بسم الله الرحمن الرحيم

جامعة النيلين

كلية علوم الحاسوب وتقانة المعلومات

بحث تكميلى لنيل درجة الماجستير فى تقانة المعلومات بعنوان :

استخدام تقنيات تنقيب البيانات لاستكشاف أنماط مؤثرات التحصيل الاكاديمي لطلاب المرحلة الثانوية

دراسة حالة (شهادة التعليم الثانوي الصف الثالث علمي من العام 2007م حتى العام 2016م)

اعداد الطالب : البدوي سعد البدوي المبارك

إشراف الدكتور:

أشرف عثمان ابراهيم

اكتوبر 2017 م

الإهداء

بسم الله الرحمن الرحيم

(وقل اعملوا فسيرى الله عملكم ورسوله والمؤمنون)

صدق الله العظيم

إلهي لا تعذبني فاني مقر بالذي قد كان مني
وما لي حيلة الا رجائي وعفوك ان عفوت وحسن ظني
فكم من زلة لي في البرايا وانت علي ذو فضل ومن
اذا فكرت في ندمي عليها عضضت اناملي وقرعت سني
يظن الناس بي خيراً واني لشر الناس إن لم تعف عني

____

لبست ثوب الرجا والناس قد رقدوا وقمت أشكو الي مولاي ما أجد
فقلت يا عدتي في كل نائبة ومن عليه لكشف الضر اعتمد
اشكو اليك امورا انت تعلمها ما لي على حملها صبر ولا جلد
وقد مددت يدي بالذل معترفا اليك يا خير من مدت اليه يد
فلا تردنها يا رب خائبة فبحر جودك يروي كل من يرد

____

وضم الإله اسم النبي الي اسمه اذا قال في الخمس المؤذن اشهد
وشق له من اسمه ليجله فذو العرش محمود وهذا محمد

____

إلى من كان دعائها سر نجاحي وحنانها بلسم جراحي إلى أغلى الحبايب
أمي الحبيبة

***

زوجتي هذا كتاب فيه أفكار سنية صغته لله قربي ابتغي الدار العلية
لك فيه حسن رأي وانتقادات ذكية وهو اهداء اليك فاقبلي مني الهدية

الى زوجتي

الي ابنائي

***

إلى إخوتي الأعزاء

***

إلى كل من كنت له متعلماً وعلمني حرفا ونهلت منه علماً

إلى أساتذة كلية علوم الحاسوب بجامعة النيلين الموقرين

الي وزاره التربيه والتعليم

إلى كل زملائي في الدراسة

الشـــــكر

قال الله تعالى : " **رَبِّ أَوْزِعْنِي أَنْ أَشْكُرَ نِعْمَتَكَ الَّتِي أَنْعَمْتَ عَلَيَّ وَعَلَى وَالِدَيَّ وَأَنْ أَعْمَلَ صَالِحًا تَرْضَاهُ وَأَدْخِلْنِي بِرَحْمَتِكَ فِي عِبَادِكَ الصَّالِحِينَ**" (النمل :١٩)

وقال الرسول صلى الله عليه وسلم "من لايشكرالناس لايشكرالله"

(الترمذي، ب.ت، ج٤:٣٣٩)

أتقدم بالشكر الجزيل لكل من وقف بجانبي يشجعني حتى إتمام هذه الدراسة و أخص بالذكر:

*** الأستاذ الدكتور/ أشرف عثمان ابراهيم الذي أشرف على هذه الدراسة ، ولكل الجهود التي بذلها معي في هذا العمل. وله خالص شكري لمساعدته الفعالة وتفاعله معي طوال فترة البحث.**

*** والشكر أجزله لأساتذة كلية علوم الحاسوب وتقانة المعلومات جامعة النيلين.**

* وشكري لأسرتي لتشجيعهم ولدعمهم المتواصل لي وكل أصدقائي والأشخاص الذين ساعدوني خلال فترة أنجازي الأطروحة، جزاهم الله عني خير الجزاء.

* كل من وقف بجانبي يساندني وأخذ بيدي حتى رأى هذا البحث النور.

ملخص البحث

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 فى امتحان الشهادة الثانوية – المساق العلمي – منذ العام ٢٠٠٧ م وحتى العام 2016 م ومن ثم الخروج بمؤشرات عامة حول الأداء الأكاديمي لدعم السياسات التعليمية لدى متخذي القرار في وزارة التربية والتعليم ، لا سيما وأن حجم البيانات وكذلك البعد الزمني الكبير نسبياً لهذه البيانات يدعم من نتائج هذا البحث. في هذه الدراسة تم اكتشاف بعض الانماط (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 التحصيل الكلى للطالب و مستوى تحصيل الطالب لبعض المواد وكذلك العلاقة بين تقدير معدل تحصيل الطالب الكلى بالتقدير المتحصل عليه فى بعض المواد . البحث محاولة لقراءة هذه النتائج وتفسيرها وعرضها والتحقق من مستواها ونوعيتها وذلك بعرضها على متخذي القرار بالوزارة. 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 تقنية العنقدة وتم استخدام خوارزمية (K–means) فى تقنية العناقيد(Clusters) والتصنيف بأشجار القرار Classification Using Decision Trees وذلك بعد عملية المعالجة الأولية Preprocessing لقاعدة البيانات وإعادة هيكلتها على شكل مستودع بيانات منطقي Logical Data Warehouse وقد استخدمت خوارزمية (K–means) في تقنية العناقيد(Clusters)، وخوارزمية (J48) في تقنية التصنيف بأشجار القرار . هذه الطرائق والخوارزميات تم تطبيقها من خلال الأداة WEKA التي تدعم الكثير من الخوارزميات والطرائق للتنقيب في البيانات.ومن نتائج البحث تم بناء مستودع بيانات منطقي وتطبيق بعض خوارزميات التنقيب في البيانات على قاعدة البيانات بوزارة التربية والتعليم الاتحادية، إضافة إلى نتائج أخرى هامة متعلقة بسجلات الطلاب مثل علاقة معدلات مرحلة الاساس وفترة الانقطاع بعد مرحلة الاساس باتجاهات الطلاب الأكاديمية، أضف إلى ذلك علاقة كثير من المواد الدراسية بتسرب الطلاب وانقطاعهم عن الدراسة، إما لصعوبة المفردات أو لخلل في الخطط الدراسية والمناهج.

Abstract

This research presents application study of the domain of knowledge discovery by using the techniques of data mining, the main object of the study is discover the major patterns of the academic student data examination of higher secondary certificate – scientific procedure, since 2007 up to 2016 to get the indicator output around the academic performance.At this study we discover some major patterns data, Some of this patterns association exist between student modified summation and the student gain level for some subjects, also relation between summation degrees with degree gained from some subject .This research attempts to read this result and interpretation, supply and verification level and its type to supplied to ministry decision maker.We choose data mining technique because it's better to used benefit of quantity data, we use different way from data mining technique to support discovery result clusters using (k-means) and classification use decision tree, after first preprocessing data for database and restriction like logical data warehouse shape, we use k-means algorithm of clusters technique and (J48) algorithm of the classification technique of decision tree, this deferent way and algorithms application through weka tool, which support more algorithms and way of data mining Last deductive abstract and suggest some recommendation which interest for decision maker.Results from this research built logical data warehouse & applying algoritm fo data mining's algorithm, in addition to the difficulties of some subjects which may from it's tough words or other mis-qualification planning.

فهرس المحتويات

فهرس الاشكال

فهرس الجداول

جدول(٥-١٠-٣)	أبرز الانماط الناتجة عن المواد – في حالة نجاح الطالب في مادة الفيزياءونجاح في مادة الرياضيات والتربية الاسلامية واللغة الانجليزية – وأثرها على نجاح الطالب	٨٧

الفصل الأول

الإطار العام للبحث

١-١ المقدمة

في عالمنا اليوم عصر الانفجار المعرفي أصبح الإنسان محاط بكم هائل من المعلومات المتنوعة في مختلف مناحي الحياة وبمختلف الصور وغدت بذلك المعلومات من أهم مكونات الحياة المعاصرة ويرجع السبب لهذا التدفق الهائل للمعلومات الي تطور وسائل النقل ومعالجة المعلومات من أجهزة اتصال وحواسيب سريعة وطرفية ذكية ووسائل نقل حديثة محدثة فهى ثورة معلوماتية بنيت علي مجموعة مكونات تتضمن كل من المعلومات و الحواسيب والاتصالات[1].

الأنظمة الالكترونية أو الأنظمة المؤتمتة أو الأنظمة المحوسبة هي تطبيقات تعتمد علي الكمبيوتر أثبتت أنها أفضل من الأنظمة التقليدية بشكل ملموس أو غير ملموس في عدة مناحي مثل تطبيق التعليم الإلكتروني.

استخدام الخدمات الالكترونية فى وقتنا الحالى اصبح من أهم الامور فى معظم المنشآت الحكومية وغير الحكومية[2].

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.

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 [1] .

في المؤسسات التعليمية والأكاديمية يمكن توظيف علمي لاستكشاف المعرفة والتنقيب في البيانات لتحسين الأداء الأكاديمي باستنباط الأنماط السائدة في بيانات هذه المؤسسات حول الطلاب وأعضاء هيئة التدريس، وارتباط كل ذلك بأهم مؤشرات الأداء كتحصيل الطلاب ومعدل بقائهم أو تسربهم ونوعية أداء أعضاء هيئة التدريس .

كما يمكن للمؤسسات التعليمية مثلا أن تتنبأ بالطلاب الذين سيتسربون، والطلاب الذين سيكون تحصيلهم العلمي ضعيف ، 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 التعليمية .

١-٢ مشكلة البحث

ان التدني في مستوي التحصيل الدراسي يعد مشكله كبيره لابد لها من حل فهي مشكله متعددة الأبعاد تارة تكون مشكله نفسيّه وتاره تكون مشكله تربويّه واخري تكون مشكله اجتماعية تتحدد مشكلة الدراسة في ايجاد العوامل المؤدية الي تدني التحصيل في درجات الطلبة في مرحلة التعليم في امتحانات الشهاده السودانية ببعض المواد .

١-٣ أسئلة البحث

١-ماهي العوامل المؤديةالي تدني التحصيل في مادتي الرياضيات والفيزياءلدي طلبة المرحله الثانويه.

٢-هل تختلف العوامل المؤدية الي تدني التحصيل في مادتي الرياضيات و الفيزياء لدى طلبة المرحلة الثانويه باختلاف جنس المعلم سواء كان ذكرا أنثى.

٣-هل تختلف العوامل المؤديةالي تدني التحصيل لدى طلبةالمرحلةالثانويةباختلاف سنوات الخبرةلدي المعلم.

١-٤ أهداف البحث

أ- التعرف علي العوامل المؤدية الي تدني التحصيل في مادتي الرياضيات والفيزياء لطلاب مرحلة الشهاده السودانيه والتعرف الي الاختلاف في العوامل المؤدية الي تدني التحصيل في مادتي الرياضيات والفيزياء لدى طلاب مرحلة الشهاده السودانيه باختلاف الجنس وخبرة المعلم .

ب-بناء مستودع بيانات منطقي لقاعدة بيانات الوزارة يساهم في اتخاذ القرار وتبصير القائمون علي الوزارة بالاخطاء وطرق معالجتها وتقييم الاعمال وتكون مرجعية .

ت-التوصل الى اسباب تدني التحصيل الاكاديمي وذلك عبر تصميم خوارزمية (K-means) للخروج بنتائج تساعد في اتخاذ القرار في الجانب الاكاديمي في وزارة التربية والتعليم.

ث-التوصل الى اسباب تدني التحصيل الاكاديمي وذلك عبر تصميم خوارزمية شجرة القرار Decision Tree لاكتشاف الأنماط Patterns السائدة في بيانات الطلبة الأكاديمية في وزارة التربية والتعليم بالسودان .

١-٥ أهمية البحث

أ- يساعد في بناء التنبؤات المستقبلية واستكشاف السلوك والإتجاهات مما يسمح بتقدير القرارات الصحيحة واتخاذها في الوقت المناسب.

ب- تساعد في اتخاذ القرار لدى القائمون بوزارة التربية والتعليم في معرفة العوامل المؤثرة في تدني التحصيل في مادتي الرياضيات والفيزياء لدى الطلبة .

ت- وضع الاختبارات التحصيلية في المؤسسات التعليمية مثلا ان تكون اختبارات شهرية.

ث- يساعد الموجهين والمشرفين علي تدريب معلمي الرياضيات والفيزياء والمعلمين في وضع الخطط العلاجية لطلابهم.

ج- عند اكتشاف موضع الخطأ أو نقطة الضعف يساهم البحث في تعديل معايير الاداء وزيادة أداء الأساتذة والمدارس بشكل عام وذلك بمعالجة الاخطاء وتقوية نقاط الضعف .

ح- تقييم أداء الطالب بالنظر الي التقدير العام يهدف الي توفير القدرة علي الإستكشاف والتركيز علي أهم المعلومات في سجل الطالب.

١-٦ حدود البحث

الحدود المكانية :

وزارة التربية والتعليم الاتحادية .

الحدود الزمانية :

ابتداءً من اكتوبر 2017م حتى يناير 2018م.

١-٧ منهجية البحث

طلاب شهادة التعليم الثانوي الصف الثالث علمي من العام 2007م حتى العام 2016م.المنهج المتبع في هذا البحث هوالمنهج المسحي وهو اسلوب بحثي تجمع فيه البيانات عن طريق طرح اسئلة على مجموعة افراد وهي طريقة بحثية شائعة الاستخدام ، ويتكون من عدة مراحل تتلخص فيما يأتي:

أ- جمع البيانات من قاعدة بيانات الطلاب (إدارة التقويم التربوى والامتحانات) الجالسون لامتحان الشهادة السودانيه .

ب-المعالجة الأولية (Preprocessing).

ت-بناء مستودع بيانات منطقي (Logical Model).

ث-اختيار التطبيق المناسب لعملية التنقيب في البيانات (WEKA).

ج- تصدير البيانات من مستودع البيانات على الشكل الذي يتطلبه التطبيق(CSV.).

ح- استكشاف الأنماط المؤثرة بتطبيق تقنيات التنقيب في البيانات على البيانات المُصدَّرة في مرحلة

خ- معالجة البيانات Processing Stage .

د- مناقشة وتقييم النتائج.

ذ- صياغة المقترحات والتوصيات للجهات المعنية وصانعي القرار.

١-٨ مصادر جمع البيانات

تتكون مصادر جمع بيانات الدراسة من المصادر الأولية والثانوية :

أولا : المصادر الأ ولية تتمثل في

- دراسة قاعدة البيانات المعمول بها حاليا فى ادراج ارقام جلوس الطلاب و استخراج شهادة المرحلة الثانوية (نظام الامتحانات) .

ثانيا : المصادر الثانوية

- نتيجة شهادة التعليم الثانوي (إدارة التقويم التربوى والامتحانات بالسودان).

١-٩ هيكلية البحث

يحتوى البحث على ست فصول الفصل الأول يحتوى على المقدمة والإطارالعام للبحث ويتضمن نبذه عن أهميةالبحث، مشكلةالبحث والمنهجيةالمتبعة فى البحث، أماالفصل الثانى يشمل الإطارالنظرى للبحث ويتكون من مبحثين المبحث الأول مكونات مستودع البيانات والمبحث الثانى مفهوم تنقيب البيانات والتقنيات المستخدمة فى تنقيب البيانات، والفصل الثالث يتكون من مبحثين المبحث الأول يتضمن التنقيب عن البيانات التعليمية والمبحث الثانى الدراسات السابقة والفصل الرابع يتكون ايضاً من مبحثين المبحث الأول التقنيات المستخدمةفى البحث: التصنيف وتقنية العنقدة وتم استخدام خوارزمية (K-means) فى تقنيةالعناقيد وخوارزمية (J48) وهى واحدة من خوارزميات شجرةالقرارفى تقنيةالتصنيف والمبحث الثانى وصف لتجهيز وتهيئة البيانات على برنامج (Weka) والصيغ التى يتعامل معها برنامج ويكا، الفصل الخامس يحتوى على تنفيذ وتحليل لبيانات نتيجة شهادة التعليم الثانوي وأخيراً الفصل السادس يحتوى النتائج والتوصيات التى خلص اليها البحث والخاتمة والمراجع والمصادرالمستخدمة فى البحث.

الفصل الثانـي

الإطار النظرى للبحث

٢-١ المقدمة

لقد ساعد الانتشار الواسع في مجال تكنولوجيا المعلومات إلى تضخم تلك المعلومات بصورة واسعة وكبيرة ؛ حيث أصبحت عنصراً مؤثراً وكبيراً في المجتمع بصفة عامة ، لذلك كانت الضرورة الهامة والأساسية هي كيفية الاستفادة منها، ولكن مع الدخول والانتشار الواسع للإنترنت وتكنولوجيا الاتصالات أصبحت الوسيلة سهلة لعملية تبادل وانتشار المعلومات؛ حيث الحاجة إلى توثيق تلك المعلومات بطريقة آلية وتجميعها في قاعدة بيانات لإمكانية البحث في محتواها مع معالجة تلك البيانات باستخدام تقنيات وأدوات فعالة تمكننا من سرعة الوصول والإسترجاع للمساعدة في الوصول إلى قرارات، من هنا ظهرت وتطورت قواعد البيانات تطوراً كبيراً منذ بدايتها حتي الآن، فمع بداية السبعينات ظهرت ما يعرف بقواعد البيانات الهرمية Hierarchical Database وصُمم هيكلها على أساس العلاقة بين السجلات التي تشكل هيكل شجري ومستويات هرمية ولهذا تعبر هذه التركيبة عن نمط العلاقات واحد – إلى كثير، وفي الثمانينات ظهرت قواعد البيانات العلائقية Relational database وهي من أكثر الأنماط استخداماً خصوصاً بعد ظهور وانتشار حزم نظم إدارة قواعد البيانات مع نظم الحاسب الشخصي، وفي التسعينات ظهرت قواعد البيانات الموجهة نحو الهدف Object–Oriented databases وهي الأكثر شيوعاً الآن لأنها تستطيع إدارة وسائط متعددة كما أنها تستخدم في تطبيقات شبكة الإنترنت؛ حيث أنها مفيدة في تخزين أنواع متعددة من البيانات، وهو ما يعرف بالجيل الرابع من قواعد البيانات، كل هذا أدي إلى بروز تحدي جديد في كيفية تحول البيانات من قواعد تخزين وبحث عن المعلومات إلى مخازن بيانات Data warehousing نستنتج منها المعرفة للمساعدة في اتخاذ القرار، لذلك كان لابد من معرفتها وخصائصها ووظائف تلك المستودعات وأهميتها في التنقيب عن البيانات Data mining واستخراجها[5] .

٢-٢مفهوم مستودعات البيانات

عبارة عن "مجموعة من بيانات دائمة تاريخية متكاملة تساعد في اتخاذ القرارات الإدارية [5] ، للمساعدة في الوصول إلى البيانات لأغراض التحليلات الزمنية واكتشاف المعرفة واتخاذ القرارات.

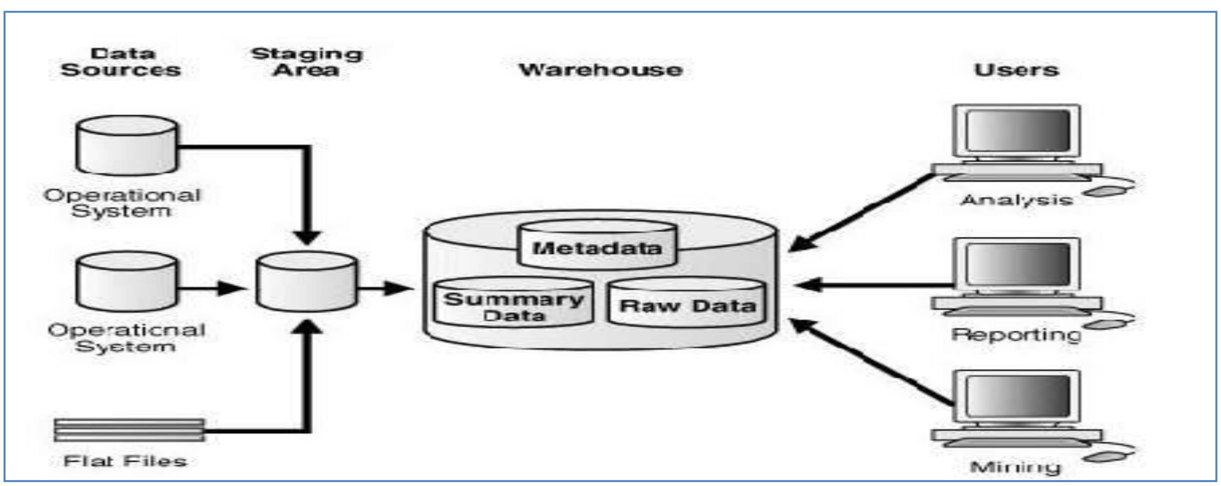
وتعرف ايضاً : بأنها مجموعة من البيانات المتكاملة والكبيرة والموضوعية التي تساعد في اتخاذ القرارات عن طريق تحليلها واستخلاص النتائج منها، وتتضمن هذه المستودعات (المخازن) بيانات من عدة مصادر مختلفة هي قواعد البيانات المنتشرة في أنحاء المؤسسة والتي تجرى عليها العمليات اليومية.

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 .

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 .

فهي مصممة خصيصاً لإستخراج واستخلاص البيانات ومعالجتها وتقديمها وتمثيلها في صورة مناسبة لهذا الغرض، وتتضمن كميات ضخمة من البيانات تكون من مصادر مختلفة، أو من عدة قواعد مختلفة من أنظمة وأماكن مختلفة، كما يعرفها البعض الآخر "على أنها نوع من قواعد البيانات تحوي كماً هائلاً من البيانات الموجهة للمساعدة على اتخاذ القرارات داخل المؤسسة [2] ، وتتميز تلك النوعية من قواعد البيانات بتطابق بنيتها الداخلية مع ما يحتاجة المستخدم من مؤشرات ومحاور التحليل، ويعرفها بندر الخيال [3] على أنها "تلك النظم التي تهتم بدراسة وتحليل وعرض بيانات المنشأة بصورة تمكن الإدارة العليا من صنع القرار"، ويتضح ذلك من معمارية بناء مستودع البيانات في شكل (٢–١) الذي يوضح هيكلها العام.



شكل(٢–١) نموذج مستودع البيانات[٣].

وقد برزت أهمية مستودعات البيانات مع وجود الشركات الكبرى؛ حيث أن كل قسم من تلك الشركات يدير قواعد بيانات خاصة به (تسويق، مالي، إدارة... الخ) والتي تتضمن كثيراً من البيانات المشتركة لذا تتطلب الحاجة إلى إيجاد بيانات متعلقة بموضوع محدد مثل العملاء على سبيل المثال من خلال البحث في قاعدة بيانات واحدة بدلاً من البحث في القواعد المختلفة، لذلك كان من الضروري تجميع البيانات في قاعدة واحدة تسمى مخزن البيانات مع الاحتفاظ بالقواعد الخاصة على أن تتم عملية تغذية دورية للمخازن إذا حدث أى تعديل أو تغير وتبديل في القواعد مع تصنيفها حسب الموضوع وذلك باستخدام برامج خاصة تسمى Middleware تستعمل بيانات دمج خاصة في كل موضوع تسمى Schema integration يحدد فيها شكل وكيفية نقل البيانات[4].

ويمكن القول بأن مستودعات البيانات هي عبارة عن قاعدة بيانات كبيرة وضخمة يمكن من خلالها الوصول إلى جميع المعلومات المتعلقة بمؤسسة أو شركة ما، ويودع في هذا المستودع البيانات الواردة من قواعد البيانات المختلفة كما يحتوي على العديد من قواعد البيانات وعلى معلومات من مصادر مختلفة، وتلك البيانات تكون في أشكال مختلفة ويمكن الوصول إليها من خلال الكمبيوتر الخادم، ويكون الوصول للمستودع غير مرئي بالنسبة للمستخدم الذي يستعمل أوامر بسيطة في التحليل واسترجاع وجمع البيانات، ويحتوى على بيانات تشير إلى كيفية تنظيمه، وأماكن تلك المعلومات، وهي تستخدم لمساعدة الإدارة في كيفية اتخاذ القرارات داخل الشركات أو المؤسسات، كما تسمح للمؤسسة في تنظيم وتحديث وتنسيق تلك البيانات، وإظهار العلاقات بين المعلومات المجمعة من الإدارت المختلفة بالمؤسسة.

٢-٣ خصائص وصفات مستودعات البيانات

يمكن إجمال خصائصها في النقاط التالية :

أ- تستخدم النموذج المتعدد الأبعاد Multidimensional model.

ب-تدعم السلاسل الزمنية Time series وتحليل التوجهات Trends analysis الذيْن يحتاجان لبيانات تاريخية لا تستطيع قواعد البيانات العادية أن توفرها.

ت-تحديث البيانات وهو يتم كل فترة بواسطة أجزاء منه تختص بهذا الأمر.

ث-استرجاع البيانات وتحليلها وهو صميم عملها وتهتم به أكثر.

ج- مستويات تجميع Aggregation وأبعاد Dimensions غير محددة.

ح- دعم معمارية العميل/الخادم وتعددية المستخدمين.

خ- الاحتفاظ بكمية ضخمة من البيانات قد تصل إلى عدة تيرابايتات.

د- أنها محددة لموضوع Subject–Oriented على سبيل المثال في منشأة طبية ما تكون موضوعاتها تتعلق بالأطباء، المرضى، الممرضون، الأدوية... الخ فهي أشياء حسية ومعنوية.

ذ- أنها متكاملة Integrated من خلال وجود علاقة بين البيانات.

ر- أنها غير قابلة للتعديل Non–Volatile أن تلك البيانات عند تحميلها في المستودعات تستخدم فقط للتحليل والدراسة والعرض.

ز- أنها مرتبطة بعامل الوقت Time–Variant وهى من أهم نقطة في تلك البيانات .

٢-٤ خطوات بناء مستودعات البيانات

إن مشروع بناء مستودع للبيانات في مؤسسة ما ليس أمراً سهلاً، بل يتطلب فريق محترف في عملية تحليل الأنظمة والأعمال، مع وجود مهارات تقنية محترفة في ذلك الفريق، وهذه الخطوات كالتالي [5] :

أ- إنشاء مساحة للبيانات Creating data preparation area وهي قاعدة بيانات ذات سعة تخزين عالية جداً تقوم بتخزين كافة البيانات القادمة من أنظمة التشغيل المختلفة لكي يتم تنقية وتعديل البيانات فيها قبل تحميلها في مستودع البيانات، ويراعى فيها أن يكون تصميم قاعدة إعداد البيانات متوافقاً مع تصميم مستودع البيانات.

ب-بناء مستودع البيانات Creating data warehouse وفيها ينشأ مستودع للبيانات والذي تحمل فيه بعد استخراجها وتنقيتها، وتصمم المستودعات دائما بحيث تسمح بوجود علاقات ذات أبعاد مختلفة.

ت-تجزئة مستودع البيانات إلى مجموعة من متاجر البيانات Data marts بحيث يكون هناك بيانات خاصة بالإدارة المالية، وأخرى خاص بإدارة الموارد البشرية أو يكون التقسيم مبنياً على فروع المؤسسة.

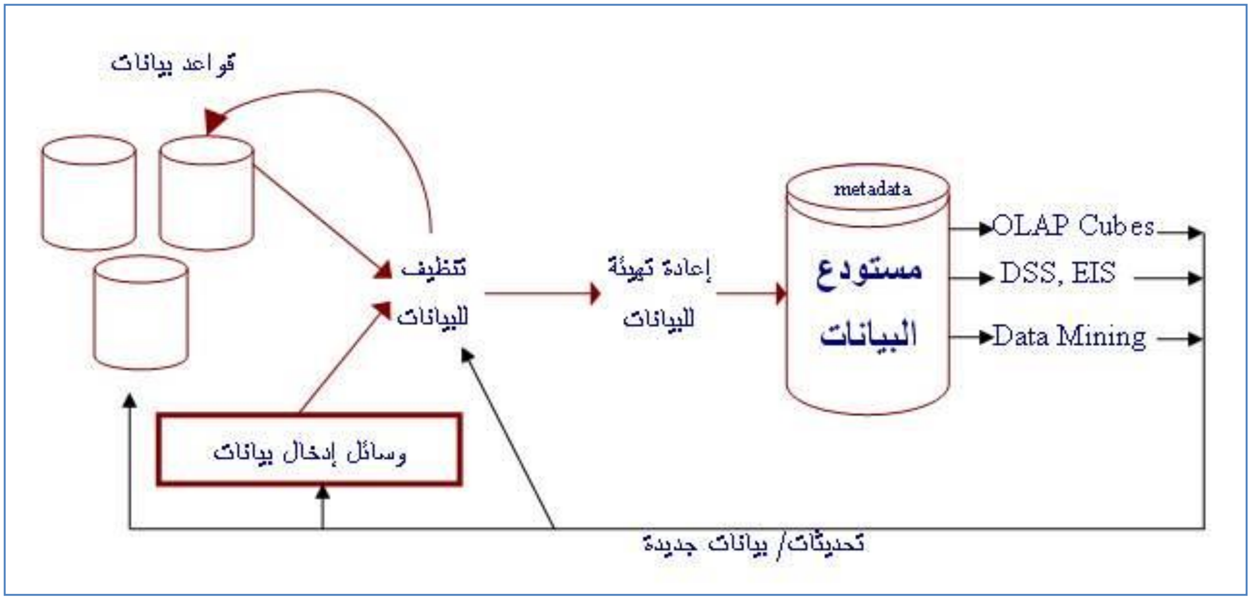
ث-دمج وتنقية ونقل البيانات Data extraction & cleaning & transformation وفي هذه المرحلة يتم جلب البيانات من مصادرها المختلفة إلى Data preparation area، ونقوم بتحويلها من صورة إلى أخرى إذا تطلب الأمر ذلك، وفي أحيان كثيرة تدمج بعض البيانات مع بعضها البعض، أو نقوم بتعريف بيانات جديدة لم تكن موجودة من قبل، بالإضافة إلى تنقية البيانات غير الصحيحة وحذف الغير مهم منها.

ج- تحميل البيانات في مستودع البيانات Loading data in data warehouse وفي تلك المرحلة تحمل من Data preparation area إلى مستودع البيانات، ويتم فيها اختبار البيانات.

ح- تحليل البيانات وإنشاء تطبيقات نظم دعم اتخاذ القرار OLAP & creating DSS applications وفي تلك المرحلة تنفذ التطبيقات الخاصة بعرض مستودع البيانات وتحليلها وتسمى

On**L**ine **A**nalytical **P**rocessing systems

وهذه التطبيقات تقوم بعرض البيانات بعدة أبعاد وتقوم باستخدام خوارزميات معقدة لتحليل البيانات، أيضاً في تلك المرحلة تستخدم أدوات التنقيب عن البيانات Data mining tools.



شكل(٢-٢) بناء مستودعات البيانات[5]

هناك بعض القضايا التي يجب مراعاتها عند بناء مستودعات البيانات وهي كالتالي[5]

١. استخلاص البيانات من عدة مصادر قد تكون غير متجانسة.

٢. تهيئة البيانات لضمان انسجامها Consistency داخل مستودع البيانات.

٣. تنظيف البيانات لضمان شرعيتها Validity، ويتم من خلال قاعدة البيانات التي أخذت منها البيانات.

٤. مراقبة وضبط حجم مستودع البيانات أثناء وبعد تحميله بالبيانات.

٥. تحديث البيانات كل فترة من الزمن.

٦. تحديد الوقت اللازم للبناء وما هي الجدوى الاقتصادية بالنسبة للمؤسسة التي سوف تستعمله ، معظم أو غالبية المؤسسات تعتمد بنية Three-tier architecture في عمل مستودعات البيانات،

يقسم المستودع إلى ثلاثة مستويات على النحو الآتي

* مستوى الأنظمة التشغيلية ومصادر البيانات المختلفة Data sources and operation systems.
* مستوى مستودع البيانات Data warehouses tier.
* مستوى تطبيقات العرض والتحليل OLAP systems.

وفي المستوى الأخير هناك الكثير من الشركات التي تقدم النظم والأدوات اللازمة لإنشاء مستودعات البيانات واستخلاص البيانات من مصادرها، وأبرزها شركة أوراكل وحلولها Oracle data mining, Oracle data warehouses builder.

٢-٥ التخزين في مستودع البيانات

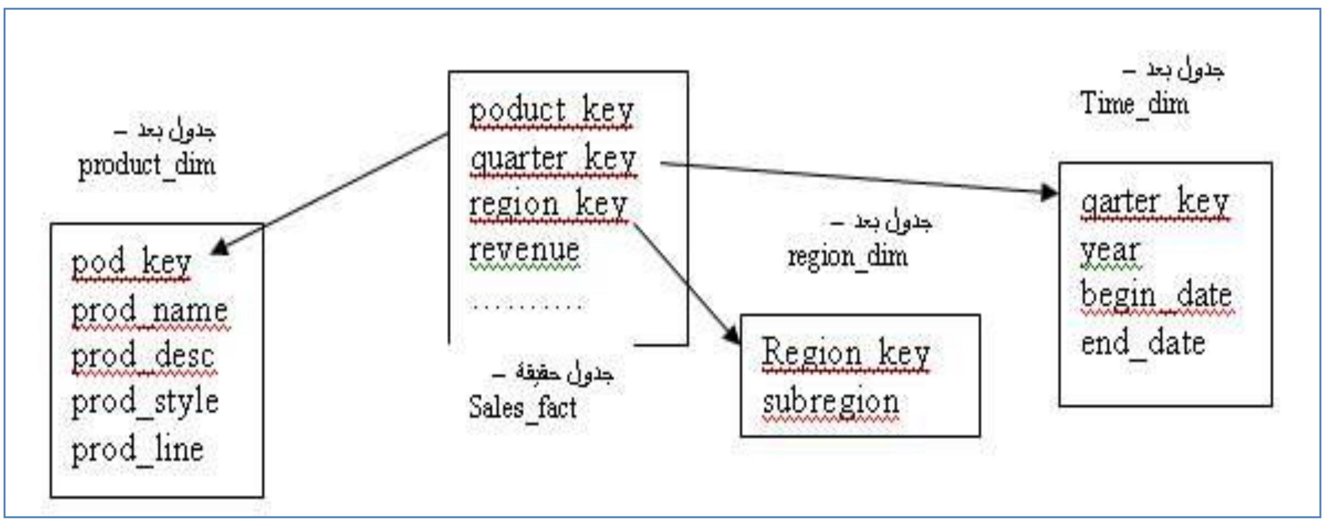
تخزين البيانات يتم بطرق متعددة الاتجاهات أو ما يعرف بالنموذج متعدد الأبعاد، وهي تستفيد من العلاقات بين البيانات لتسكينها في مصفوفات متعددة الأبعاد تسمى مكعبات البيانات Data cubes، وإذا استعملت قواعد البيانات العلائقية في مخازن البيانات وطريقة التحليل OLAP تسمى في تلك الحالة : ROLAP **R**elational، أو باستعمال قواعد بيانات متعددة الاتجاهات فتسمى Multidimensional OLAP، وعملية تخزين النموذج متعدد الأبعاد يتضمن نوعين من الجداول [5] :

* جدول البعد Dimension table وصفوفه تصف سمات البعد.
* جدول الحقيقة Fact table ويتضمن مقاييس أداء العمل، وتعرف بمؤشرات لجداول الأبعاد ويتضمن هذا الجدول البيانات.

وتوجد بنيتين شائعتين لتعريف هذه الجداول وهما

أولاً Star schema

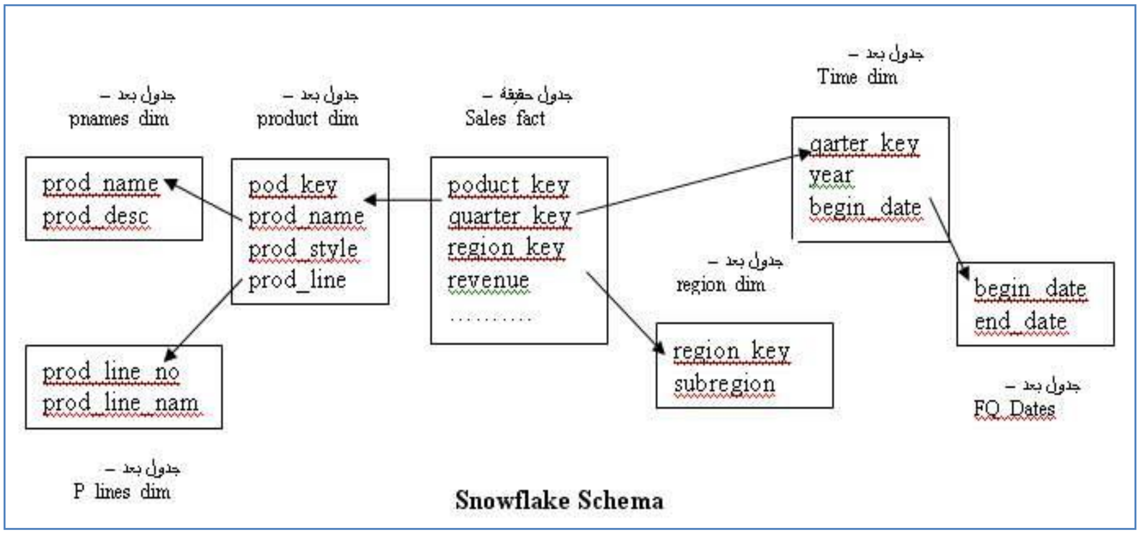
وتتكون من جدول الحقيقة مع جدول وحيد لكل بعد، ويعيبها احتياجها لمساحة كبيرة لتخزين البيانات.



شكل(٢-٣) جداول Star schema [5]

ثانياً Snowflake schema

وهى تعديل للجداول الأولى Star schema وذلك بتطبيق قوانين التطبيع Normalization عليها، ومشكلتها تضييعها لطاقة الجهاز المضيف في عمليات الربط.



شكل(٢-٤) جداول Snowflake schema [5]

٢-٦ عرض النتائج في مستودعات البيانات

عملية العرض تستخدم طريقتين لتحليل تلك البيانات [3]

الطريقة الأولى : وهى الحسابية حيث استعمال تحليل البيانات Analyze data من أجل إظهار النتائج بطريقة حسابية إحصائية (الجمع، الطرح، المعدل...الخ)

الطريقة الثانية : استخراج البيانات (Data mining) وإظهار البيانات في شكل رسم بياني أو إحصائي Diagram بحيث تحوى كل واحدة منها معلومات في مجال معين (الإدارة، التسويق، شؤون الموظفين... الخ) لتستعمل في تحديد البحث واستعمال أنظمة التحليل على تلك المخازن لإتخاذ القرارات.

٢-٧ الصعوبات التي تواجه تطبيق مستودعات البيانات

أ- الإنشاء يستغرق بعض الوقت ابتداءً من وضع الخطط حوله وحتى الانتهاء من تطبيقه.

ب-إدارة المستودع صعبة نظرا للحجم الكبير وتعقيده، ويتطلب ذلك تدريباً للقائمين عليه وخصوصاً من ناحية مراعاة جودة البيانات.

ت-التقدير لاحتياجات المستخدم قبل البدء في عملية الإنشاء.

ث-ظهور أشكال ومواقع جديدة للبيانات بعد الإنتهاء من إنشاء المستودع يزيد من صعوبة إدارته.

٢-٨ المبحث الثاني

مع وجود كميات هائله من البيانات المخزنه في قواعد البيانات الضخمه إزدادت الحاجه إلي تطوير أدوات تمتاز بالدقه لتحليل البيانات واستخراج المعلومات والمعارف منها ، ومن هنا ظهر ما يسمي بالتنقيب عن البيانات كتقنيه تهدف الي استنتاج المعرفه من كميات هائله من البيانات .

٢-٩ التنقيب في البيانات Data Mining

هي عملية البحث في قواعد البيانات (مستودعات البيانات) [3] 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 .

٢-٩-١ وللتنقيب في البيانات عدة تعريفات منها

أ- التنقيب في مجموعه ضخمه من مجلدات البيانات فضلاًعن اكتشاف العلاقه بينهما او الاجابه عن الاسئله المتخصصه التي تكون واسعه جدا عند استخدام ادوات الاستعلام التقليديه.

ب-عباره عن تحليلات لكميه كبيره من البيانات بغرض ايجاد قواعد وأمثله ونماذج يمكن ان تستخدم وتقود وتدل اصحاب القرار وتتنبأ بالسلوك المستقبلي.

ت-عباره عن تطبيق لاستخراج او اكتشاف معرفه مفيده وقابله للاستغلال من خلال مجموعه كبيره من البيانات حيث يساعد في اكتشاف المعرفه المخفيه والنماذج غير المتوقعه بالاضافه الي استكشاف قواعد جديده موجوده في قواعد بيانات كبيره.

٢-١٠ اصول نشأة التنقيب في قواعد البيانات

بدأ الباحثون من فروع مختلفه للمعرفه بالتركيزعلي تطوير ادوات اكثر فاعليه واكثر قابليه للتوسع ويمكنها ان تتعامل مع انواع متعدده الاشكال من البيانات ؛ وقد جمعهم معا هدف واحد هو مواجهه التحديات في قطاعات مختلفه. وبشكل خاص فقد طور التنقيب عن البيانات الافكار الموجوده ؛ مثل المعاينه والتقدير وخوارزميات البحث وتقنيات النمذجه واسترجاع المعلومات ونظريات التعلم المأخوذه من الذكاء الاصطناعي والتعلم الآلي [4] .

٢-١١ مفاهيم اساسيه عن التنقيب في البيانات ويضم

(On **L**ine **A**nalytical **P**rocessing**)**: OLAP

هي عمليه تحليل البيانات المتراكمه في مستودع البيانات DSS (Decision Support System)

وهي انظمه تساعد القياديين في المؤسسات والمنظمات علي اتخاذ القرارات الحاسمه والمعقده وذلك بتوفير بيانات من مستوي عالي.

(**O**n –**L**ine **T**ransaction **P**rocessing): OLTP

هي مجموعه من عمليات الاضافه والحذف والتعديل بالاضافه الي الاستعلام مع بعض التحليل الذي لا يرقي الي ان يكون مساعداً في اتخاذ القرار . ويتوفر دعم هذه العمليه في قواعد البيانات التقليديه.

٢-١٢ أهداف التنقيب في البيانات

أ- من اجل تعليل بعض الظواهر المرئيه مثلا لماذا زادت امراض السرطان في الوطن العربي

ب-من اجل التثبت من نظريه ما مثال : التثبت من النظرية التى تقول بأن الأسر الكبيرة تهتم بالضمان الصحى أكثر من الاسر الصغيرة عدداً .

ت-من اجل تحليل البيانات للوصول والحصول علي علاقات جديده وغير متوقعه.

٢-١٣ اسباب التنقيب في البيانات [3].

ان التنقيب في قواعد البيانات يهدف الي استخلاص المعلومات المخبأه فيها وهي تكنولوجيا حديثه فرضت نفسها بقوه في عصرالمعلوماتيه وفي ظل التطور التكنولوجي الكبيروإنتشاروإستخدام قواعدالبيانات. إستخدامها يوفر للمؤسسات واجهزه الامن في جميع المجالات القدره علي الاستكشاف والتركيز علي اهم المعلومات في قواعد البيانات. تركز تقنيه التنقيب علي بناء التنبؤات المستقبليه وإستكشاف السلوك والاتجاهات مما يسمح بتقدير القرارات الصحيحه وإتخاذها في الوقت المناسب. تجيب تقنيات التنقيب علي العديدمن الاسئلة،وفي وقت قياسي خاصة تلك النوعية من الاسئلةالتي يصعب الاجابةعليها، وان لم يكن مستحيلاً،بإستخدام الاحصاءالكلاسيكية،والتي ان وجدت فانهاتستغرق وقتاًطويلاًوالعديدمن إجراءات التحليل .

يعتبر استخراج البيانات جزءا مكملا من اكتشاف المعرفه في قواعد البيانات وهي العمليه الكليه التي يتم فيها تحويل البيانات الخام الي معلومات مفيده ، وعمليه اكتشاف المعرفه في قواعد البيانات تضم عدد من المراحل والتي تتمثل في Knowledge Discovery in Database (KDD) كما يلي :

٢-١٤إكتشاف المعرفه Knowledge Discovery in Database(KDD)

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 ، وعملية استكشاف المعرفة في قواعد

البيانات تتضمن عدد من المراحل التي تبدأ من جمع البيانات الخام (Raw Data) وصولاً إلى مرحلة الحصول على المعرفة الجديدة [3] .

٢-١٤-١ وفيما يلي عرض لهذه المراحل

أ- **اختيار البيانات Data Selection** وهي مرحلة اختيار البيانات المرشحة للدراسة من مستودع البيانات الكلي بحسب الغرض من الدراسة.

ب-**تهيئة البيانات Data Preprocessing** [2] وهي مرحلة تجهيز وعزل البيانات المبهمة أوالمفقودة أو التي تحتوي على شوائب من بقية البيانات كإلغاء المعلومات المتكررة، التصحيح الشكلي، معالجة البيانات الناقصة وجعلها جاهزة للتطبيق وتشمل هذه المرحلة (تنظيف البيانات ، وإزالة البيانات المفقودة ، إشتقاق البيانات ، دمج البيانات.)

ت-**تحويل البيانات Data Transformation** وهي عملية تحويل البيانات التي تم اختيارها إلى شكل ملائم للخوارزميات والتطبيقات التي ستستخدم في الدراسة . قد تشترط بعض الخوارزميات وجود البيانات على هيئة معينة قبل تطبيقها (.CSV/.ARFF)

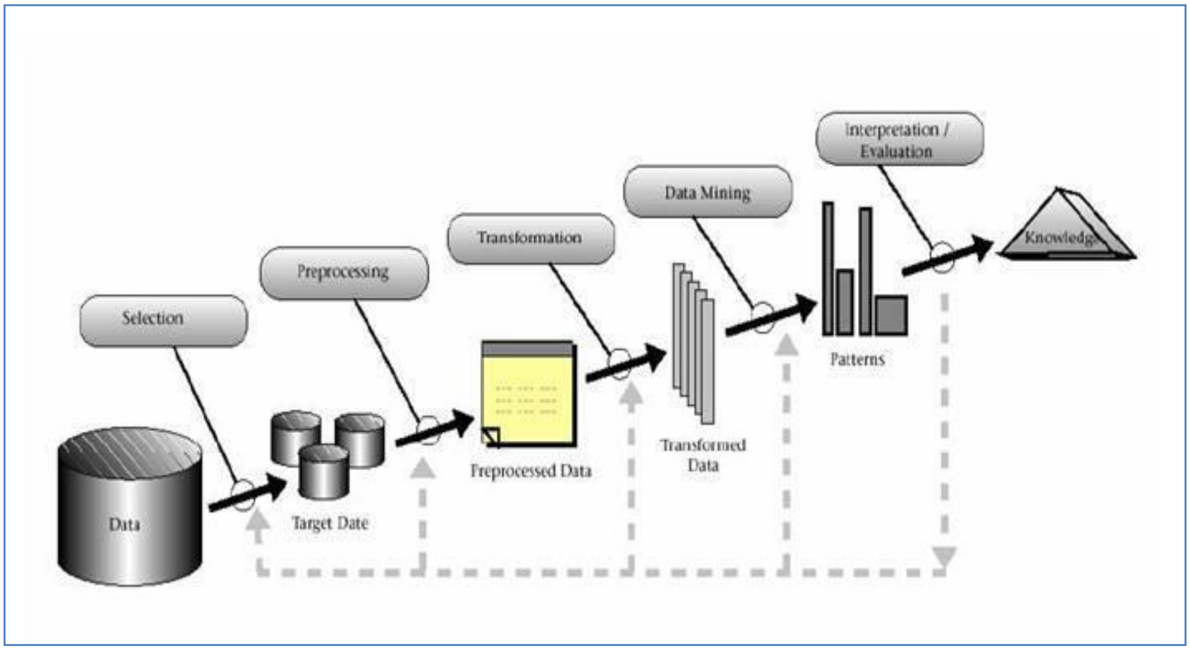
ث-**التنقيب في البيانات Data Mining** في هذه المرحلة يتم تطبيق طرق وخوارزميات ذكية وبتسلسل مناسب لاستكشاف أنماط مفيدة.

ج- **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.

ح- **تمثيل المعرفة (Knowledge Representation)**

وهي المرحله الاخيره من مراحل اكتشاف المعرفه في قواعد البيانات وهي المرحله التي يراها المستفيد، هذه المرحله اساسيه حيث تستخدم الاسلوب المرئي لمساعده المستفيد في فهم وتفسير نتائج استخراج البيانات ويمكن ان تتم في مرحلتين في آن واحد مثلا انجاز كل من مرحله تنقيه البيانات ومرحله توحيد البيانات مع بعضهما ويمكن ان تشترك مرحله اختيار البيانات مع مرحله نقل البيانات .

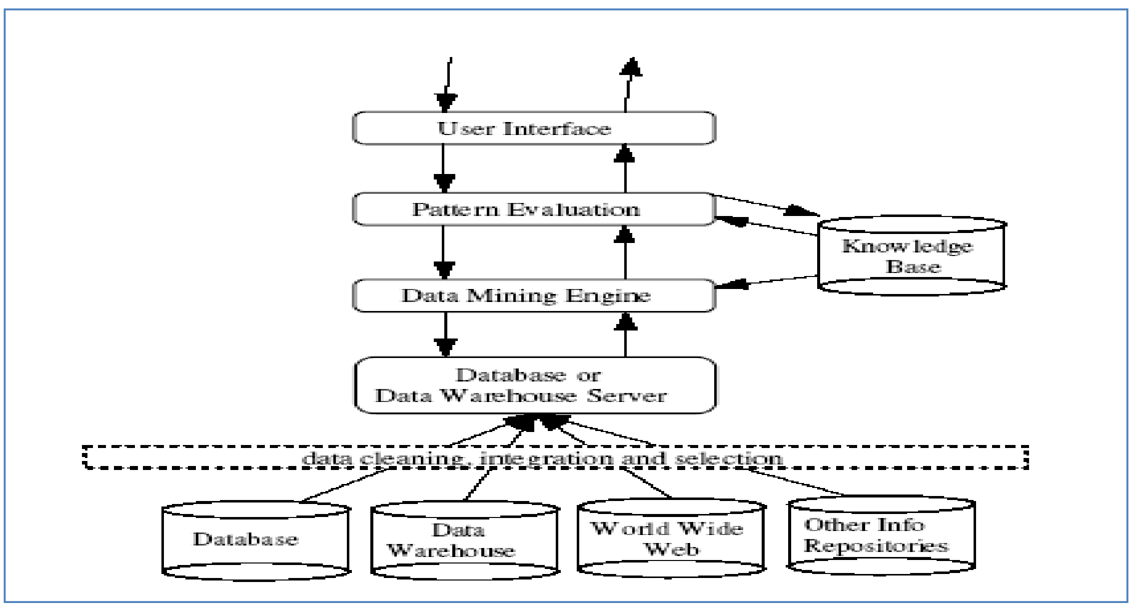
والرسم التالي يوضح التنقيب عن البيانات واكتشاف المعرفه :



شكل(٢ – ٥)خطوات استكشاف المعرفة[2]

٢-١٥ معمارية نظام التنقيب في البيانات

الشكل (6) يوضح معمارية نظام عام للتنقيب في البيانات [4] ، وسيتم فيما يلي توضيح أهم المفاهيم الواردة في الشكل.



شكل(٢ – ٦) معمارية نظام عام للتنقيب في البيانات[4]

٢-١٥-١ قاعدة المعرفة : Knowledge Base

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

٢-١٥-٢ آلة التنقيب في البيانات : Data Mining Engine

وهو ضروري لنظام الـData Mining ويشمل مجموعة من الوحدات الوظيفية مثل:

Characterization– Evolution– Outlier–Analysis–Association[4]

٢-١٥-٣ تقييم الأنماط Pattern Evaluation

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

٢-١٥-٤ واجهة المستخدم User Interface

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

أ- التفاعل مع النظام بتحديد استعلام أو مهمة معينة.

ب-التزود بمعلومات للمساعدة على تركيز البحث.

ت-تصفح قاعدة البيانات وهيكلية مستودع البيانات.

ث-تمثيل الأنماط الناتجة صورياً في عدة أشكال .

٢-١٦ التخطيط لعمليات التنقيب فى قواعد البيانات

إن تخطيط عمليات التنقيب في قواعد البيانات من الأمور المهمة للحصول على أفضل النتائج، فالتخطيط الجيد يؤدي للنتائج الجيدة. ويمكن تلخيص الخطوات الأولية للتنقيب في قواعد البيانات في ما يلي:

أ- تحديد المشكلة المراد بحثها وإيجاد الحلول لها

ب-بناء قاعدة بيانات التنقيب

ت-استكشاف البيانات

ث-تحضير البيانات للتنقيب

ج- بناء نموذج التنقيب المناسب

ح- تطبيق النموذج

خ- استخراج النتائج

٢-١٧ وسائل التنقيب عن البيانات

هنالك عدة وسائل مختلفة :

أ- التفكير وأستخلاص النتائج .

ب-الكشف عن قانون Rule Discovery .

ت-معالجة الأشارات .

ث-شبكات نووية .

ج- منحنيات غير ثابته .

٢-١٨ التحديات التى تواجه التنقيب عن البيانات

بالرغم من أن المنتجات التى تقوم بتنقيب البيانات قوية جدا ؛ الا انها لا تفى بالغرض [5] لوحدها وحتى تتمكن هذة المنتجات من النجاح فى التنقيب تحتاج لمهارات تقنية وإحصائية متخصصين لديهم القدرة العالية على التحليل الجيد وترجمة النتائج بصورة تفيد المؤسسة ؛ وتجعل عملية التنقيب فعالة وبالتالى نجد ان التحديات التى تواجه عملية التنقيب تتمثل فى الاتى

أ- عملية التنقيب تساعد فى اكتشاف الانماط والعلاقات الا انها لا تخبر المستخدم بقيمة او اهمية النمط او العلاقة التى استنبطها ؛ لان هذة هى مهمة المستخدم فالالة المستخدمة فى استخراج الذهب لا تريك كيف تقوم بتشكيلة ، فصلاحية النمط او العلاقة المكتشفة تعتمد على امكانية مقارنتها مع ظروف العالم الحقيقى فقد لا نستطيع تطبيق ما اخرجتة اداة التنقيب على ارض الواقع.

ب-عملية التنقيب تعرف الارتباطات بين المتغيرات والسلوك إلا انها لا تعرف العلاقة الغير رسمية .

٢-١٩ تقنيات التنقيب عن البيانات

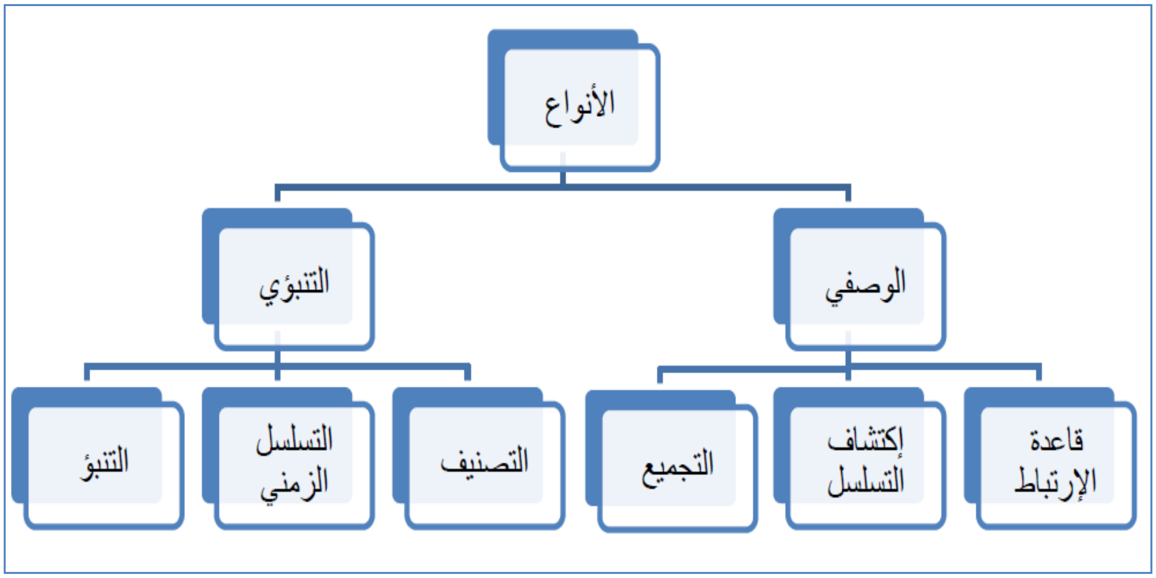
توجد عدة تقنيات تستخدم فى تنقيب البيانات و اختيار التقنية المناسبة يعتمد على طبيعة البيانات التي تتم دراستها فضلاً عن حجم البيانات وهناك نموذجان لتنقيب البيانات

٢-١٩-١ التنقيب التنبؤي (Prediction) ويضم عدة تقنيات منها

التصنيف Classification ، الانحدار Regression ، تحليل السلاسل الزمنية Time Series Analysis

٢-١٩-٢ التنقيب الوصفى (describe) ويحتوى على عدة تقنيات منها

قواعد الارتباط Association Rules ، العنقدة Clustering ، اكتشاف التسلسل Sequence Discovery المرئية Visualization التلخيص Summarization.



شكل(٢-٧) أنواع تنقيب البيانات[5]

٢-١٠ أولاَ التنقيب التنبؤي Predictive Model

وهو النموذج الذي يستخدم النتائج المعروفة المستنبطة من البيانات المختلفة لأجل التنبؤ بقيم لاحقة [3]للبيانات ويحاول إيجاد أفضل التنبؤات إعتماداً على المعطيات كمعرفة المنتج الأفضل لزبون معين بإختصار يعتمد هذا التنقيب على إستخدام معلومات قديمة لتوقع ما سيحدث في المستقبل وتكون لدى مثل هذه البيانات هدف . ويتضمن هذا النموذج أشهر تقنيات أدوات تنقيب البيانات .

٢-١٠-١ التصنيف Classification

يستخدم التصنيف بشكل واسع في حل الكثير من المشكلات خاصة تلك التي تتعلق بالأعمال [3] Business من خلال تحليل مجموعة من البيانات ووضعها على شكل أصناف أو أقسام يمكن استخدامها فيما بعد لتصنيف البيانات مستقبلاً، وهنا يكمن الفرق بين التصنيف وإنشاء العناقيد، فالتصنيف يقصد به تقسيم البيانات إلى مجاميع يتم تحديدها مسبقاً، أما إنشاء العناقيد أو ما يسمى بال 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. وطريقة الـDecision Trees وهي هيكل شجري يقدم مجموعة من القرارت التي تولد قواعد لمجموعة البيانات المصنفة (classification data) تشترط هذه التقنية وجود حقل قرار يتم تصنيف البيانات بناءً عليه .

٢-٢٠-٢ الانحدار Regression

هو تقنية تسمح بتحليل البيانات لوصف العلاقة بين متغيرين أو أكثر ، ويستخدم في تقدير قيمة أحد المتغيرين إذا عرف المتغير الآخر. و يفترض أن توضع البيانات بنوع معروف من الدوال ومن ثم يتم تحديد أفضل دالة للبيانات المعطاة .

٢-٢٠-٣ تحليل السلاسل الزمنية Time Series Analysis

إن مشاهدة البيانات عبر الزمن تنتج تحليلاً مفيدًا لأنه تتم مشاهدة سلوك البيانات عبر الزمن بشكل أساسي . وهذا يعني أن قيم الصفة المميزة للبيانات التي يتم فحصها تكون متغيرة عبر الزمن •

٢-٢٠-٤ التنبؤ Prediction

يعد التنبؤ من الأدوات التي تجذب الانتباه لأنها تتمكن من إعطاء مغزى التوقع الناجح في سياق العمل لذا فإنه يمكن النظر إلى العديد من تطبيقات تنقيب بيانات العالم الحقيقي كأنها تنبؤ بحالة بيانات مستقبلية معتمدة على بيانات سابقة وحالية•

٢-٢١ ثانياً التنقيب الوصفي Descriptive Model

وهو النموذج الذي يُعَرِّف الأنماط والعلاقات في البيانات ، وعلى العكس من النموذج التنبؤي ، فإن النموذج الوصفي يستخدم كطريقة لإستكشاف خصائص البيانات التي تتم دراستها وليس للتنبؤ بخصائص جديدة ، ويعتمد علي إعادة تنظيم البيانات ، والتنقيب في أعماقها لإستخراج النماذج الموجوده فيها كتشابه الزبائن الذي يسمح لك بإنشاء وصف بسيط عن مجموعه زبائن متشابهين ولايستوجب وجود هدف لمثل تلك التشابهات ، ويعد من أشهر تقنيات هذا النموذج من أدوات تنقيب البيانات .

٢-٢١-١ العنقدة Clustering

الهدف منها هو تحديد الاتجاهات داخل البيانات ، ويحاول هذا الاسلوب العثور على مجموعات من العناصر التي توجد عادة معا . وهي عملية تقسيم البيانات إلى مجموعات من الأصناف إعتماداً على

إشتراكها بالخواص المتشابهه وأن العنقدة هي تقسيم غير موجه للبيانات ، وهي عكس التصنيف ، كما أنها تساعد المستفيد على فهم التركيب الطبيعي لمجموعات من البيانات.

٢-٢١-٢ قواعد الارتباط Association Rules

قواعد الارتباط هي إحدى التقنيات الواعدة في الـData 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 وبمعنى آخر هي قواعد ارتباط معينة بين عدة مجموعات من البيانات في قاعدة بيانات واحدة وتعتمد طريقة قواعد الارتباط على ايجادLarge Item Set

٢-٢١-٣ اكتشاف التسلسل Sequence Discovery

إن تحليل أو اكتشاف التسلسل يستخدم لتحديد أنماط متسلسلة في البيانات وهذه الأنماط معتمدة على تسلسل زمن التأثيرات . فى هذه الطريقه يتم البحث لإكتشاف نماذج تحدث بالتسلسل إذ تكون المدخلات عبارة عن بيانات تشكل مجموعه متسلسله و كل سلسله من البيانات هى قائمه منظمه من العمليات أو المصطلحات و عندما تكون العملية عبارة عن مجموعات من المصطلحات لابد أن يحسب معها الوقت المصاحب لكل عملية ولكن مشكلة هذا النموذج تكمن فى إيجاد كل النماذج المتسلسلة مع أقل دعم يخصصه المستخدم عندما يكون الدعم لهذا النموذج هو نسبة تسلسل البيانات التى يتضمنها النموذج.

٢-٢١-٤ المرئية Visualization

إن مرئية النتائج يكون مساعدًا في تسهيل ملاحظات مخرجات خوارزميات تنقيب البيانات وفهمها[3] .

٢-٢١-٥ التلخيص Summarization

ويسمى أيضًا بالمميزات أو(الخواص) Characterization أو العمومية Generalization ويمكن تعريفها على أنها وصف الخصائص العامة للنماذج

٢-٢٢ تطبيقات تنقيب البيانات

وسائل التنقيب فى البيانات تستعمل و بنجاح فى كثير من التطبيقات الحقيقية حول العالم:

- **التجارة**

أ- الكشف عن حالات الإحتيال : تحديد المعاملات الإحتيالية.

ب–الموافقه على القروض : تحديد الاهلية الإئتمائية للعميل طالب القرض.

ت–تحليل الإستثمار : التنبؤ بمحفظة العائد على الإستثمار.

ث–التسويق و المبيعات و تحليل البيانات : تحديد العملاء المحتملين،اقامة فعالية حملة المبيعات.

• **الطب**

أ- تحليل تأثير المخدرات : من حالات المرضى لمعرفة اثار المخدرات.

ب–مكافحة الامراض وتحليل العلاقة السببية.

ت–تصنيف الامراض : تصنيف مرض السرطان.

• **السياسة**

السياسة الاجتماعية : الضرائب والفوائد .

• **التصنيع**

أ- الصناعة التحويلية فى عملية التحليل : تحديد اسباب المشاكل والتصنيع.

ب–التحليل لنتائج التجربة : ملخص لنتائج التجربة وخلق نماذج تنبؤ.

٢-٢٣ تطبيقات التنقيب فى البيانات بدأت تنمو بصورة كبيرة للاسباب الآتيه

أ- كمية البيانات الموجودة فى مخزن البيانات وسوق البيانات تنمو بصورة كبيرة ، إذ أن وجود بيئة معلوماتية كبيرة تدفع المهتمين للاستفادة منها.

ب- ظهور الكثير من ادوات التنقيب الفعالة ، شجع على ازدياد عملية التنقيب فى البيانات بكثرة.

ت- المنافسة الشديدة الموجودة فى السوق تدفع الشركات الى البحث عن طرق تساعدها على الانتاح باقل التكاليف.

الفصل الثالث

تنقيب البيانات التربوية

٣-١ المقدمة

تنقيب البيانات التربوية (Education Data Mining (EDM هو أحد التخصصات الناشئة ويعنى بتطوير أساليب لاستكشاف الانواع الفريدة من البيانات التى تؤخذ من البيئات التعليمية وأستخدام هذه الأساليب فى تحقيق فهم أفضل للطلاب والبيئات التى يتعلمون فيها ويعتبر استخراج سجلات أداءالطلاب المخزنة على الكمبيوتر من المجالات الاساسية فى هذا التخصص كما أن استخراج بيانات التسجيل يعد مجالاً رئيسياً آخر تتضمن الاستخدامات الأساسية لاستخراج البيانات التعليمية التنبؤ بأداء الطلاب ودراسة عملية التعلم بهدف التوصية بإدخال تحسينات على الممارسة التعليمية القائمة . ويمكن اعتبار استخراج البيانات واحداً من علوم التعلم ومجالاً من مجالات استخراج البيانات ، هذا ويعد تحليلات عملية التعلم من المجالات ذات الصلة به .

٣-٢ أساليب تنقيب البيانات التعليمية

هنالك ارتباط بين أنواع أساليب تنقيب البيانات التعليمية وأساليب أستخراج البيانات عموماً ولكن مع بعض الفروقات التى ترجع إلى الخصائص الفريدة للبيانات التعليمية [4] .

يصف (ريان بيكر) مجالات أستخراج البيانات التعليمية كما يلى [4].

أ- التنبؤ
ب-التصنيف
ت-الإنحدار
ث-تقدير الكثافة
ج- التجميع
ح- أستخراج العلاقات
خ- إستخراج قواعد الإرتباط
د- أستخراج الأنماط التسلسلية
ذ- إستخراج البيانات السببية .
ر- تصفية البيانات للوصول لقرار بشرى .
ز- إستكشاف النماذج .

يقول (بيكر وكالينا ياسيف) ان لاستكشاف النماذج أهمية خاصة فى إستخراج البيانات التعليمية مقارنة باستخراج البيانات بشكل عام ، فمن خلال إستكشاف النماذج ، يتم تطوير نموذج لظاهرة ما من خلال

أي عملية يمكن إثباتها بطريقة ما (تكون فى أغلب الاحيان بالتنبؤ أو هندسة المعرفة) ثم يستخدم هذا النموذج كعنصر فى تحليل آخر مثل التنبؤ أو إستخراج العلاقات [4].

٣-٣ تطبيقات تصنيف تنقيب البيانات التعليمية

للتصنيف تطبيقات كثيرة فى التكنولوجيا التعليمية الحديثه ، وفيما يلى قائمة بأهم المبادئ وأساليب التصنيف المستخدمة فى تنقيب البيانات التعليمية والتى حررها كل من كريستوبال روميرو وسبيستيان فينتورا وميكولا بتشينسكى وريان بيكر [4] .

٣-٤ المبادئ الرئيسية

أ- المصنف التمييزى أو الاحتمالى .

ب-دقة التصنيف .

ت-المطابقة المفرطة .

ث-الحدود الخطية واللاخطية للفئة .

ج- المعالجة المسبقة للبيانات .

ح- أساليب التصنيف .

خ- أشجار القرار .

د- مصنفات بايز .

ذ- الشبكات العصبية .

ر- مصنفات الجار الأقرب (K.Nearest Neighbor)

ز- أجهزة المتجهات الأعتمادية.

س-الانحدار الخطى .

ش-المقارنة .

٣-٥ عملية الإستخراج من البيانات التعليمية

فيما يلى وصف لعملية تنقيب البيانات التعليمية التى حررها كريستوبال روميرو ، سبيستيان فينتورا، ميكولا بتشسنسكى و ريان بيكر .

٣-٥-١ تتالف عملية التنقيب من ثلاث مجالات فرعية :

أ- فحص المطابقة .
ب-أستكشاف المطابقة .
ت-توسعة النموذج (مثل إطار Prom).

٣-٦ عملية تنقيب مجموعة البيانات التعليمية

أ- إعداد البيانات .
ب-الإستخراج البصرى بأستخدام تحليل المخطط المنقط .
ت-تحليل المطابقة .
ث-التحليل المنطقى الوقتى الخطى (LTL)
ج- عملية الاستكشاف بأستخدام أداء أستخراج التطابقات .

٣-٧ التطبيقات

هنالك قائمة بالتطبيقات الرئيسية لتنقيب البيانات قدمها كريستوبال روميرو وسبيستيان وفنتورا [4] . فى علم التصنيف ، فإن مجالات تطبيق أستخراج البيانات هى:

أ- تحليل البيانات وتصوير البيانات .
ب-تقديم تعليقات لدعم المعلمين .
ت-توصيات للطلاب .
ث-التنبؤ بأداء الطلاب .
ج- وضع نماذج للطلاب .
ح- اكتشاف سلوكيات الطلاب غير المرغوب فيها .
خ- تقسيم الطلاب الى مجموعات .
د- تحليل الشبكة الاجتماعية .
ذ- تطوير خرائط المفاهيم .
ر- إنشاء برامج منهجية .
ز- التخطيط ووضع الجداول الزمنية .

٣–٨ 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 التعليمية بتطبيقات التنقيب في البيانات في مجال التعليم ، وتركز على مدخلات العملية التعليمية ومخرجاتها وكيفية تأثير كل منها على الأخرى .

٣–٩–١ الدراسة الأولى استخدام تقنية التصنيف لتقييم سجل الطالب

اسم الكاتب	اسم الدراسة	تاريخ النشر
Abdullah Hussein Al-Hashiedy	"Data Mining Applications in Higher Education"	2006

قام الباحثين بدراسة في احدى جامعات الهند[2] بعمل بحث بإستخدام تقنية التصنيف وتم تطبيقها علي البيانات التعليمية أخذت من درجات طلاب الماجستير وكان عدد الطلاب 60 طالب ، وكانت البيانات تتضمن رقم الطالب ومجموع العلامات التي تحصل عليها خلال الأربعه سنوات الدراسية ، وكان الغرض من هذا البحث هو توقع أداء الطالب في الإمتحانات واظهار التحسن في أداءه الأكاديمي .

وتم إعداد البيانات (Preprocessing) علي مجموعة البيانات طبقا لقواعد الجامعة والعلامات التي حصل عليها الطالب تم تحويلها إلى الشكل الإسمي (nominal) وذلك لغرض البحث ، وتم إستخدام إحدى أدوات تنقيب البيانات وهى(Weka) وكانت العوامل المستخدمة في الخوارزميتين هي:

أ– الكفاءة في توقع نتيجة الطالب.

ب–القيم الايجابية الصحيحة والقيم الايجابية الخاطئة وقيم الROC .

ت– الزمن المأخوذ في كل خوارزمية لإنتاج الناتج .

تقنيات التصنيف المختلفة يمكن إستخدامها لتوقع نتيجة الطالب في الفصل الدراسي النهائي بحيث يكون مستند علي العلامات التي حصل عليها في الفصول الدراسية السابقة ، تقنيات التصنيف التي تم تطبيقها هي:

a. One-R Algorithm
b. Decision Table algorithm.

بعد تطبيق التقنيتين وجد أن خوارزمية (One-Rule) أفضل من خوارزمية (Decision Table) لعدة اسباب وهي أن الخوارزمية الأولى أكثر دقة من الأخرى وتؤدي أفضل منها وتأخذ وقت اقل.

٣-٩-٢ الدراسة الثانية

اسم الكاتب	اسم الدراسة	تاريخ النشر
Asma`a A. Alshargabi	"Discovering Vital Patterns from UST Students Data – Applying Data Mining Techniques"	2006

فقد قامت الباحثة فيه باستكشاف أنماط هامة في بيانات طلاب جامعة العلوم والتكنولوجيا اليمنية وذلك بتطبيق التقنيات الثلاث للتنقيب في البيانات[2] (تقنية التصنيف ، قواعد الربط و التجميع) .

- **النتائج التي تم التوصل إليها في هذه الدراسة يمكن تلخيصها فيما يلي**

أ- أن أكثر نسبة رسوب تكون في المراحل الأولى من الدراسة وأن هذا الرسوب مرتبط بانسحاب الطلاب.

ب-أن فشل الطالب في المراحل الأولى مرتبط بضعف معدلاته في المرحلة الثانوية.

ت-يرتبط الفشل لدى بعض الطلاب بتواضع القدرات الشخصية للطالب،إحدى الطرق التي اقترحتها الباحثة لعلاج هذه المشكلة هي بإضافة سنة تحضيرية للطلاب.

ث- تسرب الطلاب مرتبط بصعوبة بعض المواد وخاصة في المراحل الأولى لهم.

٣-٩-٣ الدراسة الثالثة

اسم الكاتب	اسم الدراسة	تاريخ النشر
Jing Luan	“Data Mining Applications in Higher Education”	2004

استخدمت 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[3] .

٣-٩-٤ الدراسة الرابعة

البيانات التي تطلبت للدراسة في هذا البحث تم أخذها من جامعة مشهوره في منطقة (Haryana) في الهند[3] لتَحسين أداء التعليم بتحليل البيانات واكتشاف العوامل التي تؤثر على النتائج الأكاديمية ودراسة أداء الطلاب لكي تزِيد فرص نجاح الطلاب . وتم عمل هذا البحث علي مجموعه من الطلاب وكان

عددهم30 طالب، وتمت عليها بعض خطوات تجهيز البيانات (Data Preprocessing) وهي التنظيف cleaning التحويل(Transformation) وتم تحليل قوانين الإرتباط بأستخدام أحد أدوات تنقيب البيانات وهي(Tanagra) وذلك بتطبيق قوانين الإرتباط عن طريق (Apriori Algorithm) لمقارنة أداء طلاب الدراسات العليا في مجال تطبيقات الحاسوب قبل وبعد التخرج .

بالإضافة لتحليل البيانات لمعرفة مدى تأثيرها على النتائج المتوقعة من نجاح أو فشل الطلاب وبإستخدام هذه التقنية تم التوصل الى معرفة مدى إهتمام الطلاب بمواد معينة وتصميم المناهج من حيث طريقة التدريس وتقييمها لكي يتم قياس مدى تأثيرها على الطلاب الذين لم ينجحوا لكي يصل بهم الى الحد المقنع من الاداء وكانت قيمة (confidence=16.667) وكانت قيمة (support=83.333) وكان عدد (rules=2) بالأخير تم الكشف عن عوامل مختلفة من حيث إهتمام الطالب ، تصميم المناهج ، تقييم المنهجيات التي تؤثر علي الطلاب وتسببت في اخفاقهم.

الفصل الرابع

تقنيات البحث

٤-١ المقدمة

كلما تزايدت كمية وأنواع البيانات المتوفرة أصبحت الحاجة ضرورية الى تقنيات صلدة وكفؤة ومتعددة الاستخدامات وذلك من أجل القيام بعملية تحليل هذه البيانات والإستفادة منها،وتعتبر تقنيات التصنيف من أهم التقنيات فى تنقيب البيانات .

هناك عدة تقنيات يوفرها برنامج WEKA للتنقيب في البيانات ، وسوف نسرد هنا تلك التقنيات التي تم استخدامها في هذا البحث وطريقة عمل هذه التقنيات ، حيث تم استخدام خوارزمية(K-means) ضمن تقنية العناقيدClustering، كما تم استخدام خوارزمية(J48)ضمن أشجار القرار Decision Trees وهى ضمن تقنية التصنيف Classification.

٤-٢ تقنية التصنيف Classification

التصنيف هو طريق لتصنيف أو تخصيص مؤشرات صنف -تحت إشراف معلم- تجزء النماذج الى [5] مجاميع " تدريبية " و "فحصية " تسمى بالمصنف والذى يدرب على المجاميع " التدريبية " وتستخدم مجموعة بيانات تسمى ببيانات الفحص لتقييم القدرة المعممه للمصنف classifier توضح الأمثله أدناه بعض مجالات التصنيف المختلفة وهى

أ- تصنيف المرضي إعتماداً على المرض .

ب-تصنيف أعطال الأجهزة معتمدة على السبب .

ت-تصنيف طلب سلفية إعتماداً على الدفع .

٤-٢-١ خطوات التصنيف

تصنيف البيانات هو عبارة عن عملية تتألف من خطوتين [5]

أ- الخطوة الأولى : يبنى النموذج واصفاً مجموعة من أصناف البيانات أو المبادئ التى تم تحديدها مسبقاً.

ب-الخطوة الثانية : يفرض أن كل قيد (سجل) (tuple) تابع الى صنف تم تحديده مسبقاَ ، ويكون هذا التحديد بواسطة واحدة من الصفات التى تسمى صفة مؤشر الصنف ، ويسمى قيود البيانات بأنها أمثلة example أوعينات sample تقسم هذه (الأمثله والعينات) الى مجموعتين هما

مجموعة التدريب ومجموعة بيانات الفحص ، حيث يتم بناء نماذج التصنيف فى خطوتين هما

أ- التدريب (training) يبنى النموذج من مجموعة بيانات التدريب .

ب-الفحص (test) : فحص دقة النموذج بأستخدام مجموعة بيانات الفحص .

نماذج التدريب والفحص تقسم الى الأنواع التالية :

أ- قواعد إذا نفذ if –then rule.

ب-الصيغة المنطقيةLogical formula.

ت-اشجارالقرار Decision Tree.

٤-٢-٢ الطريقة العامة لحل مشكلة التصنيف

تقنية التصنيف أو المصنف (classifier) هى طريقة منظمة systematic لبناء نماذج تصنيف من خلال إدخال مجموعة من البيانات ولها عدة تقنيات من أهمها مصنفات أشجار القرار والمصنفات القاعدية (rule–based) المعتمدة على القاعدة والشبكات العصبية (neural network) ومكنات الأسناد الموجه(support vector machines) ومصنفات بيز (byes classifier) البسيطة .

تستخدم كل تقنية من التقنيات السابقة خوارزمية تعلم algorithm learning لتحديد نموذج يلائم العلاقة بين مجموعة الصفات ومؤشر الصنف لبيانات الإدخال حيث يتم توليد النموذج من خلال خوارزمية تعلم ويجب على الإثنين (النموذج والخوارزمية) ملائمة البيانات المدخلة بصورة جيدة والتنبؤ بصورة دقيقة لمؤشرات الصنف للقيود التى لم تكن مرئية سابقاً .

لذلك فإن الهدف الرئيسى لخوارزمية التعلم هو بناء نماذج يمكن تعميمها ، أى نماذج تتنبأ بشكل دقيق بتسميات أصناف سجلات غير معروفه مسبقاً .

٤-٢-٣ انواع تقنيات التصنيف

توجد العديد من تقنيات التصنيف منها كما يلى [5]

أ- الطرق المعتمدة على شجرة القرار

ب-أستقراء القاعدة (Rule Induction)

ت-تصنيف بيز (Byse Classification)

ث-طريقة الشبكات العصبية (Neural Networks)

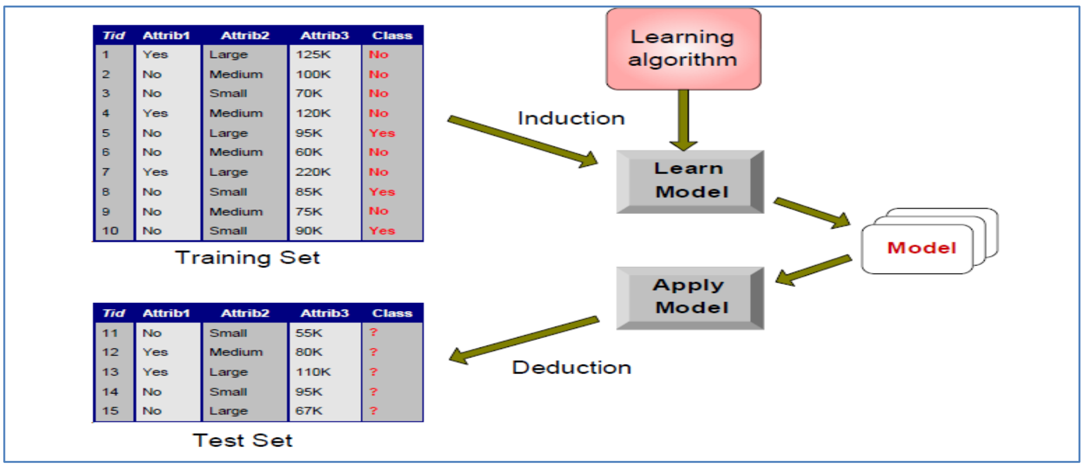
ج- خوارزمية الجار الأقرب (Nearest Neighbor)

ح- أسناد الماكنات الموجهه (support vector machines)

خ- الخوارزميات الجينية (Genetic Algorithms)

شكل(٤-١) التصنيف[5]

تتكون مشكلة التصنيف من مجموعة من القيود وكل قيد يتكون من مجموعة من الصفات وكذلك مؤشر كل صنف ، يتم بناء نموذج دقيق لكل صنف اعتماداً على مجموعة الصفات . بعد ذلك يتم أستخدام النموذج لتصنيف بيانات مستقبلية والتي تكون مؤشرات أصنافها غير معروفه .



شكل(٤-٢) خطوات بناء مصنف البيانات[5]

٤-٢-٤ مصفوفة التعارض confusion matrix [6]

يتم جدولة عدد السجلات المصنفة بشكل صحيح وعدد السجلات المصنفة بشكل خطأ على شكل مصفوفة تعرف بمصفوفة التعارض.

		Predicted class	
		Class = 1	Class = 0
Actual class	Class = 1	TP	FN
	Class = 0	FP	TN

شكل(٤-٣) مصفوفة التعارض الثنائية (confusion matrix binary) [6]

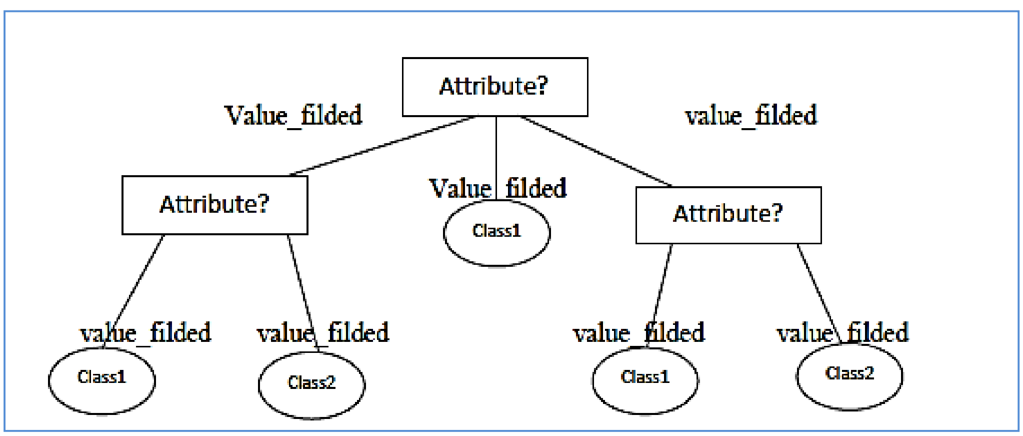
– مدخل fij في مصفوفة التعارض يشير الى عددالسجلات في الـ class i المتوقع أن تكون في الـ class j
مثلا fo1 تشير إلى عدد السجلات في الـ class 0 التي تم توقعها بشكل خطأ في الـ class 1 .
اعتماداً على المصفوفة فإن
مجموع السجلات المتوقعة بشكل صحيح هي (f11 + f00)
مجموع السجلات المتوقعة بشكل خطأ هي (f01+ f10)

٤–٢–٥ التصنيف بأستخدام أشجار القرار

شجرة القرار هى نموذج أستكشافي يظهر على شكل شجرة ، كما يعبر أسمها وبشكل دقيق ، يمثل كل فرع من فروعهاسؤالاًتصنيفياًوتمثل أوراقهاأجزاء من قاعدة البيانات تنتمى للتصنيفات التى تم بناءها. إن العقدة فى الشجرة هى عباره عن فحص لبعض الصفات ، أما الفرع فهو قيمة محتملة للصفة ويكون التصنيف حسب الخطوات التالية

أ- ابدأ من الجذر
ب-أفحص الصفة
ت-تحرك الى أسفل فرع الشجرة .

القاعدة الأساسية فى بناء شجرة القرار هى إيجاد أفضل سؤال عند كل فرع من فروع شجرة القرار بحيث يقسم هذا السؤال البيانات الى قسمين القسم الأول منها ينطبق عليه السؤال والقسم الثانى لاينطبق عليه السؤال ، وهكذا يتم من خلال سلسله من الأسئله بناء شجرة القرار بفروعها المتسلسله . بالرغم من أن شجرة القرار تستخدم فى الإستكشاف وتحضير البيانات للعمليات الأحصائية إلا إنها ايضاً تستخدم وبشكل أكثر للتنبؤ ، ومن المهم جداً عند بناء خوارزمية شجرة القرار أن يؤخذ بعين الإعتبار أن تكون قابله للتطبيق بقدر الأمكان وبشكل مثالى على كل البيانات المتوفرة .



شكل(٤–٤) شجرةالقرار التصنيف [7]

٤-٢-٦ بناء شجرة القرار

يتم بناء الشجرة القرار بمرحلتين

أ- مرحلة بناء الشجرة

يتم تجزأة بيانات التدريب بصورة مكررة الى أن تصبح هذه البيانات فى كل جزء تعود الى صنف واحد ويكون الجزء صغير بدرجة كافيه .

ب-مرحلة تشذيب شجرة القرار

يتم فى هذه المرحلة حذف الاعتمادية على الضوضاء الاحصائية أو التغييرات التى قد تكون مخصصه فقط الى مجموعة التدريب .

٤-٢-٧ مميزات أستخدام التصنيف بتقنية شجرة القرار هى كالاتى

أ- تعتبرخوارزميةشجرةالقرارطريقة سهلة لاتحتاج الى وسطاء (non parametric) لبناء نماذج التصنيف . أى أنها لاتحتاج الى فرضيات سابقة فيما يتعلق بنوع التوزيعات الاحتمالية التى يفى بها الصنف والصفات الأخرى [7] .

ب-ان التقنيات المطورة من أجل بناء شجرة القرار غير مكلفة بشكل كبير،مما يعنى أمكانية بناء نماذج سريعه حتى عندما يكون حجم مجموعة التدريب كبير جداً .

ت-من السهل نسبياً تفسير أشجار القرار خاصة اذا كانت صغيرة الحجم كما يمكن مقارنة دقة الأشجار مع تقنيات أخرى من أجل العديد من مجموعة البيانات .

ث-تعتبر أشجار القرار قوية فى حالة وجود تشويه (Noise) .

هناك عدة تقنيات يوفرها برنامج WEKA للتنقيب في البيانات، وسوف نسرد هنا تلك التقنيات التي تم استخدامها في هذا البحث وطريقة عمل هذه التقنيات ، تم استخدام خوارزمية(k-means) ضمن خوارزميات التجميع وهى ضمن تقنية العنقدة Clustering

٤-٣ تقنية العنقدة clustering

أ- تعد العنقده احدي تقنيات تعدين البيانات وهي عمليه تجزئة البيانات الاصليه الي مجاميع وتجميع البيانات ذات الخصائص المتشابهة ووضعها في عناقيد اعتماداً علي قياس المسافه فيما بينها ،

الغاية من عملية التجميع هذه هو بناء نماذج وصفية لتسهيل دراسة تلك النماذج والخروج بنتائج مهمه ومفيدة لصانعي القرار.

ب–وتتم عملية تقسيم البيانات الي مجاميع اعتماداً علي اشتراكها بالخواص المتشابهة وهي تقسيم غير موجه للبيانات علي عكس التصنيف الذي يستخدم بشكل واسع في حل الكثير من المشكلات من خلال تحليل مجموعة من البيانات ووضعها علي شكل اصناف او اقسام يمكن استخدامها فيما بعد لتصنيف البيانات المستقبلية باستخدام الطرق التي تستخدم في تصنيف البيانات.

ت– والعنقود هو مجموعة من البيانات المتشابهة فيما بينها وغير متشابهة للاعضاء المنتمية الي عناقيد اخرى.

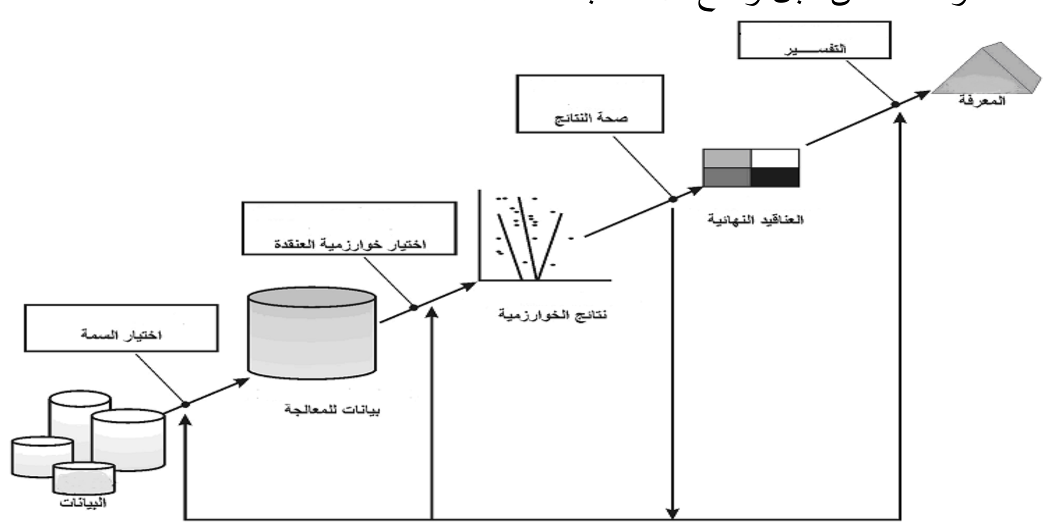
ولكي تنجح عملية تجميع البيانات (عنقدتها) يجب إجراء معالجة مسبقة للبيانات وحسب الخطوات الآتية

أ– **اختيار السمة** إن اختيار السمات (الصفات) التي تمكننا من تحليل رموز معلومات اكتر متعلقه بالمهمه المطلوب حلها .

ب– **اختيار خوارزمية العنقدة** اختيار الخوارزمية المناسبة التي تنتج من تعريف مخطط العنقدة الجيدة لمجموعة البيانات.

ت–**صحة النتائج** تتحقق صحة نتائج خوارزمية العنقدة اعتماداً علي التقنيات والمعايير المناسبة، ولأن خوارزميات العنقدة تعرف العناقيد غير معرفه مسبقاً،وبغض النظر عن الاسلوب لذا فان التجزئه النهائية للبيانات تتطلب بعض أنواع التقييم.

ث–**تفسير النتائج** إن الخبرة في مجال التطبيق تحتاج الي تكامل نتائج العنقدة مع بقية التجارب المبرهنة والمحللة من أجل وضع الاستنتاجات الملائمة.



شكل (٤–٥) يوضح استخدام تقنية العنقدة لاكتشاف المعرفة[5].

٤-٤ انواع تقنيات العنقدة:

تصنف العنقدة بناءً على المعيار أو المقياس الذي يعرِّف مدى التشابه بين عناصر البيانات المدخلة الي الخوارزمية وكذلك النظرية الاساسية التي تعتمد عليها تقنية تحليل العنقدة .

ولقد تطورت خوارزميات العنقدة بناءً على المواضيع المتعلقة بها مثل تعدين البيانات ،والتعليم الآلي ،وتمييز الأنماط ، ...وغيرهاوصُنفت تلك الخوارزميات إلي الاتية

العنقدة الهرمية Hierarchical Clustering

العنقدة التجزيئية Partitioned Clustering

عنقدة البيانات ذات الابعاد الكبيرة Clustering High Dimensional Data

العنقدة مع الصفات الوارثية Clustering With Genetic Attributes

وصنف بعض الباحثين خوارزميات العنقدة طبقاً للطريقة المعطاة في تعريف العناقيد الي الاصناف الاتية :

العنقدة الهرمية Hierarchical Clustering

العنقدة التجزيئية Partitioned Clustering

العنقدة المعتمدة علي الكثافة Density –Based Clusting

العنقدة المعتمدة علي المتعامدة Grid –Based Clustering

ولكل صنف من الاصناف المذكورة أعلاه أنواع متفرعه عنها وخوارزميات مختلفة لأيجاد العناقيد وعليه فيمكن تصنيف خوارزميات العنقدة تبعاً لنوع المتغيرات في النحو الآتي

العنقدة الإحصائية Statistical Clustering: ويعتمد هذا النوع علي مفاهيم التحليل الاحصائي ،فيعتمد علي مقاييس التشابه لتجزئة النماذج التي تكون محددة بيانات عددية.

العنقدة الأدراكية Conceptual Clustering تستخدم لعنقدة البيانات الصنفية، حيث تعنقد النماذج تبعاً للمفاهيم التي تحملها وليس لقيمتها .

العنقدة المضببة Fuzzy Clustering تستخدم تقنيات مضببة لبيانات العنقود ، إذ تفترض انه يمكن تصنيف النماذج إلي أكثر من عنقود واحد .وتستخدم البيانات الحقيقية غير المؤكدة (غير الموثوقة).

العنقدة الهشة Crisp Clustering يفترض هذا النوع تجزئه غير متداخلة ، وهذا يعني ان نقطة البيانات إما أن تنتمي إلي العنقود أو لا تنتمي.

عنقدة شبكة كوهين Kohonen Net Clustering :وتعتمد العنقدة هنا علي مفاهيم الشبكة العصبية ويتشكل العنقود من العنقدة الخارجية.

٤-٥ خوارزمية k-means

تستخدم خوارزميات التجميع علي نطاق واسع ليس فقط لتنظيم وتصنيف البيانات ولكن لضغط البيانات وبناء نموذج ترتيب البيانات .حيث أنه إذا كان بامكاننا أن نجد تجمعات من البيانات ،فأنه بالامكان بناء نموذج للمشكلة علي اساس تلك التجمعات .

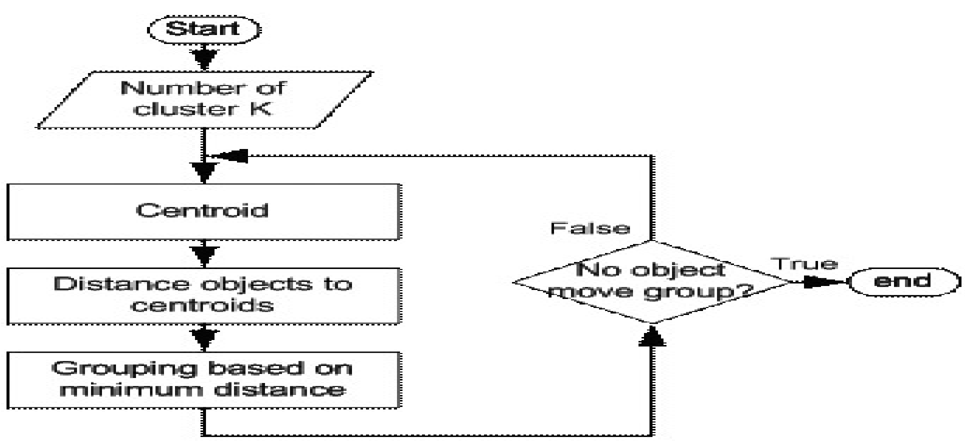
هنالك العديدمن الخوارزميات المستخدمة في عملية تجميع البيانات وسندرس أبسط خوارزمية، ألا وهي خوارزميه k-means clustering استخدمت في هذاالبحث إحدي طرق العنقدة المتمثلة بخوارزمية الـk-meansمع المنطق في تصنيف البيانات واستخدمت هذه الخوارزمية في تكوين عناقيد للبيانات الموجودة التي تختلف في مكوناتها لاكتشاف بيانات ومكونات أخري لم تظهر في البيانات الأصلية قبل تطبيق الخوارزمية.

يوجد العديد من التقنيات التي يتم من خلالها التعامل مع البيانات وأحد هذه التقنيات هي استخدام خوارزمية K-means.

اجريت العديد من الأبحاث في هذا المجال حيث استخدمت خوارزميه k-means في العنقدة عن طريق تحليل المكونات الرئيسية للبيانات الإحصائية .

تستخدم هذه الخوارزمية لتجميع عدة بيانات (أمثلة) اعتماداً علي خصائصهاالي(k)تجمع ، وتتم عملية التجميع من خلال تقليل المسافات بين مركز التجمع (centroid cluster).

يعتمد أداء هذه الخوارزمية علي المواقع الأولية لمراكز التجمعات (centroid)، ومن المستحسن تنفيذ هذه الخوارزمية عدة مرات مع اختلاف المراكز في كل مرة عن المرات السابقة .



شكل(٤-٦) خوارزمية k-means clustering[5] .

٤-٦ تطبيقات خوارزمية k-means clustering

تعتبر خوارزمية k-means clustering من أبسط خوارزميات التجميع وأكثرها فعالية ، وهنالك الكثير من تطبيقاتها أهمها

أ- التعليم بدون إشراف unsupervised learning للشبكات العصبية .

ب-تمييز النماذج pattern recognitions .

ت-التصنيف classification.

ث-معالجة الصورة image processing.

ج- الرؤية بواسطة الآلة machine vision .

ح- الذكاء الصناعي Artificial intelligent.

٤-٧ مزايا خوارزمية k-means clustering

أ- ذات فعالية عالية

ب-سهلة التنفيذ

ت-تتعامل مع القيم المستمرة والقيم المتقطعه (الاسمية)

مساوئ خوارزمية k-means clustering

أ- تحديد عدد التجمعات k يتم عشوائياًوقبل معالجة الامثلة .

ب-حساسة للحالة الأولية . يؤدي اختيار حالات أولية متعددة إلي إعطاء نتائج مختلفة للتجمعات .ونتيجة لذلك قد تقع الخوارزمية في مشكلة النهاية المحلية.

ت-شكل التجمع دائري لأنه يعتمد علي حساب المسافه

٤-٨ انواع تقنيات العنقدة

توجد العديد من تقنيات العنقدة منها كما يلى [5]

أ- خوارزمية K-means and its variants

ب-خوارزمية Hierarchical clustering

ت-خوارزمية Density-based clustering

الطرق الهرمية ينتج عن هذه الطريقة تحليل هرمي لمجموعة البيانات المعطاة وطريقة التحليل الهرمي أما أن تكون كتلية او توزيعية اعتماداً على الصيغة المطلوبة في التحليل الهرمي فالطريقة الكتلية والتي تسمى (bottom-up)تبدأمن العناصر الموجودة وتشكل مجموعات منفصلة ومن ثم تقوم تباعاً بدمج العناصر

القريبة من بعضها إلى أن يتم دمج جميع المجموعات في مجموعة واحدة أما الطريقة التوزيعية و التي تدعى (Top–down)التي تبدأ بجمع الكائنات في عنقود واحد وعند البدء بعمليات التكرار ينقسم العنقود إلى عناقيد أصغر منه حتى يصبح كل كائن من الكائنات في عنقود وحيد.

الطرق المعتمدة على الكثافة معظم أساليب العنقدة تعتمد على أساس المسافة بين الكائنات. يمكن لمثل هذه الأساليب أن توجد التجمعات الكروية الشكل وتواجه صعوبة في اكتشاف المجموعات من الأشكال العشوائية وقد تم تطوير أساليب تجميع أخرى على أساس مفهوم الكثافة الفكرة العامة منها هو النمو المتواصل للعنقود يمكن استخدامها لتصفية الضجيج (القيم الشاذة) واكتشاف عناقيد من الشكل اللعشوائي.

الطرق المعتمدة على النماذج تفترض هذه الطريقة نموذجا لكل عنقود من العناقيد والعثور على أفضل البيانات المناسبة لكل نموذج مفترض. خوارزميات هذا النوع تحدد العناقيد عن طريق إنشاء دالة الكثافة التي تعكس التوزيع المكاني لنقاط البيانات و أيضاً تعطي طريقة لتحديد عدد العناقيد تلقائياً استناداً إلى الإحصاءات القياسية وتأخذ بعين الاعتبار الضجيج والقيم الشاذة وبالتالي تعتبر من أساليب التجميع القوية.

٤–٩وصف لتجهيز وتهيئة البيانات علي برنامج WEKA

تم تنفيذالدراسة في هذا البحث اعتماداً على مراحل منهجية البحث المذكورة في الفصل الاول –(المبحث الأول) – من هذا البحث .

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 تصديرالبيانات إليه بسهولة. يتعامل البرنامج مع ثلاث صيغ من البيانات وهي إما بيانات من نوع اكسل أو بيانات على شكل ملفات نصية أو ODBC .

٤–١٠ خطوات معالجة البيانات

بالتعامل مع سجل الطالب الأكاديمي [3] نجد أنه تتوفر الكثير من المعلومات الأكاديمية المتعلقة بالطالب وهي غير مستغلة ، وفي هذا البحث تم تحديد بعض هذه البيانات والتعامل معها وقد تمت معالجتها وفقا لخطوات تنقيب البيانات وهي كالتالي

٤–١٠–١ أولاً إختيار البيانات

تم الحصول على البيانات من قبل وزارة التربية والتعليم الاتحادية والتي تخص نتيجة شهادة الثانوي – السياق العلمي ، وتتمثل هذه البيانات في نتائج طلبة الصف الثالث الثانوي وهى تحتوى على بيانات عدد ١٩٤٠٠٠

طالب وهى عبارة عن درجات الطلبة فى تسعة مواد مقررة فى شهادة التعليم الثانوية منها (٦) مواد اجبارية للجانب العلمي وهي :- التربية الاسلامية واللغة العربية والرياضيات واللغة الانجليزية والفيزياء والكيمياء و(٣) مواد اختيارية تختار منها مادة واحدة فقط اما الاحياء او الكمبيوتر او الهندسية.

٤-١٠-٢ ثانياً تهيئة البيانات

بعد إكتمال جمع البيانات من مصادرها وهو وزارة التربية والتعليم الاتحادية واستبعاد البيانات الغير مكتملة من السنوات دون العام ٢٠٠٧ م لأن هذه البيانات لابد ان تكون مكتملة لأنها تحتوي على درجات طلاب وهذه الدرجات لايمكن تعويضها بشئ آخر فلابد ان تكون حقيقية. تمت تهيئة الشكل العام للملف بحيث يناسب أداة تنقيب البيانات المستخدمة وهي (WEKA) وهي تقبل البيانات في صورة ملف (ARFF.) أو قاعدة بيانات.

٤-١٠-٣ ثالثا : تنقيب البيانات

في هذه المرحلة تم إستخدام تقنية العناقيد Clusteringوتطبيق خوارزمية(k-means) وتقنية التصنيف بأستخدام خوارزمية شجرة القرار (J48) بإستخدام برنامج (Weka).

٤-١١ الأدوات و البرامج المستخدمة فى التنقيب

هناك العديد من المنتجات والادوات والبرامج التى تقوم بعملية تنقيب البيانات والتى تنتج عن [3]شركات كبيرة ومعروفة مثل (Microsoft, SPSS , IBM) وغيرها من الشركات المعروفة ومن هذه البرامج التى تقوم بعملية التنقيب والتى تم تطبيقها واستخدام بعض التقنيات والخوارزميات الخاصة بها فى هذا البحث برنامج (WEKA) .

٤-١٢ WEKA

برنامج ويكا من البرامج الرائدة فى تنقيب البيانات وتم إنتاج هذا البرنامج فى جامعة ويكاتو WAIKATO فى هاملتون عاصمة نيوزيلاندا وهواختصار لـ **W**aikato **E**nvironment for **K**nowledge **A**nalysis
يمدنا برنامج WEKA بمجموعة من خوارزميات تعلم الآلة والتي تستخدم لإجراء مهام تنقيب البيانات.هذه الخوارزميات يمكن أن تطبق مباشرة على مجموعة بيانات أو يتم استدعاؤها من التعليمات البرمجية الخاصة بالجافا، ويحتوي أيضاًعلى أدوات تجهيز البيانات بالطرق المختلفة وكذلك أدوات تحويل قواعد البيانات مثل الخوارزميات.

طاولةالعمل تحتوي على الأدوات المستخدمة في المشاكل الرئيسية في تنقييب البيانات مثل الإنحدار والتصنيف والتجميع ، جميع الخوارزميات تأخذ مدخلاتها في شكل جدول أحادي العلاقة في صورة (ARFF.)

٤-١٢-١ صيغة ARFF (Attribute Relation File Format)

الصفوف التي تبدأ بعلامة (%) هي تعليقات ، يلي التعليقات في بداية الملف هو إسم العلاقة ، [3] الخواص الإسمية يجب ان يتبعها مجموعة من القيم التي يمكن ان تأخذها وتكون هذه القيم متضمنة في الأقواس المعرجة .القيم يمكن أن تحتوي علي فراغات في هذه الحالة لابد من ان تكون مضمنة بين علامات تقدير، الخواص الرقمية لابد ان تليها الكلمة المحجوزة (Numeric) كما هو موضح بالشكل التالي الذي يوضح مثال لملف (ARFF.) للعلاقة: (project)

```
@relation project
@attribute Sn numeric
@attribute sex numeric
@attribute Arc numeric
@attribute ENC numeric
@attribute physics numeric
```

أحد الطرق لإستخدام (Weka) هي تطبيق طريقة التعلم لمجموعة البيانات وتحميل مخرجاتها لمعرفة المزيد عن البيانات، وأيضاً يمكن إستخدام النماذج التي تم تعليمها مسبقا لتوليد التوقعات على حالات جديدة، وكذلك يمكن إستخدام مجموعة مختلفة من النماذج التي تم تعليمها مسبقاً ومقارنة اداؤها من أجل إختيار إحداها للتنبؤ.

في واجهة (Weka) التفاعلية يمكن تحديد طريقة التعلم المراد إستخدامها من القائمة ، والكثير من الطرق لديها مجموعة من المعلمات الإنضباطية التي يمكنك الوصول اليها من خلال ورقة الملكية أو من خلال محرر كائن ، ويتم إستخدام وحدة التقييم المشتركة لقياس أداء جميع المصنفات. تطبيقات نظم التعلم الفعلي هي من المصادر الأكثر قيمة التي توفرها(Weka) ، ولكن الأدوات اللازمة لتجهيز البيانات التي تسمى المرشحات تأتي في المرتبة الثانية مثل المصنفات يمكنك تحديد المرشحات من القائمة وتكييفها مع متطلباتك.

٤-١٢-٢ طريقة إستخدام (Weka)

أسهل الطرق لإستخدام (Weka) هو من خلال إستخدام [2]

أ- واجهة المستخدم الرسومية وتسمى المستكشف (Explorer)

وهذا الأخير يتيح الوصول الى جميع منشآت (Weka) بإستخدام إختيار القائمة وملأ الإستمارة . مثلا يمكن بسرعة كبيرة قراءة قاعدة البيانات من ملف (ARFF.) وتكوين شجرة القرارات منها . لكن شجرة القرارات

تعتبر فقط البداية لأنه يوجد هنالك الكثير من الخوارزميات لابد من إكتشافها ، واجهة المستكشف تساعدك على فعل هذا ، يمكنها توجيهك بعرض الخيارات في شكل قوائم.

هنالك نوعين آخرين من واجهات المستخدم (Weka)

ب- **واجهة المجرب (Experimental)**

وهي مصممة للإجابة على الأسئلة الأساسية للعملية عند الطلب على تقنيات الإرتداد والتصنيف.

ت- **واجهة تدفق المعلومات (Knowledge Flow)**

وهي تسمح بتصميم الترتيبات لمعالجة البيانات المجرية (Streamed data)

ث- **واجهة Simple CLI**

بالإضافة الى الواجهات التفاعلية توجد الوظيفة الأساسية لـ (Weka) وهي التي يمكن الوصول اليها في شكل صف بإدخال الأوامر النصية ، وهي بدورها تسمح بالوصول لكل المميزات بالنظام . عندما يعمل (Weka) لابد من الإختيار من بين هذه الأربع واجهات .

الفصل الخامس

التنفيذ وتحليل النتائج

٥-١ المقدمة

بعد استقبال برنامج WEKA للبيانات بالصيغة المطلوبة كما تم شرحه ، تصبح البيانات جاهزة لاختيار الـAttributes 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 الـAttribute المختار .
وبعد أن يتم اختيار الحقول المناسبة يتم اختيار التقنية المناسبة من تطبيق WEKA لاختيار الخوارزميات المناسبة (فى هذا البحث قمت بأختيار خوارزمية k-means وخوارزمية شجرة القرار J48) .

٥-٢ تهيئة بيانات الطلاب (Preprocessing)

إن قاعدة البيانات التي طبقت عليها تقنيات التنقيب في البيانات هي لدرجات الطلبة فى الشهادة الثانوية – المساق العلمي، وقد شملت بيانات الطلبة والطالبات لجميع انماط التعليم : النظامى ، اتحاد المعلمين ، المنازل و تعليم الكبار ، بالاضافة لجميع المواد المقرره بالصف الثالث الثانوي مع استبعاد مادة المسيحية لقلة العدد وضعف أثرها على نتائج الطلبة .

تم استبعاد بعض الحقول الغير مهمة فى الدراسة مثل اسم المدرسة واسم الطالب ونمط التعليم(نظامى ، اتحاد معلمين ، منازل ، كبار) والنوع (طالب أو طالبه) لعدم اهميتها فى البحث ، تم التركيز على درجات الطلبة خلال هذه السنوات الخاضعه للدراسة , تم تحويل الدرجات التى احرزها الطلاب الى قيم (اسمية) بالنسبة للمواد والمجموع الكلى للطالب الى تقدير يتراوح بين (الامتياز ، جيد جدا ، جيد ، مقبول ، ورسوب) بالأضافة لتحديد الطلاب الغياب فى المادة أو فى المجموع ، وذلك لاهمية القيم الاسمية (الوصفية) فى الدراسة.

المادة	ممتاز	جيد جداً	جيد	مقبول	الرسوب	الدرجة الكاملة
التربية الاسلامية	>= ٨٠ درجة	>= ٧٠درجة	>= ٦٠درجة	>= ٥٠درجة	< ٥٠ درجة	١٠٠ درجة
اللغة العربية	>= ٨٠ درجة	>= ٧٠درجة	>= ٦٠درجة	>= ٥٠درجة	< ٥٠ درجة	١٠٠ درجة
اللغة الانجليزية	>= ٨٠ درجة	>= ٧٠درجة	>= ٦٠درجة	>= ٥٠درجة	< ٥٠ درجة	١٠٠ درجة
الرياضيات	>= ٨٠ درجة	>= ٧٠درجة	>= ٦٠درجة	>= ٥٠درجة	< ٥٠ درجة	١٠٠ درجة
الفيزياء	>= ٨٠ درجة	>= ٧٠درجة	>= ٦٠درجة	>= ٥٠درجة	< ٥٠ درجة	١٠٠ درجة
الكيمياء	>= ٨٠ درجة	>= ٧٠درجة	>= ٦٠درجة	>= ٥٠درجة	< ٥٠ درجة	١٠٠ درجة

الأحياء	> =٨٠ درجة	>= ٧٠درجة	>= ٦٠درجة	>= ٥٠درجة	< ٥٠ درجة	١٠٠ درجة
الكمبيوتر	> =٨٠ درجة	>= ٧٠درجة	>= ٦٠درجة	>= ٥٠درجة	< ٥٠ درجة	١٠٠ درجة
الهندسية	> =٨٠ درجة	>= ٧٠درجة	>= ٦٠درجة	>= ٥٠درجة	< ٥٠ درجة	١٠٠ درجة
المجموع الكلى	> =٥٦٠درجة	>=٤٩٠درجة	>=٤٢٠درجة	>=٣٥٠درجة	<=٣٤٩درجة	٧٠٠ درجة

جدول(٥-١)تقدير المواد والمجموع والدرجة الكاملة لكل مادة وللمجموع(تصميم الباحث)

أخضعت الدراسة الطلبة الذين تقدموا للشهادة الثانوية من العام ٢٠٠٧ إلى العام ٢٠١٦ م .

تم تحويل بيانات من MS-Access وهو النظام المعمول به بإدارة التقويم التربوى والأمتحانات الى MS-Excel لتتناسب مع الصيغ المقبوله لبرنامج (WEKA).

غالبا ما يتم عرض البيانات في صورة (Excel Sheet) أو قاعدة بيانات ، على أي حال الطريقة الأصلية للتخزين في (Weka) هي في صورة (ARFF.) ويمكن بكل سهولة التحويل من صورة برنامج الجدولة الى صورة . (ARFF.) معظم ملفات (ARFF.) تتكون من قائمة من الحالات وقيم خواص لكل من الحالات التي تم الفصل بينها بالفاصلة .

معظم برامج الجدولة وبرامج قواعد البيانات تسمح بتصدير البيانات في ملف يأخذ هذه الصيغة (Comma Separated Value .CSV) في شكل مجموعة من السجلات مع وجود الفاصلة بين القيم . شكل البيانات عند تصدير من Access الى Excel كما هو موضح في الشكل رقم (٥-١) تحتوي قاعدة البيانات التي يدرسها هذا البحث على أكثر من ١٩٤,٠٠٠ سجل .

Year	Gender	Classifier	Islamic	I.Rating	Arabic	A.Rating	English	E.Rating	Math.	M.Rating	Physics	P.Rating	Chemistry	CH.Rating	Biology	B.Rating	Computer	C.Rating	Engineering	G.Rating	Sum.	Gen.Rat.	Evaluation
2007	Male	Regular	57	Accepted	56	Accepted	47	Fail	38	Fail	15	Fail	43	Fail	0	Absent	60	Good	0	Absent	316	Fail	Fail
2007	Male	Regular	73	V.Good	56	Accepted	57	Accepted	38	Fail	25	Fail	57	Accepted	0	Absent	60	Good	0	Absent	366	Pass	Accepted
2007	Male	Regular	77	V.Good	58	Accepted	60	Good	40	Fail	38	Fail	57	Accepted	0	Absent	70	V.Good	0	Absent	400	Pass	Accepted
2007	Male	Regular	70	V.Good	30	Fail	37	Fail	33	Fail	20	Fail	37	Fail	0	Absent	60	Good	0	Absent	287	Fail	Fail
2007	Male	Regular	73	V.Good	54	Accepted	63	Good	38	Fail	15	Fail	43	Fail	0	Absent	63	Good	0	Absent	349	Fail	Fail
2007	Male	Regular	73	V.Good	46	Fail	53	Accepted	40	Fail	23	Fail	63	Good	0	Absent	60	Good	0	Absent	358	Pass	Accepted
2007	Male	Regular	87	Excellent	46	Fail	53	Accepted	23	Fail	30	Fail	57	Accepted	0	Absent	60	Good	0	Absent	356	Pass	Accepted
2007	Male	Regular	90	Excellent	62	Good	73	V.Good	55	Accepted	20	Fail	73	V.Good	0	Absent	80	Excellent	0	Absent	453	Pass	Good
2007	Male	Regular	77	V.Good	72	V.Good	57	Accepted	48	Fail	30	Fail	70	V.Good	0	Absent	73	V.Good	0	Absent	427	Pass	Good
2007	Male	Regular	70	V.Good	16	Fail	23	Fail	30	Fail	28	Fail	27	Fail	0	Absent	33	Fail	0	Absent	227	Fail	Fail
2007	Male	Regular	67	Good	56	Accepted	47	Fail	40	Fail	18	Fail	67	Good	0	Absent	57	Accepted	0	Absent	352	Pass	Accepted
2007	Male	Regular	7	Fail	26	Fail	27	Fail	23	Fail	15	Fail	37	Fail	0	Absent	40	Fail	0	Absent	175	Fail	Fail
2007	Male	Regular	80	Excellent	50	Accepted	60	Good	45	Fail	13	Fail	60	Good	0	Absent	50	Accepted	0	Absent	358	Pass	Accepted
2007	Male	Regular	80	Excellent	54	Accepted	80	Excellent	53	Accepted	30	Fail	70	V.Good	0	Absent	67	Good	0	Absent	434	Pass	Good
2007	Male	Regular	77	V.Good	56	Accepted	50	Accepted	45	Fail	10	Fail	57	Accepted	0	Absent	70	V.Good	0	Absent	365	Pass	Accepted
2007	Male	Regular	50	Accepted	32	Fail	53	Accepted	50	Accepted	15	Fail	53	Accepted	0	Absent	77	V.Good	0	Absent	330	Fail	Fail
2007	Male	Regular	73	V.Good	56	Accepted	73	V.Good	53	Accepted	28	Fail	60	Good	0	Absent	73	V.Good	0	Absent	416	Pass	Accepted
2007	Male	Regular	80	Excellent	48	Fail	63	Good	53	Accepted	28	Fail	67	Good	0	Absent	73	V.Good	0	Absent	412	Pass	Accepted
2007	Male	Regular	57	Accepted	48	Fail	53	Accepted	43	Fail	33	Fail	57	Accepted	0	Absent	77	V.Good	0	Absent	368	Pass	Accepted
2007	Male	Regular	83	Excellent	62	Good	70	V.Good	58	Accepted	25	Fail	53	Accepted	0	Absent	73	V.Good	0	Absent	424	Pass	Good
2007	Male	Regular	87	Excellent	60	Good	70	V.Good	48	Fail	23	Fail	63	Good	0	Absent	70	V.Good	0	Absent	421	Pass	Good
2007	Male	Regular	90	Excellent	68	Good	70	V.Good	48	Fail	23	Fail	67	Good	0	Absent	77	V.Good	0	Absent	443	Pass	Good
2007	Male	Regular	80	Excellent	66	Good	73	V.Good	38	Fail	5	Fail	60	Good	0	Absent	70	V.Good	0	Absent	392	Pass	Accepted
2007	Male	Regular	73	V.Good	34	Fail	40	Fail	48	Fail	33	Fail	53	Accepted	0	Absent	57	Accepted	0	Absent	338	Fail	Fail
2007	Male	Regular	67	Good	50	Accepted	53	Accepted	23	Fail	38	Fail	43	Fail	0	Absent	53	Accepted	0	Absent	327	Fail	Fail
2007	Male	Regular	80	Excellent	50	Accepted	57	Accepted	33	Fail	23	Fail	50	Accepted	0	Absent	60	Good	0	Absent	353	Pass	Accepted
2007	Male	Regular	60	Good	36	Fail	53	Accepted	43	Fail	33	Fail	60	Good	0	Absent	63	Good	0	Absent	348	Fail	Fail
2007	Male	Regular	37	Fail	30	Fail	40	Fail	28	Fail	23	Fail	57	Accepted	0	Absent	57	Accepted	0	Absent	272	Fail	Fail
2007	Male	Regular	83	Excellent	54	Accepted	53	Accepted	28	Fail	20	Fail	63	Good	0	Absent	77	V.Good	0	Absent	378	Pass	Accepted

شكل(٥-١)بيانات الطلاب بعد التعديل والتهيئة فى الإكسل(تصميم الباحث)

Year;Gender;Classifier;Islamic;I.Rating;Arabic;A.Rating;English;E.Rating;Math.;M.Rating;Physics;P.Rating;Chemistry;CH.Rating; Biology;B.Rating;Computer;C.Rating;Engineering;G.Rating;Sum.;Gen.Rat.;Evaluation
2007;1;1;57;Accepted;56;Accepted;47;Fail;38;Fail;15;Fail;43;Fail;0;Absent;60;Good;0;Absent;316;Fail;Fail
2007;1;1;73;V.Good;56;Accepted;57;Accepted;38;Fail;25;Fail;57;Accepted;0;Absent;60;Good;0;Absent;366;Pass;Accepted
2007;1;1;77;V.Good;58;Accepted;60;Good;40;Fail;38;Fail;57;Accepted;0;Absent;70;V.Good;0;Absent;400;Pass;Accepted
2007;1;1;70;V.Good;30;Fail;37;Fail;33;Fail;20;Fail;37;Fail;0;Absent;60;Good;0;Absent;287;Fail;Fail
2007;1;1;73;V.Good;54;Accepted;63;Good;38;Fail;15;Fail;43;Fail;0;Absent;63;Good;0;Absent;349;Fail;Fail
2007;1;1;73;V.Good;46;Fail;53;Accepted;40;Fail;23;Fail;63;Good;0;Absent;60;Good;0;Absent;358;Pass;Accepted
2007;1;1;87;Excellent;46;Fail;53;Accepted;23;Fail;30;Fail;57;Accepted;0;Absent;60;Good;0;Absent;356;Pass;Accepted
2007;1;1;90;Excellent;62;Good;73;V.Good;55;Accepted;20;Fail;73;V.Good;0;Absent;80;Excellent;0;Absent;453;Pass;Good
2007;1;1;77;V.Good;72;V.Good;57;Accepted;48;Fail;30;Fail;70;V.Good;0;Absent;73;V.Good;0;Absent;427;Pass;Good
2007;1;1;70;V.Good;16;Fail;23;Fail;30;Fail;28;Fail;27;Fail;0;Absent;33;Fail;0;Absent;227;Fail;Fail
2007;1;1;67;Good;56;Accepted;47;Fail;40;Fail;18;Fail;67;Good;0;Absent;57;Accepted;0;Absent;352;Pass;Accepted
2007;1;1;7;Fail;26;Fail;27;Fail;23;Fail;15;Fail;37;Fail;0;Absent;40;Fail;0;Absent;175;Fail;Fail
2007;1;1;80;Excellent;50;Accepted;60;Good;45;Fail;13;Fail;60;Good;0;Absent;50;Accepted;0;Absent;358;Pass;Accepted
2007;1;1;80;Excellent;54;Accepted;80;Excellent;53;Accepted;30;Fail;70;V.Good;0;Absent;67;Good;0;Absent;434;Pass;Good
2007;1;1;77;V.Good;56;Accepted;50;Accepted;45;Fail;10;Fail;57;Accepted;0;Absent;70;V.Good;0;Absent;365;Pass;Accepted
2007;1;1;50;Accepted;32;Fail;53;Accepted;50;Accepted;15;Fail;53;Accepted;0;Absent;77;V.Good;0;Absent;330;Fail;Fail
2007;1;1;73;V.Good;56;Accepted;73;V.Good;53;Accepted;28;Fail;60;Good;0;Absent;73;V.Good;0;Absent;416;Pass;Accepted
2007;1;1;80;Excellent;48;Fail;63;Good;53;Accepted;28;Fail;67;Good;0;Absent;73;V.Good;0;Absent;412;Pass;Accepted
2007;1;1;57;Accepted;48;Fail;53;Accepted;43;Fail;33;Fail;57;Accepted;0;Absent;77;V.Good;0;Absent;368;Pass;Accepted
2007;1;1;83;Excellent;62;Good;70;V.Good;58;Accepted;25;Fail;53;Accepted;0;Absent;73;V.Good;0;Absent;424;Pass;Good
2007;1;1;87;Excellent;60;Good;70;V.Good;48;Fail;23;Fail;63;Good;0;Absent;70;V.Good;0;Absent;421;Pass;Good
2007;1;1;90;Excellent;68;Good;70;V.Good;48;Fail;23;Fail;67;Good;0;Absent;77;V.Good;0;Absent;443;Pass;Good
2007;1;1;80;Excellent;66;Good;73;V.Good;38;Fail;5;Fail;60;Good;0;Absent;70;V.Good;0;Absent;392;Pass;Accepted
2007;1;1;73;V.Good;34;Fail;40;Fail;48;Fail;33;Fail;53;Accepted;0;Absent;57;Accepted;0;Absent;338;Fail;Fail
2007;1;1;67;Good;50;Accepted;53;Accepted;23;Fail;38;Fail;43;Fail;0;Absent;53;Accepted;0;Absent;327;Fail;Fail
2007;1;1;80;Excellent;50;Accepted;57;Accepted;33;Fail;23;Fail;50;Accepted;0;Absent;60;Good;0;Absent;353;Pass;Accepted
2007;1;1;60;Good;36;Fail;53;Accepted;43;Fail;33;Fail;60;Good;0;Absent;63;Good;0;Absent;348;Fail;Fail
2007;1;1;37;Fail;30;Fail;40;Fail;28;Fail;23;Fail;57;Accepted;0;Absent;57;Accepted;0;Absent;272;Fail;Fail
2007;1;1;83;Excellent;54;Accepted;53;Accepted;28;Fail;20;Fail;63;Good;0;Absent;77;V.Good;0;Absent;378;Pass;Accepted

شكل(٥-٢) البيانات بعد تصديرها في ملف(CSV.)(Comma Separated Value)(تصميم الباحث)

خطوات تحويل الملف من MS-Excel الى (CSV.) (Comma Separated Value)

أ- افتح ملف الاكسيل وهو هنا باسم :Book1.xlsx .

ب-اضغط على حفظ باسم

ت–اختار له اسما وهو هنا Book1 ثم من القائمة المنسدلة نغير امتداد الملف من من xlsx. الي csv. وبذلك يكون قد تم تحويل الملف الي امتدادcsv.

ث–اضغط على حفظ .

٥–٣ وصف بيانات الطلاب (meta data)

تتكون البيانات من أثنى عشر(١٢) حقل سنوات الدراسة ، و درجات الطلاب فى المواد والمجموع الكلى للمواد ، ووصفها كالأتى بعد تحويل القيم الرقمية الى قيم وصفية (nominal) أسم الملف (Relation : Book1)

Year : numeric

I.Rating : nominal

A.Rating : nominal

E.Rating: nominal

M.Rating: nominal

P.Rating: nominal

CH.Rating: nominal

B.Rating: nominal

C.Rating: nominal

G.Rating: nominal

Gen.Rat.: nominal

Evaluation: nominal

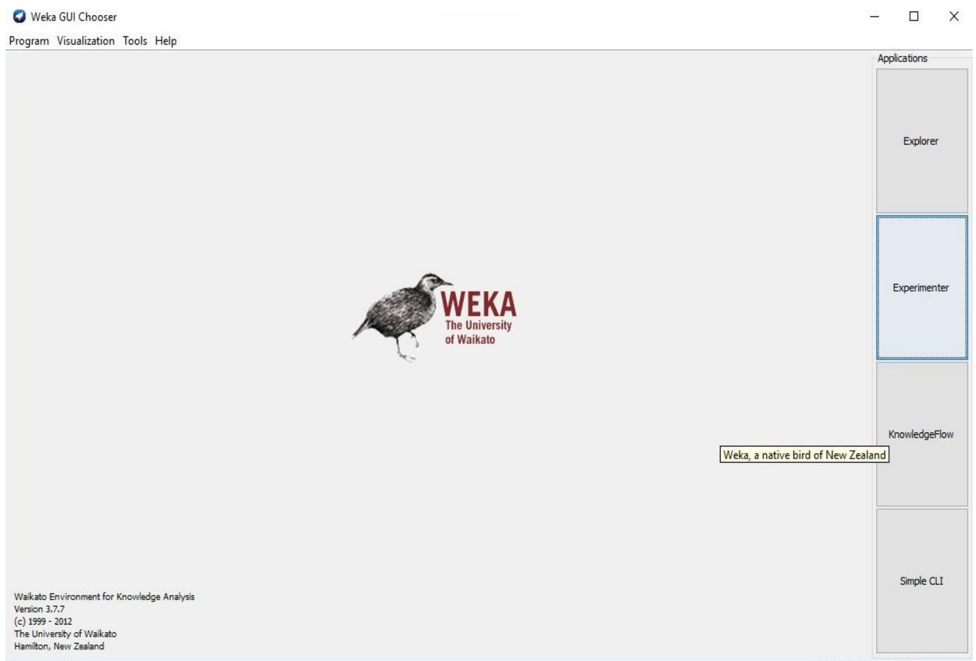
٥– ٤ الخوارزميات المستخدمة فى البحث

فى هذا البحث قمت بأختيار خوارزمية شجرة القرار J48 فى تقنية التصنيف (Classification)

وخوارزمية k–means في تقنية العناقيدClustering وشجرة القرار تعمل على تصنيف درجات الطلاب والتنبؤ بالرسوب والنجاح فى المجموع وبالاضافة الى أن الخوارزميتين تعملان على استكشاف المعرفة الكامنة في قاعدة البيانات .

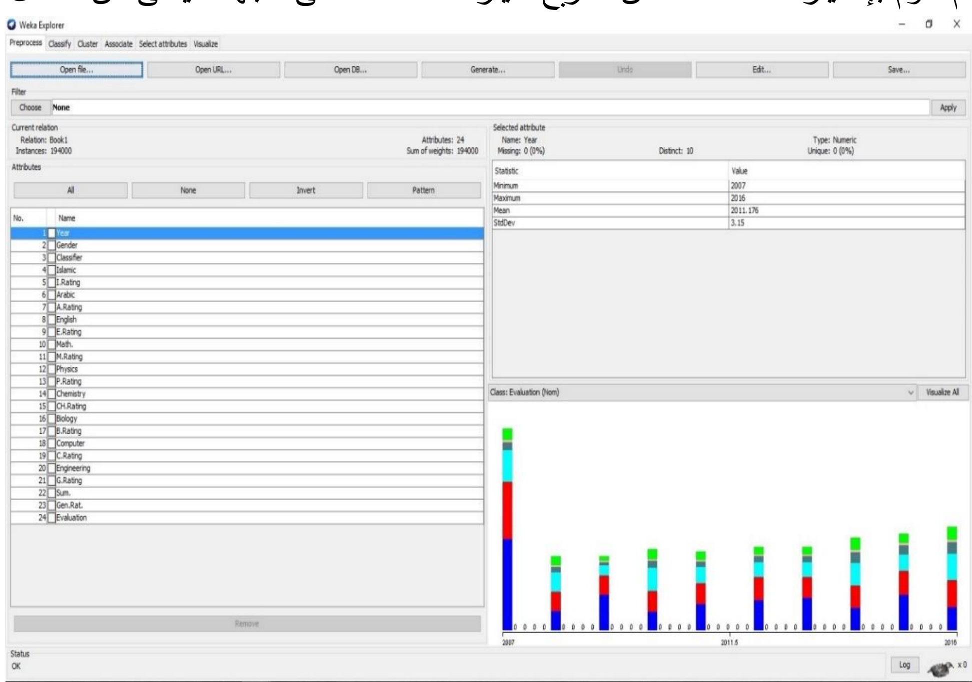
٥-٥ تحميل البيانات في المستكشف

لنبدأ تحميل البيانات في المستكشف والبداية في تحليلها ، في البدء نقوم بتشغيل برنامج (Weka) للحصول على واجهة الإختيار الرسومية كما هو موضح بالشكل التالي



شكل(٥-٣)واجهة الاختيار الرسومية

ثم نقوم بإختيار المستكشف من الأربع خيارات المتاحة على الجهة اليمنى من الشكل.

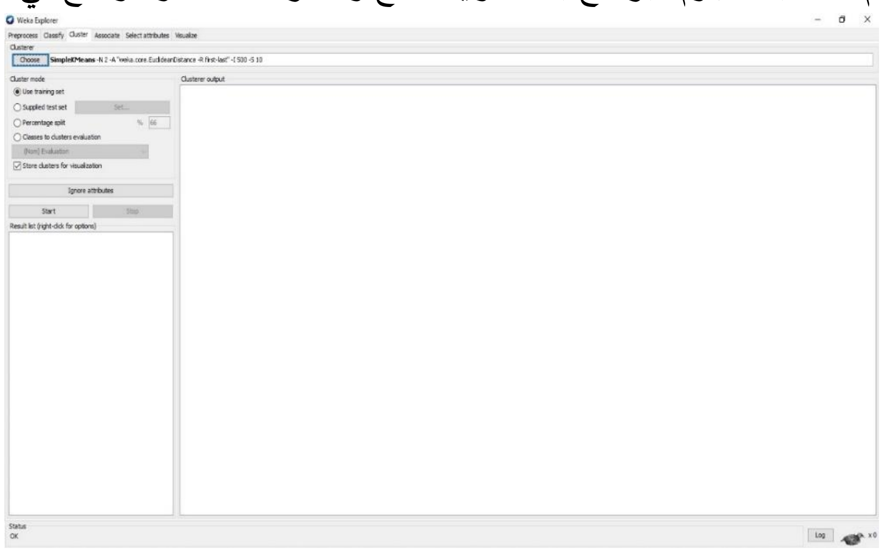


شكل(٥-٤)واجهة المستكشف في weka

في الواقع الشكل (٥-٤) يوضح ما يتم عرضه بعد القيام بتحميل البيانات . الست قوائم الأولى هي

شكل(٥-٥)ستة قوائم في واجهة المستكشف فيweka

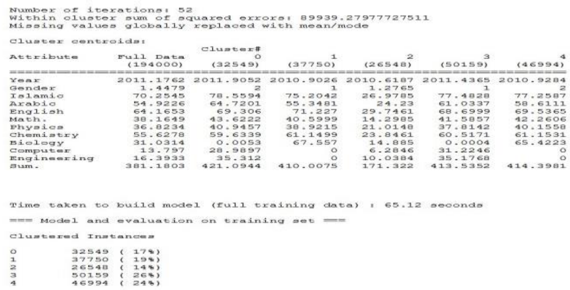
عبارة عن العمليات الأساسية التي يدعمها المستكشف ، حاليا نحن في مرحلة المعالجة المبدئية ، اضغط زر"افتح ملف" الذي من خلاله يمكن ان تختار الملف الذي تريد تحميله . إختار الملف ذو الامتداد (CSV.)بعد ان يتم تحميل البيانات نقوم بتحديد الصفات (attributes) التي نريد إجراء عملية التجميع عليها ويوجد أيضا خيارإختيارها جميعا واجراء التجميع عليها،ومن ثم بالرجوع للست قوائم الأولى نقوم بإختيارclusterثمstart،بعدهايقوم البرنامج بعمل توليدللنتائج والمخرجات كماهو موضح في الشكل(٥-٦)



شكل(٥-٦)نتائج العنقدة (cluster)

٥-٦ تنفيذ الخوارزمية k-means

خوارزميةk-means clustering تستخدم هذه الخوارزمية لتجميع عدة بيانات (أمثلة) اعتماداً علي خصائصها الي k تجمع،وتتم عمليه التجميع من خلال تقليل المسافات بين البيانات ومركز التجمع (centroid cluster) وهي ايضاً واحدة من خوارزميات العنقدة التي تعمل علي تنقيب البيانات.



شكل (٥-٧)نتائج ٥عناقيد (5clusters)بعد استخدام ١٢عمود

kMeans

Number of iterations: 52

Within cluster sum of squared errors: 89939.27977727511

Missing values globally replaced with mean/mode

Cluster centroids: Cluster#

Attribute	Full Data	0	1	2	3	4
	(194000)	(32549)	(37750)	(26548)	(50159)	(46994)
Year	2011.1762	2011.9052	2010.9026	2010.6187	2011.4365	2010.9284
Gender	1.4479	2	1	1.2765	1	2
Islamic	70.2545	78.5594	75.2042	26.9785	77.4828	77.2587
Arabic	54.9226	64.7201	55.3481	24.23	61.0337	58.6111
English	64.1653	69.306	71.227	29.7461	68.6999	69.5365
Math.	38.1649	43.6222	40.5999	14.2985	41.5857	42.2606
Physics	36.8234	40.9457	38.9215	21.0148	37.8142	40.1558
Chemistry	55.6278	59.6339	61.1499	23.8461	60.5171	61.1531
Biology	31.0314	0.0053	67.557	14.885	0.0004	65.4223
Computer	13.797	28.9897	0	6.2846	31.2246	0
Engineering	16.3933	35.312	0	10.0384	35.1768	0
Sum.	381.1803	421.0944	410.0075	171.322	413.5352	414.3981

Time taken to build model (full training data) : 65.12 seconds

=== Model and evaluation on training set ===

Clustered Instances

0	32549	(17%)
1	37750	(19%)
2	26548	(14%)
3	50159	(26%)
4	46994	(24%)

جدول(٥-٢)نتائج٥عناقيد(5clusters)بعد استخدام ١٢عمود

القيمة الكلية لعدد العناصر (194000) عنصر تم تجميعها في خمسة عناقيد كالآتي

أ- العنقود الاول يحتوى علي عدد (٣٢٥٤٩) عنصر وبنسبة التشابه بلغت ١٧ % حيث نجد أن أعلي المواد تحصيلاً مادة التربية الإسلامية بنسبه ٧٨% وكذلك مواد اللغة الانجليزية واللغة العربية.

وأقل المواد تحصيلاً في المواد الأساسية مادة الرياضيات بنسبه ٤٣% ، وأقل المواد تحصيلاً في المواد عموما هي مادة الفيزياء بنسبة تحصيل ٤٠%وكذلك نجد المواد الفرعيه (الكمبيوتر والهندسية) متشابهه في التحصيل ، ونجد أن التشابه في المواد الاساسيه (الإسلامية واللغة العربية واللغة الانجليزية) يعمل علي زياده التحصيل الكلي في المجموع كما نجدأن مادة الأحياء الاختيارية هي من اقل المواد تحصيلاً عموما نسبة لغياب من جلسوا للامتحان في هذا العنقود في هذه المادة بنسبة تحصيل ٠%.

ب-العنقود الثاني يحتوى علي عدد (٣٧٧٥٠) عنصر وبنسبة التشابه بلغت ١٩ % حيث أن أعلي تحصيل في مادة التربية الإسلامية بنسبه ٧٥% وكذلك مواد اللغة الانجليزية واللغة العربية.

وأقل المواد تحصيلاً في المواد الأساسية مادة الرياضيات بنسبه ٤٠% ، وأقل المواد تحصيلاً في المواد عموما هي مادة الفيزياء بنسبة تحصيل ٣٨%وكذلك نجد المواد الفرعيه (الكمبيوتر والهندسية) متشابهه في التحصيل فكلهم غياب، ونجد أن التشابه في المواد الاساسيه (الإسلامية واللغة العربية واللغة الانجليزية) يعمل علي زياده التحصيل الكلي في المجموع كما نجدأن مادة الأحياء الاختيارية هي من أعلي المواد تحصيلاً عموما بنسبة ٦٨%.

ت-العنقود الثالث يحتوى علي عدد (٢٦٥٤٨) عنصر وبنسبة التشابه بلغ ١٤% حيث أن أعلي تحصيل في ماده اللغة الانجليزية بنسبه ٢٩% وكذلك مواد التربية الاسلاميةواللغة العربية.

وأقل المواد تحصيلاً في المواد مادة الرياضيات بنسبه ١٤% ، وأقل المواد تحصيلاً في المواد العلمية هي مادة الفيزياءبنسبة تحصيل ٢١%وكذلك نجد المواد الفرعيه (الكمبيوتر والهندسية) متشابهه في التحصيل، ونجد أن التشابه في المواد الاساسيه (الإسلامية واللغة العربية واللغة الانجليزية) يعمل علي زياده التحصيل الكلي في المجموع كما نجدأن نسبة تحصيل مادة الأحياء الاختيارية هي ١٥%.

ث-العنقود الرابع يحتوى علي عدد (٥٠١٥٩) عنصر وبنسبة التشابه بلغت ٢٦ % حيث أعلي نسبه تحصيل في مادة التربية الإسلامية بنسبه ٧٧% وكذلك مواد اللغة الانجليزية واللغة العربية.

وأقل المواد تحصيلاً في المواد الأساسية مادة الرياضيات بنسبه ٤١% ، وأقل المواد تحصيلاً في المواد عموما هي مادة الفيزياء بنسبة تحصيل ٣٧%وكذلك نجد المواد الفرعيه (الكمبيوتر والهندسية) متشابهه في التحصيل، ونجد أن التشابه في المواد الاساسيه (الإسلامية واللغة العربية واللغة الانجليزية) يعمل علي زياده التحصيل الكلي في المجموع كما نجدأن نسبة تحصيل مادة الأحياء الاختيارية هي غياب ٠%.

ج- العنقود الخامس يحتوى علي عدد (٤٦٩٩٤) عنصر وبنسبة التشابه بلغت ٢٤ % حيث أن أعلي تحصيل في مادة التربية الإسلامية بنسبه ٧٧% وكذلك مواد اللغة الانجليزية واللغة العربية.

وأقل المواد تحصيلاً في المواد الأساسية مادة الرياضيات بنسبه ٤٢% ، وأقل المواد تحصيلاً في المواد عموما هي مادة الفيزياء بنسبة تحصيل ٤٠%وكذلك نجد المواد الفرعيه (الكمبيوتر والهندسية) متشابهه في التحصيل فكلهم غياب، ونجد أن التشابه في المواد الاساسيه (الإسلامية واللغة العربية واللغة الانجليزية) يعمل علي زياده التحصيل الكلي في المجموع كما نجدأن مادة الأحياء الاختيارية هي من أعلي المواد تحصيلاً عموما بنسبة ٦٥%.

من نواحي اخري

نلاحظ العنقود الاول والعنقود الثاني متشابهان

أ- في اعلي نسبه تحصيل (التربية الاسلامية واللغه العربيه و اللغة الانجليزية)

ب-في اقل نسبه تحصيل (الرياضيات+الفيزياء)

ت-في التحصيل المتساوي(المواد الفرعيه)

ونلاحظ العنقود الرابع والعنقود الخامس متشابهان:

أ- في اعلي نسبه تحصيل (التربية الاسلامية واللغه العربيه و اللغة الانجليزية)

ب-في اقل نسبه تحصيل (الرياضيات+الفيزياء)

ت-في التحصيل المتساوي(المواد الفرعيه)

أما العنقود الثالث فنجد نسب التحصيل متدنية ومختلف عن باقي العناقيد حيث اللغة الانجليزية هي اعلي نسبة تحصيل والرياضيات نسبة تحصيلها ادنى من الفيزياءاما المواد الفرعية فنسب تحصيلها متقاربة.

والعنقود الاول والرابع مختلفان

التحصيل عند البنات اعلي من التحصيل عند الاولاد.

والعنقود الثاني والخامس مختلفان

التحصيل عند الاولاد اعلي من التحصيل عند البنات.

٥-٧من النتائج السابقة نخلص للآتي

٥-٧-١ أن اكثر المواد تحصيلاً مادة التربية الاسلامية للاسباب الاتية

أ- تكون اكثر المواد نجاحا واكثر المواد تحصيلا لان جانب الحفظ ياخذ ٤٠% من درجات الامتحان مما يساعد الطالب علي التحصيل العالي

ب-الماده لها خمسه فروع وهذه الخمسه فروع تمكن الطالب من احراز درجات في كل فرع من الفروع بصوره جيده.

ت-اصبحت صيغه الامتحان تكاد تكون معروفة لدى الطلاب مما أكسبهم خبرة علي حل الامتحان

ث-استخدام مهارات التفكير الإبداعي في تدريس مادة التربية الاسلامية له اثره علي تحصيلهم الدراسي.

ج- التربية الاسلامية هي العقيدة الام وانها سهله لدى كثير من الطلاب.

ح- لغة القران هي اللغة العربية وانها سهله لدى كثير من الطلاب.

٥-٧-٢ اقل المواد تحصيلاً مادتي الفيزياء والرياضيات للأسباب الاتية

أ- مرور الطلبه في الصفوف الدنيا دون التمكن من مادتي الفيزياءوالرياضيات.

ب-استخدام اسلوب الحفظ المتبع في التدريس يؤدي الي تدني تحصيل الطلبه.

ت-عدم مراعاه الفروق الفرديه عند طرح المواضيع داخل المنهج يؤدي الي تدني تحصيل الطلبه.

ث-تعليم مادتي الفيزياءوالرياضيات من معلمين غير متخصصين.

ج- عدم انسجام المفاهيم الخاصه بالفيزياءوالرياضيات لمستوي الطالب.

ح- عدم التحاق معلمي الفيزياءوالرياضيات بالدورات التدريبيه المناسبه يؤدي الي تدني تحصيل الطالب.

خ- عدم توفر واستخدام الاجهزه والوسائل الحديثه يؤدي الي تدني تحصيل الطلبه.

د- ازدحام الصفوف بالطلبه يؤدي الي انخفاض تحصيل الطلبه.

٥-٧-٣ كذلك نجد ان أعلي نسبه تحصيل عند الطالبات اكثر من الطلاب للاتي

أ- النجاح عند الطالبات اكثر لان الطالبات اكثر اهتماما من الطلاب.

ب-وجود الرغبه عند الطالبات اكثر من الطلاب.

ت-قصور الطلاب في حل الواجبات اكثر من الطالبات.

```
=== Clustering model (full training set) ===

kMeans
======

Number of iterations: 15
Within cluster sum of squared errors: 60712.697046307076
Missing values globally replaced with mean/mode

Cluster centroids:
                          Cluster#
Attribute      Full Data          0          1          2          3          4
                (194000)    (58463)    (31833)    (31729)    (39845)    (32130)
================================================================================
Islamic          70.2545    86.2676    94.3009    26.6383    70.5027    60.0574
I.Rating       Excellent  Excellent  Excellent       Fail     V.Good       Good
Sum.            381.1803   418.8847   562.2256   197.0278   390.4822   303.5209
Gen.Rat.            Pass       Pass       Pass       Fail       Pass       Fail


Time taken to build model (full training data) : 10.18 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       58463 ( 30%)
1       31833 ( 16%)
2       31729 ( 16%)
3       39845 ( 21%)
4       32130 ( 17%)
```

شكل(٥-٨) نتائج ٥عناقيد (5clusters)بعد استخدام مادة التربية الاسلامية وتقديرها مقارنة مع المجموع العام والتقدير العام

kMeans

Number of iterations: 15

Within cluster sum of squared errors: 60712.697046307076

Missing values globally replaced with mean/mode

Cluster centroids: Cluster#

Attribute	Full Data	0	1	2	3	4
	(194000)	(58463)	(31833)	(31729)	(39845)	(32130)
Islamic	70.2545	86.2676	94.3009	26.6383	70.5027	60.0574
I.Rating	Excellent	Excellent	Excellent	Fail	V.Good	Good
Sum.	381.1803	418.8847	562.2256	197.0278	390.4822	303.5209
Gen.Rat.	Pass	Pass	Pass	Fail	Pass	Fail

Clustered Instances

0	58463	(30%)
1	31833	(16%)
2	31729	(16%)
3	39845	(21%)
4	32130	(17%)

جدول (٥-٣) نتائج ٥عناقيد (5clusters)بعد استخدام مادة التربية الاسلامية وتقديرها مقارنة مع المجموع العام والتقدير العام

٥-٧-٤ قراءة عناقيد التربية الاسلامية نلاحظ الآتي

أ- أن عدد ٥٨٤٦٣ طالب كان متوسط تحصيلهم ٨٦درجة وتقديرهم ممتاز كان مجموعهم ٤١٩درجة وتقدير عام ناجح.

ب-أن عدد ٣١٨٣٣ طالب كان متوسط تحصيلهم ٩٤درجة وتقديرهم ممتاز كان مجموعهم ٥٦٢درجة وتقدير عام ناجح.

ت-أن عدد ٣١٧٢٩ طالب كان متوسط تحصيلهم ٢٧درجة وتقديرهم راسب كان مجموعهم ١٩٧درجة وتقدير عام راسب.

ث-أن عدد ٣٩٨٤٥ طالب كان متوسط تحصيلهم ٧٠درجة وتقديرهم جيدجدا كان مجموعهم ٣٩٠درجة وتقدير عام ناجح.

ج- أن عدد ٣٢١٣٠ طالب كان متوسط تحصيلهم ٦٠درجة وتقديرهم جيد كان مجموعهم ٤١٩درجة وتقدير عام راسب.

```
=== Clustering model (full training set) ===

kMeans
======

Number of iterations: 12
Within cluster sum of squared errors: 53308.375571767916
Missing values globally replaced with mean/mode

Cluster centroids:
                         Cluster#
Attribute    Full Data          0          1          2          3          4
              (194000)    (27353)    (20759)    (19671)    (65009)    (61208)
===============================================================================
Arabic         54.9226    73.6368    85.8928    17.6062    59.8106    42.8568
A.Rating          Fail     V.Good  Excellent       Fail       Good       Fail
Sum.          381.1803   474.2936   570.7797   148.3316   414.1535   315.0777
Gen.Rat.          Pass       Pass       Pass       Fail       Pass       Fail


Time taken to build model (full training data) : 8.03 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       27353 ( 14%)
1       20759 ( 11%)
2       19671 ( 10%)
3       65009 ( 34%)
4       61208 ( 32%)
```

شكل(٥-٩) نتائج ٥عناقيد (5clusters)بعد استخدام مادة اللغة العربية وتقديرها مقارنة مع المجموع العام والتقدير العام

kMeans

Number of iterations: 12

Within cluster sum of squared errors: 53308.375571767916

Missing values globally replaced with mean/mode

Cluster centroids: Cluster#

Attribute	Full Data	0	1	2	3	4
	(194000)	(27353)	(20759)	(19671)	(65009)	(61208)
Arabic	54.9226	73.6368	85.8928	17.6062	59.8106	42.8568
A.Rating	Fail	V.Good	Excellent	Fail	Good	Fail
Sum.	381.1803	474.2936	570.7797	148.3316	414.1535	315.0777
Gen.Rat.	Pass	Pass	Pass	Fail	Pass	Fail

Clustered Instances

0	27353	(14%)
1	20759	(11%)
2	19671	(10%)
3	65009	(34%)
4	61208	(32%)

جدول (٥-٤) نتائج ٥عناقيد (5clusters)بعد استخدام مادة اللغة العربية وتقديرها مقارنة مع المجموع العام والتقدير العام

٥-٧-٥ قراءة عناقيد اللغة العربية نلاحظ الآتي

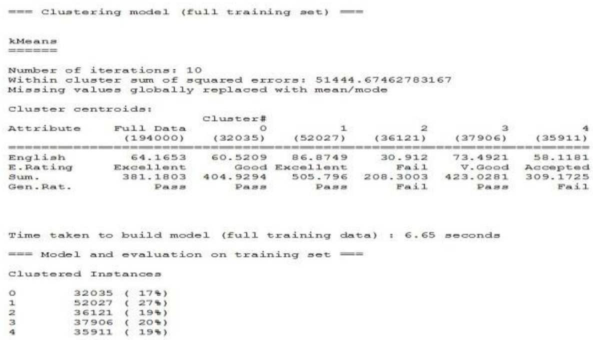
أ- أن عدد ٢٧٣٥٣ طالب كان متوسط تحصيلهم ٧٤درجة وتقديرهم جيدجداكان مجموعهم ٤٧٤درجة وتقدير عام ناجح.

ب-أن عدد ٢٠٧٥٩ طالب كان متوسط تحصيلهم ٨٦درجة وتقديرهم ممتاز كان مجموعهم ٥٧٠درجة وتقدير عام ناجح.

ت-أن عدد ١٩٦٧١ طالب كان متوسط تحصيلهم ١٨درجة وتقديرهم راسب كان مجموعهم ١٤٨درجة وتقدير عام راسب.

ث-أن عدد ٦٥٠٠٩ طالب كان متوسط تحصيلهم ٦٠درجة وتقديرهم جيد كان مجموعهم ٤١٤درجة وتقدير عام ناجح.

ج- أن عدد ٦١٢٠٨ طالب كان متوسط تحصيلهم ٤٣درجة وتقديرهم راسب كان مجموعهم ٣١٥درجة وتقدير عام راسب.



```
=== Clustering model (full training set) ===

kMeans
======

Number of iterations: 10
Within cluster sum of squared errors: 51444.67462783167
Missing values globally replaced with mean/mode

Cluster centroids:
                          Cluster#
Attribute     Full Data          0           1          2          3          4
               (194000)    (32035)     (52027)    (36121)    (37906)    (35911)
================================================================================
English         64.1653    60.5209     86.8749     30.912    73.4921    58.1181
E.Rating      Excellent       Good   Excellent       Fail     V.Good   Accepted
Sum.           381.1803   404.9294     505.796   208.3003   423.0281   309.1725
Gen.Rat.           Pass       Pass        Pass       Fail       Pass       Fail


Time taken to build model (full training data) : 6.65 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       32035 ( 17%)
1       52027 ( 27%)
2       36121 ( 19%)
3       37906 ( 20%)
4       35911 ( 19%)
```

شكل(٥-١٠) نتائج ٥عناقيد (5clusters)بعد استخدام مادة اللغة الانجليزية وتقديرها مقارنة مع المجموع العام والتقدير العام

kMeans

Number of iterations: 10

Within cluster sum of squared errors: 51444.67462783167

Missing values globally replaced with mean/mode

Cluster centroids: Cluster#

Attribute	Full Data	0	1	2	3	4
	(194000)	(32035)	(52027)	(36121)	(37906)	(35911)
English	64.1653	60.5209	86.8749	30.912	73.4921	58.1181
E.Rating	Excellent	Good	Excellent	Fail	V.Good	Accepted
Sum.	381.1803	404.9294	505.796	208.3003	423.0281	309.1725
Gen.Rat.	Pass	Pass	Pass	Fail	Pass	Fail

Clustered Instances

0	32035	(17%)
1	52027	(27%)
2	36121	(19%)
3	37906	(20%)
4	35911	(19%)

جدول (٥-٥) نتائج ٥عناقيد (5clusters)بعد استخدام مادة اللغة الانجليزية وتقديرها مقارنة مع المجموع العام والتقدير العام

٥-٧-٦ قراءة عناقيد اللغة الانجليزية نلاحظ الآتي

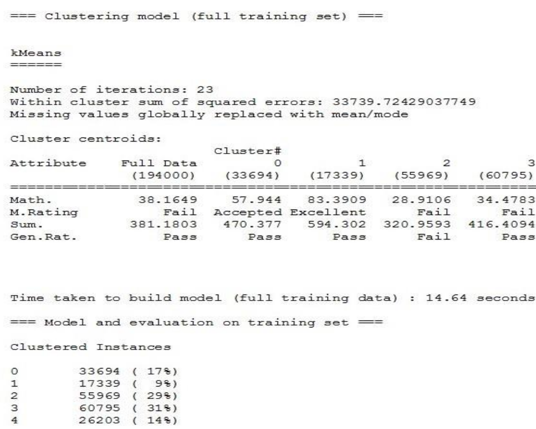
أ- أن عدد ٣٢٠٣٥ طالب كان متوسط تحصيلهم ٦١درجة وتقديرهم جيد كان مجموعهم ٤٠٤درجة وتقدير عام ناجح.

ب-أن عدد ٥٢٠٢٧ طالب كان متوسط تحصيلهم ٨٧درجة وتقديرهم ممتاز كان مجموعهم ٥٠٦درجة وتقدير عام ناجح.

ت-أن عدد ٣٦١٢١ طالب كان متوسط تحصيلهم ٣١درجة وتقديرهم راسب كان مجموعهم ٢٠٨درجة وتقدير عام راسب.

ث-أن عدد ٣٧٩٠٦ طالب كان متوسط تحصيلهم ٧٣درجة وتقديرهم جيدجدا كان مجموعهم ٤٢٣درجة وتقدير عام ناجح.

ج-أن عدد ٣٥٩١١ طالب كان متوسط تحصيلهم ٥٨درجة وتقديرهم مقبول كان مجموعهم ٣٠٩درجة وتقدير عام راسب.



```
=== Clustering model (full training set) ===

kMeans
======

Number of iterations: 23
Within cluster sum of squared errors: 33739.72429037749
Missing values globally replaced with mean/mode

Cluster centroids:
                          Cluster#
Attribute    Full Data         0          1          2          3          4
              (194000)   (33694)    (17339)    (55969)    (60795)    (26203)
==============================================================================
Math.          38.1649    57.944    83.3909    28.9106    34.4783    11.1251
M.Rating          Fail  Accepted  Excellent       Fail       Fail       Fail
Sum.          381.1803   470.377    594.302   320.9593   416.4094   172.3508
Gen.Rat.          Pass      Pass       Pass       Fail       Pass       Fail


Time taken to build model (full training data) : 14.64 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       33694 ( 17%)
1       17339 (  9%)
2       55969 ( 29%)
3       60795 ( 31%)
4       26203 ( 14%)
```

شكل (٥-١١) نتائج ٥عناقيد (5clusters)بعد استخدام مادة الرياضيات وتقديرها مقارنة مع المجموع العام والتقدير العام

kMeans

Number of iterations: 23

Within cluster sum of squared errors: 33739.72429037749

Missing values globally replaced with mean/mode

Cluster centroids: Cluster#

Attribute	Full Data	0	1	2	3	4
	(194000)	(33694)	(17339)	(55969)	(60795)	(26203)
Math.	38.1649	57.944	83.3909	28.9106	34.4783	11.1251
M.Rating	Fail	Accepted	Excellent	Fail	Fail	Fail
Sum.	381.1803	470.377	594.302	320.9593	416.4094	172.3508
Gen.Rat.	Pass	Pass	Pass	Fail	Pass	Fail

Clustered Instances

0	33694	(17%)
1	17339	(9%)
2	55969	(29%)
3	60795	(31%)
4	26203	(14%)

جدول (٥-٦) نتائج ٥عناقيد (5clusters)بعد استخدام مادة الرياضيات وتقديرها مقارنة مع المجموع العام والتقدير العام

٥-٧-٧ قراءة عناقيد الرياضيات نلاحظ الآتي

أ- أن عدد ٣٣٦٩٤ طالب كان متوسط تحصيلهم ٥٨درجة وتقديرهم مقبول كان مجموعهم ٤٧٠درجة وتقدير عام ناجح.

ب-أن عدد ١٧٣٣٩ طالب كان متوسط تحصيلهم ٨٣درجة وتقديرهم ممتاز كان مجموعهم ٥٩٤درجة وتقدير عام ناجح.

ت-أن عدد ٥٥٩٦٩ طالب كان متوسط تحصيلهم ٢٩درجة وتقديرهم راسب كان مجموعهم ٣٢١درجة وتقدير عام راسب.

ث-أن عدد ٦٠٧٩٥ طالب كان متوسط تحصيلهم ٣٤درجة وتقديرهم راسب كان مجموعهم ٤١٦درجة وتقدير عام ناجح.

ج- أن عدد ٢٦٢٠٣ طالب كان متوسط تحصيلهم ١١درجة وتقديرهم راسب كان مجموعهم ١٧٢درجة وتقدير عام راسب.

```
=== Clustering model (full training set) ===

kMeans
======

Number of iterations: 28
Within cluster sum of squared errors: 39175.300192665025
Missing values globally replaced with mean/mode

Cluster centroids:
                            Cluster#
Attribute      Full Data          0          1          2          3          4
                (194000)    (56847)    (16585)    (65504)    (35849)    (19215)
================================================================================
Physics          36.8234    31.0129    80.5256    29.0137    50.7587    16.9174
P.Rating            Fail       Fail  Excellent       Fail   Accepted       Fail
Sum.            381.1803   412.3737   595.6473   316.1757   478.9553   142.9675
Gen.Rat.            Pass       Pass       Pass       Fail       Pass       Fail


Time taken to build model (full training data) : 17.08 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       56847 ( 29%)
1       16585 (  9%)
2       65504 ( 34%)
3       35849 ( 18%)
4       19215 ( 10%)
```

شكل (٥-١٢) نتائج ٥عناقيد (5clusters)بعد استخدام مادة الفيزياء وتقديرها مقارنة مع المجموع العام والتقدير العام

kMeans

Number of iterations: 28

Within cluster sum of squared errors: 39175.300192665025

Missing values globally replaced with mean/mode

Cluster centroids: Cluster#

Attribute	Full Data	0	1	2	3	4
	(194000)	(56847)	(16585)	(65504)	(35849)	(19215)
Physics	36.8234	60.5209	80.5256	29.0137	Accepted	16.9174
P.Rating	Excellent	Fail	Excellent	Fail	V.Good	Fail
Sum.	381.1803	412.3737	595.6473	316.1757	478.9553	142.9675
Gen.Rat.	Pass	Pass	Pass	Fail	Pass	Fail

Clustered Instances

0	56847	(29%)
1	16585	(9%)
2	65504	(34%)
3	35849	(18%)
4	19215	(10%)

جدول(٥-٧) نتائج ٥عناقيد (5clusters)بعد استخدام مادة الفيزياء وتقديرها مقارنة مع المجموع العام والتقدير العام

٥-٧-٨ قراءة عناقيد الفيزياء نلاحظ الآتي

أ- أن عدد ٥٦٨٤٧ طالب كان متوسط تحصيلهم ٣١درجة وتقديرهم راسب كان مجموعهم ٤١٢درجة وتقدير عام ناجح.

ب-أن عدد ١٦٥٨٥ طالب كان متوسط تحصيلهم ٨١درجة وتقديرهم ممتاز كان مجموعهم ٥٩٦درجة وتقدير عام ناجح.

ت-أن عدد ٦٥٥٠٤ طالب كان متوسط تحصيلهم ٢٩درجة وتقديرهم راسب كان مجموعهم ٣١٦درجة وتقدير عام راسب.

ث-أن عدد ٣٥٨٤٩ طالب كان متوسط تحصيلهم ٥١درجة وتقديرهم مقبول كان مجموعهم ٤٧٨درجة وتقدير عام ناجح.

ج- أن عدد ١٩٢١٥ طالب كان متوسط تحصيلهم ١٧درجة وتقديرهم راسب كان مجموعهم ١٤٣درجة وتقدير عام راسب.

٥-٧-٩ النتائج المستخلصة من تنفيذ الخوارزمية :

أ- بينت النتائج ان هناك تشابه كبير بين بعض المواد (التربية الاسلامية واللغة العربية واللغة الانجليزية) ومادتى (الفيزياء و الرياضيات) و يعملان كعامل للنجاح والرسوب فى المجموع الكلى .

ب- بينت النتائج أحتمال نجاح الطالب فى المجموع النهائى فى حالت الرسوب فى مادتى الفيزياء والرياضيات بنسبة 29 % وهى نسبة تبين الضعف الشديد فى هاتين المادتين.

ت- بينت النتائج انه وفي حالة نجاح الطالب في المجموع العام ولكنه راسب في مادة اللغة العربية فانه يعتبر راسب في التقدير العام.

٥-٨ تنفيذ شجرة القرار (J48)

افترض أن لديك مجموعة من البيانات وتريد بناء شجرة القرارارت منها :

أولا : لابد من تهيئة هذه البيانات واطلاق المستشكف والقيام بتحميل البيانات فيه .

ثانيا : لابد من إختيار طريقة بناء شجرة القرار وبناء الشجرة وتفسير المخرجات.في المستشكف يمكنك التنقل ذهاباًوايابا ًبين النتائج التي تحصلت عليها ، تقييم النماذج التي تم بناءها على مختلف قواعد البيانات، تصور بشكل تخطيطي كلتا النماذج وقواعد البيانات انفسهم والتي تتضمن أخطاء التجميع التي يمكن أن يصنعها النموذج .

Classifier output

```
|   |   |   |   |   |   C.STATUS = Fail: Accepted (234.0/80.0)
|   |   |   |   |   |   C.STATUS = Absent
|   |   |   |   |   |   |   G.STATUS = Absent: Good (0.0)
|   |   |   |   |   |   |   G.STATUS = Pass
|   |   |   |   |   |   |   |   I.STATUS = Pass: Good (4635.0/2029.0)
|   |   |   |   |   |   |   |   I.STATUS = Fail: Accepted (27.0/1.0)
|   |   |   |   |   |   |   |   I.STATUS = Absent: Good (0.0)
|   |   |   |   |   |   |   G.STATUS = Fail: Accepted (148.0/46.0)
|   |   |   |   |   B.STATUS = Pass
|   |   |   |   |   |   I.STATUS = Pass: Good (8599.0/3845.0)
|   |   |   |   |   |   I.STATUS = Fail: Accepted (56.0/9.0)
|   |   |   |   |   |   I.STATUS = Absent: Good (0.0)
|   |   |   |   |   B.STATUS = Fail: Accepted (311.0/106.0)
|   |   |   |   CH.STATUS = Absent: Good (0.0)
|   |   |   E.STATUS = Absent: Accepted (4.0)
|   |   M.STATUS = Absent: Accepted (2.0)
|   P.STATUS = Pass
|   |   M.STATUS = Fail
|   |   |   E.STATUS = Fail: Accepted (279.0/22.0)
|   |   |   E.STATUS = Pass
|   |   |   |   CH.STATUS = Fail: Accepted (994.0/289.0)
|   |   |   |   CH.STATUS = Pass
|   |   |   |   |   I.STATUS = Pass
|   |   |   |   |   |   B.STATUS = Absent
|   |   |   |   |   |   |   C.STATUS = Pass: Good (1863.0/783.0)
|   |   |   |   |   |   |   C.STATUS = Fail: Accepted (91.0/33.0)
|   |   |   |   |   |   |   C.STATUS = Absent
|   |   |   |   |   |   |   |   G.STATUS = Absent: Accepted (1.0)
|   |   |   |   |   |   |   |   G.STATUS = Pass: Good (1809.0/763.0)
|   |   |   |   |   |   |   |   G.STATUS = Fail: Accepted (47.0/14.0)
|   |   |   |   |   |   B.STATUS = Pass: Good (4232.0/1805.0)
|   |   |   |   |   |   B.STATUS = Fail: Accepted (171.0/57.0)
|   |   |   |   |   I.STATUS = Fail: Accepted (85.0/9.0)
|   |   |   |   |   I.STATUS = Absent: Good (0.0)
|   |   |   |   CH.STATUS = Absent: Good (0.0)
|   |   |   E.STATUS = Absent: Accepted (4.0)
|   |   M.STATUS = Pass
|   |   |   I.STATUS = Pass
|   |   |   |   E.STATUS = Fail
|   |   |   |   |   CH.STATUS = Fail: Accepted (13.0/1.0)
```

Log x 0

Classifier output

```
|   |   |   |   |   CH.STATUS = Fail: Accepted (13.0/1.0)
|   |   |   |   |   CH.STATUS = Pass: Good (114.0/56.0)
|   |   |   |   |   CH.STATUS = Absent: Good (0.0)
|   |   |   |   E.STATUS = Pass
|   |   |   |   |   CH.STATUS = Fail
|   |   |   |   |   |   B.STATUS = Absent
|   |   |   |   |   |   |   G.STATUS = Absent: Good (42.0/15.0)
|   |   |   |   |   |   |   G.STATUS = Pass: Good (116.0/40.0)
|   |   |   |   |   |   |   G.STATUS = Fail: Accepted (18.0/7.0)
|   |   |   |   |   |   B.STATUS = Pass: Good (122.0/32.0)
|   |   |   |   |   |   B.STATUS = Fail: Accepted (17.0/3.0)
|   |   |   |   |   CH.STATUS = Pass
|   |   |   |   |   |   B.STATUS = Absent
|   |   |   |   |   |   |   C.STATUS = Pass: Excellent (6173.0/2993.0)
|   |   |   |   |   |   |   C.STATUS = Fail: Good (59.0/16.0)
|   |   |   |   |   |   |   C.STATUS = Absent
|   |   |   |   |   |   |   |   G.STATUS = Absent: Excellent (0.0)
|   |   |   |   |   |   |   |   G.STATUS = Pass: Excellent (6625.0/3157.0)
|   |   |   |   |   |   |   |   G.STATUS = Fail: Good (26.0/8.0)
|   |   |   |   |   |   B.STATUS = Pass: Excellent (13480.0/6106.0)
|   |   |   |   |   |   B.STATUS = Fail: Good (102.0/41.0)
|   |   |   |   |   CH.STATUS = Absent: Excellent (0.0)
|   |   |   |   E.STATUS = Absent: Excellent (0.0)
|   |   |   I.STATUS = Fail: Accepted (64.0/20.0)
|   |   |   I.STATUS = Absent: Accepted (88.0/36.0)
|   |   M.STATUS = Absent: Excellent (0.0)
|   P.STATUS = Absent: Accepted (2.0/1.0)
Gen.Rat. = Absent: Absent (3495.0)

Number of Leaves  :     55

Size of the tree :      82


Time taken to build model: 1.68 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      149605              77.116  %
```

Log x 0

Classifier output

```
Size of the tree :      82


Time taken to build model: 1.68 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      149605              77.116  %
Incorrectly Classified Instances     44395              22.884  %
Kappa statistic                          0.6865
Mean absolute error                      0.0986
Root mean squared error                  0.222
Relative absolute error                 40.3616 %
Root relative squared error             63.5433 %
Coverage of cases (0.95 level)          99.0644 %
Mean rel. region size (0.95 level)      30.6519 %
Total Number of Instances           194000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC    ROC Area  PRC Area  Class
                 1        0        1          1        1          1      1         1         Fail
                 0.916    0.13     0.636      0.916    0.751      0.694  0.925     0.646     Accepted
                 0.414    0.076    0.564      0.414    0.478      0.383  0.858     0.496     Good
                 0.97     0.068    0.534      0.97     0.688      0.692  0.963     0.535     Excellent
                 1        0        1          1        1          1      1         1         Absent
                 0        0        0          0        0          ?       0.887     0.34      V.Good
Weighted Avg.    0.771    0.046    0.712      0.771    0.73       0.7    0.944     0.734

=== Confusion Matrix ===

     a     b     c     d     e     f   <-- classified as
 81280     0     0     0     0     0 |     a = Fail
     0 35380  3160   104     0     0 |     b = Accepted
     0 18940 15428  2887     0     0 |     c = Good
     0     3   431 14022     0     0 |     d = Excellent
     0     0     0     0  3495     0 |     e = Absent
     0  1269  8336  9265     0     0 |     f = V.Good
```

Log x 0

شكل(٥-١٣)نتائج تحليل شجرة القرار

=== Summary ===

Correctly Classified Instances	149605	77.116 %
Incorrectly Classified Instances	44395	22.884 %
Kappa statistic	0.6865	
Mean absolute error	0.0986	
Root mean squared error	0.222	
Relative absolute error	40.3616 %	
Root relative squared error	63.5433 %	
Coverage of cases (0.95 level)	99.0644 %	
Mean rel. region size (0.95 level)	30.6519 %	
Total Number of Instances	194000	

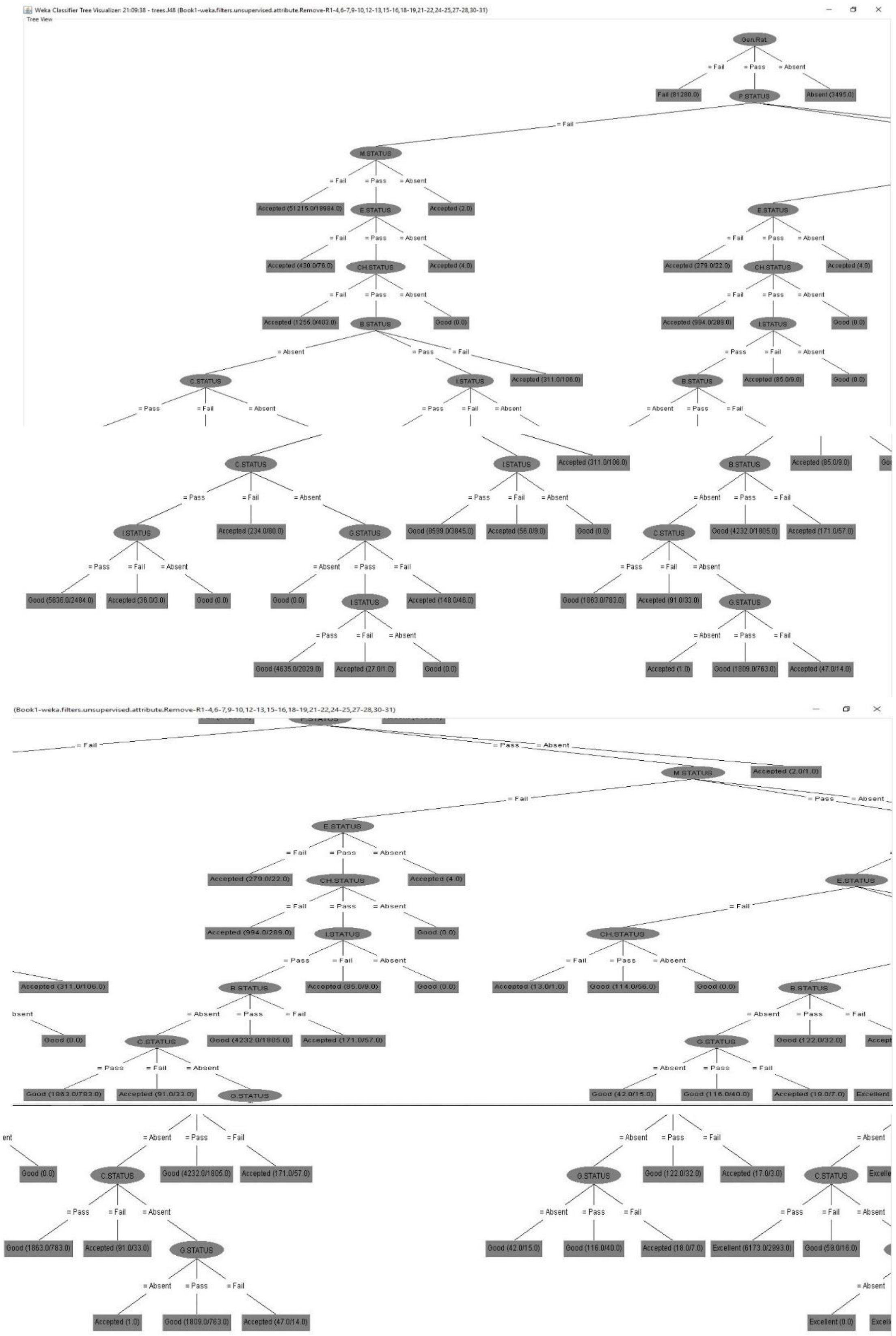
=== Detailed Accuracy By Class ===

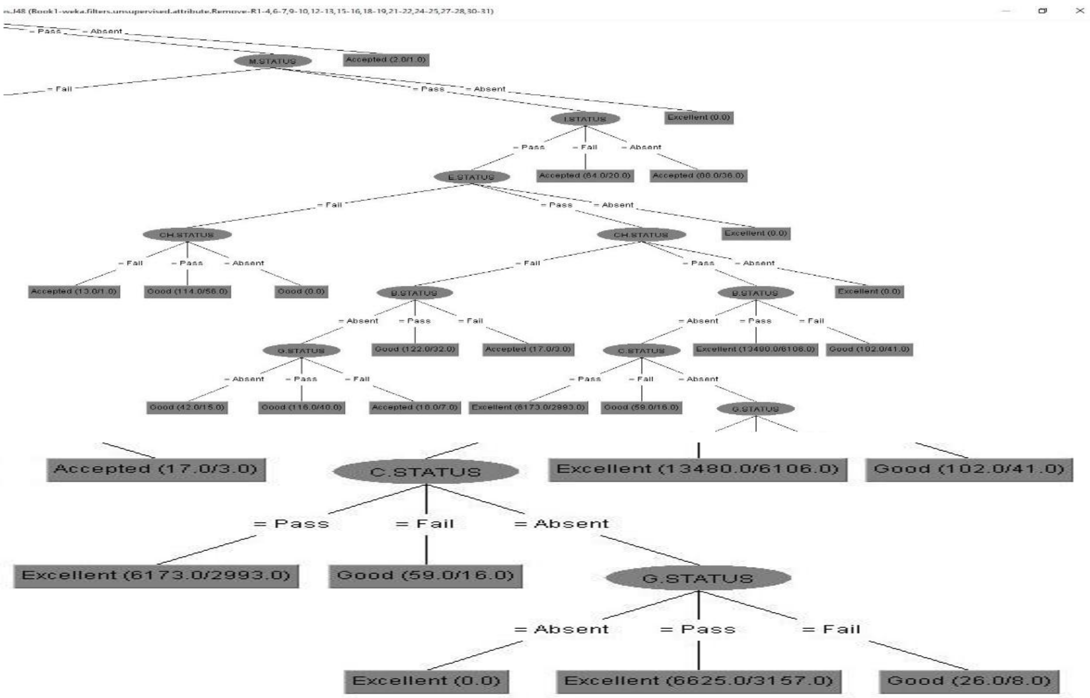
TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
1	0	1	1	1	1	1	1	Fail
0.916	0.13	0.636	0.916	0.751	0.694	0.925	0.646	Accepted
0.414	0.076	0.564	0.414	0.478	0.383	0.858	0.496	Good
0.97	0.068	0.534	0.97	0.688	0.692	0.963	0.535	Excellent
1	0	1	1	1	1	1	1	Absent
0	0	0	0	0	?	0.887	0.34	V.Good
0.771	0.046	0.712	0.771	0.73	0.7	0.944	0.734	Weighted Avg.

=== Confusion Matrix ===

a	b	c	d	e	f	<-- classified as
81280	0	0	0	0	0	a = Fail
0	35380	3160	104	0	0	b = Accepted
0	18940	15428	2887	0	0	c = Good
0	3	431	14022	0	0	d = Excellent
0	0	0	0	3495	0	e = Absent
0	1269	8336	9265	00	0	f = V.Good

جدول(٥-٨) نتائج تحليل شجرة القرار


Weka Classifier Tree Visualizer: 21:09:38 - trees.J48 (Book1-weka.filters.unsupervised.attribute.Remove-R1-4,6-7,9-10,12-13,15-16,18-19,21-22,24-25,27-28,30-31)
Tree View
Gen.Rat.
Fail (91280.0)
P.STATUS
Absent (3495.0)
M.STATUS
E.STATUS
CH.STATUS
B.STATUS
C.STATUS
I.STATUS
G.STATUS
Accepted (51215.0/18984.0)
Accepted (2.0)
Accepted (430.0/76.0)
Accepted (4.0)
Accepted (1255.0/403.0)
Good (0.0)
Accepted (311.0/106.0)
Accepted (279.0/22.0)
Accepted (994.0/289.0)
Accepted (85.0/9.0)
Good (5636.0/2484.0)
Accepted (36.0/3.0)
Accepted (234.0/80.0)
Good (8599.0/3845.0)
Accepted (56.0/9.0)
Good (4232.0/1805.0)
Accepted (171.0/57.0)
Accepted (148.0/46.0)
Good (1863.0/783.0)
Accepted (91.0/33.0)
Good (4635.0/2029.0)
Accepted (27.0/1.0)
Accepted (1.0)
Good (1809.0/763.0)
Accepted (47.0/14.0)
(Book1-weka.filters.unsupervised.attribute.Remove-R1-4,6-7,9-10,12-13,15-16,18-19,21-22,24-25,27-28,30-31)
Accepted (2.0/1.0)
Accepted (13.0/1.0)
Good (114.0/56.0)
Good (122.0/32.0)
Good (42.0/15.0)
Good (116.0/40.0)
Accepted (18.0/7.0)
Accepted (17.0/3.0)
Excellent (6173.0/2993.0)
Good (59.0/16.0)
Excellent (0.0)



شكل(٥-١٤) رسم تخطيطى لشجرة القرار توضح علاقة المواد بالنجاح والرسوب فى المجموع

٥-٩ استنتاجات شجرة القرار

أ- تبين شجرة القرار عدد الطلبة الراسبون فى المجموع (٨١,٢٨٠) طالب وعدد الطلبة الغياب كلياً عن جميع المواد (٣,٤٩٥) طالب من جملة (١٩٤,٠٠٠) طالب ، اذن عدد الطلبة الناجحون فى المجموع (١٠٩,٢٢٥) طالب .

ب-تبين تدرج المواد من الأعلى رسوباً بين الطلبة الناجحين (الفيزياء والرياضيات) نزولاً الى أقل المواد رسوباً وهى مادة التربية الاسلامية.

ت-ويتدرج طبقاً لذلك تقدير الطلبة فى المجموع النهائى بين المقبول وهو أقل من ٤٢٠ درجة وأعلى من ٣٥٠ درجة ، الى الممتاز وهو اعلى من ٥٦٠ درجة . وقد كانت النتائج كما في الجدول رقم(٥-٢) والجدول رقم (٥-٣) الذي تم استنتاجهما وتجميعهما يدوياً وبمساعدة معتبرة من برنامج WEKA .

المادة	اثرها على نجاح الطالب
الرياضيات	الرسوب (٥١,٢١٨) طالب منهم (١٨,٩٨٤) طالب احرزوا التقدير مقبول
اللغة الانجليزية	الطلبة الناجحون فى مادة الرياضيات وراسبون فى مادة اللغة الانجليزية (٤٣٠) طالب منهم (٧٦) أحرزوا تقدير مقبول والبقية بالتأكيد بتقدير أعلى من مقبول
الكيمياء	الطلبة الناجحون فى مادتى الرياضيات واللغةالانجليزية ورسبوا فى مادة الكيمياء (١٢٥٥) طالب منهم (٤٠٣) بتقدير مقبول
التربية الاسلامية	الطلبة الناجحون فى موادالرياضيات واللغة الانجليزية والكيمياء ورسبوا فى التربية الاسلامية (٣٦) طالب فقط منهم (٣) بتقدير مقبول فقط
الحاسوب	الطلبة الناجحون فى المواد السابق ذكرها ورسبوا فى مادة الحاسوب(٢٣٤) طالب منهم (٨٠) فقط بتقدير مقبول
الهندسية	الطلبة الناجحون فى المواد أعلاه ورسبوا فى مادة الهندسية (١٤٨) طالب منهم (٤٦) بتقدير مقبول
الأحياء	الطلبة الناجحون فى المواد أعلاه ورسبوا فى مادة الأحياء (٣١١) طالب منهم (١٠٦) بتقدير مقبول

جدول (٥-٩) أبرز الانماط الناتجة عن المواد-في حالة رسوب الطالب في مادة الفيزياء- وأثرها على نجاح الطالب (تصميم الباحث)

المادة	اثرها على نجاح الطالب
اللغة الانجليزية	الطلبة الراسبون فى مادة اللغة الانجليزية (٢٧٩) طالب منهم (٢٢) أحرزوا تقدير مقبول والبقية بالتأكيد بتقدير أعلى من مقبول
الكيمياء	الطلبة الناجحون فى مادة اللغةالانجليزية ورسبوا فى مادة الكيمياء (٩٩٤) طالب منهم (٢٨٩) بتقدير مقبول
التربية الاسلامية	الطلبة الناجحون فى مواد اللغة الانجليزية والكيمياء ورسبوا فى التربية الاسلامية (٨٥) طالب فقط منهم (٩) بتقدير مقبول فقط
الحاسوب	الطلبة الناجحون فى المواد السابق ذكرها ورسبوا فى مادة الحاسوب(٩١) طالب منهم (٣٣) فقط بتقدير مقبول
الهندسية	الطلبة الناجحون فى المواد أعلاه ورسبوا فى مادة الهندسية (٤٧) طالب منهم (١٤) بتقدير مقبول
الأحياء	الطلبة الناجحون فى المواد أعلاه ورسبوا فى مادة الأحياء (١٧١) طالب منهم (٥٧) بتقدير مقبول

جدول(٥-١٠-١)أبرز الانماط الناتجة عن المواد - في حالة نجاح الطالب في مادة الفيزياءورسوب في مادة الرياضيات - وأثرها على نجاح الطالب (تصميم الباحث)

المادة	اثرها على نجاح الطالب
الكيمياء	الطلبة الراسبون فى مادة الكيمياء (١٣) طالب منهم (١) بتقدير مقبول

جدول(٥-١٠-٢)أبرز الانماط الناتجة عن المواد - في حالة نجاح الطالب في مادة الفيزياءونجاح في مادة الرياضيات والتربية الاسلامية ورسوب في اللغة الانجليزية - وأثرها على نجاح الطالب (تصميم الباحث)

المادة	اثرها على نجاح الطالب
الهندسية	الطلبة الراسبون فى مادة الكيمياء (١٨) طالب منهم (٧) بتقدير مقبول
الاحياء	الطلبة الناجحون في مادة الهندسية ورسبوا في مادة الاحياء (١٧) طالب منهم (٣) بتقدير مقبول

جدول(٥-١٠-٣)أبرز الانماط الناتجة عن المواد - في حالة نجاح الطالب في مادة الفيزياءونجاح في مادة الرياضيات والتربية الاسلامية واللغة الانجليزية - وأثرها على نجاح الطالب (تصميم الباحث)

٥-١٠ فيما يتعلق بمادة اللغة العربية

أ- تبين شجرة القرار عدد الطلبة الراسبون فى اللغة العربية(٦٤,٩٣٧) طالب وعدد الطلبة الغياب كلياً في مادة اللغة العربية (٣,٤٩٥) طالب من جملة (١٩٤,٠٠٠) طالب ، اذن عدد الطلبة الناجحون فى مادة اللغة العربية(١٢٥,٥٦٨) طالب .

ب-عدد الطلبة الناجحون في مادة اللغة العربية (١٢٥,٥٦٨) منهم عدد(٣٩,٧٧٩) طالب نجحوا بتقدير مقبول، عدد(٣٧,٦٧٢) طالب نجحوا بتقدير جيد، عدد(٢٧,٣٥٣) طالب نجحوا بتقدير جيد جدا، عدد(٢٠,٧٦٤) طالب نجحوا بتقدير ممتاز.

الفصل السادس

النتائج والتوصيات

٦-١ النتائج

في هذا البحث

أ- تم بناء مستودع بيانات منطقي وتطبيق بعض خوارزميات التنقيب في البيانات على قاعدة البيانات بوزارة التربية والتعليم الاتحادية للمساهمة في توفير قاعدة معرفية لصناع القرار في الوزارة

ب-تم التوصل من خلالهذا البحث إلى استنتاجات هامة، أهمها هو الحاجة الماسة إلى بناء مستودع بيانات مترابط ومتكامل ونقي وخالي من الأخطاء للوزارة

ت-إضافة إلى نتائج أخرى هامة متعلقة بسجلات الطلاب مثل علاقة معدلات مرحلة الاساس وفترة الانقطاع بعد مرحلة الاساس باتجاهات الطلاب الأكاديمية ومستوى أدائهم خلال دراستهم الثانوية

ث-أضف إلى ذلك علاقة كثير من المواد الدراسية بتسرب الطلاب وانقطاعهم عن الدراسة، إما لصعوبة المفردات أو لخلل في الخطط الدراسية والمناهج

ج- وهناك عدة جوانب لم يتم التطرق إليها في هذا البحث إليها نظراً لعدم توفر المصادرالمطلوبة أو لعدم وجودها ضمن حدود البحثمبدئياً، وهذه يمكن تطويرها كأعمال مستقبلية لهذاالبحث مثل تحديث بيانات الطلاب، واستكمال البيانات الناقصة للطلاب في قاعدة بيانات الوزارة 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، واستخدام بعض الطرق الأحدث في التنقيب في البيانات، مثل استخدام تقنياتGenetic Algorithm وتقنيات Neural Networks

٦-٢ التوصيات

أ- تحسين قاعدة بيانات وزارة التربية والتعليم في المستقبل ومن أهمهامراجعة هيكلية قاعدة البيانات والتأكد منصحة العلاقات،ربط فروع الوزارة بقاعدة بيانات المركز الرئيسي،بناء مستودع بيانات نموذجي ومبني على أسس حديثة،استخدام تطبيق مناسب أو تطوير التطبيق الحالي ليحتوي على عواملتحققValidation صحيحة أثناء إدخالالبيانات، والتأكد من كفاءة مدخلي البيانات لتلافي الأخطاء في البيانات المدخلة

ب- لزيادة كفاءة النتائج المتحصل عليها من هذا البحث نوصي بإضافة بيانات الطلاب الذين تعزر الحصول علي بياناتهم نسبة لتواجدهم ضمن النظام اليدوى القديم ماقبل العام (٢٠٠٧) التسجيل بالدفاتر

ت- البيانات الأساسية لكل طالب يجب الحصول عليها وتخزينها، بحيث لايكون هنالك مشاكل في جمعها- عند الحوجة لتحميلها واجراء التجارب عليها.

ث- يجب تخزين البيانات بصورة جيدة لتكون في متناول اليد مما يساعد في إكمال مسيرة الأبحاث فى هذا المجال

ج- استخدام طرق عنقدة اخري مع مجموعة من البيانات للحصول علي مزيد من المعلومات.

٦-٣ الخاتمة

بعد الإطلاع على نتائج البحث وتحميلها ظهر ارتباط مواد محددة بمستويات متدنية للطلاب فيها وهى مادتى (الفيزياء والرياضيات) ومن ثم ينبغي لأصحاب القرار الوقوف عليها ومحاولة تفسير ذلك في ضوء طبيعة المادة أو مكانها في الخطة الدراسية وارتباطها بغيرها من المقررات أو بأداء بعض مدرسي هذه المواد.

أهمية سجل الطالب الأكاديمي ومدى تأثيره علي المستويات الأكاديمية للطلاب قيد الدراسة ، وبالتالي ظهرت الحوجة إلي تحميل بيانات الطلاب للإستفادة منها في صنع القرارات ، حيث أن إجراء العديد من التجارب في هىذا المجال يمكن المؤسسات التعليمية من تطوير الخطة المتبعة وزيادة الكفاءة .

وكما توضح من البحث والأبحاث السابقة أن تنقيب البيانات من الطرق الحديثة ذات الكفاءة العالية في تحميل وتجميع وتصنيف البيانات والتوصل إلى علاقات تزيد من إمكانية التحكم في اتخاذ القرار وتقديم أفضل ما يمكن للوصول للاهداف المرجوة .

٦-٤ المراجع والمصادر

٦-٤-١ المراجع العربية

[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، 2005 م.

[2] أواب الجناعى ، الحسين الحداد واخرون ،" استكشاف بعض الأ نماط المؤثرة في الأداء الأكاديمي لطلاب جامعة العلوم والتكنولوجيا " ، مجلة العلوم والتكنولوجيا المجلد (16) 2011 م

[3] بسام على ، انهار خير الدين واخرون ، "إستخدام تنقيب البيانات للتنبؤ بظاهرة تسرب طلبة كلية الإدارة والأقتصاد "،المجلة العراقية للعلوم الاحصائية العدد (18) 2010 ص ص (209_236)

[4] بندر الخيال."نبذة عن مستودعات البيانات في الحاسب الآلي". ص.٢

[5] وليد محمد الحسن العمر، عبدالباسط عبدالصمد آدم. "مستودعات البيانات". ص. ٢.

[10] 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.

[11] 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.

٦-٤-٢ المراجع الأجنبية

[6] S. A. Kumar, “Mining Of Student Academic Evaluation Records in Higher Education,” pp. 67–70, 2012.

[7] V. Kumar and A. Chadha, “Mining Association Rules in Student ’ s Assessment Data,” vol. 9, no. 5, pp. 211–216, 2012.

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

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

٦-٤-٣ مراجع الكترونية

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

[13] http://informatics.gov.sa/magazine