مجلة جامعة تشرين للبحوث والدراسات العلمية _ سلسلة العلوم الهندسية المجلد (37) العدد (2) 2015
Tishreen University Journal for Research and Scientific Studies - Engineering Sciences Series Vol. (37) No. (2) 2015

توظيف خوارزميات التنقيب في البيانات لتحليل حوادث المرور

إيهاب الديباجة*

(تاريخ الإيداع 23 / 9 / 2014. قُبِل للنشر في 4/ 3 / 2015)

□ ملخّص □

تقدم هذه الورقة البحثية مقارنة لمجموعة من خوارزميات التنقيب في البيانات Data Mining Algorithms فيما يتعلق بتحليل حوادث المرور، انطلاقاً من مرحلة إدخال البيانات، وذلك من خلال تحليل بنية التقارير الإحصائية الموجودة في فرع مرور اللاذقية وصولاً إلى مرحلة التنقيب في البيانات التي تستطيع إيجاد آلية قادرة على دراسة العوامل التي تلعب دوراً في حادث المرور بذكاء من أجل الربط وتحديد مدى العلاقة بينها وأهميتها في تسبب الحادث المروري، و ذلك بعد تصميم بنية مستودع البيانات على أساس قاعدة البيانات التي تم بناؤها لتخزين المعلومات، تم في هذا البحث ذكر مجموعة من النماذج التي تم اختبارها والتي تشكل عينة عن الاختبارات التي بنيت عليها نتائج البحث.

الكلمات المفتاحية: حوادث المرور– التنقيب في البيانات– مستودعات البيانات– قواعد بيانات معرفية– العنقدة – التصنيف – قواعد الاقتران

* ماجستير– قسم هندسة الحاسبات والتحكم الآلي– كلية الهندسة الميكانيكية والكهربائية– جامعة تشرين– اللاذقية– سورية

مجلة جامعة تشرين للبحوث والدراسات العلمية _ سلسلة العلوم الهندسية المجلد (37) العدد (2) 2015
Tishreen University Journal for Research and Scientific Studies - Engineering Sciences Series Vol. (37) No. (2) 2015

Employing Data Mining Algorithms in Traffic Accidents Analyzing

Ihab Aldibajah*

(Received 23 / 9 / 2014. Accepted 4 / 3 / 2015)

□ ABSTRACT □

In this paper we introduce a comparison for some of data mining algorithm for traffic accidents analysis.

We start by describing available data for entry by analyzing the structure of statistical reports in Lattakia traffic directorate, and proceed to data mining stage which enables us to smart study of factors that play roles in traffic accident and find its inter-relations and importance for causing traffic accident.

That comes after building data warehouse upon the database we built to store the data we gathered.

In this research we list a some of models was tested which is a sample of a many cases we checked to have the research results.

Keywords: Traffic Accidents, Data Mining, Data Warehouse, Knowledge Database, Clustering, Classification, Association Rules

*Master Degree- Computer and Automatic Control Engineering– Mechanical and Electrical Engineering Faculty – Tishreen University - Syria

مقدمة:

ينال البحث في حوادث المرور وكيفية معالجتها اهتمامًا عالميًا متناميًا، وذلك لما تسببه هذه الحوادث من استنزاف للموارد البشرية والمادية للدول، لدرجة أنه يمكن عدها من الأوبئة الفتاكة حسب تصنيف منظمة الصحة العالمية[1].

بالنسبة للدول العربية، فإن تقارير منظمة الصحة العالمية تصنف الدول العربية والإفريقية في المرتبة الأولى عالميا في مجال ضحايا حوادث السير، ونجد في سورية، أن الاتجاه العام لعدد حوادث المرور وما ينجم عنها من أضرار آخذ في الازدياد، خصوصا مع وجود تسهيلات في مجال استيراد السيارات وتصنيعها ، وعدم ملائمة البنية التحتية من طرق ونظم مرورية لهذا الكم الكبير من المركبات، مما يدعو إلى إيجاد خطط استراتيجية مدعومة بنظم اتخاذ قرار قادرة على استشراف المستقبل من خلال مؤشرات موجودة حاليا، وتحليل البيانات من أجل تحديد مراكز المشكلات، فضلا عن وجود عدد كبير من العوامل التي تسبب الحادث المروري بشكل مباشر أو غير مباشر[2].

مشكلة البحث:

من خلال الدراسات الاستكشافية والزيارات الميدانية المتعددة تم استنتاج الحاجة إلى وجود نظام معلوماتي يتناسب مع الوضع المروري في القطر وطبيعة المعلومات التي يتم تسجيلها في ضبط الحادث المعتمد مروريا، بحيث يكون قادرا على دراسة العوامل المسببة للحادث على فترة زمنية غير محددة، مع القدرة على الربط بين هذه العوامل وتقسيم الظواهر ذات العلاقة بالحادث إلى مجموعات تمتلك خواص فريدة بحيث يمكن وضع سياسات مدروسة لكل مجموعة للحد من الحوادث التي قد تتسبب بها.

أهمية البحث وأهدافه:

تكمن أهمية البحث في استخدام أدوات وتقنيات حديثة نسبيا في مجال قواعد البيانات من أجل تحليل حوادث المرور والتي أصبح عددها يتزايد بشكل مستمر في دول العالم مترافقا مع عبء اقتصادي واجتماعي كبير، وتشير توقعات منظمة الصحة العالمية إلى أن حوادث المرور سوف تكون المسبب الثالث للوفاة في عام 2020 مقارنة مع كونها المسبب التاسع في سنة 1990[3].

من المتوقع أن تسمح هذه التقنيات بتقديم مؤشرات لمتخذ القرار لبناء السياسات الإستراتيجية كما يقدم البحث في الجانب المعلوماتي اقتراحات للخوارزميات الأفضل في مجال حوادث المرور اعتمادا على مجموعة من مؤشرات قياس أداء هذه الخوارزميات.

يهدف البحث إلى تقديم بنية لنظام معلوماتي متكامل –قدر الإمكان– من أجل إدخال البيانات ذات العلاقة بالواقع المروري في سورية، من ثم تطبيق تقنيات تحليل و التنقيب في البيانات المناسبة عليها، للوصول إلى مجموعة من هذه التقنيات التي تسمح بالحصول على نتائج تقدم إلى متخذ القرار لاتخاذ القرارات المناسبة، وهذه النتائج يجب أن تقدم وفق صيغة يمكن فهمها وتفسيرها دون الحاجة للرجوع إلى التقني المختص بالتنقيب في البيانات، فضلا عن المرونة في الاستعلامات والوصول إلى تركيبات البيانات المطلوبة.

طرائق البحث ومواده:

تم اتباع منهجيات هندسية في تحليل وتصميم وتنفيذ نظم البحث، شمل ذلك هندسة البرمجيات من أجل نظم إدخال وتحليل البيانات، ومنهجيات اكتشاف المعرفة في قاعدة البيانات من أجل نظام التنقيب في البيانات.

تحديد عينة البحث وهي حوادث المرور في مدينة اللاذقية كان بسبب إمكانية تحليل النتائج بشكل أفضل وسهولة الحصول على البيانات نسبيا.

ثم تم اختيار خوارزميات التنقيب في البيانات لتغطي طيفا واسعا من عمليات اكتشاف المعرفة في قاعدة البيانات وتشكل حزمة متكاملة بدءا من الاكتشاف الداخلي للعلاقات بين البيانات في قاعدة البيانات واكتشاف التشابه بين الظواهر التي تغطيها هذه البيانات وصولا إلى تقنيات التصنيف التي تسمح بتقسيم البيانات ضمن فئات معرفة مسبقا وتعطي نوعاً من النظرة المستقبلية عن البيانات[4]، وتم تطبيق نماذج دراسة هذه الخوارزميات على برنامج SPSS Clementine V.12.

مجتمع وعينة البحث

البحث شمل محافظة اللاذقية وتم بناءَ على بيانات حوادث المرور في اللاذقية المتوفرة اختيار عينة من أربع أعوام للدراسة، بالإضافة إلى بيانات المكتب المركزي للإحصاء والتي تم الحصول عليها من المجموعات الإحصائية والمكتب المركزي للإحصاء، علما أن النظام يتعامل مع مستويات تفصيل مختلفة على أكثر من صعيد (الزمن والعناصر الأخرى المشتركة في الحادث المروري).

الدراسات السابقة:

الدراسة – Role of Road-related Mining Road Traffic Accident Data to Improve Safety Factors on Accident Severity in Ethiopia تحدثت عن تأثير حوادث المرور من الناحية الاجتماعية في أثيوبيا، وبنيت الدراسة على فرضية أن حوادث المرور هي مجموعة من العوامل التي تسبب الحوادث، تتعلق بالسائق والطريق والسيارة.. الخ، وليست حادثاً عرضياً أو عشوائياً.

هدفت الدراسة إلى إيجاد دور العوامل ذات العلاقة بالطريق من أجل إيجاد نموذج توقعي رياضي لحوادث الطرق في إثيوبيا، استخدمت برنامج ويكا من أجل تطبيق خوارزميات التنقيب في البيانات.

قدمت الدراسة مقارنة لصحة النماذج التي تم الحصول عليها باستخدام خوارزميات مختلفة من أجل تصنيف نتيجة الحادث[5].

الدراسة التي قام بها Chang and Chen عام 2005 التي اهتمت أيضا باستخدام أشجار القرار في تحليل حوادث المرور، حيث طبقا برنامجا للتنقيب في البيانات يقوم على بناء نموذج شجري لتحليل حوادث المرور على الطرق السريعة في تايوان، هما قاما ببناء شجرة انحدار ونموذج تصنيف يربط بين حوادث المرور والمتغيرات الهندسية للطرق السريعة. سميت الأداة Classification And Regression Tree Analysis (CART) وبينت دراساتهما أنها ذات فعالية من أجل تحديد تواتر حوادث المرور على الطرق السريعة[6].

أما في مجال العنقدة وتقسيم البيانات، فقد استخدم Hung and Wong في عام 2002، مزيج من تحليل العناقيد والتحليل الانحداري ونظم المعلومات الجغرافية من أجل جمع مقدار كبير من البيانات ودراسة تأثير الحوادث في الطرق في هونغ كونغ، هدفت الدراسة إلى توفير المعلومات والمعرفة للسلطات من أجل تحديد الموارد اللازمة لتحسين

شروط السلامة للحد من الحوادث الخطرة، وبينت تلك البيانات تحسن تقدير القيم المتوقعة لعدد الحوادث باستخدام نظم المعلومات الجغرافية والنظم الأخرى مقارنة مع استخدام البيانات التاريخية فقط في عملية التحليل[7].

كما نجد الدراسة التي قام بها Srisuriyachai في عام 2007 باستخدام خوارزمية K-Means للعنقدة و naïve bayes للتصنيف من أجل تحليل حوادث المرور في بانكوك. نتيجة البحث كانت تعبيراً عن مجموعة من فئات حوادث المرور والتي يمكن استخدامها من أجل تقييم حوادث المرور في المنطقة المدروسة[8].

لوحظ من دراستنا المرجعية غياب وجود دراسة عربية عن التنقيب في البيانات في مجال حوادث المرور، أما الدراسات الأخرى التي اطلعنا عليها فهي تقوم بدراسة خوارزمية معينة أو أحد جوانب التنقيب في البيانات، حيث أن الحاجة تبرز لوضع تصور لمنظومة متكاملة لتسجيل بيانات الحوادث والتنقيب فيها، وهذه المنظومة يجب أن تتناسب مع كل بلد حسب طبيعة البلد والمجتمع والبنية التحتية لمنظومة النقل (نقل عام، نقل خاص، شوارع، نوع المركبات،...)، ونوع التقارير والبيانات المتوفرة عن الحوادث، وهو ما يستدعي تطوير هذا النوع من الدراسات في سورية.

الإطار النظري للبحث [9,10,11,12,13,14,15,16,17]

1- تطور النظرة العلمية حول أسباب حوادث المرور[9]

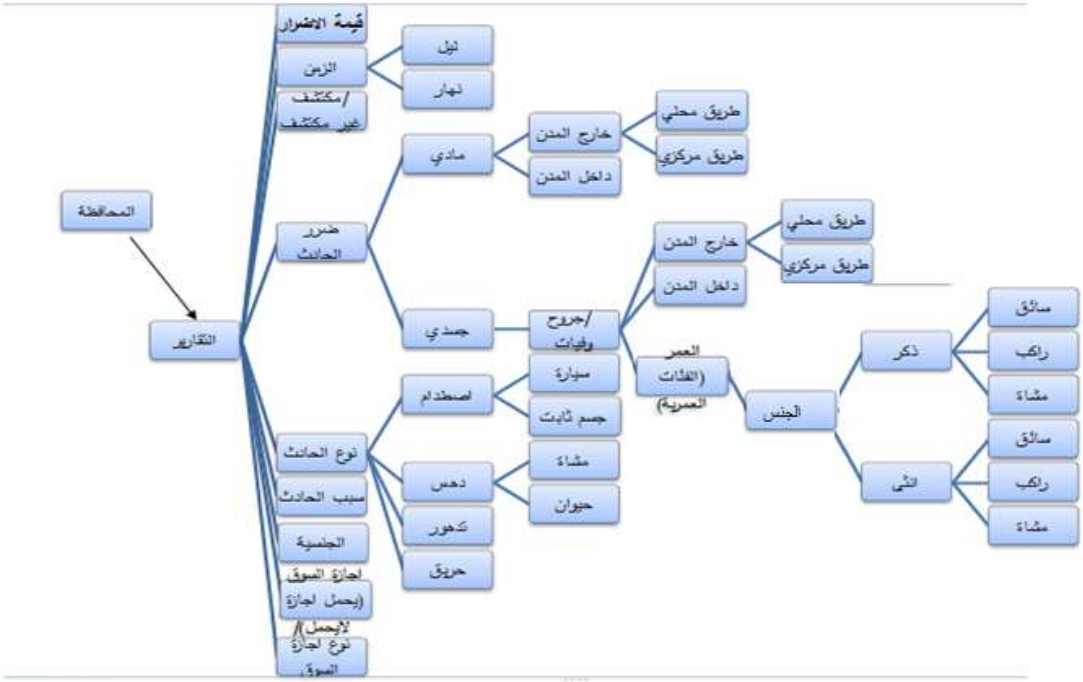
مع مرور الوقت ازداد الاعتراف بقصور النظرة أحادية السبب لحادث المرور مما ساهم في ظهور النظرة متعددة الأسباب في الفترة الواقعة بين الستينيات و الثمانينيات وقد ساهم في تكوين هذه النظرة بدء انتشار الحواسب الالكترونية واستعمالها لتحليل قواعد لبيانات الضخمة المتعلقة بحوادث المرور.

وحسب هذه النظرة لعبت عدة عوامل دورا في وقوع حادث المرور على شكل تأثير متبادل بين العوامل البشرية والبيئية والعوامل المتعلقة بالعربات و الطرق. يمكن التلخيص

تعتمد طريقة معالجة بيانات الحوادث في هذا البحث على مبدأ الدراسة الوبائية لحوادث المرور epidemiological study of road traffic accident وهي تستند إلى نفس الخلفية العلمية للأبحاث الوبائية في مجال الصحة العامة، حيث يتم في هذا النوع أخذ معطيات عن كل حادث تسمح بالتحليل الإحصائي لتحديد فئات الأشخاص أو المركبات الأكثر تعرضا للحوادث في ظروف معينة، ويتم التحري عن الخصائص المميزة لهذه الفئات من اجل توجيه جهود خاصة ونوعية لتحسن ظروف السلامة المتعلقة بها، هذا النوع من الدراسات هو الذي سمح على سبيل المثال بكشف مدى خطورة ظاهرة حوادث المرور لفئات العمر الشابة تحديدا وكذلك ظاهرة تزايد حوادث المرور لدى فئات السائقين المعمرين في مجتمعات الشمال الأكثر تطورا[9].

2- البنى المعلوماتية المستخدمة للنظام

تدخل في حادث المرور حسب النظرة المتعددة الأسباب كما ذكر سابقا مجموعة من العوامل وهذه العامل تجتمع معا لتشكل البيئة المحيطة بالحادث، وحسب النظام المستخدم في إدارة فرع المرور، ومن خلال دراستنا لنماذج التقارير الإحصائية لحوادث المرور في محافظة اللاذقية، أمكننا هيكلة معلومات التقارير على الشكل(1):



الشكل (1) هيكلية عناوين تقارير حوادث المرور حسب النماذج الإحصائية المعتمدة في فرع مرور اللاذقية

تم إسقاط هذه الهيكلية على مكعب فائق *Hyper Cube* لمستودع بيانات المشروع يتكون من الأبعاد المبينة في الجدول (1)، وعلى أساس هذه الأبعاد تم بناء قاعدة البيانات الخاصة بتخزين معلومات التقارير الإحصائية التي حصلنا عليها من فرع مرور اللاذقية، حيث يمثل كل بعد جدولا تعريفيا *Definition Table* ترتبط معه بيانات الحادث:

الجدول (1) الأبعاد التي تم تحديدها لمكعب البيانات اعتمادا على تحليل التقارير الإحصائية لحوادث المرور

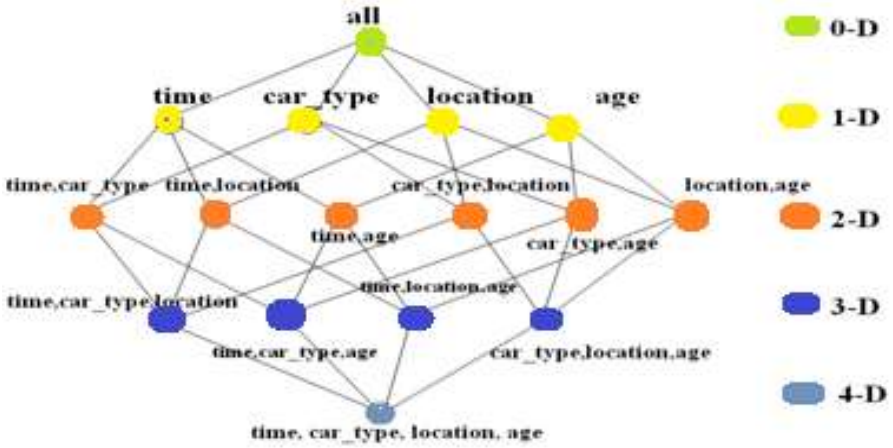
المكان(مدينة الحادث). يتضمن هرمية : منطقة المدينة – المدينة	الأضرار الناتجة عن الحادث وتم تقييمه إلى الضرر الرئيسي والضرر الفرعي.	التاريخ. يتضمن هرمية: السنة– الفصل – الربع – الشهر – اليوم (في نظامنا يتم اعتماد الوصول في الهرمية إلى مستوى الشهر فقط)	فئة السيارات المشتركة في الحادث (تم تقييمه إلى النوع الرئيسي للسيارات والنوع الفرعي).
نوع الحادث (تم تقييمه إلى النوع الرئيسي للحادث والنوع الفرعي).	الطريق وتم تقييمه إلى فئة الطريق ومعلومات الطريق.	مدة الحصول على الشهادة.	أعمار ضحايا الحوادث.
شهادة السوق.	مستخدم الطريق.	الجنسية.	سبب الحادث.

بنية مستودع البيانات المستخدمة هي *Snow Flake Schema* حيث تم تطبيق قوانين التقييس *Normalization* على الأبعاد من خلال وجود جداول تصنيف رئيسية مرتبطة مع جداول فرعية، مقارنة مع النموذج النجمي *Star Schema* والذي يتكون بشكل أساسي من جدول للحقيقة *fact table* متصل مباشرة مع جداول الأبعاد.

مثلا، الجدول الخاص ببعد أضرار الحادث، تم تقييسه ليصبح جدول نوع الضرر الرئيسي وجدول نوع الضرر الفرعي. إن اختبار هذه البنية سوف يزيد من تعقيد الاستعلامات والزمن اللازم لعمليات الربط بين الجداول، ولكن من ناحية أخرى سوف يوفر المرونة المطلوبة لتوسيع نظام تخزين بيانات حوادث المرور مستقبلا، مثلا يمكن توسيع مفهوم ضرر الحادث في النظام من خلال الربط مع جدول جديد يخزن معلومات التكلفة الاقتصادية التقديرية لكل نوع ضرر للحادث.

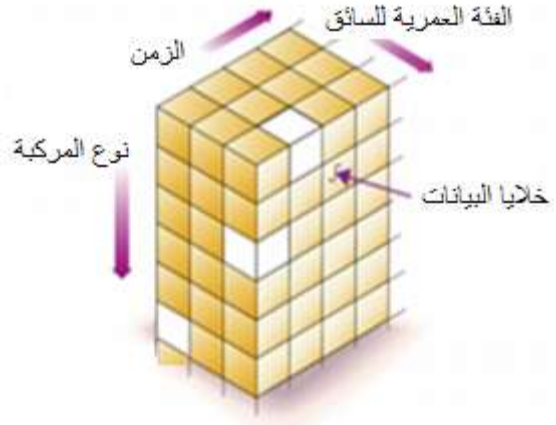
يبين الشكل (2) مثالا عن شبكة المناظير التي يمكن توليدها من مكعب ذي أربعة أبعاد وهي عمر السائق و موقع الحادث ونوع السيارة والزمن، يمكن إدخال كامل العوامل التي تدخل في الحادث مثل نوع الحادث وغيرها لتكون جزءا من المكعب، وتمثل كل نقطة في هذه الشبكة قيمة مجمعة *Aggregated Value* يستطيع نظامنا لتحليل البيانات تصديرها إلى منظومة التنقيب في البيانات لتطبيق الخوارزميات المختلفة عليها.

على سبيل المثال، تبين العقدة *all* العدد الكلي لحوادث المرور بغض النظر عن العوامل التي تلعب دورا في الحادث، أما العقدة *car_type* فهي عدد حوادث المرور حسب نوع السيارة، العقدة *time,location* تمثل عدد حوادث المرور في كل موقع ومن أجل كل فترة زمنية.



الشكل (2) المناظير المختلفة التي يمكن توليدها من مكعب لتحليل الحوادث بأربعة أبعاد

إن نظام التنقيب في البيانات المصمم من قبلنا سوف يسمح بالنظر إلى البيانات من مختلف الأبعاد من اجل الحصول على معلومات شاملة عن ظاهرة حوادث المرور.



الشكل (3) المناظير المختلفة التي يمكن توليدها من مكعب لتحليل الحوادث بأربعة أبعاد

فمثلا في نظامنا حسب الهيكلية المعتمدة في تقارير حوادث المرور التي ذكرناها، وحسب تصميم قاعدة البيانات والتي تتكون من جدول الحقيقة وعدد أبعاد 12 مبينة في الجدول (1)، فإن عدد المناظير التي يمكن توليدها للبيانات هو $4096=2^{12}$ منظور للبيانات.

3- الخوارزميات المستخدمة في النظام[10,11,12,13,14,15,16,17]

في مجال التنقيب في البيانات، يمكننا الحديث عن 3 تقنيات أساسية وهي:

- التصنيف *Classification*.
- العنقدة *Clustering* التي تسمى أحيانا التجزئة *Segmentation*.
- التنقيب في قواعد الاقتران *Association rules mining*.

التصنيف هو مهمة إسناد الكائنات (*object*) إلى واحد أو أكثر من الفئات مسبقة التعريف، ومن الأمثلة على مهام التصنيف ضبط رسائل البريد الإلكتروني المزعجة (*spam*) استنادا إلى ترويسة الرسالة ومحتواها، وتصنيف الخلايا على أنها خبيثة (*malignant*) أو حميدة (*benign*) استنادا إلى نتائج فحوص الرنين المغناطيسي *Magnetic Resonance Imaging* (*MRI*)، وتصنيف المجرات بناء على شكلها.

العنقدة تقسم البيانات إما إلى جماعات (عناقيد) لها معنى *(meaningful)*، أي لها تمثيل طبيعي في الواقع، مثلا تقسيم مجموعة من حوادث المرور وفق الطبيعة الزمانية أو المكانية للحادث بشكل أساسي، و إما إلى جماعات مفيدة *(useful)*، أي مجموعات يمكن تمثيلها وتلخيصها *Data Summarization* من خلال عينة معينة، وهنا تكون هذه العينة هي مركز هذه العناقيد *Cluster Center*، مثلا التعبير عن مجموعة من حوادث المرور من خلال عينة محددة من هذه الحوادث تماثل العينة الإجمالية من ناحية الصفات الأساسية، أو إلى كلا التصنيفين السابقين.

أما حالات الاقتران فهي تعد مفيدة من أجل اكتشاف علاقات مهمة مخفية في مجموعات بيانات ضخمة. يمكن تمثيل العلاقات المكتشفة بشكل قواعد اقتران *(association rules)* أو مجموعات من البنود المتكررة *(frequent items)*[10].

تمت دراسة مجموعة من الخوارزميات الخاصة بكل تقنية وإجراء مقارنات بينها لإيجاد الخوارزمية الأنسب وفق مجال البحث وهو تحليل حوادث المرور.

3-1- خوارزميات التصنيف [11,12]

بالنسبة لخوارزميات التصنيف تمت دراسة خوارزميات *CART,CHAID,SVM*

CART هي اختصار لـ أشجار التصنيف والانحدار *Classification and Regression Tree*.

CHAID هو اختصار لـ *Chi-squared Automatic Interaction Detector*. وهو طريقة إحصائية عالية الكفاءة في التقسيم إلى قطاعات *segmentation* أو للتحكم بنمو الشجرة.

خوارزمية *SVM* اختصار لـ *Support Vector Machine* و تعمل بشكل جيد مع البيانات التي لها عدد كبير من الأبعاد. هناك جانب أخر لهذه الطريقة هو أنها تمثل حد القرار باستخدام مجموعة جزئية من أمثلة التدريب تعرف بمتجهات الدعم (*support vectors*).

الشكل (4) يبين التمثيل البياني لتوابع مقاييس عدم النقاء *Entropy, Gini, Misclassification error*، حيث تصف معايير عدم النقاء مقدار تماثل السجلات ضمن العقدة. تم اختيار معيار *GINI* لدراسة عدم نقاء العقدة

في شجرة التصنيف، ونجد أن عدم النقاء يبلغ أقصى قيمة بالنسبة لهذا المقياس في النقطة (0.5,0.5) والتي تعني أن السجلات الموجودة في هذه العقدة بنسبة معينة من المحتمل أن تنتمي لعقدة أخرى بالنسبة نفسها.

المعيار *GINI* عند العقدة *t* من أجل فئات مستهدفة هي *i* و *j*، يعطى وفق المعادلة *[11]*:

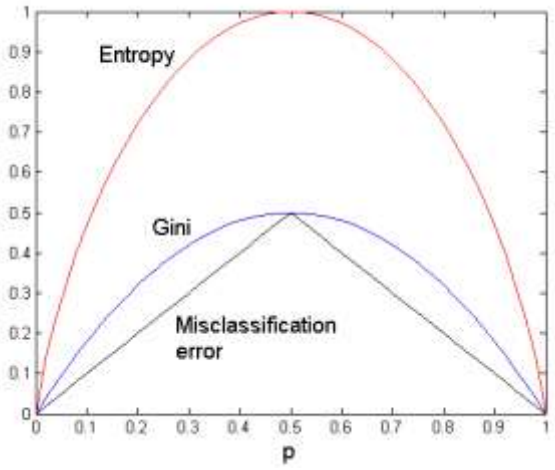
$$g(t) = \sum_{j \neq i} p(j \mid t) p(i \mid t) \qquad (1)$$

حيث *i*، *j* هي تصنيفات الحقل المستخدم للتقسيم، و

$$p(j \mid t) = \frac{p(j,t)}{p(t)} \qquad (2)$$

$$p(j,t) = \frac{\pi(j) N_j(t)}{N_j} \qquad (3)$$

$$p(t) = \sum_j p(j,t) \qquad (4)$$



الشكل (4) التمثيل البياني لمقاييس عدم النقاء

كما تم اختيار معيار ملخص الربح *gain summary* من أجل تقييم أداء الخوارزمية[12] .

يقدم ملخص الربح إحصائيات وصفية للعقد النهائية من الشجرة، وإذا كان الحقل المستهدف بعملية التصنيف هو حقل مستمر *(scales)* حينها سوف يبين ملخص الربح المتوسط الموزون (التثقيل) للقيمة المستهدفة لكل عقدة نهائية.

$$g(t) = \sum_{i \in t} w_i f_i x_i \qquad (5)$$

حيث w_i هو الوزن للسجل *i* ويأخذ القيمة 1 إذا لم يتم تحديد وزن للسجل، f_i هي قيمة حقل التكرار للسجل *i*، x_i هي قيمة الحقل الهدف للسجل *i*.

وإذا كانت الحقل المستهدف هو رمزي (فئوي) حينها سوف يبين ملخص الربح النسبة المئوية المثقلة للسجلات في الفئة المستهدفة التي تم اختيارها:

$$g(t,j) = \frac{\sum_{i \in t} f_i x_i(j)}{\sum_{i \in t} f_i} \qquad (6)$$

حيث $x_i(j)=1$ إذا كان السجل x_i في الفئة المستهدفة *j*، وإلا ستكون قيمتها صفر، وإذا تم تحديد مقدار الفائدة *profit* في الشجرة، عندئذ سيكون الربح هو متوسط قيمة الفائدة لكل عقدة نهائية :

$$g(t)=\sum_{i\in t} f_i P(x_i) \qquad (7)$$

حيث $P(x_i)$:هي قيمة الفائدة المرتبطة مع القيمة الهدف والموجودة في السجل x_i .

3-2- خوارزميات قواعد الاقتران[13]

بالنسبة لخوارزميات قواعد الاقتران تم دراسة خوارزمية *Apriori* وخوارزمية *GRI*.

تم اختيار معيار الدعم و الثقة مع عدد القواعد المستنتجة كمؤشرات لأداء الخوارزميات وكمعيار يتم من خلاله اقتصاص *(trim)* فضاء البحث الأسي استنادا إليها وهو ما يسمى بـ"التشذيب استنادا إلى الدعم" *(Support- based Pruning)* [13].

يعرف الدعم بالقانون (8):

$$s=\text{Support}=\frac{\text{number of transactions containing both A and B}}{\text{total number of transactions}} \qquad (8)$$

والثقة:

$$c=\text{Confidence}=\frac{\text{number of transactions containing both A and B}}{\text{number of transactions containing A}} \qquad (9)$$

يستخدم المقياس *J* الكمي لتحديد أهمية القاعدة في خوارزمية *GRI*، بدلا من المعيارين السابقين المستخدمين من أجل خوارزمية *Apriori*، ويعرف من خلال القانون (10):

$$J=p(x)\left[p(y\mid x)\ln\frac{p(y\mid x)}{p(y)}+[1-p(y\mid x)]\ln\frac{1-p(y\mid x)}{1-p(y)}\right] \qquad (10)$$

p(y) هي احتمال أن يطابق الجزء الأول من الشرط لمثال ما من قاعدة البيانات.

p(x) هي احتمال إن يطابق الجزء الذي يلي *then* من القاعدة المثال من قاعدة البيانات.

p(y|x) هي الاحتمال الشرطي للجزء اللاحق من الشرط مشروطا بالجزء الأول منه.

3-3 - خوارزميات العنقدة [14,15,16,17]

بالنسبة لخوارزميات العنقدة تم دراسة خوارزميات *K-means,two-steps,kohenen*.

K-Means والتي تعد من أبرز خوارزميات العنقدة، وهي تقنية عنقدة جزئية استنادا إلى نموذج الأصل تحاول إيجاد عدد يحدده المستخدم (*K*) من العناقيد الممثلة بواسطة مركز ثقلها[14]، وخوارزمية *two-step* التي تعتبر

خوارزمية لتحليل العناقيد تم تصميمها للتعامل مع مجموعات كبيرة جداً من البيانات، وتستطيع التعامل مع كل من المتغيرات المستمرة والفئوية، كما تتميز بقدرتها على التحديد التلقائي لعدد العناقيد مما يسمح بتحديد تركيز امثلي لنقاط البحث في المشكلة[15].

أما خوارزمية *Kohonen*، فهي نوع خاص من نماذج الشبكات العصبية التي تنفذ "التعلم غير المراقب" *Unsupervised Learning*، فهي تأخذ أشعة المدخلات وتنجز نوعاً من العنقدة المنظمة مكانياً (فضائياً)[16].

تم استخدام المعاينة للمسافات الإقليدية بين العناصر من أجل دراسة توزع السجلات على العناقيد وتحديد مدى مطابقة التوزيع الذي تم الحصول عليه مع التوزيع الحقيقي للقيم.

في كل تكرارية لخوارزمية العنقدة يتم توزيع كل سجل على العنقود ذي الأقرب مركزاً ويقاس هذا القرب باستخدام مربع المسافة الإقليدية:

$$d_{ij} = \left\| X_i - C_j \right\|^2 = \sum_{q=1}^{Q} (x_{qi} - c_{qj})^2 \qquad (11)$$

حيث x_i هي شعاع حقول الإدخال المرمزة *Encoded input fields* للسجل i، C_j هي شعاع مركز العنقود للعنقود j، Q عدد حقول الإدخال المرمزة، x_{qi} هي قيمة حقل الإدخال المرمز ذي الترتيب q للسجل ذو الترتيب i، C_{qj} هي قيمة حقل الإدخال المرمز ذات الترتيب q للسجل ذي الترتيب j.

في حالة اعتماد الفضاء الإقليدي لقياس المسافة فإننا نستخدم مجموع مربعات الأخطاء *SSE (sum of squared error)* باعتباره التابع الهدف الذي يقيس جودة العنقدة[17].

$$SSE = \sum_{i=1}^{k} \sum_{x \in C_i} d(x, c_i)^2 \qquad (12)$$

d: المسافة الإقليدية القياسية بين كائنين في الفضاء الإقليدي.

X: كائن.، C_i: العنقود رقم i.، c_i: مركز العنقود رقم i.، K: عدد العناقيد.

النتائج والمناقشة:

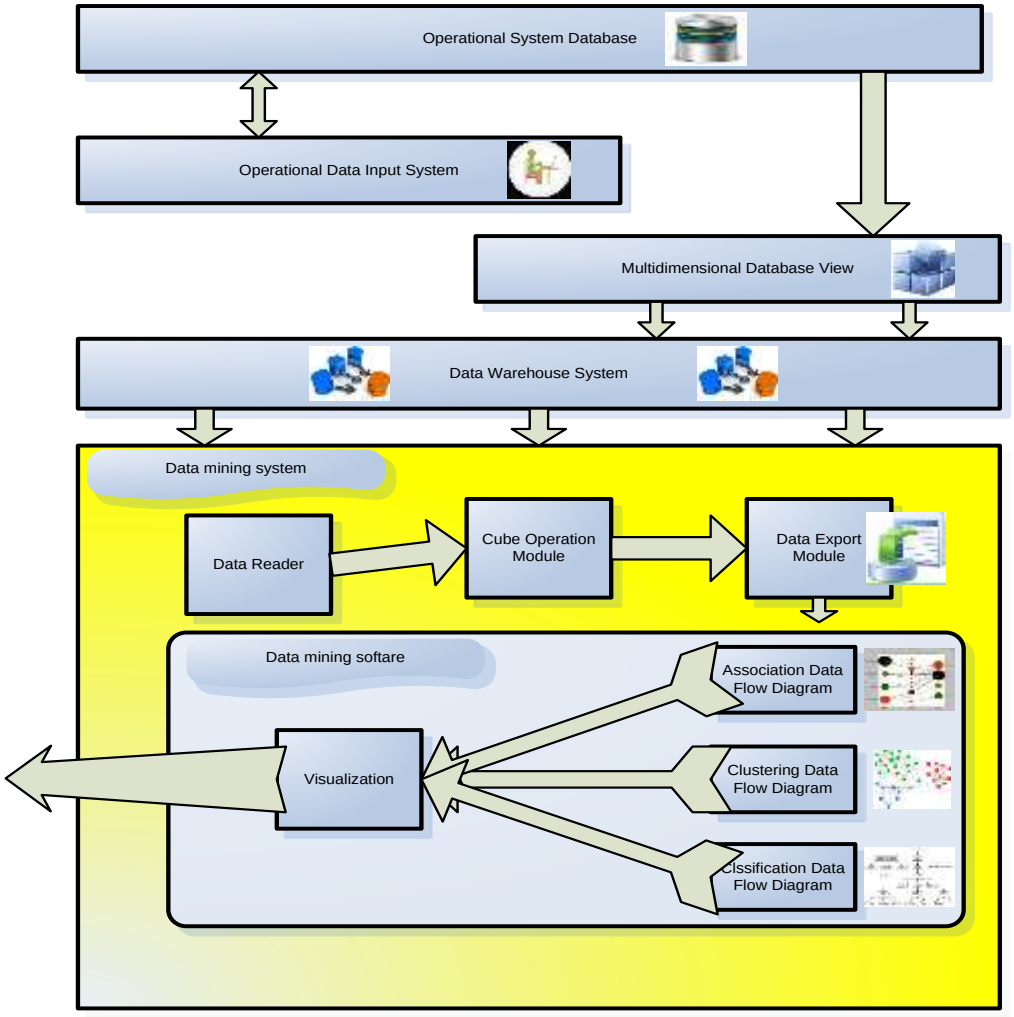
تم استخدام برمجية *SPSS Clementine V.12* من أجل نظام التنقيب في البيانات، حيث قمنا بتصميم نماذج متعددة للخوارزميات المدروسة، وتمرير البيانات إليها من نظام التعامل مع البيانات متعددة الأبعاد.

حجم عينة البيانات كان 6000 سجل لحوادث المرور في اللاذقية والتي تم إدخالها خلال الفترة المدروسة، تم تقسيمها بالنصف من أجل التدريب والاختبار.

1- المخطط الصندوقي للنظام

يبين الشكل (5) المخطط الصندوقي للنظام الذي قمنا بتصميمه بناء على ما ذكرناه في فقرات البحث، وهو يبدأ من تصميم قاعدة البيانات *Operational System Database* بناء على الأبعاد التي تم تحديدها وفق هيكلية العناوين في تقارير حوادث المرور الإحصائية، والتي تتصل مع نظام إدخال البيانات العملياتية إلى النظام *Operational Data Input System*، يقوم نظام التعامل مع البيانات متعددة الأبعاد بالحصول على البيانات من قاعدة البيانات وفق مستويات التجميع المطلوبة أي حسب إحدى مناظير قاعدة البيانات ومن ثم تصدير النتيجة إلى مستودع البيانات.

في مستودع البيانات يتم تمرير هذه البيانات إلى نماذج التنقيب في البيانات المختلفة التي تقوم بمعالجة البيانات ومن ثم إظهارها بشكل مرئي ومفهوم على شكل مخططات بيانية أو قواعد للدارس أو متخذ القرار.

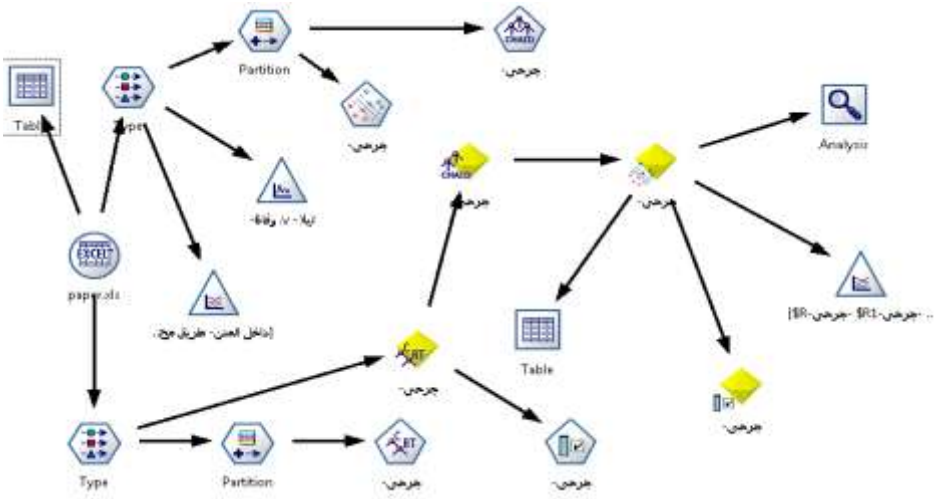


الشكل(5) المخطط الصندوقي للنظام

2- نتائج تطبيق خوارزميات التصنيف Classification Algorithms Results

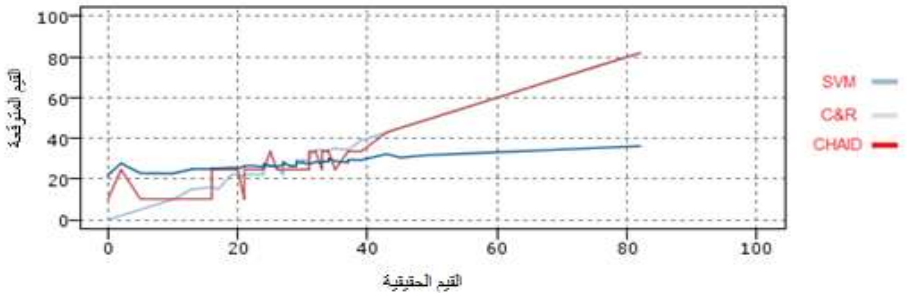
تم تطبيق عدد من نماذج تدفق البيانات وذلك لدراسة خوارزميات التصنيف، يبين الشكل (6) أحد النماذج التي تم تصنيفها من خلال النظام لتصنيف أعداد الجرحى بناء على مكان الحادث (داخل المدن – خارج المدن / طريق مركزي – طريق فرعي/) وزمن الحادث (ليلا – نهارا).

في النموذج المدروس والمبين في الشكل (6)، تم إدخال البيانات من خلال قاعدة البيانات لحوادث المرور في مدينة اللاذقية والتي أشرنا إليها سابقا، ومن ثم تم تمرير البيانات إلى نظام التعامل مع البيانات متعددة الأبعاد، ثم إلى عناصر تقوم بتحليل نوعية البيانات لتحدد طريقة نمذجتها (حقول مستمرة – فئوية –..)، ومن ثم تم إخضاعها لمجموعة من عمليات التقسيم بين عينات التدريب وعينات الاختبار، بالإضافة لعمليات فلترة قبل تمريرها على الخوارزميات المطلوبة، وأخيرا تم تمرير نتائج الخوارزميات على عناصر العرض *Display Components* لتمثيل النتائج بشكل بياني.



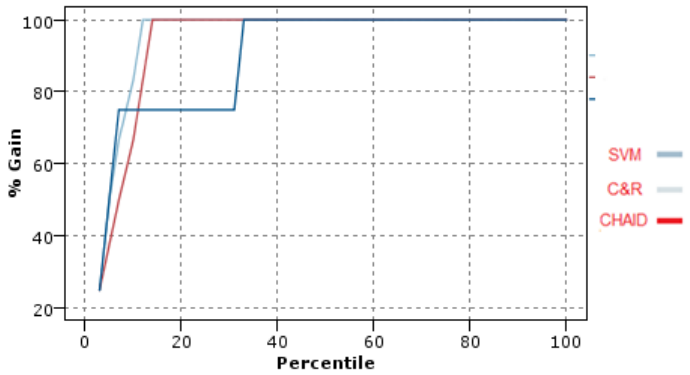
الشكل (6) أحد نماذج تدفق البيانات لخوارزميات التصنيف

يبين تطبيق خوارزميات التصنيف السابقة أن خوارزمية *CART* و *CHIAD* حققت نتيجة أفضل من خوارزمية *SVM* وذلك من خلال دراسة الارتباط الخطي *linear correlation* بين القيم المتوقعة والقيم الحقيقية، وتظهر الخطوط البيانية اقتراب الارتباط بينهما إلى الحالة المثالية في كثير من الأحيان مقابل ابتعاد القيم المتوقعة لخوارزمية *SVM* عن القيم الحقيقية وخصوصا من أجل القيم الكبيرة ويتضح ذلك في الشكل (7).



الشكل (7) مخطط الارتباط بين القيم الحقيقية والقيم المتوقعة من قبل المصنف

كما يبين الشكل (8) الذي يعرض مخطط الربح *gain chart* أن النموذج الخاص بالخوارزميتين يصل إلى حالة الاستقرار بشكل مبكر أكثر من خوارزمية *SVM* مع التقدم في حجم عينة التدريب، وغالبا يكون ذلك بعد تجاوز الربع الأول من حجم عينة التدريب.



الشكل (8) مخطط الربح للنموذج السابق

يبين الجدول (2) تحليل نتائج الخوارزميات من أجل النموذج المدروس والمولدة من قبل نظام *SPSS Clementine* ونلاحظ أن القيم الخاصة بالخطأ وهي:

- متوسط الخطأ *Mean Error*
- مقدار الخطأ ألأعظمي *Maximum Error*
- مقدار الخطأ الأدنى *Minimum Error*
- متوسط الخطأ المطلق *Mean Absolute Error*
- الانحراف المعياري للأخطاء *Standard Deviation*
- الارتباط الخطي للنموذج المولد *Linear Correlation*

تميل إلى تحقيق نتائج أفضل لصالح *CART* و *CHIAD*.

الجدول (2) قيم الخطأ في النموذج السابق من أجل الخوارزميات الثلاثة المدروسة

	القيم من أجل خوارزمية CART	القيم من أجل خوارزمية CHAID	القيم من أجل خوارزمية SVM
Minimum Error	-16.0	-16.2	-24.285
Maximum Error	11.0	18.8	50.088
Mean Error	0.372	0.791	0.47
Mean Absolute Error	2.419	4.549	7.9
Standard Deviation	4.546	7.082	12.184
Linear Correlation	0.949	0.868	0.833
Occurrences	43	43	43

بالإضافة إلى ما سبق، فإن الملاحظة التفاعلية *interactive observation* للنماذج المدروسة للمؤشر *index* والمبينة في الشكل (9) والشكل (10) تبين أنه يحقق نتائج متقاربة من أجل مجموعة التدريب لخوارزمية *CART* و*CHAID*، أما من أجل مجموعة الاختبار فإن خوارزمية *CART* حققت نتائج أفضل بشكل عام، ويظهر ذلك من خلال القيم المرتفعة للمؤشر *index* وهو ما يسمح بإيقاف تفصيل الشجرة عند ارتفاعات أقل مع دقة مقبولة للمصنف، حيث أن المؤشر *Index* هو النسبة المئوية للاستجابة *Response* (وتعرف الاستجابة على أنها هي النسبة المئوية للسجلات في العقدة الحالية والتي تقع ضمن فئة العقدة. ويشار إليه أحيانا بالمصطلح *Hits*) في العقدة الحالية. ويعبر عنها كنسبة مئوية من النسبة المئوية للاستجابة من أجل كامل قاعدة البيانات[17].

على سبيل المثال، إذا كانت الاستجابة لمجموعة من السجلات هي 300% ، فإن احتمال أن تنتمي هذه السجلات إلى هذه العقدة فعلا، هو ثلاثة أضعاف النسبة، وذلك مقارنة مع سجلات قاعدة البيانات ككل.

Training Sample

Nodes	Node: n	Node (%)	Mean	Index (%)
6	1.00	2.33	21.00	340.75
5	1.00	2.33	18.00	292.08
11	9.00	20.93	12.89	209.14
13	1.00	2.33	12.00	194.72
12	3.00	6.98	10.67	173.08
14	6.00	13.95	10.17	164.97
8	1.00	2.33	5.00	81.13
1	21.00	48.84	0.00	0.00

الشكل (9) قيم المؤشر *Index* لخوارزمية *CART* من أجل أحد النماذج المدروسة

Training Sample

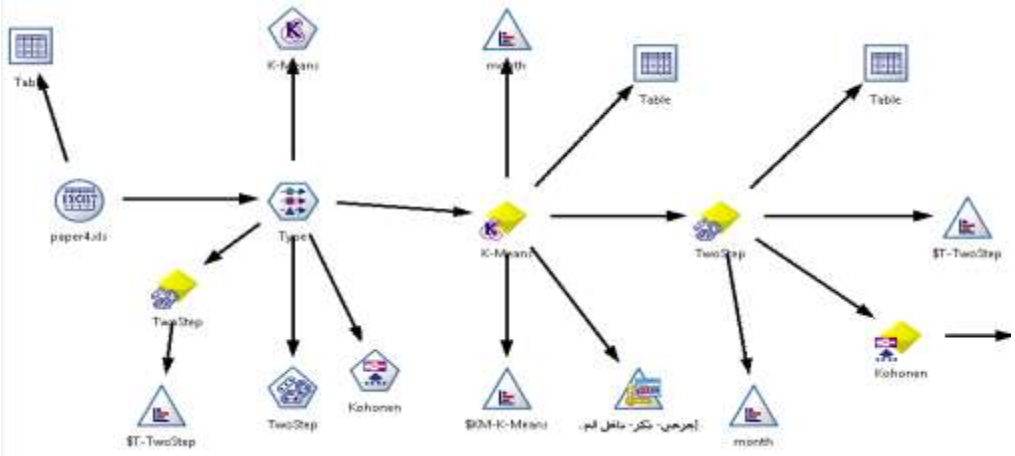
Nodes	Node: n	Node (%)	Mean	Index (%)
3	1.00	2.33	21.00	340.75
5	20.00	46.51	11.95	193.91
6	1.00	2.33	5.00	81.13
1	21.00	48.84	0.00	0.00

الشكل (10) قيم المؤشر *Index* لخوارزمية *CHIAD* من أجل نفس النموذج

3- نتائج تطبيق خوارزميات العنقدة Clustering Algorithms Results

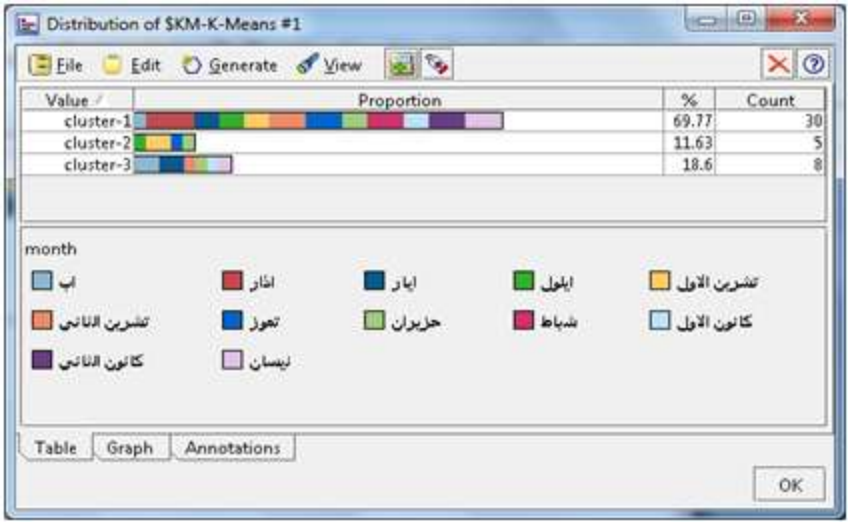
بشكل مشابه لخوارزميات التصنيف تم بناء عدة نماذج للعنقدة باستخدام خوارزميات *Two -step, K-mean, Kohenen*، وتم دراسة توزع البيانات ضمن العناقيد ومحاولة التقريب بين الأسس الطبيعية للتوزع والتقسيم الحاصل للبيانات في كل عنقود. بعد اختبار عدة عوامل كمقياس، ارتأينا أخذ عامل الزمن وتحديدا الأشهر لدراسة توزع الأشهر على العناقيد واختبار قدرة الخوارزميات على معرفة الظواهر الشهرية لحوادث المرور.

الشكل (11) يبين تدفق البيانات في أحد نماذج الاختبار التي تم تصميمها لدراسة خوارزميات العنقدة باستخدام نظام *SPSS Clementine*، حيث المدخلات هي مكان الحادث وجنس السائق ونتيجة الحادث والزمن والتي تم تصديرها من نظام التعامل مع البيانات متعددة الأبعاد، ويربط النموذج بين الخوارزميات الثلاث المدروسة، ونستخدم عقدة الرسم البياني لدراسة توزع الحوادث زمنيا على العناقيد.



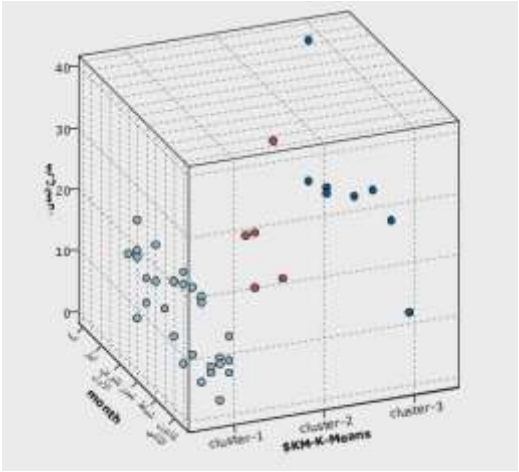
الشكل (11) نموذج العنقدة لسجلات تتضمن زمن الحادث ونتيجته ومكانه وجنس السائق

الشكل (12) يبين محتويات العناقيد على شكل مخطط لتوزع قيم حقل معين ضمن هذه العناقيد، اعتمدنا في بحثنا بشكل عام دراسة توزع الأشهر ضمن العناقيد ومقارنتها مع ترتيب الأشهر في السنة، بالإضافة إلى مقارنة نسبة توزع الأشهر في العناقيد التي يولدها النموذج.



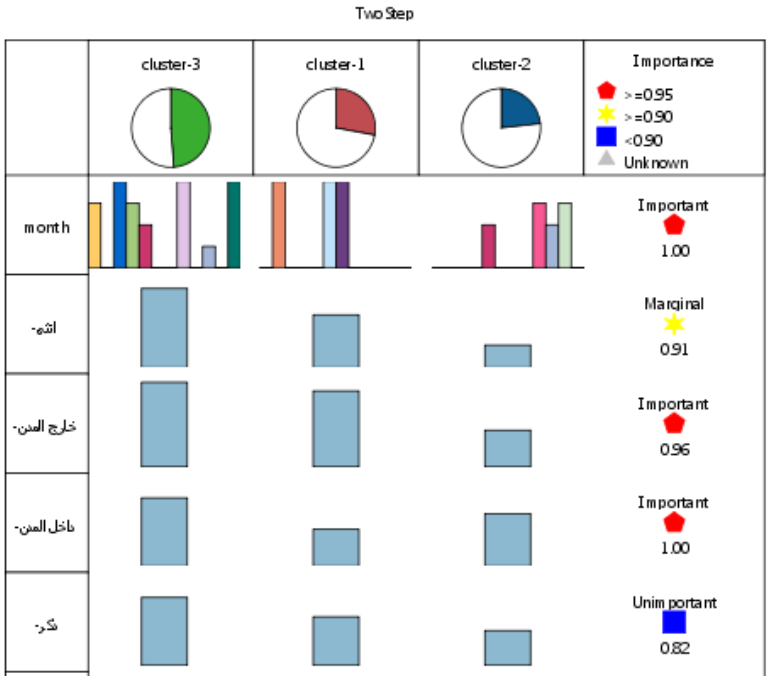
الشكل (12) **توزع الأشهر في العناقيد من أجل خوارزمية** *K-Means*

تسمح النظرة الثلاثية الأبعاد إلى محتويات العناقيد كما هو مبين في الشكل (13) بالاطلاع على البيانات المحتواة في العناقيد من أجل أكثر من حقل بيانات، ويمكن أن تعكس التجمعات في مكعب البيانات التوزيع الطبيعي للظاهرة من أجل العوامل المدروسة.



الشكل (13) **نظرة ثلاثية الأبعاد على سجلات العناقيد وفق مكان الحادث وزمن الحادث**

ونلاحظ أن مخرجات النظام قادرة على تحديد أهم العوامل التي تلعب دورا في تشكيل وتحديد العضوية ضمن العناقيد، وبالتالي يمكننا من خلال النظام تجميع البيانات وتحديد العوامل الأساسية التي يجب التركيز عليها لدى دراسة هذه البيانات، فمن الشكل رقم (14)، نرى أن زمن ومكان الحادث يلعب دورا في تحديد أعداد الحوادث أكثر من جنس السائق وذلك من خلال قيم المؤشر *Importance* التي قام برنامج التنقيب عن البيانات بتحديدها لدى تحليله لعينة البيانات.



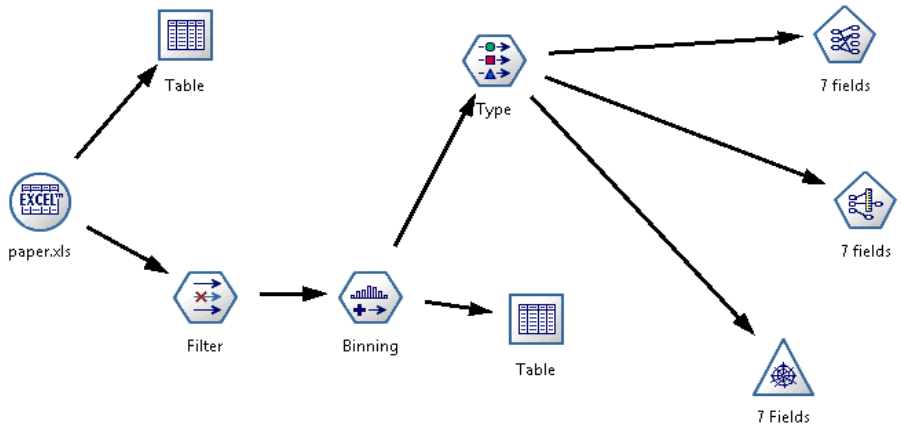
الشكل (14) أهمية العوامل المدروسة في توزيع البيانات على العناقيد من أجل أحد النماذج المدروسة

من خلال دراسة نتائج نماذج الاختبار، والتي بينا أحدها في الشكل (12)، تبين لنا أن خوارزمية *Two Step* وخوارزمية *kohonen* حققت أداء أفضل من خوارزمية *K-Means* حيث استطاعت بشكل عام إنشاء عناقيد أكثر نقاء للبيانات، إذ قامت تلقائيا بتجميع البيانات وفق الأشهر المتقاربة زمنيا ومن أجل سنوات مختلفة، فضلا عن أن قدرة الخوارزمية على تحديد العدد الأمثلي للعناقيد يسمح بالحصول على تجميع أكثر دقة للظاهرة المدروسة وبالتالي معالجة أكثر تركيزا للظاهرة وأسبابها.

على الرغم من أن خوارزمية *Kohonen* تعطي عدد أكبر من العناقيد مقارنة مع العدد التلقائي الذي تعطيه خوارزمية *Two- step*، إلا أنه وحسب التجربة فقد استطاعت إعطاء أشهر أكثر نقاء واستطاعت التعرف على الأشهر بنسبة كاملة في عدد لا بأس به من الحالات.

4- نتائج تطبيق خوارزميات التنقيب عن قواعد الاقتران AR Algorithms Results

تم بناء عدة نماذج لتحليل قواعد الاقتران من أجل تشكيلة مختلفة من البيانات. النموذج المبين في الشكل (15)، كانت مدخلاته الممررة من نظام التعامل مع البيانات متعددة الأبعاد هي ضرر الحادث والفئة العمرية لمستخدم الطريق، بالإضافة إلى الزمن، والعقد النهائية تمثل الخوارزمية المستخدمة.



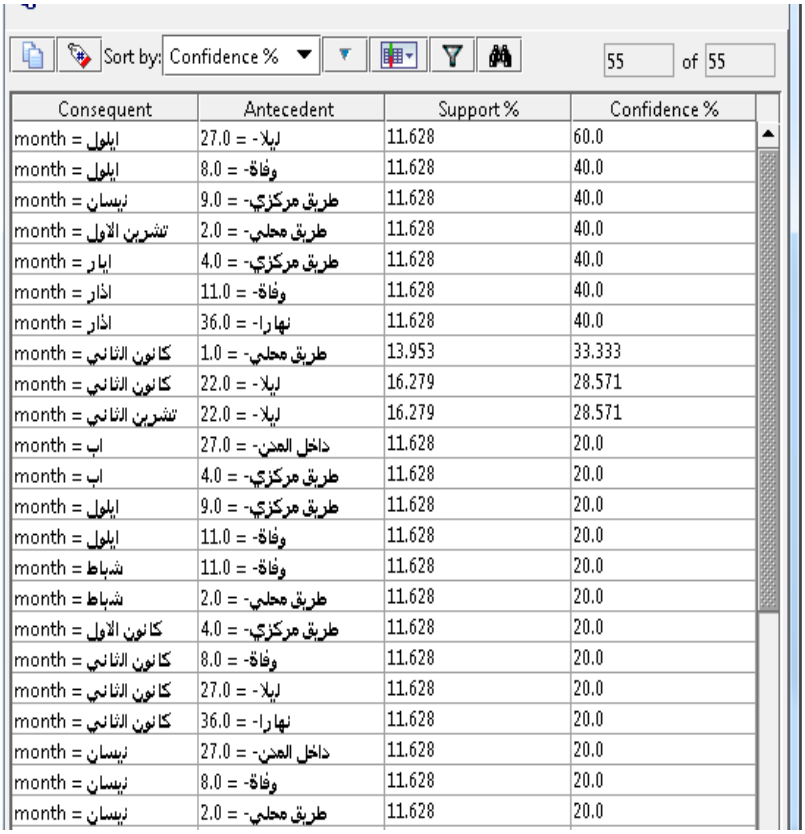
الشكل (15) نموذج دراسة حالات الاقتران

الإعدادات الأولية للخوارزميات، وهي الإعدادات الافتراضية في برمجية *SPSS Clementine* والتي وجدنا أنها كانت مناسبة لحجم العينة المدروسة في البحث، كانت كالتالي:

- الحد الأدنى للدعم *minimum support* : 10%.
- الحد الأدنى للثقة *minimum confidence*: 10%.
- العدد الأعلى للقواعد *maximum number of rules*: 200 قاعدة.
- العدد ألأعظمي للأجزاء السابقة من الشرط *maximum number of antecedents*: 5

الشكل (16) يبين عينة من نتائج تطبيق النموذج السابق، وهو يظهر نسبة الدعم والثقة للقاعدة، ويظهر طرفي القاعدة. بالاطلاع على النتائج المذكورة، وجد أنه:

- استطاعت خوارزمية *Apriori* اكتشاف 30 قاعدة ضمن البارامترات المحددة، مقارنة مع خوارزمية *GRI* التي استطاعت إيجاد 177 قاعدة.
- بالنسبة لخوارزمية *Apriori* فإن 5 قواعد كانت بثقة مقدارها 40% والباقي بثقة مقدارها *11.7* تقريبا مقارنة مع خوارزمية *GRI* التي استطاعت إيجاد 18 قاعدة بثقة *40*% ودعم *11.7* تقريبا



Sort by: Confidence % 55 of 55

Consequent	Antecedent	Support %	Confidence %
month = ايلول	27.0 = -ليلا	11.628	60.0
month = ايلول	8.0 = -وفاة	11.628	40.0
month = نيسان	9.0 = -طريق مركزي	11.628	40.0
month = تشرين الاول	2.0 = -طريق محلي	11.628	40.0
month = ايار	4.0 = -طريق مركزي	11.628	40.0
month = اذار	11.0 = -وفاة	11.628	40.0
month = اذار	36.0 = -نهارا	11.628	40.0
month = كانون الثاني	1.0 = -طريق محلي	13.953	33.333
month = كانون الثاني	22.0 = -ليلا	16.279	28.571
month = تشرين الثاني	22.0 = -ليلا	16.279	28.571
month = اب	27.0 = -داخل العش	11.628	20.0
month = اب	4.0 = -طريق مركزي	11.628	20.0
month = ايلول	9.0 = -طريق مركزي	11.628	20.0
month = ايلول	11.0 = -وفاة	11.628	20.0
month = شباط	11.0 = -وفاة	11.628	20.0
month = شباط	2.0 = -طريق محلي	11.628	20.0
month = كانون الاول	4.0 = -طريق مركزي	11.628	20.0
month = كانون الثاني	8.0 = -وفاة	11.628	20.0
month = كانون الثاني	27.0 = -ليلا	11.628	20.0
month = كانون الثاني	36.0 = -نهارا	11.628	20.0
month = نيسان	27.0 = -داخل العش	11.628	20.0
month = نيسان	8.0 = -وفاة	11.628	20.0
month = نيسان	2.0 = -طريق محلي	11.628	20.0

الشكل (16) عينة من نتائج تطبيق النموذج السابق باستخدام خوارزمية *Apriori*

بشكل مشابه، ومن خلال النماذج المتعددة التي تم تطبيقها على كلتا الخوارزميتين، وجدنا أن خوارزمية *GRI* قادرة على اكتشاف عدد أكبر من القواعد ذات الثقة الأعلى والأكثر دعما في قاعدة البيانات، وبما أن الخوارزمية لا تقبل حقول خرج إلا من النوع الفئوي، فإننا ننصح باستخدام تقسيم المجالات المستمرة *Binning range fields* إلى فئات في حال التعامل معها كمخرجات للخوارزمية.

الاستنتاجات والتوصيات:

- ضرورة اعتماد نموذج علمي مناسب لتسجيل بيانات حوادث المرور.
- ضرورة تأهيل الكادر البشري في مديريات المرور بشكل يسمح بتسجيل بيانات وضبوط حوادث المرور بشكل علمي ودقيق.
- الإسراع باعتماد نظام معلوماتي لتخزين بيانات حوادث المرور، إذ أنه وبتاريخ تنفيذ البحث لا يوجد أي نظام حاسوبي لتسجيل وأرشفة حوادث المرور في اللاذقية أو في إدارة المرور في دمشق أو في وزارة النقل، والبيانات الورقية الموجودة ليست الأنسب للتطبيقات المعلوماتية.
- بنية مستودعات البيانات وفق النموذج *Snowflake* هي الأنسب لبناء تطبيق مرن وقابل للتطوير بسرعة وفق المتطلبات الجديدة للنظام لأنه يوفر المرونة المطلوبة لتوسيع نظام تخزين بيانات حوادث المرور مستقبلا من خلال احتوائه على جداول رئيسية وعدد غير محدود من الجداول الفرعية التي يمكن أن تصف البيانات.

- الخوارزميات *CART* و *CHIAD* وفق الاختبارات هي الأفضل لتصنيف قيم وإحصاءات حوادث المرور مقارنة مع خوارزمية *SVM*.
- خوارزمية *Two Step* حققت في معظم الاختبارات نتائج أفضل في تحليل قيم حوادث المرور من خوارزمية *K-Means* وننصح باستخدامها بسبب قدرتها على التحديد التلقائي لعدد العناقيد الأمثلي، كما أن خوارزمية *Kohonen* تقدم طريقة مميزة في تنظيم السجلات في فضاء ثنائي الأبعاد، وبمساعدة خوارزمية *Two Step* تقدم فهما أفضل لتوزع البيانات ضمن السجلات.
- خوارزمية *GRI* ذات نتائج أفضل من خوارزمية *Apriori* حيث اكتشفت عدد اكبر من القواعد المهمة ذات الوثوقية العالية.
- ننصح بالعمل على إنشاء مركز وطني لدعم القرار على مستوى رئاسة مجلس الوزراء مع اعتماد بنية تحتية معلوماتية وطنية يتم فيها إدخال كامل المعلومات المتوفرة من كافة الوزارات من أجل تحليلها ودراستها.

المراجع:

[1] *SUKHAI ,ANESH., P JONES,ANDY., HAYNES ,ROBIN., "Epidemiology And Risk Of Road Traffic Mortality In South Africa." South African Geographical Journal 91 (1) 4 – 15 (2009).*

[2] الإدارة العامة للمرور في سورية – التقارير الإحصائية المنشورة عن حوادث المرور للأعوام 2008 – 2011.

[3] *World Health Organaizaton, www.who.org, "World report on road traffic injury prevention summary" (2004).*

[4] *Alex, A.Freitas., "A Survey of Evolutionary Algorithms for Data Mining and Knowledge Discovery." Postgraduate Program in Computer Science,Pontificia Universidade Catolica do Parana Rua Imaculada Cnceicao,1155 (2011).*

[5] *Beshah ,Tibebe., Hill ,Shawndra., "Mining Road Traffic Accident Data to Improve Safety:Role of Road-related Factors on Accident Severity in Ethiopia"*

Department of Information Science1, Operations and Information Management Department2, Addis Ababa University, Ethiopia1 , University of Pennsylvania, Philadelphia, PA2(2010).

[6] *Chang, L. , W. Chen ., "Data mining of tree-based models to analyze freeway accident frequency." Journal of Safety Research 36: 365-375(2005).*

[7] *Ng, K-S, Hung., W-T.,Wong W-G., "An algorithm for assessing the risk of traffic* accidents*", Journal of Safety Research, 33 387-410(2012).*

[8] Srisuriyachai, S., "Analysis of road traffic accidents inNakhon Pathom province of *Bangkok using data mining". Graduate Studies. Bangkok, Mahidol University (2007).*

[9] Nilambar,Jha., D.K. Srinivasa, GautamRoy., S. Jagdish, *"EPIDEMIOLOGICAL STUDY OF ROAD TRAFFIC ACCIDENT CASES: A STUDY FROM SOUTH INDIA", Jawaharlal Institute of Post Graduate Medical Education & Research, Pondicherry*.

[10] Ozer , Patrick., "Data Mining Algorithms for Classification". *BSc Thesis Artificial Intelligence, Radboud University Nijmegen(2009).*

[11] Therneau ,Terry M., Atkinson, Elizabeth J. , "An Introduction to Recursive Partitioning Using the RPART Routines". *Mayo Foundation (2014).*

[12] Storkey, Amos., "Learning from Data: Decision Trees". *University of Edinburgh (2014).*

[13] G.JAYALAKSHMI DR.K.NAGESWARA RAO . "MINING ASSOCIATION RULES FOR LARGE TRANSACTIONS USING NEW SUPPORT AND CONFIDENCE MEASURES ". *V.R.SiddharthaEngineering College,Kanuru,Vijayawada,A.P,India (2009).*

[14] Alsabti, Khaled., Ranka, Sanjay., Singh, Vineet.,"An Efficient K-Means Clustering Algorithm". *Syracuse University, University of Florida, Hitachi America, Ltd.*

[15] *Dumalo, Menchita F., Oh, Byung-Joo., " Implementation Of Two Step Clustering Algorithm With Self Organizaing Maps For Semantic Integration Of Heterogeneous Data Sources", Hannam University (2009).*

[16] *Brocki, Lucas., "Kohonen Self-Organizing Map for the Traveling Salesperson Problem", Polish–Japanese Institute of Information Technology, (2008).*

[17] *Lee, Won-Chan., "Classification Consistency and Accuracy for Complex Assessments Using Item Response Theory", Center for Advanced Studies in Measurement and Assessment(2008).*