استفاده از تکنیک داده کاوی در خوشه بندی و شناسایی الگوهای تصادفات جاده ای

چکیده

خوشه بندی داده های تصادفات جاده ای با استفاده ار الگوریتمهای داده کاوی منجر به استنتاج مجموعه قوانینی می شود که می تواند توسط پلیس راهنمایی و رانندگی و سازمانهایی نظیر راه و ترابری، حمل و نقل ، پایانه ها و مهندسین بزرگراه برای بهبود ایمنی راهها مورد استفاده قرار گیرد. در این تحقیق برای پیاده سازی تکنیک داده کاوی از مدل استاندارد CRISP- DM استفاده شده است. داده های مورد بررسی مربوط به ۱۴۹۶۰ تصادف در محور کرج - چالوس طی سالهای ۹۳-۹۰ می باشد. برای خوشه بندی متغیرهای تصادف از الگوریتم K-means، Kohonen و Two Step استفاده شده است .به منظور اطمینان از میزان همبستگی متغیرهای جاده ای با شدت تصادفات از مدل همبستگی Apriori و برای تحلیل شدت تصادفات در راههای برون شهری از مدل درخت دسته بندی و رگرسیون (C&R tree) استفاده شده است. نتایج حاصل از مدل خوشه بندی نشان می دهد که متغیرهای جاده ای موثردر بروز تصادفات به ترتیب اهمیت شامل هندسه محل، جهت حرکت راه، خط کشی جاده ، وجودمانع دید، وجود نقص راه، نوع شانه راه، شرایط سطح راه، تعمیرات محل و نوع رویه راه می باشند. نتایج حاصل از به کارگیری مدل همبستگی ضمن تایید نتایج خوشه بندی نشان می دهد که در ۹۵٪ موارد تصادفات در این محور طی ۳ سال گذشته منجر به نوع تصادف خسارتی شده است. نتایج به دست آمده از الگوریتم درخت دسته بندی و رگرسیون ضمن دسته بندی متغیرها نشان می دهد که هر متغیر چند درصد منجر به نوع تصادف جرحی، فوتی و خسارتی می شود که مهمترین عامل مانع دید و کم اهمیت ترین نوع رویه راه می باشد.

واژگان کلیدی: تصادفات جاده ای، داده کاوی، استاندارد CRISP- DM، خوشه بندی، همبستگی، طبقه بندی

Using data mining techniques in clustering and identify patterns of road accidents

abstract

Clustering of road accident data by using data mining algorithms lead to derive a set of rules that can be used by traffic police and organizations such as transportation terminals and highway engineers to improve road safety. In this study to implement data mining techniques, Standard Model CRISP- DM is used.The required data includes 14960 accidents of karaj- chalos road during 90-93year. For clustering, K-means, Kohonen and Two Step algorithms is used. In order to ensure data correlation with the severity of road accidents we use Apriori corelation model andfinally to analyze the severity of accidents on roads classification and regression tree model (C & R tree) is used. The results of the clustering model indicates that the variables contributing in road accident including geometry, accident location, direction of the way, lining the road, the shoulder of the road, road surface conditions, location and type of repairs procedures respectively.The results of the correlation model indicate that among tree accident type (injury, death and damage) the most percentage of accident type is regarded as damage.Classification and Regression Tree algorithm results show the percentage degree of each variables that lead to three accident type that the most important is existance of vision obstruct.

Key words: Road accidents, data mining, Standard CRISP- DM, clustering, correlation, classification

۱.مقدمه

در دهه های اخیر توانایی بشر برای تولید و جمع آوری داده ها به سرعت افزایش یافته است . عواملی نظیراستفاده گسترده از توانایی فناوری اطلاعات، تجهیزات آزمایشگاهی، پیشرفت در جمع آوری داده ها و سیستمهای سنجش از راه دور ماهواره ای، در این تغییرات نقش مهمی داش ته اند. این رشد انفجاری در داده های ذخیره شده باعث پیدایش فناوری جدیدی شده تا این حجم داده را به اطلاعات و دانش تبدیل کند. در این میان داده کاوی سازمان ها را قادر می سازد تا ازسر مایه داده هایشان به درستی بهره برداری نمایند و از این ابزار برای پشتیبانی تصم یم گیری استفاده کنند(محمودی و همکاران ،۱۳۹۲). داده کاوی همچنین پردازش بهینه تصمیم گیری را در سازمانها تسهیل می کند و از طریق استخراج دانش با ارزش از داده ها تصمیم گیری را برای مدیران سازمانها تسهیل می کند. بنابراین ضروری است تا برای استفاده از این ابزار در سازمانها اهمیت بیشتری قائل شد تا در نهایت به فرآیند تصمیم گیری بهینه مدیران منجر شود (شهرابی ،۲۰۰۷). امروزه موضوع تأمین تردد ایمن در سطح شبکه راههای درون شهری و برون شهری یکی ازاصول اساسی حاکم برمهندسی راه، ترافیک و برنامه ریزی حمل و نقل است. عدم وجود ایمنی به خصوص در جاده های برون شهری باعث بروزحوادث ناگواری میگردند. هرساله بیش از۱میلیون نفردرسوانح جاده ای کشته میشوندکه۷۰درصد آنها مربوط به کشورهای درحال توسعه میباشد(تانگ و مک دونالد، ۲۰۰۸)[۱]. بر اساس پیش بینی پروژه بار بیماری سازمان جهانی بهداشت، سوانح ترافیکی می تواند به عنوان سومین علت مرگ و ناتوانی در سال ۲۰۲۰رتبه بندی شود(سازمان بهداشت جهانی، ۲۰۰۵)[۲]. در نتیجه بررسی علل جاده ای مربوط به بروز تصادفات به منظور بهبود ایمنی باید بیشتر مورد بررسی قرار گیرد.

ایمنی راهها از دغدغه های اصلی صنعت حمل ونقل کشور محسوب میگردد و هم اکنون هزینه های بسیاری برای مهار روند افسار گسیخته تصادفات و تلفات جاده ای صرف میگردد. به طور حتم استفاده از

[۱].Tang & Mc Donald
[۲]World health Organization

کارهای علمی در رابطه با بهبود ایمنی راهها درایجاد سیستم و برنامه ریزی های بهتر و کارساز مؤثر خواهند بود. به طور کلی بر اساس مطالعاتی که انجام شده است عوامل موثر بر تصادفات به ۳ دسته تقسیم می شوند: عوامل انسانی ۹۷٪، عوامل مربوط به جاده ۷۰٪ و عوامل مربوط به وسیله نقلیه ۳۱٪. که این عوامل به صورت زنجیره وار به یکدیگر متصل می باشند(تانگ ، ۲۰۰۸)[۱]. درتحقیقات مختلف نقش هر کدام از این سه عامل متفاوت در نظر گرفته شده است. عده ای از محققان معتقدند که عامل انسانی بیشتر از سایرین نقش دارد ولی عده ای دیگر نیز چنین اعتقادی ندارند. در مباحث مربوط به تصادفات جاده ای، مبحث جاده نقش بسیار مهمی را ایفا میکند به این دلیل که آیتمهای دیگر مرتبط با تصادف در ایران با سایر کشورها تقریبا برابری می کند اما آیتم جاده خیلی مشابه با سایرکشورها نیست. هدف این تحقیق ارائه روشی مکانمند مبتنی بر داده کاوی به روش مدل خوشه بندی [۲] جهت خوشه بندی علل جاده ای موثر در بروز تصادفات، مدل همبستگی[۳]جهت اطمینان از میزان همبستگی عوامل جاده ای بانوع تصادف و مدل درخت دسته بندی و رگرسیون [۴] جهت تحلیل شدت تصادفات در راههای برون شهری است . داده کاوی باکشف الگوهای پنهان موجود در داده های تصادفات می تواند نقش مؤثری درکاهش تلفات ناشی از تصادفات در راههای برون شهری داشته باشد (آندرسون،۲۰۰۹). مدلهای خوشه بندی برتشخیص گروههای از رکوردهای مشابه و نام گذاری آن رکوردها با توجه به خوشه ای که به آن تعلق دارند تمرکز دارند. ارزش این مدل به واسطه توانایی آن در کشف گروه بندی های مناسب داده ها و ارائه توصیف مفیدی از آن گروهبندی هاست. مدلهای همبستگی نیز به میزان صحت رابطه میان متغیرها و نوع تصادف می پردازد. مدل درخت دسته بندی و رگرسیون یک مدل داده کاوی کارا و بدون پیش فرض در خصوص رابطه بین متغیرهای مستقل و متغیر هدف است و از روشهایی است که به طور گسترده در کاربردهای مختلف مهندسی استفاده شده است . هدف اصلی این پژوهش تحلیل شدت تصادفات و تعیین

[۱] Tong
[۲] Clustering
[۳] Association
[۴] Classification and Regression tree

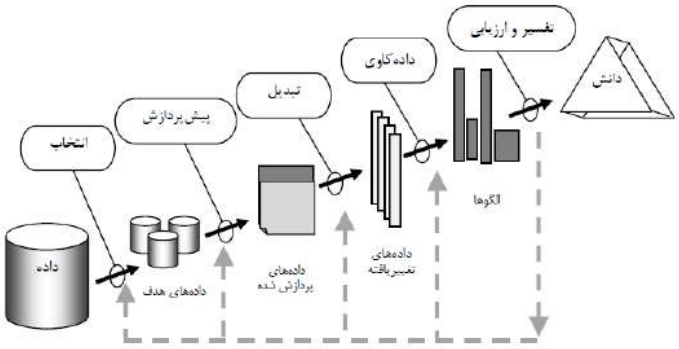
عوامل موثر برآن(به خصوص عوامل جاده ای) در محور کرج – چالوس با استفاده از مدلهای خوشه بندی، همبستگی و طبقه بندی می باشد. در واقع این پژوهش به دنبال پاسخ به این سوال اساسی است که چه عواملی بر شدت تصادفات جاده ای در محورفوق موثر ه‌ستند؟

۲. پیشینه و مبانی نظری پژوهش:

در این بخش ابتدا به تعریف کشف دانش از پایگاه داده، داده کاوی و خوشه بندی پرداخته می شود

کشف دانش از پایگاه داده:

کشف دانش از پایگاه داده در واقع فرآیند تشخیص الگوها و مدلهای موجود در داده هاست. الگو و مدلهایی که معتبر، بدیع، بالقوه، مفید و کاملا قابل فهم هستند (فیاد و شاپیرو و اسمیت ،۱۹۹۶).[۱] به عبارت دیگر، هدف این فرآیند یافتن الگوها و یا مدلهای جالب موجود در پایگاه داده است که در میان حجم عظیمی از داده ها مخفی هستند (غضنفری ، ۱۳۸۷). فرآیند کشف دانش از پایگاه داده شامل مراحل کلی انتخاب، پیش پردازش، تبدیل ، به کارگیری روش های داده کاوی و ارزیابی محصولات حاصل از داده کاوی مطابق با شکل۱می باشد.



شکل۱ : فرآیند کشف دانش از پایگاه داده

[۱] FiAD, Shapiro & Smitt

داده کاوی :

داده کاوی یکی از گامهای فرآیند کشف دانش از پایگاه داده می باشد که شامل به کار گیر ی، تحلیل داده و الگوریتمهای کشف شده می باشد که با پذیرش محدودیتهای محاسباتی، الگوهای خاصی را تولید می کند(فیاد،۱۹۹۶).

خوشه بندی :

خوشه بندی به عمل تقسیم جمعیت ناهمگن به تعدادی از زیر مجموعه ها یا خوشه های همگن گفته می شود. در خوشه بتدی هیچ دسته از پیش تع یین شده ای وجود ندارد و داده ها صرفا بر اساس تشابه،گروهبندی می شوند و عناوین هر گروه نیز توسط کاربر تعیین می گردد (شهرابی،۱۳۹۳). خوشه بندی به این شکل انجام می شود که رکوردهایی که بیشترین شباهت را به هم دارند در یک خوشه قرار می گیرند . در نتیجه داده های موجو د در خوشه های متفاوت کمترین شباهت را به یکدیگر خواهند داشت. هدف در همه الگوریتمهای خوشه بندی کمینه کردن فاصله درون خوشه ای و بیشینه نمودن فاصله بین خوشه ای است. عملکرد خوب یک الگوریتم خوشه بندی زمانی محرز می شود که تا حد امکان خوشه ها را از هم دور کند و بع لاوه رکوردهای موجود در یک خوشه بیشترین شباهت را به یکدیگر دارا باشند . طبق تحقیقی که در سال ۲۰۱۲ انجام شده است یکی از کاربردهای خوشه بندی، کشف الگو در میان داده هاست (لیا و همکاران،۲۰۱۲).[1]به طور کلی حادثه ترافیکی یا تصادف به رویدادی اطلاق می شود که در آن حداقل یک وسیله نقلیه با یک وسیله نقلیه دیگر یا با یک شخص یا یک علرضه برخورد کند که معمولا آسیب مالی جانی یا خسارتی در پی خواهد داشت.

[1] Lia and et all

پیشینه پژوهش:

در ساختار پیشینه تحقیق به پژوهشهایی که در داخل و خارج از کشور که در زمینه بکارگیری الگوریتم های داده کاوی در مباحث مختلف پییش بینی و کاهش تصادفات جاده ای انجام شده است می پردازیم.

در سال ۱۳۸۷ در تحقیقی که توسط محمد سرایی و دیگران انجام شده است با استفاده از الگوریتم درخت طبقه بندی و رگرسیون [۱] به تحلیل داده هایی که درباره تصادفات جاده ای در ناحیه میرلند انگلستان در سال ۲۰۰۰جمع آوری شده است پرداخته شده است. نتایج نشان می دهد متغیرهایی چون گروه سنی، شرایط رانندگی،اندازه ماشین و عوامل مورد انتظاری چون زمان ساعت شلوغی دربروز تصادفات موثر می باشند . علی توکلی کاشانی در سال ۱۳۸۸ طی تحقیقی با استفاده از الگوریتم طبقه بندی CART [۲] به بررسی و تشخیص مهمترین عوامل موثر در میزان شدت جراحات رانندگان در جاده های روستایی پرداختند . نتایج به دست آمده نشان داده است که عدم استفاده از کمربند ایمنی ، سبقت نامناسب و سرعت زیاد مهمترین عوامل در میزان شدت جراحات وارده هستند . در تحقیقی که توسط شان هیل [۳] در سال ۲۰۱۰ انجام شده است با استفاده از الگوریتم kohonen به بررسی وبه کارگیری تکنیکهای داده کاوی برای مرتبط کردن ویژگیهای جاده ای ثبت شده و شدت تصادفات در اتیوپی پرداخته شده است. بر خلاف نتایج تحقیقات گذشته که بر ویژگیهای راننده در علل بروز تصادفات تاکید داشتند این تحقیق عوامل مربوط به جاده (محل تصادف: مناطق بازار یا نزدیک مدرسه، نوع سطح جاده یا جنس آن، شرایط سطح جاده، شرایط آب و هوا، شرایط نور، جهت جاده و....) را عامل اصلی تصادفات در نظر گرفته است . در تحقیقی توسط سینگ چانگ و لی ین [۴] (۲۰۱۱) با استفاده از تکنیک مدلسازی k-means به کشف تاثیرات عوامل غیر رفتاری که شامل ویژگیهای هندسی بزرگراه، عوامل ترافیکی که شامل حجم ترافیک در روز و شرایط محیطی که شامل بارش سالیانه می باشد پرداخته شده است . نتایج پژوهش نشان می دهد عوامل فوق که عوامل خطر نامیده می شوند در بروز شدت تصادفات نقش موثری دارند.

---

[۱] C & R tree
[۲] C&R TREE
[۳] Shan Hill
[۴] Hsing-chung & Li-yen Chang

طی تحقیقی در سا ل ۲۰۱۲ توسط رساک و دیوید [۱] به استفاده الگوریتم k-means به منظور تشخیص نقش عوامل انسانی در بروز و شدت برخوردها و تصادفات جاده ای پرداخته شده است. نتایج پژوهش نشان می دهد بستن کمربند ایمنی، داشتن گواهینامه، سن و جنس در بروز تصادفات نقش موثری دارند . در تحقیقی توسط شین چن [۲](۲۰۱۳) به بررسی عوامل موثر بر بروز تصادفات در منحنی های جاده با استفاده از تکنیکهای داده کاوی پرداخته شده است. نتایج نشان می دهد عوامل موثر در بروز تصادفات به ترتیب عوامل انسانی، جاده ، وسیله نقلیه و عوامل مربوط به راننده می باشند. این عوامل به ترتیب در بروز تصادفات در منحنی ها تاثیر دارند . در تحقیقی توسط مایو چونگ و مارسین پاپکی[۳](۲۰۱۳) با استفاده از الگوریتم طبقه بندی C.۵ [۴] برای جراحات وارده به رانندگان ۵ طبقه در نظر گرفته شده است که شامل بدو ن جراحت، جراحت ممکن، جراحت بدون ناتوانی، جراحت با ناتوانی و جراحات مرگبار می باشد . نتایج نشان می دهد مهمترین عوامل بروز تصادف مرگبارشامل عدم استفاده از کمربند ایمنی، شرایط نور و استفاده از الکل می باشد. همچنین سرعت واقعی وسیله نقلیه هنگام تصادف در تعیین سط ح صدمات وارده بسیار مهم است . با توجه به تحقیقات ذکر شده در حوزه کاربرد داده کاوی در تحلیل داده های مکانی ترافیکی، خلاصه ای از فعالیتهای انجام شده در این زمینه در جدول۲ ذکر می شود.

[۱] Rasak & david
[۲] shin Huey Chen
[۳] Miao M. Chong & Marcin Paprzycki
[۴] C.5

جدول۲: خلاصه ای از تحقیقات انجام شده در تحلیل داده های ترافیکی

ردیف	نام تحقیق	سال	روش تحقیق	نرم افزار	منطقه مورد مطالعه	الگوریتم مورد استفاده	نتایج
۱	استفاده از تکنیکهای داده کاوی در تحلیل داده هایی تصادفات ترافیک جاده ای	۱۳۸۷	طبقه بندی		میرلند انگلستان	C & r tree VIM	تایید عواملی مانند شرایط رانندگی ،اندازه ماشین و زمان ساعت شلوغی در بروز تصادفات اثر دارند
۲	تشخیص مهمترین عوامل موثر در میزان شدت جراحات رانندگان در جاده های روستایی	۱۳۸۸	طبقه بندی	-	جاده های روستایی دوطرفه ایران	CART	عدم استفاده از کمربند ایمنی ، سبقت نامناسب و سرعت زیاد مهمترین عوامل در میزان شدت جراحات وارده هستند.
۳	به کارگیری تکنیکهای داده کاوی برای مرتبط کردن ویژگیهای جاده ای ثبت شده و شدت تصادفات	۲۰۱۰	خوشه بندی	متلب	اتیوپی	kohonen	عوامل مربوط به جاده (محل تصادف: مناطق بازار یا نزدیک مدرسه، نوع سطح جاده یا جنس آن، شرایط سطح جاده، شرایط آب و هوا،شرایط نور، جهت جاده را عوامل اصلی تصادفات در نظر گرفته است
۴	تحلیل تناوب تعداد تصادفات در آزادراه با استفاده از multivariate daptive regression splines	۲۰۱۱	خوشه بندی			k-means	عوامل غیر رفتاری مانند ویژگیهای هندسی بزرگراه،عوامل ترافیکی شرایط محیطی که شامل بارش سالیانه می باشد به عنوان عوامل خطر در بروز تصادفات موثر هستند.
۵	نقش عوامل انسانی در تصادفات و شدت برخوردهای جاده ای	۲۰۱۱	خوشه بندی	متلب	انگلیس	k-means	عامل انسانی موثرترین عامل در بروز تصادفات جاده ای است.نبستن کمربند ایمنی،نداشتن گواهینامه رانندگی، سن و جنس از عوامل موثر در بروز تصادفات هستند
۶	تحلیل تصادفات با استفاده از درخت تصمیم و شبکه های عصبی	۲۰۱۳	درخت تصمیم		فنلاند	C.۵	مهمترین عوامل بروز تصادف مرگبار :عدم استفاده از کمربند ایمنی، شرایط نور و استفاده از الکل می باشد .سرعت واقعی وسیله نقلیه هنگام تصادف در تعیین سطح صدمات وارده بسیار مهم است
۷	بررسی الگوها و فاکتورهای موثر بر بروز تصادفات در منحنی های جاد	۲۰۱۲	خوشه بندی			Kohonen	عوامل جاده،محیطی،عوامل مرتبط با راننده و وسیله نقلیه به طور ترکیبی در بروز تصادفات موثر می باشد
۸	پژوهش حاضر	۲۰۱۵	خوشه بندی طبقه بندی همبستگی	clementine	محور کرج - چالوس	k-means/ Kohonen Two step Apriori C & R tree	عوامل جاده ای موثر در بروز تصادفات به ترتیب شامل هندسه محل، جهت حرکت راه، خط کشی جاده، وجودمانع دید، وجود نقص راه ، نوع شانه راه ، شرایط سطح راه ، تعمیرات محل و نوع رویه راه می باشند

با مرور پژو هشهای انجام شده می توان دریافت که از تکنیک داده کاوی در بررسی عوامل جاده ای موثر بر بروز تصادفات تحقیقات اندکی انجام شده است. در حالیکه همانگونه که در مقدمه بیان شد شناسایی عوامل جاده ای در بروز تصادفات حائز اهمیت است. هم‌چنین در اکثر تحقیقات از روشهای خوشه بندی یا طبقه بندی به تنهایی استفاده شده است. در حالیکه در پژوهش حاضر از ۳ مدل خوشه بندی برای دسته بندی عوامل جاده ای موثر در بروز تصادفات، مدل طبقه بندی به منظور تحلیل شدت و نوع تصادفات در راههای برون شهری و مدل همبستگی به منظور اطمینان از میزان همبستگی متغیرهای جاده ای با شدت و نوع تصادف استفاده شده است .کشف این قوانین می تواند توسط پلیس راهنمایی و رانندگی و سازمان راه و ترابری برای بهبود ایمنی راهها مورد استفاده قرار گیرد.

سوالهای تحقیق:

۱- مهم‌ترین عوامل جاده ای اثرگذار بر افزایش شدت و نوع تصادفات در جاده کرج-چالوس چیست؟

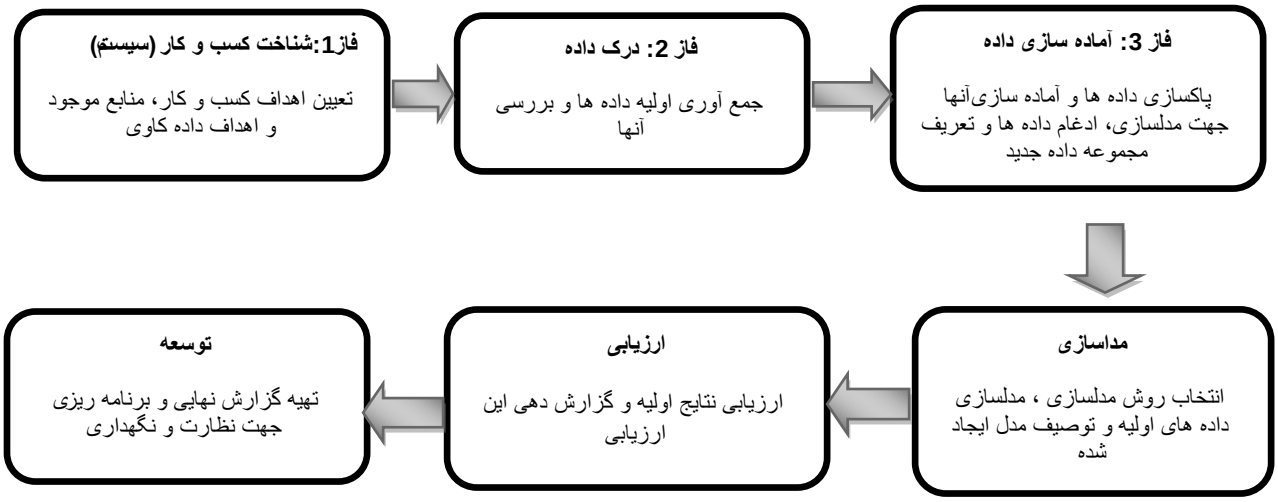
۲- میزان همبستگی عوامل با شدت تصادفات چقدر است؟

۳ . روش شناسی تحقیق:

همانطور که گفته شد هدف اصلی تحقیق بررسی عوامل جاده ای موثر در بروز تصادفات در جاده های برون شهری با استفاده از تکنیکهای دا ده کاوی می باشد . در این تحقیق از مدل استاندارد CRISP-DM[۱] که یکی از روشهای بسیار قوی در د اده کاوی می باشد و توسط شرکتهای دایملر کرایسلر[۲]، اس پی اس اس [۳] و ان سی آر [۴]توسعه یافته است استفاده شده است. این متدولوژی از گامهای شناخت سیستم، در ک داده، آماده سازی داده ها، مدل سازی، ارزیابی و توسعه سیستم مطابق شکل ۲ تشکیل شده است(شهرابی،۱۳۹۰). برای خوشه بندی متغیرها از ۳ الگوریتم خوشه بندی

[۱] CRISP-DM
[۲] Daimler Chrtsler
[۳] SPSS
[۴] NCR

k-means, kohonen و two step برای طبقه بندی متغیرها از الگوریتم خوشه بندی درخت رگرسیون و طبقه بندی(CART) و برای بررسی همبستگی متغیرها با نوع تصادف از الگوریتم همبستگی Apriori استفاده شده است.



شکل۲: مراحل کریسپ

برای داشتن تحقیقی موفق در زمینه داده کاوی، باید متخصص داده کاوی از توان و تجربه متخصص کسب و کار در تمام فرآیند داده کاوی بهره مند شود(رادفر، نظافتی و یوسفی اصل ، ۱۳۹۳).

الگوریتم k-means

این الگوریتم یکی از ساده ترین و البته مشهورترین الگوریتمهای یادگیری غیر هدایت شده است .درk-means عملا مجموعه داده ها به تعداد خوشه های از پیش تعیین شده تقسیم می شوند .ایده اصلی در این الگوریتم تعریف k مرکز برای هریک از خوشه هاست. بهترین انتخاب برای مراکز خوشه ها در الگوریتم k-means قرار دادن آنها در فاصله

هرچه بیشتر از یکدیگر است. پس از آن هر رکورد در مجموعه داده به نزدیکترین مرکز خوشه تخصیص می یابد .از نقاط قوت این است که معمولا سریع ترین روش برای خوشه بندی مجموعه داده های بزرگ است.

الگوریتم TwoStep

این الگوریتم از یک روش خوشه بندی دو مرحله ای استفاده می کند . مرحله اول با یک بار گذر از داده ها، آنها را در مجموعه قابل قبولی از زیرشاخه ها فشرده می کند. قدم دوم از یک روش خوشه بندی سلسله مراتبی، به منظور ادغام تکاملی این زیر خوشه ها به خوشه های بزرگتر، بهره می برد (شهرابی،۱۳۹۲). یکی از مزایای این الگوریتم اجرا بر روی مجموعه داده های بزرگ است (شهرابی،۱۳۹۲). و آنها را با کاآیی زیاد اداره می کند (علیزاده ،۱۳۹۲). از دیگر نقاط قوت ان این است که قادر به مدیریت داده هایی با انواع مختلف فیلدهاست.

الگوریتم Kohonen

این الگوریتم که با نام الگوریتمهای خودسازمانده نیز معروف هستند، از نوعی شبکه عصبی به منظور خوشه بندی مجموعه داده ها به خوشه های مجزا استفاده میکند . هنگامی که شبکه به طور کامل آموزش دید، رکوردهای شبیه هم در نقشه خروجی مجاور هم قرار می گیر ند، درحالیکه رکوردهایی که متفاوتند دور از هم واقع می‌شوند(علیزاده،۱۳۹۲). هریک از ۳ الگوریتم خوشه بندی با معیارهای زمان اجرا، تعداد خوشه ها و شاخص سیلوئیت[۱] بر روی داده ها اعمال می شوند تا بهترین الگوریتم انتخاب شود .

شاخص تراکم و جدایی سیلوئیت با مقادیر ضعیف،متوسط و خوب نشان داده می شود. میانگین مقدار شاخص سیلوئیت برای ارزیابی اعتبار خوشه بندی و همچنین برای تصمیم گیری در مورد انتخاب تعداد کلاسهای بهینه مورد استفاده قرار می گیرد که این میزان بر اساس دوری و نزدیکی مشاهدات و خوشه ه ا به یکدیگر محاسبه می شوذ .مقدار S(i)با استفاده از فرمول زیر قابل محاسبه است:

[۱] Silhouette

$$S(i)=\frac{(b(i)-a(i))}{\max(a(i),b(i))} \quad (۱)$$

A(i) میانگین فاصله بین مشاهده i با سایرمشاهدات در یک خوشه مشابه و b(i) میانگین فاصله مشاهده به تمام مشاهدات در خوشه های دیگر می باشد .بر اساس فرمول بالا مقدار S(i) بین ۱- و ۱+ قرار دارد. اگر S(i) به ۱+ نزدیکتر باشد، به این معنی است که خوشه بندی نمونه خوب صورت گرفته است و خوشه پیشنهاد شده برای نمونه مورد نظر مناسب می باشد .ولی اگر S(i)به۱- نزدیکتر باشد به این معنی است که خوشه بندی نمونه، به خوبی انجام نشده و خوشه پیشنهاد شده برای داده مورد نظر نامناسب می با شد.برای تعیین میزان همبستگی بین نتایج حاصل از الگوریتمهای خوشه بندی نیز از الگوریتم همبستگی Apriori استفاده شده است.

الگوریتم Apriori

این الگوریتم بطور گسترده ای برای استخراج قوانین وابستگی استفاده می شود . هدف این الگوریتم پیدا کردن همه وابستگی‌ها در میان آیتم‌هاست، با استفاده از مجموعه مشخص تراکنش ها (هر تراکنشی شامل مجموعه ای از آیتم ها) با minimum support و minimum confidence مشخص، سعی در پیدا کردن همه قوانین وابستگی دارد که شرایط minimum support, minimum confidence را برآورده می‌کند.

مدل درخت دسته بندی و رگرسیون

درخت تصمیم یکی از رایج ترین تکنیکهای داده کاوی است که سادگی و کارآمدی آن باعث شده تا علی رغم مشکلاتی که در اجرای الگوریتم با متغیرهای دارای نویز یا صفات فاقد مقدار وجود دارد، به شکل گسترده ای در کاربردهای مختلف و مسائل مربوط به یادگیری ماشینی استفاده شود.درخت دسته بندی و رگرسیون الگوریتمی از درخت تصمیم است که دارای هر دو قابلیت دسته بندی و رگرسیون است (هان و همکاران، ۲۰۰۶). ساختار مدل درخت دسته بندی و رگرسیون به گونه ای است که در ابتدا همه داد ه ها در اولین گره(گره ریشه) قرار می گیرند و

سپس بر اساس متغیر جداکننده ای که بیشترین همگنی و خلوص را برای هر شاخه ایجاد میکنددر ریشه اشعاب ایجاد می شود. عمل تقسیم بندی متغیرها در شاخه های درخت آن قدر ادامه می یابد که داده های موجود در هر گره بیشترین همگنی را برای تعلق به یک دسته خاص داشته باشند .گره هایی که در انتهای درخت قرار می گیر ند گره نهایی یا برگ نامیده می شوند به گره هایی که مابین گره ریشه و گره های نهایی هستند گره میانی گفته می شود

۴.یافته های تحقیق

در این بخش ضمن معرفی داده و نمونه مورد بررسی، یافته های حاصل از پیاده سل زی روش پیشنهادی مورد تجز یه و تحلیل قرار میگیرد.

۱-داده و منطقه مورد مطالعه : منطقه مورد مطالعه محور کرج- چالوس به طول تقریبی ۱۶۰ کیلومتر(۶۰ مابل) می باشد.از ۱۶۰ کیلومتر طول مسیر کرج - چالوس ،۷۲ کیلومتر آن در حوزه جغرافیایی استان البرز قرار دارد (شکل۳) . جاده چالوس، با نام رسمی جاده۵۹، یکی ازمهمترین جاده ها برای مردم تهران و کرج است که از میان شهر کرج در استان البرز شروع و به شهر چالوس در مازندران وصل می شود. این مسیر از راههای اصلی برون شهری پرخطر کشور بوده که بسیار مستعد تصادف است. با ساخت این جاده بود که دسترسی مهندسان سد سازی بر دره های تو درتوی رودخانه کرج آسان شد و دریاچه زیبای سد امیر کپر را برای تامین آب آشامیدنی پایتخت نشینان و اراضی کشاورزی پایین دست بر دل کوه و دره نشاند .در تکمیل جاده، تونل کندوان به مسافت یک هزارو ۸۸۶ اردیبهشت ماه سال ۱۳۱۴ شمسی با عرض بین پنج الی هفت متر و ارتفاع شش متر ساخته شد، تونلی که مسیر جاده را ۱۳ کیلومتر کاهش داد و پس از آن تونل های دیگرروزنه دل کوه شدند. تردد خودرویی دراین جاده در روزهای عادی بطورمیانگین بین ۳۰ تا ۴۰ هزار دستگاه خودرو است که درایام پایانی هفته ترددها ۲۰ درصد افزایش می یابد. پس از جمع آوری و انتقال داده ها به پایگاه داده، پیش پردازشهای لازم بر روی داده های تصادفات قبل از فرآیند داده کاوی انجام می گیرد تا از ایجاد خوشه ها و الگوهای نامناسب جلوگیری شود. این پیش پردازشها شامل نرمال سازی متغیرهای ورودی،

هم مقیاس کردن و حذف داده های تکراری و ناقص می باشد . همچنین برخی از متغیرهای ورودی با هم ترکیب می شوند تا حافظه و حجم پردازشهای لازم به منظور تحلیل شدت تصادفات کاهش یابد.

طول : ۱۶۰ کیلومتر(۶۰ مایل)

ابتدای محور: شهر کرج در استان البرز

انتهای محور: چالوس در استان مازندران

سهم استان البرز از موقعیت جغرافیابی محور:۷۲کیلومتر

تردد خودرویی: در روزهای عادی بطورمیانگین بین ۳۰ تا ۴۰ هزار دستگاه خودرو ،درایام پایانی هفته افزایش ۲۰ درصد ترددها

شکل۳:موقعیت جغرافیایی محور کرج-چالوس

۲-متغیرهای مورد استفاده در تحقیق:

داده های مربوط به تصادفات با استفاده از فرمهای کام ۱۱۴ استخراج می شود. مجموعه داده ه‌ای کام ۱۱۴، توسط فاوا ناجا با همکاری راهور طراحی و به وسیله پلیس راهور مدیریت میشود. این مجموعه داده ها، مرجع رسمی عمومی تمام تصادفات جاده ای ایران است که به وسیله افسران کارشناس تصادف، اطلاعات مورد نظر در این سیستم ثبت میشود. نسخه کنونی این دادگان تصادفات، نسخه ویرایش شده ای از کام ۱۰۷ تا کام ۱۱۳ است و همچنین اطلاعات مربوط به بیش از یک میلیون تصادف را در بردارد و بر دو نوع خسارتی و جرحی می باشد . داده های تصادفات توسط افسر پلیس برای هر رویداد تصادف تکمیل می شودو درآخر هر ماه با توجه به وضعیت سلامتی مجروحان

تصادف (برای مثال افزایش تعداد متوفیان تصادف در روزهای پس از حادثه) اطلاعات بروز می شوند . مطالعات پیشین نشان می دهد داده های تصادفات از سالی به سال دیگر متفاوت و دارای تغییر پذیری آماری قابل توجهی هستند. در این تحقیق داده های مربوط به تصادفات در محور کرج- چالوس بین سالهای ۹۳-۹۰ برای مطالعه ۱۴۹۶۰تصادف به کار گرفته شد ه اند. در جدول۲ به تعریف این متغیرها پرداخته می شود . متغیرنوع تصادف، متغیر هدف(متغیر وابسته) است که خود شامل ۳ سطح خسارتی، جرحی و فوتی می باشد و ۱۰ متغیر دیگر متغیرهای پیش بینی کننده هستند از آنجا که متغیرهای مورد بررسی شامل عوامل مکانی راه و عوارض مجاور آن، ترافیک و همچنین تعداد زیادی از متغیرهای ثبت شده در رکوردهای تصادف هستند، این حجم از داده ورودی فرآیند پیش بینی شدت تصادف را پیچیده می کند. از طرفی چنانچه متغیرهای ورودی مدل وابسته بوده یا ارتبـاطی با خروجی نداشته باشند صحت یا قابلیت اطمینان پیش بینی یا دسته بندی تحت تاثیر قرار خواهد گرفت(میتچل،۱۹۹۷). استفاده از متغیرهای وابسته ضمن افزایش مدت زمان لازم جهت ایجاد مدل، کارایی فرآیند پیش بینی را کاهش داده و تفسیر نتایج به درستی صورت نخواهد گرفت لذا قبل از فرآیند پیش بینی، وابستگی متغیرها مورد بررسی قرار گرفته و با حفظ صحت، تعداد متغیرهای ورودی به حداقل ممکن کاهش می یابد.

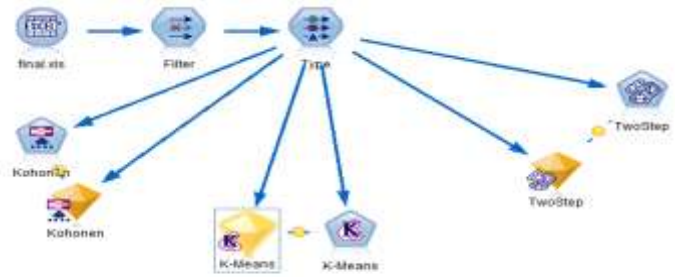
جدول ۲ : تعریف متغرهای تصادف

متغیر	تعریف متغیر
نوع تصادف	خسارتی ، جرحی ، فوتی
مانع دید	تل خاک، ساختمان ،کیوسک، شیب، طوفان ،شن مه/دود ،نور خورشید، نور چراغ ،وسیله نقلیه مقابل ،وسیله در حال حرکت ،وسیله متوقف، کولاک، درخت بوته قوس قائم، ندارد
نوع رویه راه	شنی و خاکی ، آسفالت
تعمیرات راه	در حال تعمیر با علائم کافی ، در حال تعمیر بدون علائم کافی، در حال تعمیر هست، در حال تعمیر نیست
شرایط سطح راه	یخبندان و برفی، شنی و خاکی، روغنی و کثیف،خشک ، تر
هندسه محل	مستقیم، سربالایی/سر پایینی ،پیچ، مسطح ،مستقیم ،مسطح
خط کشی	ممتد، مقطع، ندارد
نوع شانه	شانه ندارد ، شانه خاکی ، شانه آسفالته
جهت حرکت راه	دوطرفه مجزا ،شیب عرضی و طولی غیر استاندارد، قوس با زاویه تند، ، یک طرفه، کم عرض بودن معبر، دوطرفه غیر مجز

نقص راه	اختلاف سطح بین آسفالت و شانه، فقدان حفاظ ایمنی کنار معبر،فقدان شانه خاکی و پارکینگ، نقص روشنایی معبر،قوس با زاویه تند، لغزندگی سطح جاده، نشست جاده ای، نقص خط کشی معبر، نقص رو شنایی معبر، نقص علائم عمودی ، کم عرض بودن معبر،شیب عرضی و طولی غیر استاندارد، نقص رویه آسفالت، نقص علائم افقی، وجود مانع دست انداز، ندارد
علت تامه	انحراف به راست .انحراف به چپ، باز کردن ناگهانی در، تجاوز به چپ ناشی از سرقت، تغییر مسیر ناگهانی، حرکت در خلاف جهت، دور زدن در محل ممنوع ، عبور از محل ممنوع ،عدم توانایی در کنترل وسیله نقلیه، عدم توجه به جلو، عدم رعایت حق تقدم ،عدم رعایت فاصله، نقص مقررات حمل بار، نقص فنی مستمر وسیله نقلیه ،حرکت با دنده عقب، تغییر مسیر ناگهانی، تخطی از سرعت مطمئنه ،تجاوز از سرعت‌مقرر

۳-تجزیه و تحلیل و ارزیابی:

پس از پاکسازی داده های مورد نظر در این مرحله از تکنیک خوشه بندی استفاده شده است. بدین منظور از نسخه نرم افزار IBM SPSS Modeler استفاده شده است. در این نرم افزار سه نوع الگوریتم kohonen، k-means و Two step برای مدل سازی خوشه بندی وجود دارد. هر یک از این ۳ الگوریتم بر روی داده ها اعمال شده است تا بهترین الگوریتم انتخاب شود. در این مرحله از جدول اکسلی که حاوی متغیرهای مربوط به تصادفات جاده ای می باشد استفاده شده است. برای ساخت جریان داده ابتدا فایل اکسلی که برای این کار ساخته شده است به عنوان منبع داده وارد می شود . بغیر از صفت سال بقیه صفتها فیلتر شده اند و سپس با استفاده از گره نوع، ویژگیهای صفت تعیین شده است و پس از آن هر سه الگوریتم ذکر شده روی داده ها اجرا می شود. شکل ۳ مراحل ساخت جریان داده را نشان می دهد.



شکل ۳: نمودار جریان داده برای خوشه بندی داده ها

برای انتخاب الگوریتم مناسب جهت خوشه بندی متغیرها زمان اجرا ، تعداد خوشه ها و شاخص سیلوئت استفاده شده است. بدین معنا که هرچه زمان اجرا کمتر، تعداد خوشه ها نتناسب با نظر خبره و شاخص سیلوئیت بیشتر باشد الگوریتم مناسب تر است. مطابق نظر خبره تعداد خوشه ها بین ۳ تا ۵ مناسب است. جدول ۳ مقایسه معیارهای فوق در هر یک از ۳ الگوریتم خوشه بندی را نشان می دهد.

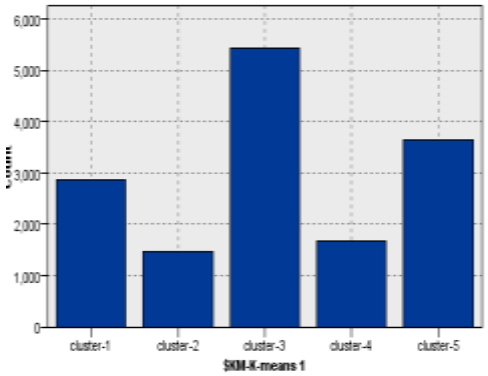
جدول ۳ : مقایسه معیارهای انتخاب در الگوریتمهای خوشه بندی مورد نظر

معیار / نام الگوریتم	زمان اجرا	تعداد خوشه ها	شاخص سیلوئیت
K MEANS	۱<	۵	۰.۱۸۴
TWO STEP	<۱	۳	۰.۱۱۹
KOHONEN	<۱	۱۲	۴۹۰/۰

مطابق جدول ۳ ارآنجائیکه الگوریتم kohonen از نظر هر سه معیار ضعیف تر می باشد انتخاب نمی شود . با توجه به اینکه زمان اجرا در دو الگوریتم دیگر یکسان است اما شاخص سیلوئیت در الگوریتم k- means بیشتر است در نتیجه الگوریتم k- means انتخاب شده است. سپس الگوریتم پذیرفته شده در بخش مدلسازی تحقیق روی داده ها اعمال شده و نتایج حاصل از آنها در قالب نمودار و جدول ارائه شده است.

جدول ۴ : نتایج خوشه بندی با الگوریتم k- means

نام خوشه به ترتیب اولویت	سایز خوشه به درصد
۳	۵۰/۲۶
۵	۱۵/۶۴
۱	۱۳/۴۷
۴	۱۰/۲۳
۲	۱۰/۴

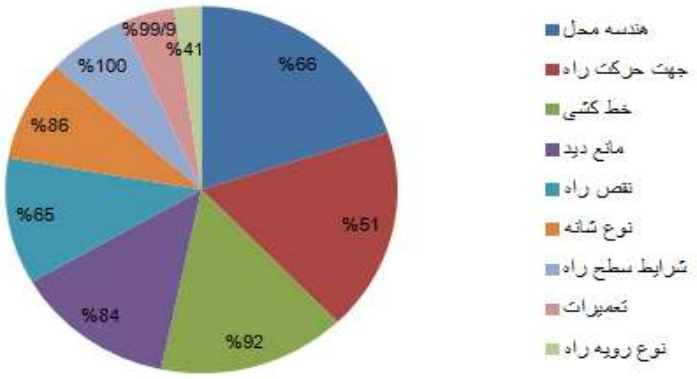


شکل۴: تفکیک خوشه ها

با توجه به نتایج به دست آمده خوشه ۳ با فراوانی ۵۰/۲۶درصد به عنوان خوشه بهینه شناخته شده و خوشه ۲ با فراوانی ۱۰/۴ درصد به عنوان ضعیف ترین خوشه در نظر گرفته شده است . در نتیجه در ادامه تحلیلها بر اساس متغیرهای موجود در خوشه ۳ می باشد.در جدول ۵ و شکل۴ متغیرهای موجود در خوشه ۳ که در مرحله مدلسازی قبل به دست آمده است، به ترتیب اهمیت نشان داده می شوند.

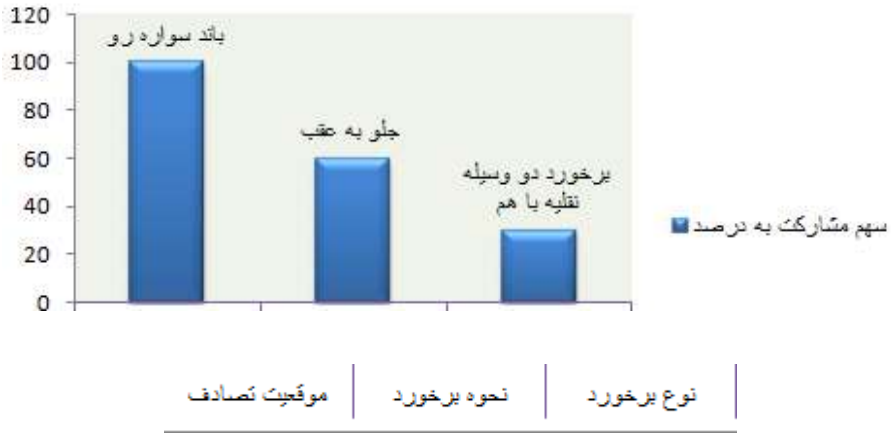
جدول۵: متغیرهای پیش بینی کننده به ترتیب اهمیت

متغیر	موقعیت	سهم متغیرها به درصد
هندسه محل	مستقیم ، مسطح	۶۶
جهت حرکت راه	دو طرفه غیر مجزا	۵۱
خط کشی	مقطع	۹۲
مانع دید	ندارد	۸۴
نقص راه	ندارد	۶۵
نوع شانه	ندارد	۸۶
شرایط سطح راه	خشک	۱۰۰
تعمیرات	در حال تعمیر نیست	۹۹/۹
نوع رویه راه	آسفالت	۴۱



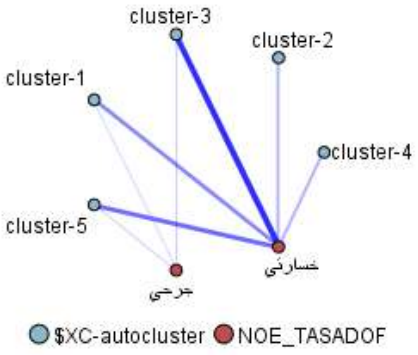
شکل۴: توزیع دمتغیرهای پیش بینی کننده به ترتیب اهمیت

با توجه به جدول ۵ عوامل موثر در بروز تصادفات به ترتیب اهمیت شامل هندسه محل، جهت حرکت راه، وجود خط کشی، مانع دید، نقص راه، نوع شانه راه، شرایط سطح راه، تعمیرات محل و نوع رویه راه می باشد . به عنوان مثال در ۶۶٪ موارد تصادفاتی که منجر به نوع خسارتی می شوند در هندسه محل مستقیم، مسطح اتفاق می افتد .با توجه به اینکه موقعیت متغیرهای مانع دید، نقص راه و نوع شانه راه گزینه ندارد می باشد لذا اهمیت متغیرهای دیگر در بروز تصادف افزایش می یابد.در مرحله بعد، مدلسازی با الگوریتمهای خوشه بندی برای متغیر نوع داده به کار برده شد. این متغیرها شامل موقعیت تصادف ، نحوه برخورد و نوع برخورد می باشند. نتایج حاصل از خوشه بندی در شکل۵ نشان داده شده است.



شکل۵: نتایج خوشه بندی نوع داده

همانطور که مشاهده می شود بیشترین آمار تصادف در باند سواره رو، بیشترین برخورد ماشینها با یکدیگر از جلو وسیله نقلیه به عقب وسیله نقلیه دیگر و بیشترین درصد نوع برخورد مربوط به برخورد دو وسیله نقلیه با یکدیگر می باشد . سپس به منظور بصری سازی داده ها گراف خوشه ها مطابق شکل ۶ نشان داده شده است.



شکل ۶: گراف خوشه بندی متغیرهای جاده ای تصادفات

همانطور که مشاهده می شود خوشه ۳ که بهترین خوشه به دست آمده می باشد ، بیشتر از خوشه های دیگر منجر به نوع خسارتی می شود و خوشه های ۵، ۱، ۲ و ۴ به ترتیب در اولویتهای بعدی قرار دارند. پس از خوشه بندی داده ها

برای پی بردن به میزان همبستگی بین متغیرها و نوع تصاد ف که شامل ۳ طبقه خسارتی، جرحی و فوتی می باشد از الگوریتم همبستگی Apriori استفاده شده است.

قواعد ارتباط شبکه ای

قواعد ارتباط شبکه ای اولین بار توسط اگراوال [۱](۱۹۹۳) معرفی شد. پس از آن بسیاری از محققان این الگوریتم را توسعه دادند. یک قاعده ارتباط شبکه ای به صورت یک عبارت $x \rightarrow y$ بیان می شود که در آن x مجموعه ای از اقلام و y یک قلم می باشد به این معناست که اگر مجموعه اقلام x رخ دهد انگاه قلم y با احتمال مشخصی رخ خواهد داد. وظیفه اصلی قواعد ارتباط شبکه ای شناسایی الگوهای پنهان میان اقلام موجود در پایگاه داده های بزرگ می باشد. در الگوریتم قواعد ارتباط شبکه ای استاندارد جهت یافتن قواعد جذاب و کارا برای هر قاعده دو مقدار عددی پشتیبانی و اطمینان به صورت زیر محاسبه می شود.

پشتیبانی(sup) : نسبت تعداد تراکنشها با همه اقلام موجود در قاعده به تعداد کل تراکنشها.

اطمینان (conf): نسبت تعداد تراکنشها با همه اقلام موجود در قاعده به تعداد تراکنشها با اقلام موجود در جمله شرطی).

در واقع، مقدار پشتیبانی یک قاعده نمایانگر میزان فراوانی، و مقدار اطمینان نمایانگر قدرت آن می باشد. جهت انتخاب قواعد با میزان فراوانی و قدرت مطلوب، دو مقدار عددی حداقل پشتیبانی (minsup) و حداقل اطمینان (minconf)به عنوان مقادیر آستانه توسط تحلیل گر تعیین می گردد . اگر برای یک قاعده دو شرط Sup>minsup و conf >minconf برقرار شد آن قاعده به عنوان یک قاعده ارتباط شبکه ای مطلوب انتخاب می شود در غیر اینصورت رد می گردد. با توجه به در نظر گرفتن مقادیر آستانه و برقراری شرط $x \rightarrow y$ قواعد مطرح شده در جدول ۷ به عنوان قاعده شبکه ای مطلوب در نظر گرفته شده است.

[۱] Agraval

در این مرحله برای ساخت جریان داده ابتدا فایل کسلی که برای این کار ساخته شده است به عنوان مربع داده وارد می شود بغیر از صفت سال بقیه صفتها فیلتر شده اند و سپس با استفاده از گره نوع، ویژگیهای صفت تعیین شده است و پس از آن الگوریتم Apriori روی داده ها اجرا می شود. جهت درک بهتر قواعد به دست آمده آنها را به صورت یک جدول نشان می دهیم به طوری که با با توجه به ۳ طبقه (خسارتی ، جرح ی، فوتی) در نظر گرفته شده برای نوع تصادف بیشترین فراوانی مربوط به خسارتی می باشد که این مورد نتیجه به دست آمده در مرحله خوشه بندی را نیز تایید می کند.

جدول۷: قواعد ارتباط شبکه ای نهایی

Sup(%)	Conf(%)	متغیر	موقعیت	نوع تصادف
۹۹/۳	۸۸/۵	شرایط سطح راه	خشک	خساری
۹۰/۹	۱۲/۹	هندسه محل	مستقیم/سربالایی/ سرپائینی	خسارتی
۹۰/۸	۲۰/۵	نوع شانه	خاکی	خسارتی
۹۱/۸	۲۱/۵	جهت حرکت راه	دوطرفهغیر مجزا	خسارتی
۹۰/۶	۷۴/۴	نقص راه	ندارد	خسارتی
۹۱/۵	۱۲/۲	خط کشی	مقطع	خسارتی
۹۰/۶	۲۵/۶	تعمیرات راه	در دست تعمیر نیست	خسارتی

جدول۷ نشان می دهد با توجه به دوقاعده پشتیبانی و اطمینان تا چه حد می توان به نتایج به دست آمده از میزان همبستگی بین متغیرها و نوع تصادف اطمینان کرد . مطابق خروجیهای به دست آمده از اجرای این الگوریتم در مورد متغیر مانع دید ۹۱ درصد می توان بهعدم وجود مانع دید در مسیر مورد مطالعه اطمینان کرد . همچنین درباره شرایط سطح راه ۸۸/۵ درصد می توان به خشک بودن سطح راه اطمینان کرد. نتایج به دست آمده در این مرحله نتایج حاصل از خوشه بندی در مرحله پیش را نیز تایید می کند . پس از انجام محاسبات هزینه دسته بندی اشتباه و هرس شاخه های غیر ضروری که با توجه به دانش متخصصان ایمنی راه و کارشناسان پلیس راه انجام گرفته است درخت دسته

بندی و رگرسیون بهینه با استفاده از داده های آموزشی به منظور تحلیل نوع تصادفات ایجاد می شود . در این قسمت متغیر نوع تصادف متغیر وابسته می باشد که خود دارای ۳ طبقه خسارتی، جرحی و فوتی است.۸ متغیر دیگر متغیرهای مستقل می باشند که شامل وجود مانع دید، شرایط سطح راه، هندسه محل، خط کشی جاده، نوع شانه راه، جهت حرکت راه، نقص راه و نوع رویه راه می باشد. پس از مشخص کردن نوع متغیرها با ایجاد جریان داده از فایل اکسل از الگوریتم درخت دسته بندی و رگرسیون برای طبقه بندی عوامل استفاده می کنیم . در جدول ۸ خلاصه ای از نتایج به دست آمده در این مرحله نشان داده شده است.

جدول۸ : خلاصه ای از توزیع نوع تصادف با متغیرهای کلیدی

	نوع تصادف					
متغیرهای مستقل	جرحی		خسارتی		فوتی	
مانع دید						
تل خاک، ساختمان ،کیوسک، شیب، طوفان ،شن مه/دود ،نور خورشید، نور چراغ ،وسیله نقلیه مقابل ،وسیله در حال حرکت ،وسیله متوقف، کولاک	%۵.۸۸۲	۱۷	۹۴.۱۳۸	۲۷۳	%۰.۰۰۰	۰
درخت بوته قوس قائم، ندارد	%۸.۱۵۶	۷۷۰	۹۰.۰۶۵%	۸۵۰۳	%۱.۷۷۹	۱۶۸
شرایط سطح راه						
تر ،روغنی و کثیف ،یخبندان و برفی	%۵.۰۸۹	۶۶	%۹۴.۱۴۰	۱۲۲۱	%۰.۷۷۱	۱۰
خشک شنی و خاکی	%۸.۳۳۵	۷۲۶	%۸۹.۸۳۹	۷۸۲۵	%۱.۸۲۵	۱۵۹
هندسه محل						
مستقیم، سربالایی/سر پایینی ،پیچ، مسطح	%۶.۶۱۰	۱۹۴	%۹۲.۲۶۶	۲۷۰۸	%۱.۱۲۴	۳۳
مستقیم ،مسطح	%۸.۴۱۹	۵۸۷	%۸۹.۷۵۹	۶۲۵۸	%۱.۸۲۲	۱۲۷
خط کشی						
کل	٪۷/۸۹۳	۶۷۳	٪۹۰/۳۲	۷۶۹۸	٪۱/۷۸	۱۵۲
نوع شانه راه						
شانه آسفالته،شانه ندارد	%۸.۳۲۴	۶۴۱	%۹۰.۰۱۴	۶۹۳۲	%۱.۶۶۲	۱۲۸
شانه خاکی	%۷.۱۰۰	۱۴۱	%۹۱.۲۳۹	۸۱۲	۱.۵۵۲%	۳۳
جهت حرکت راه						
دوطرفه مجزا ،شیب عرضی و طولی غیر استاندارد، قوس با زاویه تند، ، یک طرفه، کم عرض بودن معبر	%۷.۶۹۵	۵۷۰	%۹۰.۷۷۹	۶۷۲۴	%۱.۵۲۶	۱۱۳
دوطرفه غیر مجزا	%۹.۱۲۶	۲۱۰	%۸۸.۷۸۷	۲۰۴۳	%۲.۰۸۶	۴۸
نقص راه						
اختلاف سطح بین آسفالت و شانه، فقدان حفاظ ایمنی	%۷.۸۷۸	۷۶۳	%۹۰.۳۲۳	۸۴۳۸	%۱.۷۹۸	۱۶۸

کنار معبر،فقدان شانه خاکی و پارکینگ، نقص روشنایی معبر،قوس با زاویه تند، لغزندگی سطح جاده، نشست جاده ای، نقص خط کشی معبر، نقص روشنایی معبر، نقص علائم عمودی ،						
شیب عرضی و طولی غیر استاندارد، نقص رویه آسفالت، نقص علائم افقی، وجود مانع دست انداز، ندارد	%۱۶.۵۰۰	۳۳	۸۰.۵۰۰%	۱۶۵	%۱.۰۰۰	۲
نوع رویه راه						
کل	%۸.۱۴۲		۸۹.۸۲۳%		%۲.۰۳۵	

مطابق جدول۸ هریک از متغیرها به صورت یک درخت که دارای یک ریشه و دو شاخه می باشد نشان داده شده اند . مهمترین عامل مانع دیددر نظر گرفته شده است. درشاخه اول که شامل وجود متغیرهای تل خاک، ساختمان ،کیوسک، شیب، طوفان، شن، مه ، نور خورشید، نور چراغ، وسیله نقلیه مقابل ، وسیله در حال حرکت ، وسیله متوقف و کولاک دسته بندی شدند از تعداد۱۷ تصادف، ۵/۸۸۲٪ تصادفات منجر به نوع جرحی، از ۲۷۳ تصادف، ۹۴/۱۳۸٪ منجر به نوع خسارتی و ۰٪ منجر به نوع فوتی می شود و در شاخه دوم که شامل وجود متغیرهای درخت، بوته، قوس قائم و عدم وجودمانع دید می باشد، ۸/۱۴۲٪ تصادفات منجر به نوع جرحی، ۸۹/۸۵۲۲٪ منجر به نوع خسارتی و ۲/۰۳۵ ٪ منجر به نوع فوتی می شود . نتایج همچنین نشان می دهد که در ۹۹ درصد موارد متغیرهای تصادف منجر به نوع تصادف خسارتی شده اند که تایید کننده نتایج حاصل از مرحله قبل نیز می باشد.

بر اساس نظر بریمن (۲۰۰۱) به منظور توسعه مدل C&R tree، مجموعه داده باید به طور تصادفی به دو زیر مجموعه[۱] آموزشی [۲]و تست [۳] تقسیم شوند. در این تحقیق میزان داده های تخصیص داده شده به زیر مجموعه های آموزشی و آزمایش به ترتیب ۷۰ و ۳۰ می باشد. جدول ۹ صحت پیش بینی مدل را که با تقسیم تعداد داده ها ی پیش بینی شده صحیح به تعداد کل داده ها ی مشاهده شده برای داده های تست و آموزش محاسبه شده است را نشان می دهد.

[۱] Subset
[۲] Training
[۳] Testing

جدول ۹ : صحت پیش بینی مدل برای ۳ کلاس

صحیح پیش بینی شده	داده های آزمایش	صحیح پیش بینی شده	داده های آموزش	
٪۹۱/۷۶	۳۲۷۶	٪۹۱/۶	۹۰۸۳۶	صحیح
٪۸/۲۴	۲۹۴	٪۸/۴	۹۰۲	اشتباه
	۳۵۷۰		۹۱۷۳۸	تعداد کل

همانطور که مشاهده می شود از تعدادکل ۹۱۷۳۸ داده آموزشی، ۹۱ درصد توسط نرم افزار منطق یابی شدند به عبارتی پیش بینی خود نرم افزار هم مطابق نتایج به دست آمده می باشد اما نرم افزار منطق ۹۰۲ مورد راپیدا نکرده است که ۸/۴ ٪ را شامل می شود که مقدار قابل توجهی نیست و می تواند به خاطر علل مختلفی باشد . همچنین از تعداد کل ۳۵۷۰ مورد داده های آزمایشی ۹۱/۷۶٪ به طور صحیح پیش بینی شدند پس نتایج به دست آمده در مرحله قبل نیز تایید می شوند.

.۵. بحث و نتیجه گیری:

همانطور که در بخشهای پیشین اشاره شد . یکی از اهداف این تحقیق تعیین عوامل موثر بر شدت تصادفات (به خصوص عوامل جاده ای)در راههای برون شهری است . در این تحقیق پس از اجرای الگوریتمهای خوشه بندی روی مجموعه داده های مختلف مربوط به عوامل مکانی (جاده ای) موثر در بروز تصادفات و ارزیابی مدلها و یافتن الگوریتم بهینه برای رسیدن به اهداف تحقیق، به خوشه بندی مکانی داده های جاده ای تصادفات پرداخته شد. با توجه به معیارهای زمان اجرا ،تعداد خوشه ها و شاخص سیلوئیت الگوریتم k- means به عنوان الگوریتم بهینه انتخاب شد . مطابق نتایج به دست آمده از اجرای این الگوریتم ۵ خوشه به دست آمد که خوشه ۳ از بیشترین اهمیت برخوردار است . در این خوشه متغیرهای پیش کننده به ترتیب اهمیت شامل هندسه مح ل(۶۶%)، جهت حرکت راه (۵۱%)، خط کشی جاد ه (۴۲٪)، وجودمانع دید(۹۲%) ، وجود نقص راه (۸۴%)،نوع شانه راه (۶۵%)، شرایط سطح راه(۸۶%) ، تعمیرات محل(۱۰۰%) و نوع رویه راه (۹۹%)می باشند. همانطور که مشاهده می شود ۳ عامل اول از اهمیت بیشتری برخوردارند .همچنین در خوشه بندی متغیرهای نوع داده که شامل موقعیت تصا دف، نوع برخورد و نحوه برخورد می باشد ، بیشترین آمار تصادف در باند سواره رو (۱۰۰٪)، بیشترین برخورد ماشینها با یکدیگر از جلو وسیله نقلیه به عقب وسیله نقلیه

دیگر(٪۳۰) و بیشترین درصد نوع برخورد مربوط به برخورد دو وسیله نقلیه با یکدیگر (٪۶۰)می باشد. سپس برای بررسی میزان همبستگی نتایج حاصل ازالگوریتمهای خوشه بندی از الگوریتم همبستگی Apriori استفاده شد. طبق این الگوریتم، در ۳ طبقه نوع تصادف(خسارتی ، جرحی و فوتی)، ۹۵ ٪ تصادفات منجر به نوع خسارتی می شود. همچنین با توجه به ۲ معیار sup و conf در این الگوریتم نتایج به دست آمده در مرحله خوشه بندی نیز تایید شد . پس از انجام محاسبات هزینه دسته بندی اشتباه و هرس شاخه های غیر ضروری که با توجه به دانش متخصصان ایمنی راه و کارشناسان پلیس راه انجام گرفته است درخت دسته بندی و رگرسیون بهینه با استفاده از داده های آموزشی به منظور تحلیل نوع تصادفات ایجاد می شود . متغیر نوع تصادف متغیر وابسته می باشد که خود دارای ۳ طبقه خسارتی، جرحی و فوتی است. ۸ متغیر دیگر متغیرهای مستقل می باشند که به ترتیب اهمیت شامل وجود مانع دید، شرایط سطح راه ، هندسه محل، خط کشی جاده ، نوع شانه راه ، جهت حرکت، راه نقص راه و نوع رویه راه می باشند. مطابق این جدول مهمترین عامل مانع دید در نظر گرفته شده است . درشاخه اول مانع دید که شامل وجود متغیرهای تل خاک، ساختمان ،کیوسک، شیب، طوفان، شن، مه ، نور خورشید، نور چراغ ، وسیله نقلیه مقابل، وسیله در حال حرکت ، وسیله متوقف و کولاک می باشد، از تعداد۱۷ تصادف، ۵/۸۸۲٪ تصادفات منجر به نوع جرحی، از ۲۷۳ تصادف، ۹۴/۱۳۸٪ منجر به نوع خسارتی و ۰٪ منجر به نوع فوتی می شوند. در شاخه دوم که شامل وجوددرخت، بوته، قوس قائم، وعدم وجود مانع می باشد، ۸/۱۴۲٪ تصادفات منجر به نوع جرحی، ۸۹/۸۵۲۲٪ منجر به نوع خسارتی و ۲/۰۳۵ ٪ منجر به نوع فوتی می شوند. نتایج همچنین نشان می دهد که در ۹۵ درصد موارد متغیرهای تصادف منجر به نوع تصادف خسارتی شده اند .در انتها برای صحت پیش بینی مدل برای ۳ طبقه ازداده های تست و آموزش استفاده شد. نتایج نشان می دهد از تعدادکل ۹۱۷۳۸ داده آموزشی، ۹۱ درصد توسط نرم افزار منطق یابی شدند به عبارتی پیش بینی خود نرم افزار هم مطابق نتایج به دست آمده می باشد اما نرم افزار منطق ۹۰۲ مورد راپیدا نکرده است که ۸/۴ ٪ را شامل می شود که مقدار قابل توجهی نیست و می تواند به خاطر علل مختلفی باشد .همچنین از تعداد کل ۳۵۷۰ مورد داده های آزمایشی ۹۱/۷۶٪ به طور صحیح پیش بینی شدند.

این تحقیق نشان داد که با استفاده از روشهای ناپارامتری و ابزارهای داده کاوی می توان از میان حجم عظیمی از داده های تصادفات، الگوهایی مفید استخراج کرد و شدت تصادفات را تحلیل کرد. هرچند روشهای داده کاوی هریک دارای مزایا و معایبی هستند که بسته به کاربرد مورد نظر در فرآیندهای کلاسه بندی و پیش بینی می بایست مورد ارزیابی قرار گیرند. الگوریتمهای خوشه بندی و درخت تصمیم از قوی ترین و پرکاربردترین الگوریتمهای داده کاوی است که به منظور تحلیل حجم زیادی از داده های تصادفات و کشف دانش قابل استفاده است. در مقایسه با اندک مطالعات انجام شده جهت تحلیل شدت تصادفات با استفاده ترکیبی از الگوریتمهای داده کاوی این تحقیق با استفاده از الگوریتمهای خوشه بندی همبستگی و مدل درخت دسته بندی در تحلیل و تفسیر داد ه های تصادفات مهم ترین عوامل جاده ای موثر بر بروز تصادفات را شناسایی و میزان همبستگی بین عوامل ذکر شده و نوع تصادف را نیز مشخص کرده است . مدل پیشنهادی به دلیل قابلیت نمایش گرا فیکی الگوها رابطه پیچیده بین متغیرهای ورودی و نوع تصادفات را نمایش داد. تعیین چگونگی و حوزه تاثیر عوامل مکانی بر قطعات راه می تواند نقش موثری در پیشگیری از وقوع تصادفات آتی و کاهش شدت تصادفات در بخش های مختلف راه داشته باشد. یافته های این تحقیق در شناسایی عوامل تاثیر گذار بر سطوح مختلف شدت تصادف می تواند کارشناسان پلیس راهنمایی و رانندگی و متخصصان ایمنی راه را در اتخاذ تصمیمات لازم به منظور کاهش شدت تصادفات جاده ای یاری کند.

در نهایت در راستای تکمیل این مطالعه در پژوهشهای آتی پیشنهاد می شود که با توسعه روشی مشابه به منظور تحلیل تصادفات در بخشهای تصادف خیز برون شهری و درون شهری موجبات کاهش نرخ و شدت تصادفات را فراهم کنند. همچنین عوامل موثر دیگر در بروز تصادفات که شامل عوامل انسانی ، خودرو ، عوامل محیط و فاکتورهای زمانی می باشند مورد بررسی قرار گیرند.

سپاسگزاری

بدین وسیله مراتب قدردانی خود را از مرکز تحقیقات کاربردی پلیس راهور ناجا جهت قرار دادن داده های مورد استفاده در این تحقیق اعلام می داریم.

منابع:

- ابراهیم خانی، سمیه، افضلی، مهدی، وشکوهی، علی (۱۳۹۰). پیش بینی و بررسی عوامل تصادفات جاده ای با استفاده ازالگوریتمهای داده کاوی. فصلنامه دانش انتظامی زنجان، ۳۲(۲)، ۳۲-۲۶
- ابریشمی، سمیه ، خان تیموری، علی رضا ، و مقصودی، بهروز (۱۳۹۰). بررسی عوامل تصادفات جاده ای با استفاده از الگوریتمهای داده کاوی. *پنجمین کنفرانس داده کاوی ایران*،دانشگاه صنعتی امیرکبیر،تهران.
- افضلی، مهدی ، شکوهی، علی ، وابراهیم خانی، سمیه (۱۳۹۰). بررسی کاربرد فنون داده کاوی در تحل یل و استخراج الگوهای تصادفات جاده ای مبتنی بر سیستم اطلاعات مکانی، فصلنامه مهندسی حمل و *نقل*،سال سوم،۹۴-۶۵ .
- پاک گوهر، علیرضا ، سرهنگ دوم صادقی کیا، عباس (۱۳۸۷).تحلیل داده های آماری تصادفات را نندگی به وسیله درخت تصمیم. *فصلنامه مطالعات* مدیریت ترافیک،سال سوم،شماره۸، ۴۵-۳۲ .
- پیروتی، مسعود. (۱۳۸۹).بررسی تاثیر پارامترهای اقلیمی بر تصادفات جاده ای (مورد مطالعه: جاده سردشت-ارومیه)، پایان نامه کارشناسی ارشد،دانشگاه تبریز.
- زینلی، سایه، و حسینعلی، فرهاد (۱۳۸۸).کاربرد فنون داده کاوی مکانی در تحلیل تصادفات جاده ای در تقاطعها، *همایش ملی مهندسی عمران کاربردی و دستاوردهای نوین.*
- سلامی، مهرداد. (۱۳۸۹). تحلیل فضایی تصادفات محورهای ارتباطی استان خراسان جنوبی با استفاده از GIS ، پایان نامه کارشناسی ارشد، دانشگاه پیام نورتهران.
- سلطانی، فاطمه .(۱۳۹۲). تحلیل فضایی سوانح رانندگی در مباد ی ورودی شهرها با استفاده از GIS (مورد مطالعه :مبادی ورودی شهر زنجان)، پایان نامه کارشناسی ارشد ،دانشگاه زنجان.
- شهرابی، جمال (۱۳۹۲). *داده کاوی با* کلمنتاین. تهران: انتشارات امیرکبیر.
- کرمی، شهرام .(۱۳۸۲).تحلیل تصادفات جاده ای با رویکرد اقلیمی با استفاده از GIS(جاده فیروزکوه-ساری)،پایان نامه کارشناس ارشد،دانشگاه تربیت مدرس.

- عفتی، ،میثم، رجبی، ،محمدعلی، و شعبانی، شاهین(۱۳۹۱) . توسعه یک سیستم دانش مبنای مکانمند جهت پیش بینی تصادفات در مسیرهای برون شهری. مجله مهندسی حمل و نقل ، سال سوم ،۲۴-۱۴.
- عفتی، ،میثم، رجبی، ،محمدعلی، و شعبانی، شاهین(۱۳۹۱) . تحلیل شدت تصادفات در راههای دوخطه – دوطرفه بین شهری، فصلنامه مطالعات پژوهشی راهور، سال سوم،۱۳۰-۱۰۳
- میرزاقلی، مرتضی. (۱۳۹۲).تعیین تاثیرات محیطی و مشخصات هندسی راه بیرجند- قائن بر میزان تصادفات جاده ای با استفاده از الگوریتمهای د اده کاوی، پایان نامه کارشناسی ارشد،موسسه آموزش عالی غیر انتفاعی شمال.

- Akomolafe, K. (۲۰۰۹). "Enhancing road monitoring and safety through the use of geo spatial technology" Interna-tional Journal of Physical Sciences, ۴, ۳۴۳-۳۴۸.
- Akomole, O. (۲۰۰۴). Predicting possibilities of Road Accidents occurring, using Neural Network. M. Sc. Thesis, Department of Computer Science, University of Ibadan.
- Anderson, T.K. (۲۰۰۹). Kernel density estimation and K-means clustering to profile road accident hotspot. Accident Analysis & Prevention, ۴۱(۳), ۳۵۹–۳۶۴. • Aworemi,
- Aworemi, J.R., Ibraheem Adegoke, A. and Segun Oluwaseun, O. (۲۰۱۰). Analytical study of the causal factors of road traffic crashes in southwestern Nigeria. Educational Research, ۱(۴), ۱۱۸-۱۲۴.
- Dr.Geetha Ramani &Shanthi Anderson) ۲۰۱۱). Classification of Vehicle Collision Patterns in Road Accidents using Data Mining Algorithms,International Journal of Computer Applications۴(۱۳),۱۷-۲۵ .
- Gelfand, S.G., Ravishanker, C.S., & Delp, E.J. (۱۹۹۱) .An iterative Growing and Pruning Algorithm for Classification Tree Design, ۱۳,۱۶۳-۱۷۴.
- Jha, M., McCall, C. and Schonfeld, P. (۲۰۰۱). Using GIS, genetic algorithms and visualization in highway development. Journal of Computer - aided Civil and Infrastructure Engineering, ۱۶(۶), ۳۹۹ – ۴۱۴.
- K. GEURTS, G. WETS, T. BRIJS, AND K. VANHOOF, Clustering and profiling traffic roads by means of accident data, in Proceedings of the European Transport Conference ۲۰۰۳, Strasbourgh (France), October ۸-۱۰, ۲۰۰۳.
- Malgundkar, T., Rao, M. & Mantha S. S. (۲۰۱۲). "GIS driven urban traffic analysis based on ontology", International Journal of Managing Information Technology (IJMIT), ۴,۱۴-۲۵.

- Mohan, D., Tiwari, G., Khayesi, M. & Nafukho. F. M. (۲۰۰۶). Road traffic injury prevention, training manual, World Health Organization.), ۲۳, .۵۰-۶۴
- Ossenbruggen, P., pendharkar, J., &Ivan, J. (۲۰۰۱). Roadway safety in rural and small urbanized areas. Accid. Anal, vol.۳۳۴, ۴۸۵-۴۹۸.
- Tavakoli kashani, a., shariat-mohaymany, d., & ranjbari, a. (۲۰۱۰). a data mining approach to identify key factors of traffic injury severity ,preliminary communication safety and security of traffic,۵۸ , ۳۲-۵۴.
- Venkatadri.M. & Lokanatha C. (۲۰۱۱). "A review on data mining from past to the future", International Journal of Computer Applications, ۱۵, ۱۹–۲۲.
- Ng, K-S, Hung, W-T and Wong W-G (۲۰۰۲) An algorithm for assessing the risk of traffic accidents, Journal of Safety Research, ۳۳, ۳۸۷-۴۱.