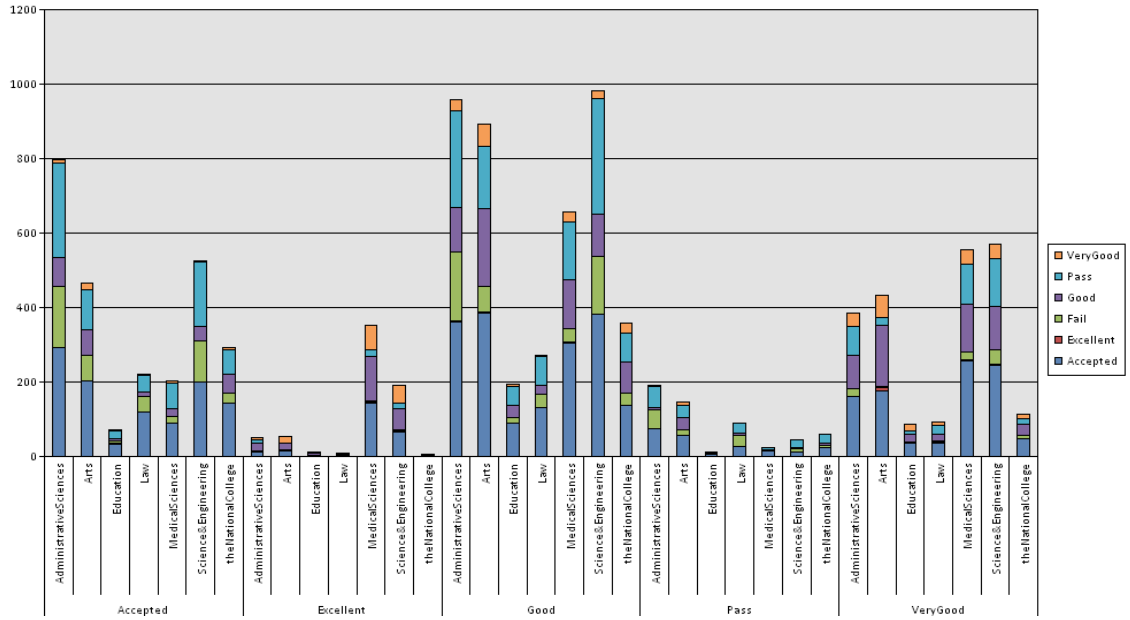
- **علاقة معدل الثانوية باختيار التخصص ومستوى الطالب**

اتضح من خلال نتائج الخوارزميات المستخدمة على بيانات الطلاب ومعدلاتهم في المرحلة الثانوية ومعدلهم النهائي في الجامعة مصنفين بحسب معدلات تخرجهم في الثانوية كما يلي:

- ممتاز Excellent: الملاحظ أن معظم الطلاب المتخرجين بمعدل ثانوية ممتاز، يفضلون الالتحاق بكلية الطب.

- جيد جدا Very Good: الطلاب الحاصلين على معدل جيد جداً في الثانوية، يستطيعون تحقيق امتياز أو جيد جداً أوجيد كحد أدنى في العلوم الإنسانية، والملاحظ أنه في الكلية الطبية يحصلون على معدل مقبول وفي الهندسة معدل ضعيف وحالات رسوب كثيرة.
- جيد Good: الطلاب الحاصلين على معدل جيد بالثانوية يستطيعون تحقيق امتياز في العلوم والإنسانية. وبهذا المعدل يحقق الطالب في كلية العلوم والهندسة معدلاً متدنياً جداً .
- مقبول (Accepted (Pass: اتضح من خلال تحليل النتائج أن الطلاب الحاصلين على معدل مقبول في الثانوية، يتجه معظمهم إلى العلوم الإنسانية.

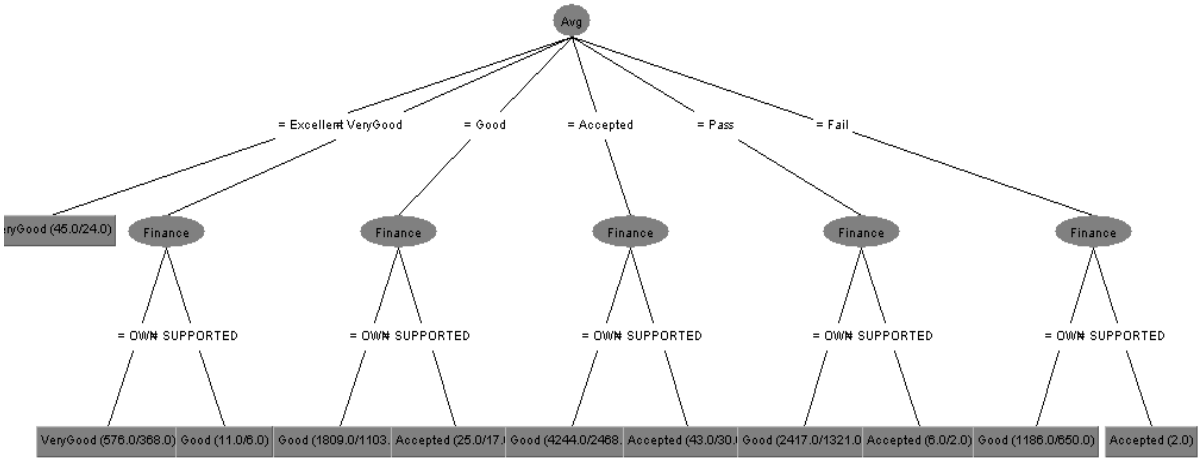
يوضح الشكل (11) كثرة الملتحقين بكلية العلوم الطبية للطلاب الحاصلين على معدل ممتاز Excellent، كما يوضح توزع الحاصلين على معدل جيد جدا Very Good بين كليتي الطب والهندسة، وإن التحقوا بالعلوم الإنسانية (يشار إليها في الشكل بـ Arts) فإنهم يحققون تقدير امتياز.



الشكل (11): علاقة معدلات الثانوية وتأثيرها في مختلف الكليات

- **علاقة المنح الدراسية بمستوى الطالب الأكاديمي**

بالنسبة للطلاب المدعومين (منحة) سواء من الجامعة أو من جهات أخرى، اتضح أن من تخرج منهم من الثانوية بمعدل جيد فما فوق، يميل مستواهم إلى التحسن أو الحصول على معدلات أفضل خلال سنوات الدراسة الجامعية. بخلاف الطلاب الذين تقل معدلاتهم عن (جيد) فإن مستواهم إن لم يكن مماثلاً فإنه ينخفض وذلك ما يتضح من الشكل (12)، ويمكن القول إجمالاً أن الطلاب الذين يستحقون المنح الدراسية هم من يحصلون على معدل جيد فما فوق في الثانوية العامة.



الشكل (12): علاقة المنح الدراسية بمستوى الطالب

- **علاقة المواد بمستوى الطالب أو تسربه**

من الجدير بالذكر هنا أنه لم يتم في هذا البحث دراسة جميع تخصصات الجامعة ولكن تم اختيار التخصصات ذات العدد الأكبر للطلاب، فكانت التخصصات التالية وهي قسم هندسة حواسيب – كلية العلوم والهندسة، وقسم علوم حاسوب – كلية العلوم والهندسة، قسم طب بشري – كلية العلوم الطبية، قسم المحاسبة – كلية العلوم الإدارية والإنسانية، وقسم الدراسات الإسلامية – كلية العلوم الإدارية والإنسانية. وهي عينات كافية وممثلة حيث أن أعداد الطلاب فيها كبيرة.

كما هنا أن ظهور ارتباط مواد محددة بمستويات متدنية للطلاب فيها ومن ثم تسربهم ينبغي لأصحاب القرار الوقوف عليه ومحاولة تفسير ذلك في ضوء طبيعة المادة أو مكانها في الخطة الدراسية وارتباطها بما قبلها من المقررات أو بأداء بعض مدرسي المواد. وقد كانت النتائج كما في الجدول (1) الذي تم استنتاجه وتجميعه يدوياً حيث أن مثل هذا الجدول لا يمكن أن ينتج عن برنامج WEKA.

جدول (1): أبرز المعارف الناتجة عن المواد في تخصصات مختلفة

المادة	التخصص	النتيجة
هياكل متقطعة	هندسة حاسوب	ترتبط نتائج اختبارات هذه المادة بعلاقة مع فشل الكثير من الطلاب، وبالتالي توقفهم في الفصل الدراسي أو انقطاعهم من الدراسة أو الخروج من الجامعة.
إلكترونيات 2	هندسة حاسوب	ترتبط بعلاقة مع حالات كثيرة من حالات فصل الطلاب أكاديمياً من القسم وذلك لرسوبهم فيها كما ترتبط بانسحابهم من الفصل الدراسي أو الانقطاع عن الدراسة.
جبر خطي	علوم حاسوب	ترتبط هذه المادة بانقطاع الطلاب عن الدراسة أو انسحابهم.
إلكترونيات 1	هندسة حاسوب	ترتبط هذه المادة بانسحاب الطلاب من قسم هندسة حاسوب والانتقال إلى قسم علوم حاسوب.
ذكاء اصطناعي 1	علوم حاسوب	مستوى الطلاب في هذه المادة متدني (مقبول) وتتبعها حالات كثيرة من الانقطاع عن الدراسة.
معالجات ولغة تجميع	علوم حاسوب	الفشل فيها مرتبط بانسحاب ما يقارب 60% من الطلاب.
الرعاية الصحية الأولية وطب الأسرة	طب بشري	ظهر هذا النمط لكثرة تكرار حدوثه، وهو انسحاب كثير من الطلاب الذين يرسبون بهذه المادة، لتتضح نتيجة غير مسبوقة، حيث لم تُسجل أي حالات طبيعية للمادة، فكل الراسبين في هذه المادة انحصروا في (انقطاع عن الدراسة – تأجيل المادة – انسحاب من الفصل).
الفيزياء الطبية	طب بشري	ترتبط بانقطاع الطلاب عن الدراسة وبعض حالات التأجيل للمادة وحالات إعادة للمادة وانسحاب من الفصل.
مبادئ الاقتصاد الجزئي	محاسبة	ظهر أن الحصول على معدلات متدنية أو الفشل في هذه المادة مرتبط بانقطاع الطلاب عن الدراسة.
رياضة بحتة	محاسبة	انحصرت حالات جميع الطلاب الراسبين في هذه المادة في (منقطع، باقي للإعادة، مؤجل) ولا وجود لحالات أخرى. قد تكون هذه المادة العلمية البحتة تسببت بمضايقة طلاب قسم المحاسبة، كونهم ابتعدوا عن الأقسام العلمية لضعفهم في مثل هذا النوع من المواد، وقد يكون السبب صعوبة مفرداتها.
آداب العالم والمتعلم	دراسات إسلامية	المعدل العام لطلاب القسم في هذه المادة هو مقبول ونسبة كبيرة منهم ينقطعون عن الدراسة.
النحو 1	دراسات إسلامية	انحصرت معدلات جميع من يدرسون مادة النحو (1) في المعدلات (ضعيف، مقبول، جيد) فقط ، رغم أن معظم الطلاب يعيدون امتحان هذه المادة.

10. الاستنتاجات

في هذا البحث تم بناء مستودع بيانات منطقي وتطبيق بعض خوارزميات التنقيب في البيانات على قاعدة البيانات بجامعة العلوم والتكنولوجيا اليمنية للمساهمة في توفير قاعدة معرفية لصناع القرار في الجامعة. وقد تم التوصل من خلال هذا البحث إلى استنتاجات هامة، أهمها هو الحاجة الماسة إلى بناء مستودع بيانات مترابط ومتكامل ونقي وخالي من الأخطاء للجامعة، إضافة إلى نتائج أخرى هامة متعلقة بسجلات الطلاب مثل علاقة معدلات الثانوية العامة وفترة الانقطاع بعد الثانوية باتجاهات الطلاب الأكاديمية ومستوى أدائهم خلال دراستهم الجامعية وهو ما يحتم على الجامعة إعادة النظر في سياسة القبول والتسجيل، أضف إلى ذلك علاقة كثير من المواد الدراسية بتسرب الطلاب وانقطاعهم عن الدراسة، إما لصعوبة المفردات أو لخلل في الخطط الدراسية والمناهج. وهناك عدة جوانب لم يتم التطرق إليها في هذا البحث إليها نظراً لعدم توفر المصادر المطلوبة أو لعدم وجودها ضمن حدود البحث مبدئياً، وهذه يمكن تطويرها كأعمال مستقبلية لهذا البحث مثل، تحديث بيانات الطلاب، واستكمال البيانات الناقصة للطلاب في قاعدة بيانات الجامعة والتي قد تكون فعّالة في استنتاج مزيد من الأنماط والعلاقات، مثل مكان الميلاد، والحالة الصحية، والحالة الاجتماعية، دراسة بيانات الكادر التدريسي، ومدى تأثيرهم على مستوى الطلاب، إضافة بيانات الفروع والمنتسبين إلى الدراسة، واستخدام بعض الطرق الأحدث في التنقيب في البيانات، مثل استخدام تقنيات Neural Networks Genetic Algorithm . وأخيراً فهناك توصيات لتحسين قاعدة بيانات جامعة العلوم والتكنولوجيا في المستقبل، أهمها، مراجعة هيكلية قاعدة البيانات والتأكد من صحة العلاقات، ربط فروع الجامعة بقاعدة بيانات المركز الرئيسي، بناء مستودع بيانات نموذجي ومبني على أسس حديثة، استخدام تطبيق مناسب (أو تطوير التطبيق الحالي) ليحتوي على عوامل تحقق Validation صحيحة أثناء إدخال البيانات، والتأكد من كفاءة مدخلي البيانات لتلافي الأخطاء في البيانات المدخلة.

11. المراجع

[1] أحمد محمد شيخ العطاس، "مستودعات بيانات موزعة للمساعدة في اتخاذ القرار– تطبيق على جامعة العلوم والتكنولوجيا "– رسالة ماجستير في تخصص علوم الحاسوب– جامعة العلوم والتكنولوجيا، 2005م.

[2] محمد عمر بابطاط، وآخرون، "تصميم نظام لدعم القرارات في مبيعات المنتجات الغذائية"، مشروع تخرج بكالوريوس – قسم علوم الحاسوب– جامعة العلوم والتكنولوجيا، 2005م

[3] مجلة المعلوماتية http://informatics.gov.sa/magazine

[4] Jiawei han & Micheline Kawber, "Data Mining Concepts and Techniques", Second Edition, University of Illinois at Urbana, 2003.

[5] Asma`a A. Alshargabi, "Discovering Vital Patterns from UST Students Data - Applying Data Mining Techniques", Master thesis in computer science, University of Science and Technology, Yemen, 2006.

[6] Abdullah Hussein Al-Hashiedy, "Data Mining Applications in Higher Education", Master thesis in Computer Information Systems, Arab Academy of Financial and Banking Sciences, Yemen Branch, 2006.

[7] Alex Berson, Stephen Smith and Kurt Thearling, "Building Data Mining Application for CRM", McGraw Hill, 2000

[8] http://kdd.ics.uci.edu/(UCI Knowledge Discovery in Databases Archive), May 2005.

[9] Michael J. A. Berry, Gordon Linoff, "Data Mining Techniques- For Marketing, Sales, and Customer Support", Wiley, 1997, 1st Edition.
[10] Michael J. A. Berry, Gordon Linoff, "Mastering Data Mining",Wiley, 2001, 2nd Edition.
[11] Tom M. Mitchell, Gordon Linoff, "Machine Learning", McGraw Hill, 1997.
[12] Jing Luan, "Data Mining Applications in Higher Education", 2004.
[13] Two Crows Corporation, "Introduction to Data Mining and Knowledge Discovery", 1999, 3rd Edition.
[14] A. Salazar, J. Gosabez, I. Bosch, R. Miralles, L. Vergara, "A Case Study of Knowledge Discovery on Academic Achievement, Student Desertion, and Student Retention", IEEE, 2004.
[15] Glenn A.Growe, "Comparing Algorithms and Clustering Data: components of the data mining process", 1999.
[16] Aijun An, ShakilKhan, Xiangii Huang, "Objective and Subjective Algorithms for Grouping Association Rules", IEEE, 2003.
[17] George Konstantinou Lekas, "Data Mining The Web: The Case of City University Log's Files", 2000.
[18] Bradly P., Fayyad U. ,and Reina C., "Scaling Clustering Algorithms to Large Databases ", 1998.
[19] Fayyad U., Shapiro G., and Smyth P., "From Data Mining to Knowledge Discovery in databases", 1996.
[20] Witten, I.H. and Frank, E., "Data Mining: Practical machine learning tools and techniques", Morgan Kaufmann, 2005, 2nd Edition.