بسم الله الرحمن الرحيم

جامعة السودان للعلوم والتكنولوجيا

كليـة علـوم الحـاسـوب وتـقـانة المعلومـــات

مقارنة خوارزميات التعلم الالي علي بيانات حوادث المرور

Comparing The Machine Learning Algorithms By Using Traffic Accident Dataset

بحث تكميلي مقدم كمطلوب لنيل درجة بكالوريوس الشرف في علوم الحاسوب

أكتوبر 2017 م

بسم الله الرحمن الرحيم

جامعة السودان للعلوم والتكنولوجيا

كليـة علـوم الحـاســـــوب وتـقـانـة المعلومـات

مقارنة خوارزميات التعلم الالي علي بيانات حوادث المرور

Comparing the Machine Learning Algorithms By Using Traffic Accident Dataset

أكتوبر 2017 م

بحث تكميلي مقدم كمطلوب لنيل درجة بكالوريوس الشرف في علوم الحاسوب

إعداد الطلاب

عبد الله عمر إبراهيم

محمد معاوية محمد

إشراف :أ.علي الأميـن

آيـة

قال تعالى : (**فاذكروني أذكركم واشكروا لي ولا تكفرون) البقرة(152)**

صدق الله العظيم

الحمد لله

الحمد لله الذي أنزل على عبده الكتاب .أظهر الحق بالحق وأخزى الأحزاب.وأتمَّ نوره، وجعل كيد الكافرين في تباب.أرسل الرياح بشرًى بين يدي رحمته وأجرى بفضله السحاب.وأنزل من السماء ماء، فمنه شجر، ومنه شراب.جعل الليل والنهار خلفه فتذكر أولو الألباب.نحمَده تبارك وتعالى على المسببات والأسباب .وصلي اللهم وسلم وبارك علي جناب الحبيب المحبوب حبيبك المصطفي وعلي اله وصحبه وسلم سلاما تاماً اللهم علمنا ما جهلنا وانفعنا بما علمتنا وزدنا علما .

الإهداء

إلي الوالد الحبيب

- إلي قدوتي الأولي ونبراسي الذي ينير دربي.
- إلي من أعطاني ولم يزل يعطيني بالحدود.
- إلي من رفعت رأسي عليا افتخارا به.
- ولاكني لا املك إلي أن ادعوا الله عز وجل أن يبقيه ذخرا لنا.
- أطال الله عمرك فيما يحب ويرضي.

إلي الوالدة العزيزة

- فيا من علمتني أبجدية الحروف.
- ويا من علمتني الصمود مهما تبدلت الظروف.
- اخط لكي كلمات مدها حبر دمي.
- كلمات ملؤها شكر وعرفان
- كلمات تتردد علي كل لسان
- نعم إنها أمي الغالية..

الشكر والعرفان

لابد لنا ونحن نخطو خطوتنا الاخيره في الحياة الجامعية من وقفه ونعود إلي أعوام قضيناها في رحاب الجامعة مع أساتذتنا الكرام الذين قدموا لنا الكثير باذلين بذلك جهودا كبيره في بناء الغد لتبعث الأمة من جديد . وقبل إن نمضي نقدم اسمي آيات الشكر والامتنان والتقدير والمحبة إلي الذين حملوا أقدس رسالة في الحياة.

"كن عالما .. فإن لم تستطع فكن متعلما، فإن لم تستطع فأحب العلماء ، فإن لم تستطع فلا تبغضهم"

إلي الذين مهدوا لنا طريق العلم والمعرفة...

إلي جميع أساتذتنا الأفاضل...

وأخص بالتقدير والشكر:

الأستاذ :علي الأمين.

وكذلك نشكر كل من ساعد علي إتمام هذا البحث وقدم لنا العون ومد لنا يد المساعدة وزودنا بالمعلومات اللازمة لإتمام هذا البحث.

المستخلص

إن استخدام تقنية التنقيب في البيانات يوفر للمؤسسات القدرة علي فهم جميع البيانات و التركيز علي أهم المعلومات الموجودة في قواعد البيانات,وترتكز تقنيات التنقيب علي البيانات علي بناء التنبؤات المستقبلية واكتشاف المعرفة و السلوك والاتجاهات ، مما يسمح باتخاذ القرارات الصحيحة في الوقت المناسب.

نموذج تحليل بيانات حوادث المرور هو نموذج يقوم علي تحليل بيانات الحوادث معتمدا علي بعض العوامل .

تعتمد فكرة المشروع علي تنقيب البيانات حيث يهدف المشروع إلي تحليل مجموعة من بيانات حوادث المرور لولاية الخرطوم وتدريب بعض خوارزميات تعلم اللالة ومقارنة نتائج تلك الخوارزميات مع بعضها ،حيث استخدم في هذا البحث خوارزميات تعلم الآلة في تحليل البيانات.

Abstract

The use of data mining technology provides access to information and data in databases. Data mining techniques are based on future predictions, knowledge discovery and behavior, allowing for the right decisions to be made in a timely fashion.

Traffic Accident Data Analysis Model is a model based on the analysis of accident data based on certain factors.
The idea of the project is based on data mining. The project aims at analyzing a set of traffic accident data for the state of Khartoum and the training of some algorithms of learning algorithms and comparing the results of these algorithms with each other.

فهرس المصطلحات

شرح المصطلح	المصطلح
False positive	FP
True negative	TN
False negative	FN
True positive	TP

فهرس الجداول

فهرس الاشكال

فهرس الصور

فهرس المحتويات

الباب الاول

المقدمة

1.1 المقدمة

تمثل مشكلة الحوادث المرورية من المشكلات الاجتماعية والاقتصادية والصحة العامة المرتبطة بالتنمية . حيث تسبب الحوادث في مقتل أكثر من مليون شخص ،كما يصاب أكثر من 15 مليون شخص بجروح من جراء تلك الحوادث علي الطرق كل عام ، تتحمل الدول النامية والدول ذات الاقتصاديات المحدودة العبء الأكبر حيث تمثل الحوادث احد قضايا التنمية التي تؤثر تأثيرا غير مناسبا علي الفقراء في الدول المنخفضة الدخل والمتوسطة الدخل . وتستنزف الحوادث المرورية عادة من 1 إلي 3 في المائة من إجمالي الناتج المحلي لأي دول.

وينجم عن الحوادث المرورية وفاة (1200000) شخص سنويا كما يصاب 50 مليون شخصا بالإعاقة بسبب حوادث المرور ، تعد حوادث المرور السبب الرئيسي الثاني للوفيات للفئة العمرية من 5 إلي 29 سنة ، كما تعد السبب الرئيسي الثالث في الوفيات في الفئة العمرية ما بين 30 و44 سنة ،وتقدر منظمة الصحة العالمية إن عدد الوفيات سيزداد بنسبة 80% في الدول النامية وذات الدخول المتدنية بحلول 2020 م إذا لم تتخذ إجراءات فورية للتصدي لهذه الحوادث وأسبابها سيفقد العالم يوميا أكثر من (3000)شخص من جراء حوادث المرور أما علي صعيد إقليم الشرق الأوسط في المنظمة الصحة العالمية والذي يشمل معظم الدول العربية ، فإنه يتوفي أكثر من (130) ألف شخص سنويا. [1]

أصبحت الحوادث المرورية تمثل وبشكل كبير هاجسا وقلقا لكافة أفراد المجتمع ، وأصبحت واحدة من أهم المشكلات التي تستنزف الموارد المادية والطاقات البشرية وتستهدف المجتمعات في أهم مقومات الحياة والذي هو العنصر البشري باضا فه إلي ما تكبده من مشاكل اجتماعية ونفسية وخسائر مادية ، مما أصبح لزاما إيجاد حلول ومقترحات ووضعها موضع التنفيذ للحد من هذه الحوادث أوعلي اقل تقدير معالجة أسبابها والتخفيف من أثارها السلبية[2].

1.2 مشكلة البحث

بالرغم من تطور التقنيات في تحليل البيانات إلا أننا نجد انه لا توجد أي دراسة علمية منهجية تستخدم للحد من هذه الحوادث ووضع استراتجيات مناسبة للسلامة المرورية ،ونجد أن معظم الطرق المستخدمة في تحليل بيانات حوادث المرور إحصائية.

1.3 فرضيات البحث

كلما زادت كمية البيانات المستخدمة في التحليل كلما كانت نسبة الخطأ اقل ودقة تنبؤ المصنف عالية.

1.4 أهمية البحث

تعد هذه الدراسة من الدراسات القليلة التي تنبه لموضوع أهمية تحليل المعطيات المرورية .وتمثل أهمية البحث في معرفة أسباب الحوادث وتقليل الحوادث. لذلك لابد من الاستفادة القصوى من بيانات حوادث المرور المأخوذة من الإدارة العامة للمرور.

1.5 أهداف البحث

يهدف هذا البحث إلي-:

- تحليل بيانات حوادث المرور بطريقة منهجية علمية.
- جعل الالة قادرة علي التعلم.
- تحليل عميق للبيانات لكي تساعد في حل المشكلة.

1.6 منهجية البحث

- يتم في هذا البحث استخدام المنهج الوصفي والاعتماد علي بيانات حوادث المرور المقدمة من الإدارة العامة للمرور.
- كما تتم الإشارة إلي الأسلوب الاستدلالي في تحليل معطيات المرور من اجل تحليل بيانات الحوادث وما يرتبط بها من عوامل مثل رخصة السائق وحالة السائق والفئة العمرية وزمان الحادث ومكان الحادث ونوعية المركبة.
- استخدام تقنية machine learning في التحليل.

1.7 حدود البحث

بيانات حوادث المرور لولاية الخرطوم من يناير إلي ابريل2017 م.

1.8 هيكلية البحث

يتضمن هذا البحث بالاضافه إلي هذا الفصل الفصول.

الباب الثاني : يتضمن نبذه عامه عن الإطار النظري ل مفهوم تنقيب البيانات وأساليبها , والدراسات السابقة.

الباب الثالث : يتضمن المنهجية المستخدمة في البحث.

الباب الثاني

الإطـــار النظري والـدراســـــات

الســـــابقــة

الفصل الاول

1.9 الإطار النظري

1.9.1 المقدمة

يتميز عصرنا الراهن عصر الانترنت والاقتصاد الرقمي بوجود كميات كبيرة للبيانات حتى أصبح من المستحيل علي المحللين استخلاص معلومات ذات معني باللجوء فقط إلى المداخل التقليدية للتحليل التمهيدي للبيانات.

مع وجود كميات كبيرة من البيانات المخزنة في قواعد البيانات ومخازن البيانات إزدات الحاجة إلي تطوير أدوات تمتاز بالقوة لتحليل البيانات واستخراج المعلومات والمعارف منها ، من هنا ظهر مايسمى بالتنقيب في البيانات كتقنية تهدف إلي استخراج المعرفة من كميات هائلة من البيانات.

وتعتبر تقنية التنقيب في البيانات من التقنيات الحديثة التي فرضت نفسها في بقوة في عصر المعلوماتية ،واستخدامها يوفر للشركات والمنظمات والمؤسسات الحكومية في جميع المجالات القدرة علي اكتشاف المعلومات الموجودة في قواعد البيانات ،كما ترتكز تقنية التنقيب في البيانات علي بناء التنبؤات المستقبلية مما يسمح باستخدام القرارات الناسبة في الوقت المناسب.

وتعتبر مرحلة استكشاف المعرفة من البيانات من أكثر المراحل تعقيدا. والمرتبطة إلي حد بعيد بعملية تطوير أخري مهمة جدا هي مستودعات البيانات ،حيث إن الكثير من الشركات والمنظمات تستخدم عملية التنقيب في قواعد البيانات بشكل منهجي ومنظم بوصفها تشكل جوهر العمل الذي يعتمد علية في تفعيل النشاط وتحقيق الميزة التنافسية.

1.9.2 مفهوم التنقيب في البيانات

ظهر مصطلح التنقيب في البيانات في منتصف التسعينات في الولايات المتحدة الأمريكية، كنوع يجمع ما بين الإحصاء وتكنولوجيات(قواعد البيانات ،الذكاء الاصطناعي ،التعلم الآلي) .

توجد عدة تعريفات لهذا المفهوم حيث يمكن تعريفها بأنها" الاستكشاف الآلي والمؤتمن لأنماط شائعة وغير جلية مخفية في قاعدة بيانات معينة وأيضا تعرف بأنها,[3]" إجراءات تحليل دقيقة وذكية، تفاعلية و تسلسلية، تسمح لمسيري النشاطات عند استخدام هذه الإجراءات باتخاذ قرارات والقيام بأعمال ملائمة في صالح النشاط المسؤولين عنة والمؤسسة التي يعملون فيها [4]" وإنها أيضا" عبارة عن تحليل لمجموعات كبيرة الحجم من البيانات المشاهدة للبحث عن علاقات محتملة وتلخيص للبيانات في أشكال جديدة لتكون مفيدة ومفهومة لمستخدمها .[5]" من خلال التعريفات السابقة يمكن القول أن

التنقيب في قواعد البيانات يهدف إلي استخلاص المعلومات المخبأة فيها،واستخدامه يوفر للمؤسسات في جميع المجالات علي استكشاف والتركيز علي أهم المعلومات في قواعد البيانات بالإضافة إلي كثرة البيانات الموجودة والمخزنة في ما يسمي بقواعد البيانات أصبحت تقنية التنقيب في البيانات محل تساؤل من عديد من الباحثين للاستفادة منها ، ومع زيادة انتشار مستودعات التخزين الضخمة ما يسمي (*data warehouses*)،أصبح من الضروري إيجاد تقنيات وطرق ووسائل لاستخلاص المعلومات والمعرفة من هذه البيانات المكدسة واستغلالها في حل المشاكل واتخاذ القرارات ، باستخدام تطبيقات الحاسوب الحديثة والتي تعتبر تكنولوجيا حديثة ذكية قائمة علي جعل الحاسوب" يفكر كما يفكر الإنسان ويفعل كما يفعل الإنسان " جاءت فكرة الكشف والتنقيب علي هذه البيانات بطرق ذكية للمساعدة في حل المشاكل واتخاذ القرارات وتعتبر خطوة من خطوات استكشاف المعرفة من قواعد البيانات.

1.9.3 أهمية أسلوب تنقيب البيانات

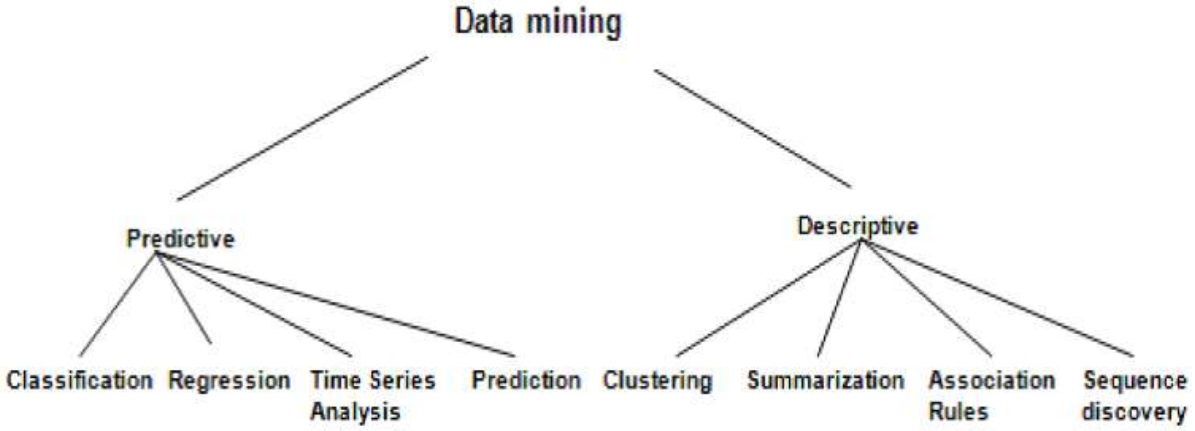
- عملية تحليلية للاستكشاف والبحث في بيانات ضخمة وهائلة لاستخراج أنماط مفيدة وإيجاد العلاقات ومدي الارتباط بين عناصرها.
- من اجل تحليل البيانات للحصول علي علاقات جديدة غير متوقعة.
- نقل عالم الأنظمة الواقعي إلى عالم افتراضي يمارس فيه متخذي القرار التحليل واختبار الفرضيات على شاشات الحاسوب ذات القدرة الرسمية العالية الدقة إلى أن يصلوا إلى ما يطمحون إليه من فهم وقناعات قبل اتخاذ القرارات بشأن مايدرسون من أنظمة.
- التنبؤ ومن ثم استنتاج إجابات مقدرة تقديرا إحصائيا لكميات البيانات.

1.9.4 أهداف أسلوب تنقيب البيانات

- إن التنقيب في قواعد البيانات يهدف إلي انتزاع واستخلاص أنماط مفيدة، وهي تكنولوجيا حديثة، أصبحت مهمة في ظل التطور السريع وانتشار استخدام قواعد البيانات.
- استخدامها يوفر للمؤسسات وأجهزة الأمن في جميع المجالات القدرة علي استكشاف، والتركيز علي أهم المعلومات في قواعد البيانات.
- تركز تقنيات التنقيب علي بناء التنبؤات المستقبلية واستكشاف السلوك والاتجاهات، مما يسمح بتقدير القرارات الصحيحة واتخاذها في الوقت المناسب.
- تجيب تقنيات التنقيب علي العديد من الأسئلة ، وفي وقت قياسي ، بخاصة تلك النوعية من الأسئلة التي يصعب الإجابة عليها ،إن لم يكن مستحيلا ،باستخدام تقنيات الإحصاء الكلاسيكية ، والتي كانت وان وجدت فإنها تستخدم وقتا طويلا والعديد من الإجراءات.

1.9.5 نماذج التنقيب في البيانات

- النماذج التنبؤية (predictive data mining): يحاول إيجاد أفضل التنبؤات اعتمادا علي المعطيات ويستخدم هذا التنقيب علي المعلومات القديمة لتوقع ما سيحدث في المستقبل.
- النماذج الوصفية: (Descriptive data mining) تعتمد علي إعادة تنظيم البيانات والتنقيب في أعماقها لاستخراج المؤشرات الموجودة فيها.



شكل(1-2)

وتنقسم إلي نوعان :

- النماذج التنبؤية
- النماذج الوصفية

1.9.6 أدوات التنقيب في البيانات

- السلاسل الزمنية
- .التصنيف
- التنبؤ
- التلخيص
- التجزئة
- تحليل الارتباط
- الكشف عن التغيرات أو الانحرافات

1.9.7 مراحل عملية التنقيب في البيانات

يمكن تلخيص مراحل و خطوات عملية التنقيب في البيانات كما يلي:

- .فهم طبيعة الأعمال

يعتبر المطلب الأول لاكتشاف المعرفة هو فهم المشاكل و المسائل التي تواجهها الأعمال . و بمعني أخر، كيف يمكن تحقيق المنفعة الأعظم من التنقيب في البيانات، مما يتطلب وجود صيغة واضحة ومحددة لأهداف الأعمال.

- فهم البيانات

تعتبر مسالة معرفة ماهية وطبيعة البيانات عامل مهم في نجاح عملية التنقيب في البيانات و اكتشاف المعرفة حيث أن معرفة البيانات بصورة جيدة تعني مساعدة المصممين على استخدام الخوارزميات أو الأدوات المستخدمة للمسائل المحددة بدقة عالية.

وهذا يقود إلى تعظيم فرص النجاح بالإضافة إلى رفع الفعالية و الكفاءة لنظام اكتشاف المعرفة .ولاتحتاج عملية التنقيب في البيانات إلى تجميع البيانات في مستودع البيانات، أما إذا كاف مستودع البيانات موجود في الأنظمة، فمن الأفضل عدم احتكار المستودع بشكل مباشر لغرض التنقيب في البيانات.

و يمكن تلخيص الخطوات الضرورية لعملية فهم البيانات كالآتي:

- تجميع البيانات

و هي الخطوة الموجهة نحو تحديد مصدر البيانات في الدراسة بما في ذلك استخدام البيانات العامة الخارجية مثل) المرور و الضرائب وغيرها.(

- توصيف البيانات

وهي الخطوة التي تركز على توصيف محتويات الملف الواحد من الملفات أو الجداول.

- جودة البيانات و تحقيقها

هذه الخطوة تحدد ما إذا كان تقليل أو إهمال بعض البيانات غي الضرورية أو كونها رديئة الجودة وقد لا تصلح في الدراسة .لان النموذج الجيد يحتاج إلي بيانات جيدة مما يتوجب أن تكون البيانات صحيحة وذات مضمون دقيق.

- التحليل الاسترشادي للبيانات

تستخدم الأساليب مثل الإظهار المرئي أو التصور أو عملية التحليل المباشر التي تودي إلي إجراء التحليل الأولي للبيانات تعتبر هذه الخطوة مهمة وضرورية لأنها تركز على تطوير الفرضيات المتعلقة بالمشكلة قيد الدراسة.

- تهيئة البيانات

وتشمل الخطوات التالية:

- الاختيار وتعني اختيار المتغيرات المتوقعة و حجم العينة.
- صياغة المتغيرات و تحويلها

حيث يجب دائما أن تصاغ المتغيرات الجديدة لبناء النماذج الفعالة.

- تكامل البيانات حيث أن مجاميع البيانات في دراسة التنقيب عن البيانات من الممكن خزنها في قواعد بيانات متعددة لأغراض التي تكون بحاجة توحيدها في قاعدة بيانية واحدة.
- تصميم و تنسيق البيانات حيث تتعلق هذه الخطوة في إعادة ترتيب حقول البيانات كما يتطلب في نموذج التنقيب في البيانات.

- صياغة نماذج الحل و ثبوتها

إن بناء و صياغة نموذج الحل السليم و الدقيق يتم من خلال عملية الخطأ و الصواب، حيث كثيرا ما تحتاج مثل هذه العملية إلى مساعدة المتخصصين في التنقيب عن البيانات بهدف اختبار وفحص مختلف البدائل للحصول على أفضل نموذج لحل المشكلة قيد الدراسة.

- التقييم و تعليل نتائج النموذج

حالما يتم صياغة النموذج و التحقق من ثباته ك صدقة ، تجري مباشرة عملية التحقق من ثبات حزمة البيانات التي يتم تغذيتها بواسطة النموذج وبما أن نتائج هذه البيانات معروفة ، لذا فان النتائج المتوقعة تقارن مع النتائج الفعلية في ثبات حزمة البيانات قيد التشغيل وتودي هذه المقارنة أو المفاضلة إلي التحقق من دقة النموذج

- نشر و توزيع النموذج

حيث تشتمل هذه الخطوة على نشر و توزيع النموذج داخل المنظمة لمساعدة عملية صنع القرار . وان النموذج الصالح يجب أيضا أن يحقق الرضا لدي المستفيدين طالما أن اختيار النموذج لابد أن يتم من خلا الدراسة الاسترشادية أو نموذج مصغر من الدراسة الشاملة.

1.9.8 تعلم الآلة

هو أحد فروع الذكاء الاصطناعي التي توفر القدرة على التعلم لأجهزة الحاسوب .تهتم بتصميم وتطوير الخوارزميات والتقنيات التي تسمح للحاسوب بامتلاك خاصية" التعلم ."هناك مستويين من التعلم :الاستقرائي والإستنتاجي .يقوم التعلم الاستقرائي باستنتاج قواعد وأحكام عامة من البيانات الضخمة.

يتداخل" علم الآلة "مع علم إحصاء الحوسبة .Computational Statistics ويهتم بصنع التنبؤات من خلال استخدام الحاسب، يرتبط علم التحسين الرياضي Mathematical Optimization ، الذي يركز على اختيار البديل الأفضل من بين العديد من البدائل المتاحة،كما يوفر الكثير من الوسائل والنظريات والتطبيقات لتعلم الآلة[6].

1.9.9 أنواع التعلم

- التعلم المراقب.
- التعلم الغير المراقب.

سيتطرق هذا البحث علي نوع من أنواع تعلم الآلة وهو التعلم المراقب الذي يهدف إلي ربط المدخلات والمخرجات وذلك عن طريق إعطاء أمثلة للمدخلات (input) والمخرجات (output) المرغوبة للآلة من قبل المعلم.

من أهم الأدوات المستخدمة في عملية التعلم المراقب هي أداة التصنيف وهو النوع الأكثرُ استخداما في تعلم الآلة .في هذا النوع يكون الدخل مصنفا إلى نوعين أو أكثر .وهدفُ عملية التعلم إنتاجُ نموذجٍ يستطيع تصنيفَ أي دخل جديد إلى نوع أو أكثر من الأنواع المعرفة سابقًا . مثال على هذا النوع، عمليةُ تصنيف البريد الإلكتروني وعملية التعرف على الوجوه.

الفصل الثاني

1.10 الدراسات السابقة

1.10.1 المقدمة

سنتناول في هذا الباب الدراسات والبحوث التي قدمت في مجال التنقيب علي البيانات في تحليل بيانات المرور، وتتبع كيفية تتحاياها للبيانات والإستفاده منها في كثير من الاستخدامات وفي حل الكثير من المشاكل.

- S. Krishnaveni، (2011) ، العمل مع بعض نماذج التصنيف للتنبؤ بالإصابات التي حدثت في حادث سير في نيجيريا ومقارنتها[7]

يستخدم هذا البحث على نهج (neural networks) القائمة في حين أن تحليل (decision trees) البيانات يمكن استخدامها للعمل على الحد من المجزرة على الطرق السريعة.تم تصنيف البيانات في بيانات مستمرة وفئوية حيث تم تحليل البيانات المستمرة باستخدام تقنية (neural networks) الاصطناعية والبيانات الفئوية باستخدام تقنية .(decision trees) وأظهرت النتائج أن نهج (decision trees) تفوق على (neural networks) بمعدل خطأ أقل ومعدل دقة أعلى . يعتمد هذا البحث على ثلاثة أسباب مهمة للحوادث بسبب انفجار الإطارات، وفقدان السيطرة والإكثار من السرعة.

- Naina et. Al. (2016) استخدم نموذج التصنيف، مع Dichotomiser 3 التكرارية وخوارزمية .(decision trees) كما تقارن الخوارزمية الحالية مع تعزيز خوارزمية C 4.5 مع استخدام أداة Weka أنها أظهرت أن من المفيد عندما نستخدم مجموعات البيانات الكبيرة والنتائج هي مؤثرة جدا[8].

- K. Geetha 2015هذه الدراسة تعمل على بيانات حوادث المرور من .Tamilnadu city والهدف الرئيسي من هذه الدراسة هو تقليل عدد حوادث الطرق .تتم إدارة بيانات حوادث المرور في شكل نص أو تنسيقات رقمية بطريقة غير مصنفة[9].
- قدم معاوية وأسماء هذه الدراسة في السودان 2015 بعمل تحليل بيانات الحوادث والهدف من الدراسة حوادث الطرق وعلاقتها مع حوادث الشباب واستخدم الشبكة العصبية في البحث[10].
- Miao Chongهذه الدراسة تعمل علي بيانات حوادث المرور في الولايات المتحدة مأخوذة من (NASS) والهدف من البحث جعل لغة الآلة ذكية في كشف نوعية الهدف اعتمادا علي بعض العوام

جدول يوضح المقارنة بين الدراسات السابقة

الدراسة	السنة	الأدوات والتقنيات	نقاط ضعف الدراسة
1	2011	neural networks	لا يوجد
2	2016	Wekaوخوارزمية (decision trees)	قلة البيانات
3	2015		لا يوجد
4	2015	الشبكة العصبية	لايوجد
5	2015	neural networks-support vector machine-	لايوجد
البحث المقترح	2017	Weka(support vector machine-شجرة القرار naïve–Bayes)	فقدان بعض البيانات

الجدول رقم(2-2)

الباب الثالث

منهجية البحث

الفصل الأول

1.11 منهجية البحث

1.11.1 المقدمة

هذا القسم يتحدث عن المنهجية المتبعة في البحث من فهم المشاكل التي تواجهها المؤسسة وكذلك في فهم البيانات بصورة جيدة ومعرفة ماهية البيانات , ويوضح كيفية تجميع البيانات، ويوضح معني كل متغير من تلك البيانات كما تم استخدام منهجية ال CRISP-DM في التحليل.

- طبيعة الأعمال

من خلال دراستنا إلي البيانات المأخوذة من الإدارة العامة للمرور نجد أن عملية التحليل المستخدمة في البيانات عبارة عن تحليل تقليدي للبيانات أي أنها تستخدم الطريقة الإحصائية في تحليل بياناتها وهنا تكمن المشكلة لأنها لا تستفيد من البيانات لذلك لابد من إيجاد بديل في عملية التحليل للاستفادة من هذه البيانات بصورة كبيرة , لذلك تمكنت هذه الدراسة من إيجاد الحل والاستفادة من البيانات بصورة كبيرة حيث استخدمت تعلم الآلة في التحليل .

- فهم البيانات

من الخطوات الضرورية لعملية فهم البيانات يمكن تلخيصها في الأتي-:

- تجميع البيانات

لقد تم تجميع هذه البيانات من الإدارة العامة للمرور من سجلات الحوادث لولاية الخرطوم وذلك من الفترة من يناير إلي ابريل 2017 م.

- توصيف البيانات

تتكون البيانات من 12 متغير سنقوم بوصف كل متغير علي حدة .

- الأقسام : ونقصد بها أقسام محليات ولاية الخرطوم وهي بيانات نصية.

#	اسم المحلية
1	الخرطوم
2	الجبل
3	امبدة
4	شرق النيل
5	كرري
6	بحري
7	امدرمان

جدول رقم (1-1-3)

اليوم : ونقصد بها أيام الأسبوع.

#	اليوم
1	الجمعة
2	السبت
3	الأحد
4	الاثنين
5	الثلاثاء
6	الأربعاء
7	الخميس

جدول رقم(2-1-3)

الشهور : ونقصد بها الشهور التي سجل فيها الحدث.

#	الشهر
1	يناير
2	فبراير
3	مارس
4	ابريل

جدول رقم(3-1-3)

نوع الحادث : يقصد به نوع الحدث الذي تم تسجيله وينقسم إلي(1 حالة وفاة ,2 تعني أذي جسيم).

عمر المتهم

#	عمر المتهم
2	الأعمار من11 سنة إلي20 سنة
3	الأعمار من 21 سنة إلي 30 سنة
4	الأعمار من 31 سنة إلي 40 سنة
5	الأعمار من 41 سنة إلي 50 سنة
6	الأعمار من 51 فما فوق

جدول رقم(4-1-3)

زمن الحادث: ويقصد به زمن وقوع الحادث الذي تم تسجيله وينقسم إلي(ص :صباحا ,م :مساءا).

الجنس : ويقصد به نوع الشخص ذكر أوانثي.

رخصة السائق : ويقصد بها الرخصة التي يستخدمها السائق.

#	نوع الرخصة
أجنبية	يحمل رخصة أجنبية
م خ	يحمل رخصة ملاكي الخرطوم
ع خ	يحمل رخصة عمومي الخرطوم
ع و	يحمل رخصه عمومي ولائي
م و	يحمل رخصه ملاكي ولائي
غير محدد	لايمكن تحديد إذا كان يحمل رخصة
لايملك	لا يملك رخصة قيادية

جدول رقم(5-1-3)

نوع المركبة - : ويقصد بها نوع العربة التي تسببت في الحادث ومن أمثلتها (موتر ,صالون ,حافلة , ركشة ,وجميع الموديلات).

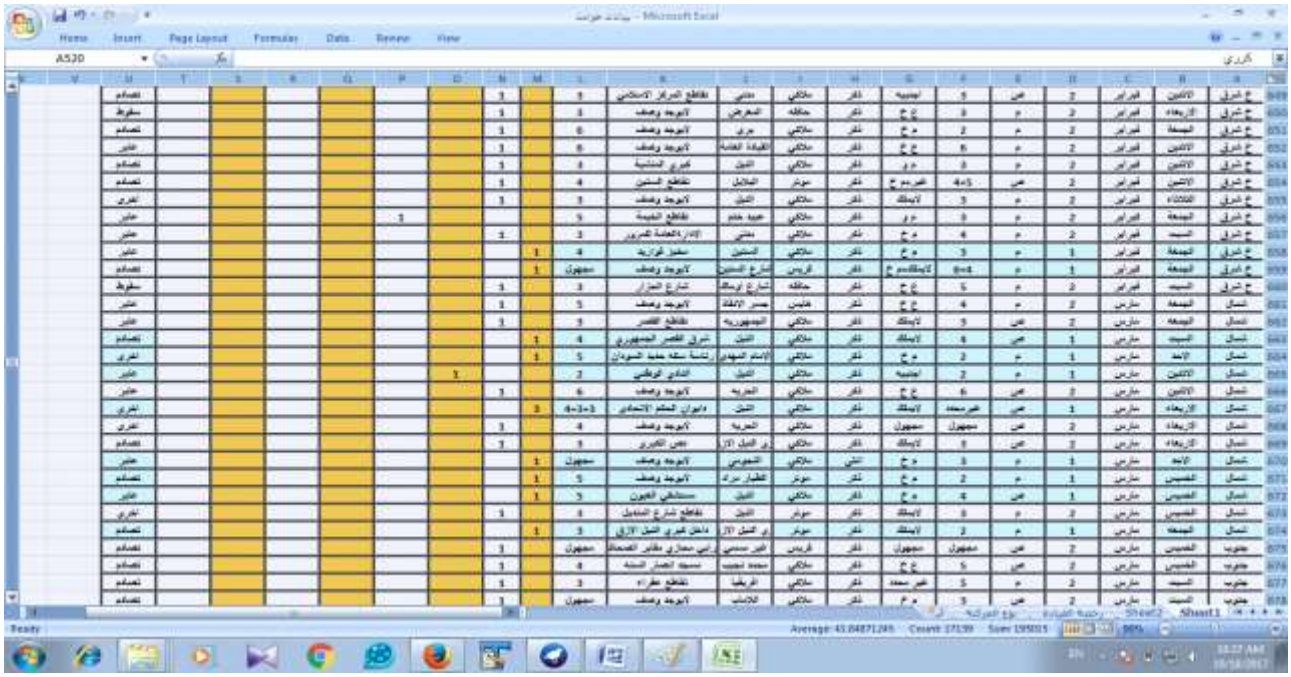
الشارع : ويقصد به الطريق الذي وقع فيه الحادث مثل(شارع مدني ,شارع عبيد ختم ,عمر الصول , وغيرها).

الموقع : يقصد به موقع الحادث.

عمر المصاب : هي نفس أعمار المتهمين بالاضافه إلي("1" تعني الأعمار من 10 فما تحت).

#	عمر المتهم
2	الأعمار من11 سنة إلي20 سنة
3	الأعمار من 21 سنة إلي 30 سنة
4	الأعمار من 31 سنة إلي 40 سنة
5	الأعمار من 41 سنة إلي 50 سنة
6	الأعمار من 51 فما فوق
1	الأعمار من 10 فما تحت

جدول رقم**(6-1-3)**



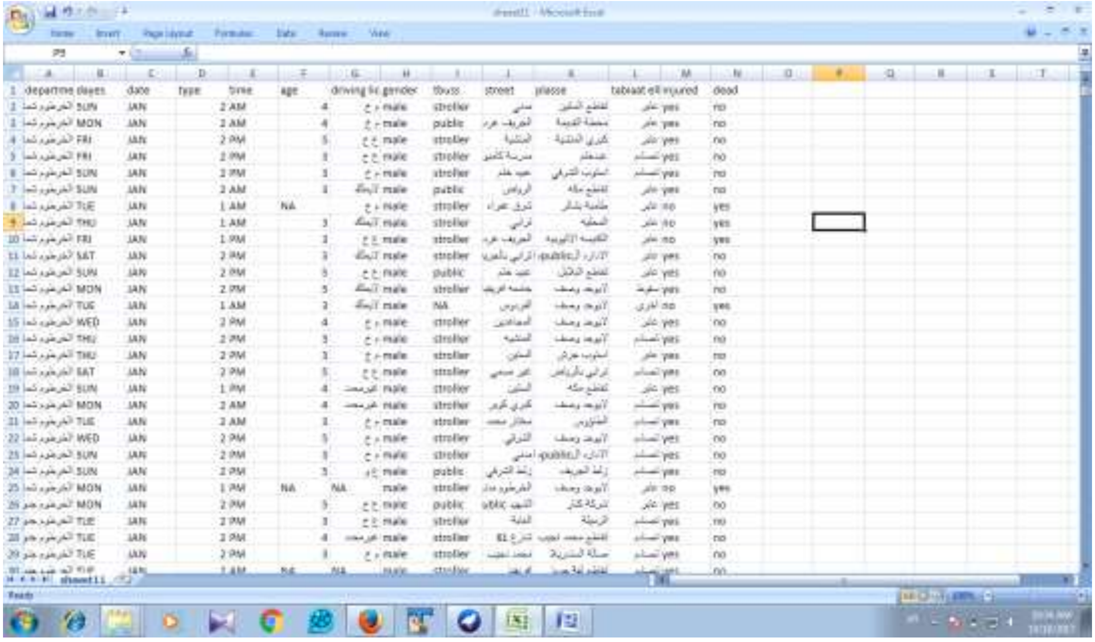
صورة رقم(7-1-3)

- جودة البيانات و تحقيقها

يحتاج النموذج الجيد إلي بيانات جيدة مما يتوجب أن تكون البيانات صحيحة وذات مضمون دقيق لذلك توجب علينا حذف بعض البيانات التي تعتبر غير ضرورية أو مهمة وقمنا بدمج بعض البيانات وذلك للتقليل من البيانات في عملية التصنيف كما موضح في شكل رقم.(7-1-3)

- تهيئة البيانات

تم اختيار المتغيرات من البيانات وتنسيقها وإعادة ترتيب كل الحقول وجعل البيانات جاهزة لعملية التحليل كما موضح في الصورة رقم **(8-2-3)** .


departme daues date type time age driving lic gender tbuss street plasse tabiaat ell injured dead

صورة رقم (3-2-8)

القسم الثاني

1.12 التصنيف

1.12.1 المقدمة

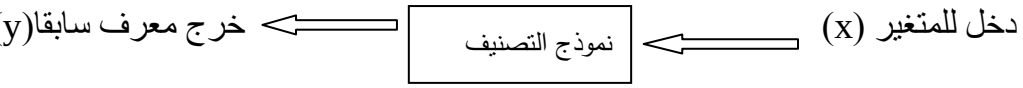
ينقسم تعلم الآلة إلي عدة أنواع منها التعلم الخاضع للإشراف ويسمى أيضًا التعلم التنبؤي في هذا النوع من التعلُّم يتم تدريب الآلة باستخدام دخل معروفِ الخرج مسبقًا ، مثلا مجموعةٍ من رسائل البريد الإلكتروني المصنفة مسبقًا إلى مهمة أو غير مهمة، والمطلوب تعلم كيفية ربط الدخل مع الخرج ليصبح بالإمكان مستقبلًا التنبؤ بالخرج من أجل أي دخل جديد .يندرج تحت هذا النوع أنواعٌ فرعية من التعلم بحسب الخرج المطلوب من نظام تعلم الآلة ومن أمثلة هذا النوع هو التصنيف, ومن أهم التقنيات المستخدمة فيه وهي شجرة القرار,naïve Bayes. support vector machine

1.12.2 التصنيف

وهو النوع الأكثرُ استخدامًا في تعلم الآلة. في هذا النوع يكون الدخل مصنفا إلى نوعين أو أكثر. وهدفُ عملية التعلم إنتاجُ نموذجٍ يستطيع تصنيفَ أي دخل جديد إلى نوع أوأكثر من الأنواع المعرفة سابقًا.

1.12.3 تقنية التصنيف

التصنيف هو مهمة تعليم دالة معينة f لربط مجموعة الخصائص (x) بفئة (class) معرفة مسبقا تسمي (y) تعرف الدالة (f) بنموذج التصنيف.



الشكل (3-2-1)

تتم عملية بناء المصنف بخطوتين رئيسيتين هما التعليم (خطوة بناء نموذج التصنيف) والتصنيف (استخدام النموذج ليتنبأ بفئات البيانات الغير معروفة).

أ- يتم تدريب خوارزمية التصنيف (التعلم) علي بيانات التدريب المحتوية علي سجلات معروفة لبناء المصنف الذي يستخدم لفحص البيانات التي تحتوي علي البيانات .

ب- يتم اداء المصنف بحساب عدد السجلات المتوقعة المصنفة بشكل صحيح والسجلات المصنفة بشكل خاطئ فيما يسمي بمصفوفة التضارب يتم تقييم اداء المصنفات بالحصول علي اعلي دقة واقل نسبة خطا عند تطبيقها علي بيانات الاختبار.

يتم جدولة عدد السجلات المصنفة بشكل صحيح وعدد السجلات المصنفة بشكل خاطئ علي شكل مصفوفة تسمي مصفوفة التعارض .

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

الشكل(2-2-3)

كل مدخل f_{ij} في مصفوفة التعارض يشير إلي عدد السجلات في class I المتوقعة أن تكون في ال class فمثلا f01 تشير إلي عدد السجلات في ال class 0 التي تم توقعها بشكل خاطئ في ال. class 1

اعتمادا علي المصفوفة فان:

مجموع السجلات المتوقعة بشكل صحيح هي.(f11+f00)

مجموع السجلات المتوقعة بشكل خطأ هي.(f10+f01)

1.12.4 تقنيات التصنيف

يتضمن التصنيف عدة تقنيات رئيسية سيناقش هذا البحث ثلاثة تقنيات :

1. **شجرة القرار**

هنالك عدة خوارزميات في شجرة القرار في هذا البحث سوف يستخدم خوارزمية j48 .

2. **Naïve Bayes**

سوف يستخدم خوارزمية naïve Bayes

3. **Super vector machine**

سوف يستخدم خوارزمية support vector machine

	Decision Trees	Naïve Bayes	SVM
Accuracy in general	**	*	****
Speed of learning with respect to number of attributes and the number of instances	***	****	*
Speed of classification	****	****	****
Tolerance to missing values	***	****	**
Tolerance to irrelevant attributes	***	**	****
Tolerance to redundant attributes	**	*	***
Tolerance to highly Interdependent attributes (e.g. parity problems)	**	*	***
Dealing with discrete/binary/continuous attributes	****	***(not continuous	**(not discrete)
Tolerance to noise	**	***	**
Dealing with danger of overfitting	**	***	**
Attempts for incremental learning	**	****	**
Explanation ability/transparency of knowledge/classifications	****	****	*
Model parameter handling	***	****	*

الجدول رقم **(3-2-3)**

الجدول رقم (3-2-3) يوضح بعض المقارنات للخوارزميات و علي هذا الاساس سوف يتم اختيار الخوارزمية الافضل في اداة عملية التصنيف.

الفصل الثالث

1.13 التدريب والتقييم

1.13.1 مقدمة

في هذا القسم يتم تدريب البيانات عن طريق إدخالها في weka وبعد ذلك يتم تقييم اداء المصنف.

1.13.2 تدريب الخوارزميات علي البيانات

تدريب خوارزمية supper vector machine علي بيانات حوادث المرور :

- استخدمت أداة weka 3.8.1 لتدريب خوارزمية support vector machine علي بيانات الحوادث لعمل (train) لبناء مصنف البيانات وتم اختباره علي بيانات الاختبار (test) ثم تم تقييم أداة المصنف من خلال حساب دقة التصنيف ونسبة الخطأ في التصنيف ويمكن حساب دقة المصنفات بحساب الأتي :-
- نسبة الضبطية (precision) نسبة سجلات الclass المصنفة بشكل صحيح إلي السجلات المصنفة في ال class .
- نسبة الاستدعاء (recall) نسبة سجلات الclass المصنفة بشكل صحيح إلي عدد إلي السجلات في ال class
- الايجابية الكاذبة (FP) أي يتم تصنيف البيانات علي أنها كاذبة وهي في الحقيقة صحيحة
- السلبية الكاذبة (fn) تصنيف علي أنها صحيحة وهي في الحقيقة مهددات .
- الايجابية الصادقة (TP) تصنيف السجلات علي أنها صحيحة وهي في الحقيقة صحيحة.
- السلبية الصادقة (fn) تصنيف السجلات علي أنها مهددات وهي في الحقيقة مهددات.

يتم حساب دقة المصنف اعتمادا علي معدل الكشف ومعدل الإنذار الكاذب .

استخدم هذا البحث بيانات حواث المرور وكانت نتائج التدريب كما يلي :

Predicated class			
Actual class		Class dead	**Class injured**
	Class dead	851	**0**
	Class injured	0	**367**

بعد ادخال البيانات إلي weka واختيار التدريب المناسب وهنا تم اختيار متجة الة الدعم تم الحصول علي الصورة رقم (3,3,1).

```
Time taken to build model: 0.48 seconds

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

Correctly Classified Instances        1188               97.5369 %
Incorrectly Classified Instances        30                2.4631 %
Kappa statistic                          0.9238
Mean absolute error                      0.0246
Root mean squared error                  0.1569
Relative absolute error                  7.433  %
Root relative squared error             38.5741 %
Total Number of Instances             1218

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.993    0.090    0.977      0.993    0.985      0.925    0.951     0.975     injured
                 0.910    0.007    0.971      0.910    0.939      0.925    0.951     0.902     dead
Weighted Avg.    0.975    0.073    0.975      0.975    0.975      0.925    0.951     0.960

=== Confusion Matrix ===

   a   b   <-- classified as
 956   7 |   a = injured
  23 232 |   b = dead
```

صورة رقم (3,3,1)

تدريب خوارزمية naïve Bayes علي بيانات حوادث المرور

- استخدمت أداة weka 3.8.1 لتدريب خوارزمية naïve Bayes علي بيانات الحوادث لعمل (train) لبناء مصنف البيانات وتم اختباره علي بيانات الاختبار (test) ثم تم تقييم أداة المصنف من خلال حساب دقة التصنيف ونسبة الخطأ و معدل الانذار الكاذب في التصنيف ويمكن حساب دقة المصنفات بحساب الأتي :-
- نسبة الضبطية (precision) نسبة سجلات الclass المصنفة بشكل صحيح إلي السجلات المصنفة في ال class .
- نسبة الاستدعاء (recall) نسبة سجلات الclass المصنفة بشكل صحيح إلي عدد إلي السجلات في ال class
- الايجابية الكاذبة (FP) أي يتم تصنيف البيانات علي أنها كاذبة وهي في الحقيقة صحيحة
- السلبية الكاذبة (fn) تصنيف علي أنها صحيحة وهي في الحقيقة مهددات .
- الايجابية الصادقة (TP) تصنيف السجلات علي أنها صحيحة وهي في الحقيقة صحيحة.
- السلبية الصادقة (fn) تصنيف السجلات علي أنها مهددات وهي في الحقيقة مهددات.

يتم حساب دقة المصنف اعتمادا علي معدل الإنذار الكاذب .

استخدم هذا البحث بيانات حواث المرور وكانت نتائج التدريب كما يلي :

Predicated class			
Actual class		Class dead	**Class injured**
	Class dead	851	**1**
	Class injured	61	**367**

بعد ادخال البيانات إلي weka واختيار التدريب المناسب وهنا تم اختيار خوارزمية naïve bayes تم الحصول علي الصورة رقم (3,3,2).

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

Correctly Classified Instances        1186               97.3727 %
Incorrectly Classified Instances        32                2.6273 %
Kappa statistic                          0.919
Mean absolute error                      0.0306
Root mean squared error                  0.1544
Relative absolute error                  9.2373 %
Root relative squared error             37.9496 %
Total Number of Instances             1218

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.991    0.090    0.976      0.991    0.984      0.920    0.953     0.975     injured
               0.910    0.009    0.963      0.910    0.935      0.920    0.953     0.943     dead
Weighted Avg.  0.974    0.073    0.974      0.974    0.973      0.920    0.953     0.968

=== Confusion Matrix ===

   a   b   <-- classified as
 954   9 |   a = injured
  23 232 |   b = dead
```

الصورة رقم (3,3,2)

تدريب خوارزمية j48 علي بيانات حوادث المرور :

استخدمت أداة weka 3.8.1 لتدريب خوارزمية j48 علي بيانات الحوادث لعمل (train) لبناء مصنف البيانات وتم اختباره علي بيانات الاختبار (test) ثم تم تقييم أداة المصنف من خلال حساب دقة التصنيف ونسبة الخطأ في التصنيف ويمكن حساب دقة المصنفات بحساب الأتي :-

- نسبة الضبطية (precision) نسبة سجلات الclass المصنفة بشكل صحيح إلي السجلات المصنفة في ال class .
- نسبة الاستدعاء (recall) نسبة سجلات الclass المصنفة بشكل صحيح إلي عدد إلي السجلات في ال class
- الايجابية الكاذبة (FP) أي يتم تصنيف البيانات علي أنها كاذبة وهي في الحقيقة صحيحة
- السلبية الكاذبة (fn) تصنيف علي أنها صحيحة وهي في الحقيقة مهددات .
- الايجابية الصادقة (TP) تصنيف السجلات علي أنها صحيحة وهي في الحقيقة صحيحة.

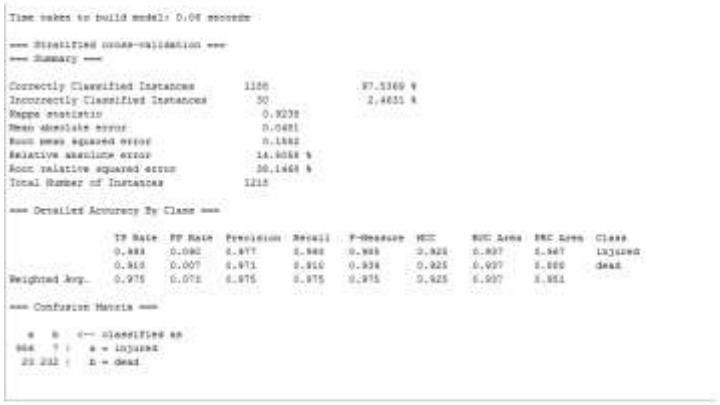
- السلبية الصادقة (fn) تصنيف السجلات علي أنها مهددات وهي في الحقيقة مهددات.

يتم حساب دقة المصنف اعتمادا علي معدل الإنذار الكاذب .

استخدم هذا البحث بيانات حوادث المرور وكانت نتائج التدريب كما يلي :

		Predicated class	
Actual class		Class dead	**Class injured**
	Class dead	956	**7**
	Class injured	23	**232**

بعد ادخال البيانات إلي weka واختيار التدريب المناسب وهنا تم اختيار لخوارزمية j48 وقد تم الحصول علي الصورة رقم (3,3,3).



صورة رقم (3,3,3)

1.13.3 تقييم خوارزميات التصنيف

1. التقييم بالنسبة لخوارزمية j48

		Predicated class	
Actual class		Class dead	**Class injured**
	Class dead	956	**7**
	Class injured	23	**232**

accuracy	recall	FP	Fn	TP	TN	Precision
97.5%	99.2%	9%	0.7%	99.2 %	81.8%	97.6%

2. التقييم بالنسبة لخوارزمية supper vector machine

			Predicated class
Actual class		Class dead	**Class injured**
	Class dead	851	**0**
	Class injured	0	**367**

Accuracy	recall	FP	Fn	TP	TN	Precision
100%	70%	0%	0%	100%	100%	100%

3.التقييم بالنسبة لخوارزمية naïve bayes

Predicated class			
Actual class		Class dead	**Class injured**
	Class dead	850	**1**
	Class injured	61	**306**

accuracy	recall	FP	Fn	TP	TN	Precision
94.9%	99.9%	16.6%	0.11%	99.8%	26.2%	93.3%

1.13.4 مقارنة الخوارزميات في التقييم

	accuracy	recall	FP	FN	TP	TN	Precision
Naïve Bayes	94.9%	99.9%	16.6%	0.11%	99.8%	26.2%	93.3%
J48	97.5%	99.2%	9%	0.7%	99.2 %	81.8%	97.6%
Supper vector machine	100%	70%	0%	0%	100%	100%	100%
The pest algorithm	Supper vector machine	Naïve Bayes	Supper vector machine	Supper vector machine	Supper vector machine	Supper vector machine	Supper vector machine

جدول رقم (3-3-1)

الباب الرابع

النتائج
والتوصيات

النتائج والتوصيات

2.1 النتائج

من خلال مقارنة الخوارزميات التي تم التوصل اليها في الجدول رقم () نجد ان

- تم تحليل البيانات بصورة منهجية وأصبحت جاهزة لاستنتاج علاقات أخري.
- اثبت خوارزمية supper vector machine كفاءتها في تحليل البيانات بمعدل دقة 100%ومعدل خطأ.0%

2.2 التوصيات

نسبةً للإمكانيات المتاحة والزمن لم يتمكن الدارسون من عمل تحليل عميق للبيانات نسبة لمحدودية المتغيرات. لذا نوصي الدارسون ب:

- المحاولة بقدر الإمكان علي الحصول علي البيانات المفقودة التي يمكن أن توثر في عملية بناء النموذج مثل(حالة السائق والإضاءة ونوع الموديل) .
- زيادة عدد البيانات.
- تجربة عدة خوارزميات في عملية التحليل.

2.3 الخاتمة

ومالنا في الختام إلا أن نقول الحمد لله الذي هدانا و ما كنا لنهتدي لولا أن هدانا الله ،أملين أن يستمر البحث حتى يصبح للجهة المعنية نظاما خاصا للتنقيب في البيانات ، وان يستفيد من بعدنا من بحثنا في ما أصبنا وان يستفيدوا مما أخطانا فيه.

المصادر والمراجع

لجنة الامم المتحدة: تحسين السلامة المرورية علي الصعيد العالمي) وضع الاهداف الاقليمية والوطنية للخد من هذة الحوادث المرورية علي [] الطرق. 2012

بشير عباس, العلاق , الادارة الرقمية: المجلات والتطبيقات, مركزالامارات للدرسات الاستراتيجية, ابوظبي2005 [2]

[3] Bazsalica M., Naim P., Data mining pour le Web, éd. Eyrolles, Paris, 2001

عبد الستار العلي، عامر إبراهيم قنديلجي، غساف العمرم، المدخل إلى إدارة المعرفة،دار الدستَة للنشر و التوزيع و الطباعة، الطبعة الاولي [4]

, عمان.2006

[5] Hand d., Mannila H., Smyth R., Principles of Data Mining, MIT Press, London.

[6] Krishnaveni ans Dr. M. Hemalatha, “A perspective analysis of traffic Accident Using Data Mining Techniques”, International Journal of computer Application

[7]Naina Mahajan and Bikram Pal Kaur, “Analysis of Factors of Road Traffic Accidents using Enhanced Decision Tree Algorithm “International Journal of Computer Applications (0975 – 8887) Volume 135 – No.6, February 2016

[8]manner K. Geetha and C. Vaishnavi, “Analysis on Traffic Accident Injury Level Using Classification”, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 2, February 2015, ISSN: 2277 128X.

[[9]دراسة حوداث الطرق والسلامة المرورية, السودان2015

[10] Traffic Accident Analysis Using Machine Learning Paradigms USA 2004.

[11]http://www.alyaum.com/article/4054373