

شناسایی عوامل موثر و بررسی تصادف‌های ترافیکی با استفاده از رویکردهای داده‌کاوی (مطالعه موردی آزادراه تهران-قم)

بهزاد مسلم، دانشجوی کارشناسی ارشد، گروه مهندسی صنایع، دانشکده فنی و مهندسی، دانشگاه آزاد اسلامی واحد علوم و تحقیقات تهران، تهران، ایران

فرزاد موحدی سبحانی (مسئول مکاتبات)، استادیار، گروه مهندسی صنایع، دانشکده فنی و مهندسی، دانشگاه آزاد اسلامی واحد علوم و تحقیقات تهران، تهران، ایران

عباس سقایی، دانشیار، گروه مهندسی صنایع، دانشکده فنی و مهندسی، دانشگاه آزاد اسلامی واحد علوم و تحقیقات تهران، تهران، ایران

E- mail : fmovahedi@iaou.ac.ir

پذیرش: ۱۳۹۶/۱۲/۱۳

دریافت: ۱۳۹۶/۰۸/۰۴

چکیده

هدف اصلی این پژوهش شناسایی عوامل کلیدی و بررسی الگوریتم‌های مختلف داده‌کاوی در تصادفات ترافیکی در ایران، بخصوص در جاده‌های برون شهری است. تصادفات ترافیکی برون شهری یکی از منابع اصلی جهت تجزیه و تحلیل و بررسی شدت حوادث رانندگی و علل موثر بر آنها است. در ادامه مجموعه قوانینی که می‌تواند در شناسایی عوامل و تاثیر آنها در کاهش تصادفات موثر باشد استخراج خواهد شد. ۵۰۹۹ رکورد از داده‌های جمع‌آوری شده از محور تهران- قم در استان تهران مورد استفاده قرار گرفت. برای دستیابی به اهداف این پژوهش از تکنیک‌های مختلف داده‌کاوی استفاده گردید. به همین منظور از روش ارزیابی انتخاب مبتنی بر همبستگی برای شناسایی عوامل موثر و انتخاب بردار ورودی استفاده گردید. سپس ۶ الگوریتم داده‌کاوی، بیزین ساده، لجیستیک، پرسپترون چندلایه، کلاس‌بندی از طریق رگرسیون، قوانین استنتاجی (پارت) و درخت تصمیم‌گیری جی‌اچ‌ا، برای پیش‌بینی دقت مدل‌های مورد ارزیابی با استفاده از نرم افزار داده‌کاوی و کامورد استفاده قرار گرفتند؛ همچنین از الگوریتم اپریوری به همراه دو مدل جی‌اچ‌ا و پارت جهت استخراج قوانین استفاده شد. نتایج حاصل از استخراج قوانین نشان داد که دصور عامل تصادف در صحنه تصادف، نوع برخورد، مانع دید، موقعیت تصادف، شرایط سطح راه، هندسه محل تصادف و علل مستقیم بیان شده توسط پلیس برای تصادف از مهم‌ترین عواملی بودند که در قوانین استخراج شده از مجموعه قوانین به آنها اشاره شده و بیشترین تعداد تکرار را داشتند. نتایج نشان دادند که الگوریتم‌های پرسپترون و پارت بهترین عملکرد را در میان سایر الگوریتم‌ها جهت پیش‌بینی در اختیار داشتند.

واژه های کلیدی : پیش‌بینی، تصادفات ترافیکی، داده‌کاوی، مدل سازی

۱. مقدمه

استنتاجی (پارت) و درخت تصمیم‌گیری جی ۴۸ با استفاده از نرم افزار داده‌کاوی وکا استفاده شده است. در بخش بعدی این مقاله، پیشینه پژوهشی درباره پیش‌بینی شدت تصادفات ترافیکی با روش‌های کلاس‌بندی بررسی شده است. بخش سه، مقدمه‌ای بر مدل‌های کلاس‌بندی، الگوریتم‌های مورد استفاده و روش‌های ارزیابی و همچنین توضیحاتی در مورد داده‌های مورد مطالعه ارائه می‌دهد. سپس، نتایج و تحلیل‌های مربوطه در بخش چهار ارائه شده است. در نهایت، نتیجه‌گیری بر اساس نتایج مطالعه توصیف می‌شود.

۲. پیشینه پژوهش

۱-۲ پیشینه نظری

داده‌های به دست آمده از تصادفات ترافیکی و با توجه به بحث این تحقیق، به طور خاص تصادفات ترافیکی برون شهری یکی از منابع اصلی جهت تجزیه و تحلیل و بررسی شدت حوادث رانندگی و علل موثر بر آنها است. این مجموعه داده‌ها می‌تواند شامل ده‌ها پارامتر مختلف باشد که بر اساس رویکردها و تکنیک‌های مختلف می‌توان آنها را مورد بررسی قرار داد. با بررسی ادبیات و پژوهش‌های صورت گرفته در این زمینه مشاهده شد که تنها مقالات [Castro and Kim, 2016] به بررسی الگوریتم جی ۴۸ پرداختند. البته [de Oña, López and Abellán, 2013, Deb and Liew, 2016, Tao, et al., 2016] نیز با الگوریتم سی ۴/۵، که مشابه با الگوریتم جی ۴۸ است اقدام به بررسی کردند. در این میان مدل درخت کلاس‌بندی و رگرسیون (کارت) با بیشترین اقبال از سوی نویسندگان روبرو بوده است [Anvari, Tavakoli, Kashani and Rabieyan, 2017, Chang and Wang, 2006, de Oña, López and Abellán, 2013, Jung, Qin and Oh, 2016, Kashani and Mohaymany, 2011, Kumar and Toshniwal, 2017, Mohaymany, Kashani and Ranjbari, 2010, Pakgozar, et al., 2011, Tavakoli, Kashani, Rabieyan and Besharati, 2014]. در کنار

با پیچیده‌تر شدن جوامع انسانی، نقش حمل و نقل برای جوامع بشری روز به روز اهمیت بیشتری پیدا می‌کند. یکی از مهم‌ترین وظایف در ایمنی جاده‌ها، تعیین دلایل تصادفات ترافیک جاده‌ای است. سهم قابل توجهی از شبکه جاده‌های برون شهری ایران شامل مسیرهای دو طرفه و دوبانده است. ۲۱۷۹۷ کیلومتر از جاده‌های اصلی برون شهری، ۳۰ درصد از شبکه‌های جاده‌ای ایران را تشکیل می‌دهد و بیش از ۹۰ درصد مسافران ایرانی، حالت حمل و نقل جاده‌ای را با توجه به سهولت و قیمت تمام شده، برای سفرهای خود انتخاب می‌کنند. طبق آمار منتشر شده توسط پژوهشکده حمل و نقل وزارت راه و ترابری، آمار تلفات جاده‌ای در ایران ۲۰ برابر کشورهای صنعتی و ۵ برابر کشورهای هم‌تراز با ایران است [Kashani and Besharati, 2017]. به همین دلیل تصادفات جاده‌ای در کشور ما یک تهدید جدی به شمار می‌رود و نیازمند توجه جدی و سریع در این راستا است. بنابراین نیازمند راهکاری برای رفع این مشکل و به دنبال پاسخی برای رفع این مساله حیاتی که سالانه صدمات جبران‌ناپذیری به بدنه اجتماعی و اقتصادی کشور وارد می‌سازد هستیم. این پژوهش شکاف تحقیقاتی در زمینه دقت و صحت روش‌های مورد استفاده و عدم مقایسه روش‌های مختلف به منظور انتخاب روشی مناسب‌تر جهت انجام کلاس‌بندی و کشف الگوهای مناسب از آنها و همچنین شکاف در ارائه قوانینی جامع در حوزه تصادفات که بتواند مفید باشد را شناسایی کرده و اهداف خود را برای پوشش این شکاف‌های علمی بر مبنای یافتن مهم‌ترین خصیصه‌های موثر و سپس تحلیل و بررسی دقت مدل‌های داده‌کاوی برای پیش‌بینی تصادفات برون شهری در ایران قرار داده است. به همین منظور از ۶ الگوریتم داده‌کاوی، بیزین ساده، لجیستیک، پرسپترون چندلایه، کلاس‌بندی از طریق رگرسیون، قوانین

اعمال شده است که جای کار بیشتری را نشان می‌دهد. جدیدترین تحقیقات این حوزه نشان دهنده روند صعودی در توجه به این موضوع مهم را نشان می‌دهد. الگوریتم‌های جی‌۸، پارت، بیزین ساده و پرسپترون چندلایه در [Taamneh, Alkheder and Taamneh,2017]، کارت در [Anvari, Tavakoli Kashani and Rabieyan,2017]، کارت، بیزین ساده و بردار حمایت ماشینی در [Kumar and Toshniwal,2017] و چاید و شبکه بیزین در [Prati, Pietrantoni and Fraboni,2017] جدیدترین الگوریتم‌های مورد استفاده در تحقیقات اخیر هستند.

۲-۲- پیشینه تجربی

تامنه و همکارانش (۲۰۱۷) به بررسی تصادفات ترافیکی و یافتن دلایل آن با توجه به مشخصات رانندگان، جاده و تصادفات پرداخته‌اند. آنها با بررسی ۵۹۷۳ مورد اطلاعات تصادفات در شهر ابوظبی امارات در سال‌های ۲۰۰۸ تا ۲۰۱۳ و با استفاده از روش‌های مختلف داده‌کاوی و استفاده از الگوریتم‌های مختلف کلاس‌بندی همچون درخت تصمیم‌گیری، قوانین استنتاجی، شبکه‌های عصبی و چندین روش دیگر به دنبال یافتن عوامل موثر بر تصادفات بودند. ایشان برای رسیدن به این نتایج از نرم افزار داده کاوی وکا استفاده کرده‌اند. نتایج آنها شامل عوامل مهم دخیل در مرگ و میر تصادفات بود که عواملی چون سن، جنسیت، ملیت راننده، سال تصادف، وضعیت تلفات و نوع برخورد را ارائه کردند [Taamneh, Alkheder and Taamneh, 2017].

انوری و همکارانش به شناسایی مهم‌ترین عوامل در احتمال مقصر بودن موتورسیکلت سواران توسط داده‌کاوی با استفاده از مدل‌های درخت کلاس‌بندی پرداختند. آنها داده‌هایی از ایران را که در سال ۲۰۱۱ از پلیس دریافت کرده بودند با استفاده از مدل درخت کلاس‌بندی و رگرسیون تحلیل کرده و به تعیین عوامل مقصر بودن یا نبودن موتورسوار در تصادف

این‌ها [de Oña, López and Abellán,2013, Kwon, Rhee and Yoon,2015, Montella, et al.,2011] نیز با مدل‌های دیگری از درخت تصمیم‌گیری بررسی‌های خود را انجام دادند. الگوریتم پارت که می‌تواند قوانین ایجاد کند تنها در [Taamneh, Alkheder and Taamneh,2017] مورد بررسی قرار گرفته است. استفاده از الگوریتم بیزین به دو دسته تقسیم شده‌اند [Kumar and Toshniwal,2017, Kwon, Rhee and Yoon,2015, Taamneh, Alkheder and Taamneh,2017] از بیزین ساده و [Castro and Kim,2016, Prati, Pietrantoni and Fraboni,2017] از شبکه بیزین استفاده کردند.

استفاده از شبکه‌های عصبی و به خصوص الگوریتم پرسپترون چندلایه که الگوریتم کارآمد و البته زمانبری است در [Castro and Kim,2016, Taamneh, Alkheder and Taamneh,2017] مورد استفاده قرار گرفته است. الگوریتم کاربردی رگرسیون لجستیک در این موضوع نسبت به سایر الگوریتم‌ها کاربرد کمتری دارد ولی همچنان در [Kwon, Rhee and Yoon,2015, Pakgozar, et al.,2011, Yau, Lo and Fung,2006] مورد استفاده قرار گرفته است. با وجود آنکه در بیشتر پژوهش‌ها با استفاده از الگوریتم‌های درخت تصمیم‌گیری اقدام به ارائه قوانین کردند اما تنها در [Geurts, Thomas and Wets,2005, Montella,2011, Montella, et al.,2011] از الگوریتم‌های استخراج قوانین استفاده شده است.

هدف ما بررسی تصادفات در جاده‌های برون شهری در ایران است. تحقیقات پیشین انجام شده در این منطقه همگی از الگوریتم کارت استفاده کرده‌اند [Anvari, Tavakoli Kashani and Rabieyan,2017, Kashani and Mohaymany,2011, Mohaymany, Kashani and Ranjbari,2010, Pakgozar, et al.,2011, Tavakoli Kashani, Rabieyan and Besharati,2014]. همانطور که مشخص شد علاوه بر مکان و موضوع که بررسی کمی روی آن صورت پذیرفته است، الگوریتم‌های محدودی نیز

از این الگوها در سایر کشورها نیز بهره برد. این پژوهش استفاده از چندین روش کاوش داده‌ها را در انجام کارهای آینده پیشنهاد می‌دهد [Jung, Qin and Oh, 2016].

توکلی‌کاشانی و همکارانش به بررسی رویکردهای داده‌کاوی در عوامل موثر بر شدت تصادفات سرنشینان موتور سیکلت پرداخته‌اند. آنها با استفاده از روش درخت کلاس‌بندی و رگرسیون، داده‌های ایران را در یک دوره ۴ سال مورد بررسی قرار دادند و دریافته‌اند نوع منطقه، سطح مورد استفاده و ناحیه آسیب دیده بدن بیشترین تاثیر را بر جراحات موتورسواران دارند. همچنین دقت پیش‌بینی مدل ساخته شده بهبود زیادی نسبت به تحقیقات پیشین خود داشت و استفاده از کلاه ایمنی تاثیر بسیار زیادی در کاهش آسیب‌ها داشته است [Tavakoli, Rabieyan and Besharati, 2014].

توکلی‌کاشانی و شریعت‌مهمی، شدت جراحات ترافیکی در جاده‌های دوبانده و دوطرفه برون شهری ایران را بر اساس مدل‌های درخت کلاس‌بندی و طی مدت سه سال تجزیه و تحلیل کردند. آنها دریافته‌اند که حرکات ماریج و نبستن کمربند ایمنی مهم‌ترین عوامل موثر بر شدت آسیب است. همچنین نتایج به دست آمده نسبت به تحقیقات پیشین بهبود یافته است [Kashani and Mohaymany, 2011].

حقیقی و غلام‌نژاد در پژوهشی ابتدا شرایط عبور از عرض خیابان در یک مدرسه داخل شهری درکناره خیابان به وسیله ویدیوگرافی مورد ارزیابی و تجزیه و تحلیل قرار گرفته، سپس با استفاده از تکنیک‌های ریاضی با استفاده از متغیرهای استخراج شده مدل‌سازی تعارض و ریسک ایمنی عبور پرداخته شده است. در همین راستا ارزیابی‌های آماری عبور دانش‌آموزان از عرض خیابان نیز مورد مطالعه قرار گرفته است. نتایج این تحقیق نشان می‌دهد که ۱۰ متغیر نظیر اندازه گروه عبور، سرعت نزدیک شدن، زمان انتظار قبل از عبور، زمان توقف و انتظار روی خط، متوسط زمان عبور، سرعت متوسط عبور، توجه به وسایل نقلیه نزدیک شونده، عقب‌برگشتن از

پرداختند. نتایج نشان داد که نوع تصادف و سن راننده در احتمال مقصر بودن موثر بود. همچنین تصادف جلو به عقب بیشترین میزان تصادف در این زمینه را داشت. در نهایت پیشنهاد آموزش بیشتر به رانندگان و نصب علائم هشدار دهنده ارائه گردیده بود [Anvari, Tavakoli Kashani and Rabieyan, 2017].

کومار و توشنیوال، به تجزیه و تحلیل تصادفات ترافیکی وسایل نقلیه با دو چرخ، در هند پرداختند. آنها با تجزیه و تحلیل عوامل موثر بر شدت تصادفات به وسیله سه الگوریتم درخت تصمیم‌گیری رگرسیون و کلاس‌بندی، بیزین ساده و بردار حمایت‌ماشینی دریافته‌اند که برای ۱۳ منطقه مورد بررسی عوامل متفاوتی وجود دارد. همچنین مدل درخت رگرسیون و کلاس‌بندی دقت بهتری نسبت به دو الگوریتم دیگر داشت [Kumar and Toshiwal, 2017].

پرانی و همکارانش، با استفاده از تکنیک‌های داده‌کاوی اقدام به پیش‌بینی شدت تصادفات دوچرخه سواران کردند. آنها با بهره‌گیری از دو الگوریتم درخت تصمیم‌گیری چایلد و شبکه بیزین دریافته‌اند برای درخت تصمیم‌گیری چایلد، نوع جاده، نوع تصادف و سن دوچرخه سوار مهم‌ترین عوامل بودند. و برای شبکه بیزین نوع تصادف، نوع جاده و نوع سرنشینان وسیله به عنوان مهم‌ترین عامل شدت تصادفات دوچرخه سواران به دست آمد [Prati, Pietrantonì and Fraboni, 2017].

جانگ و همکارانش به بررسی بهبود استراتژی‌های پلیس جهت بهبود ایمنی عابران پیاده با استفاده از مدل‌های درخت کلاس‌بندی پرداختند. با استفاده از داده‌ها و تحلیل آنها با استفاده از روش‌های درخت کلاس‌بندی به دنبال راهکاری برای کمک به پلیس کره جنوبی بودند. نتایج این تحقیق نشان داد که سن عابران پیاده و نوع حرکت آنها دو عامل اولیه در شدت جراحات بودند. همچنین در ادامه تجزیه و تحلیل کلی و الگوهای خاص جهت استفاده پلیس تهیه گردید که می‌توان

۱-۳ تعریف داده‌ها

برای این پژوهش، داده‌های جمع‌آوری شده توسط سازمان راه‌داری و حمل‌ونقل جاده‌ای ایران طی سال‌های ۸۹ تا ۹۲ (۲۰۱۰-۲۰۱۳) مورد استفاده قرار گرفت. این داده‌ها شامل ۵۰۹۹ رکورد تصادفات ترافیکی صورت گرفته در محور تهران-قم در استان تهران طی یک دوره ۴ ساله بود.

۲-۳ تعریف مدل‌ها

در این بخش برای مدل‌های مورد بررسی تعریف کوتاهی ارائه شده است.

- درخت تصمیم‌گیری جی ۴۸: درخت‌های تصمیم‌گیری از طریق جداسازی متوالی داده‌ها به گروه‌های مجزا ساخته می‌شوند و هدف در این فرآیند افزایش فاصله بین گروه‌ها در هر جداسازی است. هر گره داخلی در این درخت‌ها یکی از متغیرهای ورودی را نشان می‌دهد و دارای تعداد شاخه‌ای برابر با تعداد مقادیر ممکن است که متغیر ورودی است. هر گره برگ دارای مقدار مشخصه هدف است.
- قوانین استنتاجی (پارت): این روش یک فرآیند تکرار شونده است که رویکرد تقسیم و تسخیر را دنبال می‌کند. در این روش در هر تکرار، یک زیر مجموعه از مجموعه داده‌های آموزش با استفاده از یکی از الگوریتم‌های درخت تصمیم‌گیری برای تولید قوانین استفاده می‌شود.

- بیزین ساده: بیزین ساده، یک الگوریتم طبقه‌بندی بر اساس قضیه بیزین است، با فرض ساده‌ای که هر دو متغیر ورودی مستقل است. اگر چه این فرض بیش از حد ساده شده است اما به همین صورت، این الگوریتم به طور موثر در بسیاری از مشکلات پیچیده در دنیای واقعی به خصوص طبقه‌بندی استفاده می‌شود.

- شبکه عصبی پرسپترون چندلایه: این روش یک تکنیک یادگیری تحت نظارت است که برای طبقه‌بندی و رگرسیون استفاده می‌شود. پرسپترون چند لایه، یک شبکه عصبی مصنوعی ایجاد می‌کند که متشکل از گره‌های چندگانه در سه

مسیر، نحوه عبور، حاشیه ایمنی و نوع وسیله نقلیه بر ایمنی دانش‌آموزان مؤثر بوده که در این میان توجه به وسیله نقلیه در حال عبور از همه بیشتر و زمان انتظار پیش از عبور از هر خط کمترین تاثیر را در تصمیم‌گیری‌های عبور از عرض خیابان داشته اند [Haghghi and GholamNejad, 2016].

۳. روش شناسی پژوهش

این پژوهش با هدف کشف عملکرد تکنیک‌های داده‌کاوی در پیش‌بینی تصادفات ترافیکی در ایران صورت گرفته است. به علاوه هدف بعدی شناسایی و بیان عوامل مهم و موثر بر تصادفات قرار داشت. در نهایت مجموعه قوانین استخراج شده که می‌تواند در شناسایی عوامل موثر در تصادفات و کاهش آن‌ها موثر باشند ارائه شده است. برای دستیابی به اهداف این پژوهش نخست از روش ارزیابی انتخاب مبتنی بر همبستگی^۱ برای شناسایی عوامل موثر بر تصادفات و انتخاب بردار ورودی استفاده گردید. سپس برای بررسی دقت مدل‌های پیش‌بینی از الگوریتم داده‌کاوی بیزین ساده^۲، لجیستیک^۳، پرسپترون چندلایه^۴، کلاس‌بندی از طریق رگرسیون^۵، قوانین استنتاجی (پارت)^۶ و درخت تصمیم‌گیری جی ۴۸^۷ با استفاده از نرم افزار داده‌کاوی وکا استفاده شد. همچنین از الگوریتم اپروری^۸ به همراه الگوریتم جی ۴۸ و پارت جهت استخراج قوانین استفاده شدند. در این بخش روش‌های مورد استفاده، داده‌های مورد استفاده و نحوه جمع‌آوری آنها، ساخت مدل‌های کلاس‌بندی و در نهایت استخراج دانش مورد نیاز توضیح داده خواهد شد. این گام‌ها را می‌توان در ۴ بخش تعریف کرد: جمع‌آوری داده‌ها، پیش‌پردازش داده‌ها، ساخت و ارزیابی مدل‌های کلاس‌بندی، کشف دانش مورد نیاز. در این پژوهش متغیر شدت تصادف به عنوان متغیر هدف در مدل به کار برده شده است و سایر متغیرهای تصادف به عنوان متغیرهای ورودی استفاده شده اند.

همچنین میزان همبستگی بین آن‌ها، ارزش یک زیر مجموعه از ویژگی‌ها را ارزیابی می‌کند. زیرمجموعه ویژگی‌هایی که با کلاس همخوانی زیادی دارند، در مقایسه با آن‌ها که همخوانی کمتری دارند، ترجیح داده می‌شوند.

۳-۳ پیش پردازش داده‌ها

پیش از انجام عملیات داده‌کاوی و ارائه مدل‌ها، داده‌ها مورد ارزیابی قرار گرفتند. هر رکورد تصادف شامل ۵۷ خصوصیت درباره تصادفات که شامل داده‌های پرت و خالی و داده‌های کد شده در کنار داده‌های کیفی بود. پس از پیش پردازش ستون‌های مربوط به کد، ستون‌های خالی و نامرتب از مجموعه حذف گردیدند. در نهایت مشخص گردید تنها ۲۴ خصیصه قابلیت اطمینان و اتکا جهت بررسی انتخاب خصیصه‌های موثر بر تصادفات را داشتند. بر اساس متغیر هدف، نوع تصادفات به ۳ دسته فوتی، جرحی و خسارتی تقسیم بندی شده است.

۳-۴ روش شناسی و مدل

نخست به دنبال انتخاب و کاهش خصوصیت‌های درگیر در تصادفات هستیم. به همین منظور از روش ارزیابی انتخاب مبتنی بر همبستگی برای شناسایی عوامل موثر بر تصادفات و انتخاب بردار ورودی استفاده شد. برای جستجو در این روش از الگوریتم جستجوی حریصانه^۹ استفاده شد. این الگوریتم یک جستجوی پیش‌رانه^{۱۰} یا پسرانه^{۱۱} را از طریق خصوصیات زیر مجموعه‌ها انجام می‌دهد.

در مرحله بعد ۶ الگوریتم، بیزین ساده، لجیستیک، پرسپترون چندلایه، کلاس بندی از طریق رگرسیون، قوانین استنتاجی (پارت) و درخت تصمیم‌گیری جی ۴۸ که هر یک از خانواده متفاوتی هستند انتخاب گردیدند تا علاوه بر انتخاب روش‌های مختلف انواع مختلف مدل‌های آماری، درخت‌ها، مدل‌های تجربی و سایرین مورد بررسی جامع قرارگیرد. با استفاده از نرم افزار داده‌کاوی وکا، اثر هر یک از روش‌ها توسط دو روش مورد بررسی قرار گرفت. نخست داده‌ها توسط

یا چند لایه (لایه ورودی، لایه خروجی و یک یا چند لایه پنهان) است. متغیرهای ورودی بر روی متغیرهای خروجی با استفاده از یک یا چند لایه پنهان نگاشت می‌شوند. در نتیجه استفاده از یک الگوریتم انتشار پس‌رو در آموزش شبکه‌های تولید شده توسط پرسپترون چندلایه، با موفقیت برای حل بسیاری از مشکلات دشوار مورد استفاده قرار می‌گیرد. پرسپترون چند لایه دارای قابلیت جداسازی داده‌هایی است که به صورت خطی قابل جدا شدن نیستند.

• رگرسیون لجستیک: رگرسیون لجستیک، یک مدل آماری رگرسیون برای متغیرهای وابسته دو سویی است. این مدل را می‌توان به عنوان مدل خطی تعمیم یافته‌ای که از تابع لجبت به عنوان تابع پیوند استفاده می‌کند و خطایش از توزیع چندجمله‌ای پیروی می‌کند، به حساب آورد.

• کلاس بندی از طریق رگرسیون: درختان طبقه بندی و رگرسیون، روش‌های یادگیری ماشین برای ساخت مدل‌های پیش بینی از داده‌ها هستند. این مدل‌ها توسط تقسیم بندی بازگشتی فضای داده‌ها و تخصیص یک مدل پیش‌بینی ساده در هر پارتیشن بهینه شده اند. درختان رگرسیون برای متغیرهای وابسته هستند که مقادیر مستقل یا دستورالعمل را با خطای پیش بینی، معمولاً با تفاوت مربع بین مقادیر مشاهده شده و پیش بینی شده اندازه گیری می‌کنند.

• اperiوری: اperiوری یک الگوریتم کلاسیک برای یادگیری قوانین وابستگی است. اperiوری روی پایگاه‌های داده شامل تراکنش‌ها ساخته شده است. ورودی این الگوریتم مجموعه‌ای از مجموعه آیت‌ها است. این الگوریتم تلاش می‌کند تا زیرمجموعه‌هایی از آیت‌ها را که حداقل بین چند مجموعه آیت مشترک است بیابد. اperiوری یک الگوریتم پایین به بالا است، آن‌گونه که در هر مرحله یک آیت به زیرمجموعه‌های مکرر اضافه می‌شود.

• انتخاب ویژگی مبتنی بر همبستگی: این الگوریتم با در نظر گرفتن توانایی پیش بینی فردی هر یک از ویژگی‌ها و

جدول ۱. ماتریس اشتباهات

مشاهدات	پیش‌بینی	
	مثبت	منفی
مثبت	درست-مثبت (TP)	اشتباه-منفی (FN)
منفی	اشتباه-مثبت (FP)	درست-منفی (TN)

بر اساس ماتریس اشتباهات شاخص دقت مدل^{۱۹}، را تعریف می‌کنیم که بیانگر تفاوت میان نتیجه واقعی و نتیجه پیش‌بینی شده است. برای این منظور مواردی که به درستی کلاس‌بندی و محاسبه شده‌اند را نسبت به کل موارد پیش‌بینی شده درصد بندی می‌کند. فرمول ۱ نحوه محاسبه آن را نشان می‌دهد:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \quad (1)$$

شاخص‌های نرخ مثبت صحیح^{۲۰} و نرخ مثبت اشتباه^{۲۱} برای اساس ماتریس اشتباهات به ترتیب از طریق فرمول‌های ۲ و ۳ محاسبه شده و برای محاسبه شاخص نمودار منحنی عملکرد سیستم^{۲۲} در ادامه استفاده می‌شوند.

$$\text{True Positive Rate (TPR)} = \frac{TP}{TP + FN} \quad (2)$$

$$\text{False Positive Rate (FPR)} = \frac{FP}{FP + TN} \quad (3)$$

شاخص دیگری که علاوه بر دقت مدل مورد ارزیابی قرار می‌گیرد منحنی عملکرد سیستم است که محورهای آن را نرخ مثبت صحیح و نرخ مثبت اشتباه تشکیل می‌دهند. برای رسم نمودار، مقادیر نرخ مثبت صحیح و نرخ مثبت اشتباه، نقاط برش بین ۰ و ۱ را محاسبه می‌کنند. سپس منحنی بر روی نقاط محاسبه شده برازش داده می‌شود. ناحیه زیر منحنی بین ۰/۵ تا ۱ است و بر همین اساس می‌توان عملکرد مدل‌ها را مورد توجه قرار داد. اگر ناحیه زیر منحنی بزرگتر از ۰/۹ باشد یک مدل بسیار عالی داریم. اگر ناحیه بین ۰/۸ تا ۰/۹ باشد با یک مدل خوب مواجه هستیم و اگر بین ۰/۷ و ۰/۸ باشد با یک مدل قابل قبول روبرو هستیم.

مجموعه داده تمرینی^{۱۲} مورد ارزیابی اولیه و تمرین قرار گرفتند که ارزیابی و ارزش هر مدل مورد بررسی مشخص گردد. سپس توسط روش تقسیم در صدی^{۱۳}، داده‌ها به دو مجموعه ۷۰ درصدی برای تمرین و ۳۰ درصدی برای آزمون تقسیم شده و مورد ارزیابی قرار گرفتند. داده‌های به دست آمده از هر دو روش و با توجه به هر الگوریتم انتخابی مقایسه خواهند شد. در نهایت جهت استخراج قوانین از الگوریتم اپریوری به همراه الگوریتم جی‌۸/۴ و پارت استفاده می‌شود و قانون‌های برتر مشخص می‌گردند.

۳-۵ روش ارزیابی مدل

عملکرد مدل کلاس بندی شده توسط یک ماتریس اشتباهات^{۱۴} تعریف می‌شود که نمونه‌های درست و نادرست کلاس‌بندی را برای هر کلاس نشان می‌دهد. اندازه‌گیری‌هایی که برای ارزیابی عملکرد یک کلاس‌بندی نیاز است توسط ماتریس اشتباهات محاسبه می‌شوند. جدول ۱ این ماتریس را نشان می‌دهد. این ماتریس دارای ۴ مقدار زیر است:

- درست-مثبت (TP)^{۱۵}: نمونه‌هایی که درست هستند و به عنوان درست کلاس‌بندی شده‌اند. این مقدار بیانگر میزان موفقیت پیش‌بینی است.

- درست-منفی (TN)^{۱۶}: نمونه‌هایی که درست هستند و به عنوان اشتباه کلاس‌بندی شده‌اند. این مقدار بیانگر به طور صحیح مردود شده است.

- اشتباه-مثبت (FP)^{۱۷}: نمونه‌هایی که اشتباه هستند و به عنوان درست کلاس‌بندی شده‌اند. این مقدار بیانگر یک هشدار غلط است یا به بیان دیگر بیانگر خطای نوع ۱ است.

- اشتباه-منفی (FN)^{۱۸}: نمونه‌هایی که اشتباه هستند و به عنوان اشتباه کلاس‌بندی شده‌اند. این مقدار بیانگر مقادیر از دست رفته یا به بیان دیگر بیانگر خطای نوع ۲ است.

۳-۶ استخراج قوانین

برای درک عوامل اصلی که بر شدت تصادفات تاثیرگذارند، مجموعه قوانینی را توسط الگوریتم های اپریوری، جی ۴۸ و پارت استخراج می کنیم. از مجموعه داده ها به عنوان یک مجموعه آموزشی برای الگوریتم فوق استفاده کرده و در نهایت قانون هایی که دارای درجه اهمیت بالاتری هستند را انتخاب می کنیم.

شبه کد مربوط به روند انجام تحقیق به شرح زیر است :

۱. تعریف و جمع آوری داده ها
۲. پیش پردازش داده ها
 ۱. ۲. یکپارچه سازی داده ها
 ۲. ۲. حذف داده های نامناسب، غیر قابل اعتماد و دارای نویز از مجموعه داده ها
۳. روش شناسی پژوهش
 ۱. ۳. انتخاب خصیصه ها بر اساس مدل انتخاب ویژگی مبتنی بر همبستگی (CFS)
 ۲. ۳. انتخاب خصیصه هدف
 ۳. ۳. محاسبه مدل های داده کاوی (جی ۴۸)، پارت، بیزین ساده، پرسپترون، لجستیک، کلاس بندی از طریق رگرسیون) برای مجموعه داده تهران-قم
 ۱. ۳. ۳. روش داده های تمرینی
 ۲. ۳. ۳. روش تقسیم درصدی (۷۰ درصد آموزش، ۳۰ درصد آزمون)
۴. ارزیابی نتایج و شاخص ها
 ۱. ۴. شاخص ماتریس ماتریس اشتباهات
 ۲. ۴. شاخص نرخ مثبت های صحیح (TPR)
 ۳. ۴. شاخص نرخ مثبت های اشتباه (FPR)
 ۴. ۴. شاخص دقت مدل (Accuracy)
 ۵. ۴. شاخص منحنی مشخصه عملکرد سیستم (ROC)
۵. استخراج قوانین
۶. پایان

۴-۱ یافته های پژوهش

۴-۱-۱ شناسایی عوامل و انتخاب خصیصه

نتایج حاصل از اعمال روش ارزیابی انتخاب مبتنی بر همبستگی با استفاده از الگوریتم جستجوی حریصانه برای شناسایی عوامل موثر بر تصادفات و انتخاب بردار ورودی در جدول ۲ در زیر آورده شده است. ۸ خصوصیت، عامل تصادف، نوع برخورد، مانع دید، موقعیت تصادف، شرایط سطح راه، هندسه محل، علل مستقیم و نوع تصادف که متغیر هدف است، انتخاب گردیدند.

جدول ۲. خصوصیت های انتخاب شده با استفاده از روش

ارزیابی انتخاب مبتنی بر همبستگی

متغیر ها	نوع	توضیحات
عامل تصادف	مستقل	توصیف عامل تعیین کننده تصادف
نوع برخورد	مستقل	توصیف نحوه برخورد ها در تصادف و نوع تصادف
مانع دید	مستقل	توصیف وجود مانع و نوع مانع یا عدم وجود مانع
موقعیت تصادف	مستقل	توصیف مکان تصادف در جاده یا حاشیه جاده
شرایط سطح راه	مستقل	شرایط سطح راه با توجه به نوع جاده و وضعیت هوا
هندسه محل	مستقل	هندسه جغرافیایی و مکانی تصادف
علل مستقیم	مستقل	علل مستقیم موثر بر تصادف
نوع تصادف	وابسته	تعیین شدت تصادف بر اساس خسارتی، جرحی، فوتی

۴-۲ پیش بینی دقت مدل ها

نتایج حاصل از بررسی داده ها توسط ۶ مدل مطرح شده به ترتیب در زیر آورده شده است. نتایج پیش بینی برای الگوریتم جی ۴۸ در جدول ۳ ارائه شده است. دقت مدل برای الگوریتم جی ۴۸ بر اساس داده های تمرینی به ترتیب برای خسارتی، جرحی و فوتی، ۰/۹۹۵، ۰/۴۰۱ و ۰/۱۳۳ به دست آمد. همچنین بر اساس روش تقسیم درصدی به ترتیب ۰/۹۹۶،

شناسایی عوامل موثر و بررسی تصادف‌های ترافیکی با استفاده از رویکردهای داده‌کاوی ...

تقسیم در صدی به ترتیب ۰/۹۹۶، ۰/۴۷۳، ۰/۰۳۷ بود. نتایج پیش بینی در جدول ۴ آورده شده است. از نتایج مشخص است که دقت مجموع در روش نمونه‌گیری و در روش تقسیم درصدی ۰/۹۴۹ و ۰/۹۵۴ است. همچنین سطح زیر نمودار عملکرد سیستم به ترتیب برای داده‌های تمرینی و تقسیم درصدی ۰/۹۲۴ و ۰/۹۳۲ بود.

۰/۴۳۲ و ۰/۰۳۷ به دست آمد. ارزیابی دقت مجموع با نمونه‌گیری توسط داده‌های تمرینی ۰/۹۴۹ و برای تقسیم درصدی ۰/۹۵۲ بود که در مجموع بهبود را نشان می‌دهد. اما ارزیابی سطح نمودار در داده‌های تمرینی ۰/۹۲۵ بود که پس از آزمون تقسیم درصدی به ۰/۹۱۶ رسید. دقت الگوریتم پارت بر اساس داده‌های تمرینی به ترتیب برای خسارتی، جرحی و فوتی به مقدار ۰/۹۹۵، ۰/۴۰۹ و ۰/۱۶۳ به دست آمد و برای

جدول ۳. دقت پیش بینی مدل جی ۴۸

الگوریتم	نمونه گیری	جراحت مشاهده شده	موارد صحیح طبقه بندی شده	موارد اشتباه طبقه بندی شده	دقت (Accuracy)	سطح زیر منحنی (AUC)	زمان (ثانیه)
جی ۴۸	خسارتی	۴۷۲۱	۲۳	۰/۹۹۵	۰/۹۲۶	ایجاد:	
	داده های تمرینی	۱۰۳	۱۵۴	۰/۴۰۱	۰/۹۱۱	۰/۱۶	
	فوتی	۱۳	۸۵	۰/۱۳۳	۰/۹۳۳	تمرین:	
	مجموع	۴۸۳۷	۲۶۲	۰/۹۴۹	۰/۹۲۵	۰/۰۲	
پارت	تقسیم درصدی (تمرین ۷۰٪ آزمون ۳۰٪)	۱۴۲۳	۶	۰/۹۹۶	۰/۹۱۸	ایجاد:	
	خسارتی	۳۲	۴۲	۰/۴۳۲	۰/۹۰۸	۰/۱۷	
	جرحی	۱	۲۶	۰/۰۳۷	۰/۸۰۶	آزمون:	
	فوتی	۱۴۵۶	۷۴	۰/۹۵۲	۰/۹۱۶	۰/۰۱	

جدول ۴. دقت پیش بینی مدل پارت

الگوریتم	نمونه گیری	جراحت مشاهده شده	موارد صحیح طبقه بندی شده	موارد اشتباه طبقه بندی شده	دقت (Accuracy)	سطح زیر منحنی (AUC)	زمان (ثانیه)
پارت	خسارتی	۴۷۱۸	۲۶	۰/۹۹۵	۰/۹۲۴	ایجاد:	
	داده های تمرینی	۱۰۵	۱۵۲	۰/۴۰۹	۰/۹۰۷	۰/۴۱	
	فوتی	۱۶	۸۲	۰/۱۶۳	۰/۹۳۲	تمرین:	
	مجموع	۴۸۳۹	۲۶۰	۰/۹۴۹	۰/۹۲۴	۰/۰۲	
جی ۴۸	تقسیم درصدی (تمرین ۷۰٪ آزمون ۳۰٪)	۱۴۲۳	۶	۰/۹۹۶	۰/۹۳۳	ایجاد:	
	خسارتی	۳۵	۳۹	۰/۴۷۳	۰/۹۱۸	۰/۳۹	
	جرحی	۱	۲۶	۰/۰۳۷	۰/۸۸۶	آزمون:	
	فوتی	۱۴۵۹	۷۱	۰/۹۵۴	۰/۹۳۲	۰/۰	

بهزاد مسلم، فرزاد موحدی سبحانی، عباس سفایی

جدول ۵. دقت پیش بینی مدل بیزین ساده

الگوریتم	نمونه گیری	جراحت مشاهده شده	موارد صحیح طبقه بندی شده	موارد اشتباه طبقه بندی شده	دقت (Accuracy)	سطح زیر منحنی (AUC)	زمان (ثانیه)
بیزین ساده	تقسیم	خسارتی	۴۷۰۳	۴۱	۰/۹۹۱	۰/۹۱۰	ایجاد: ۰/۰۲
	درصدی	جرحی	۶۲	۱۹۵	۰/۲۴۱	۰/۸۸۹	تمرین: ۰/۰۶
	(تمرین ۷۰، آزمون ۳۰)	فوتی	۳۸	۶۰	۰/۳۸۸	۰/۹۱۱	
	مجموع	مجموع	۴۸۰۳	۲۹۶	۰/۹۴۲	۰/۹۰۹	
لجستیک	تقسیم	خسارتی	۱۴۲۱	۸	۰/۹۹۴	۰/۹۱۷	ایجاد: ۰/۰۲
	درصدی	جرحی	۱۷	۵۷	۰/۲۳۰	۰/۸۳۹	تمرین: ۰/۰۶
	(تمرین ۷۰، آزمون ۳۰)	فوتی	۸	۱۹	۰/۲۹۶	۰/۹۰۵	
	مجموع	مجموع	۱۴۴۶	۸۴	۰/۹۴۵	۰/۹۱۳	

جدول ۶. دقت پیش بینی مدل لجستیک

الگوریتم	نمونه گیری	جراحت مشاهده شده	موارد صحیح طبقه بندی شده	موارد اشتباه طبقه بندی شده	دقت (Accuracy)	سطح زیر منحنی (AUC)	زمان (ثانیه)
لجستیک	تقسیم	خسارتی	۴۷۱۸	۲۶	۰/۹۹۵	۰/۹۲۹	ایجاد: ۱/۰۸
	درصدی	جرحی	۱۱۲	۱۴۵	۰/۴۳۶	۰/۹۱۲	تمرین: ۰/۰۶
	(تمرین ۷۰، آزمون ۳۰)	فوتی	۱	۹۷	۰/۰۱۰	۰/۹۴۴	
	مجموع	مجموع	۴۸۳۱	۲۶۸	۰/۹۴۷	۰/۹۲۸	
پرسپترون چندلایه	تقسیم	خسارتی	۱۴۲۴	۵	۰/۹۹۷	۰/۹۲۲	ایجاد: ۰/۰۲
	درصدی	جرحی	۲۵	۴۹	۰/۳۳۸	۰/۹۰۰	تمرین: ۰/۰۶
	(تمرین ۷۰، آزمون ۳۰)	فوتی	۴	۲۳	۰/۱۴۸	۰/۸۳۷	
	مجموع	مجموع	۱۴۵۳	۷۷	۰/۹۵۰	۰/۹۲۰	

جدول ۷. دقت پیش بینی مدل پرسپترون چندلایه

الگوریتم	نمونه گیری	جراحت مشاهده شده	موارد صحیح طبقه بندی شده	موارد اشتباه طبقه بندی شده	دقت (Accuracy)	سطح زیر منحنی (AUC)	زمان (ثانیه)
پرسپترون چندلایه	تقسیم	خسارتی	۴۷۲۱	۲۳	۰/۹۹۵	۰/۹۳۲	ایجاد: ۵۳/۱۱
	درصدی	جرحی	۹۱	۱۶۶	۰/۳۵۴	۰/۹۱۷	تمرین: ۰/۰۶
	(تمرین ۷۰، آزمون ۳۰)	فوتی	۳۲	۶۶	۰/۳۲۷	۰/۹۴۳	
	مجموع	مجموع	۴۸۴۴	۲۵۵	۰/۹۵۰	۰/۹۳۱	
پرسپترون چندلایه	تقسیم	خسارتی	۱۴۲۳	۶	۰/۹۹۶	۰/۹۴۰	ایجاد: ۵۳/۵۵
	درصدی	جرحی	۳۰	۴۴	۰/۴۰۵	۰/۹۲۴	

شناسایی عوامل موثر و بررسی تصادف‌های ترافیکی با استفاده از رویکردهای داده‌کاوی ...

تمرین ۷۰،	فوتی	۵	۲۲	۰/۱۸۵	۰/۹۱۹	آزمون:
آزمون (۳۰)	مجموع	۱۴۵۸	۷۲	۰/۹۵۳	۰/۹۳۸	۰/۰۳

جدول ۸. دقت پیش بینی مدل کلاس بندی از طریق رگرسیون

الگوریتم	نمونه گیری	جراحت مشاهده شده	موارد صحیح طبقه بندی شده	موارد اشتباه طبقه بندی شده	دقت (Accuracy)	سطح زیر منحنی (AUC)	زمان (ثانیه)
کلاس بندی از طریق رگرسیون	تقسیم در صدی (تمرین ۷۰، آزمون ۳۰)	خسارتی	۴۷۰۸	۳۶	۰/۹۹۲	۰/۹۱۴	ایجاد:
	داده های تمرینی	جرحی	۱۰۵	۱۵۲	۰/۴۰۹	۰/۹۰۰	۱/۰۹
	تمرینی	فوتی	۰	۹۸	۰/۰۰۰	۰/۸۵۵	تمرین:
	مجموع	مجموع	۴۸۱۳	۲۸۶	۰/۹۴۴	۰/۹۱۲	۰/۰۳
طریق رگرسیون	تقسیم در صدی (تمرین ۷۰، آزمون ۳۰)	خسارتی	۱۴۲۱	۸	۰/۹۹۴	۰/۹۱۶	ایجاد:
	درصدی	جرحی	۳۱	۴۳	۰/۴۱۹	۰/۹۰۴	۰/۷۲
	تمرین ۷۰، آزمون ۳۰	فوتی	۱	۲۶	۰/۰۳۷	۰/۸۱۰	آزمون:
	مجموع	مجموع	۱۴۵۳	۷۷	۰/۹۵۰	۰/۹۱۴	۰/۰

داده های تمرینی برای کلاس فوتی ولی در مجموع دقت مدل برای داده های تمرینی برابر با ۰/۹۴۴ و برای تقسیم درصدی ۰/۹۵۰ بود. همچنین سطح زیر نمودار عملکرد سیستم برای داده های تمرینی ۰/۹۱۲ و برای تقسیم درصدی ۰/۹۱۴ به دست آمد

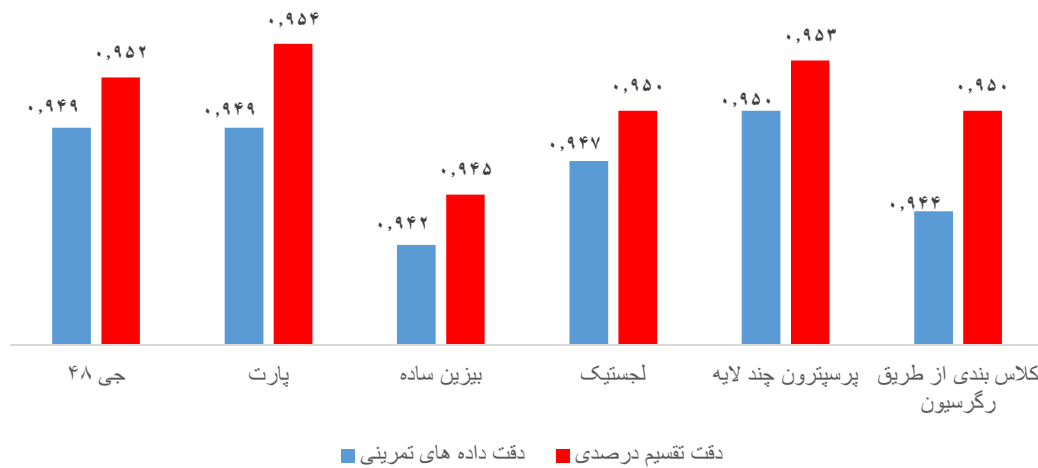
شکل ۱ نشان می دهد که دقت مجموع برای مدل پرسپترون چندلایه بر اساس روش داده های تمرینی عملکرد بهتری را نسبت به سایر مدل های مورد بررسی از خود نشان داد. اما در روش تقسیم درصدی، روش پارت بیشترین دقت پیش بینی را از خود نشان داد و سپس با اختلاف کمی پرسپترون چند لایه بهترین عملکرد را از خود نشان داد. از بررسی این نمودار می توان دریافت که همه مدل ها دارای دقت پیش بینی بسیار خوبی هستند اما در میان همین روش ها الگوریتم بیزین ساده هم در داده های تمرینی و هم در تقسیم درصدی عملکرد ضعیف تری نسبت به سایرین از خود نشان داد هر چند نکته قابل توجه برای این روش میزان واریانس کم میان دو روش مورد ارزیابی است که میتوان از آن اینگونه برداشت کرد که این مدل برای بررسی و مقایسه مدل ها مدل مناسبی خواهد بود

نتایج پیش بینی برای الگوریتم بیزین ساده در جدول ۵ ارائه شده است. نتایج دقت مجموع برای این الگوریتم نشان می دهد که داده های تمرینی ۰/۹۴۲ بوده و برای تقسیم در صدی بهبود یافته و به ۰/۹۴۵ رسیده است. این بهبود در نتایج بررسی در سطح زیر نمودار هم مشاهده می شود و از ۰/۹۰۹ در داده های تمرینی به ۰/۹۱۳ در تصمیم در صدی رسیده است. نتایج به دست آمده برای الگوریتم لجستیک در جدول ۶ آورده شده است. پیش بینی های این الگوریتم برای داده های تمرینی دارای دقت ۰/۹۴۷ و برای تقسیم درصدی دارای دقت ۰/۹۵۰ است. سطح زیر منحنی در داده های تمرینی ۰/۹۲۸ بود و در تقسیم درصدی ۰/۹۲۰ به دست آمد. نتایج حاصل شده از الگوریتم پرسپترون چندلایه در جدول ۷ آورده شده است. دقت مجموع در داده های تمرینی ۰/۹۵۰ و در تقسیم درصدی ۰/۹۵۳ است. نتایج حاصل از زیر سطح نمودار برای داده های تمرینی ۰/۹۳۱ به دست آمد و برای تقسیم درصدی با بهبود روبرو شده و ۰/۹۳۸ به دست آمد. نتایج حاصل شده از الگوریتم کلاس بندی از طریق رگرسیون در جدول ۸ ارائه شده است. علارغم عدم تشخیص مناسب در

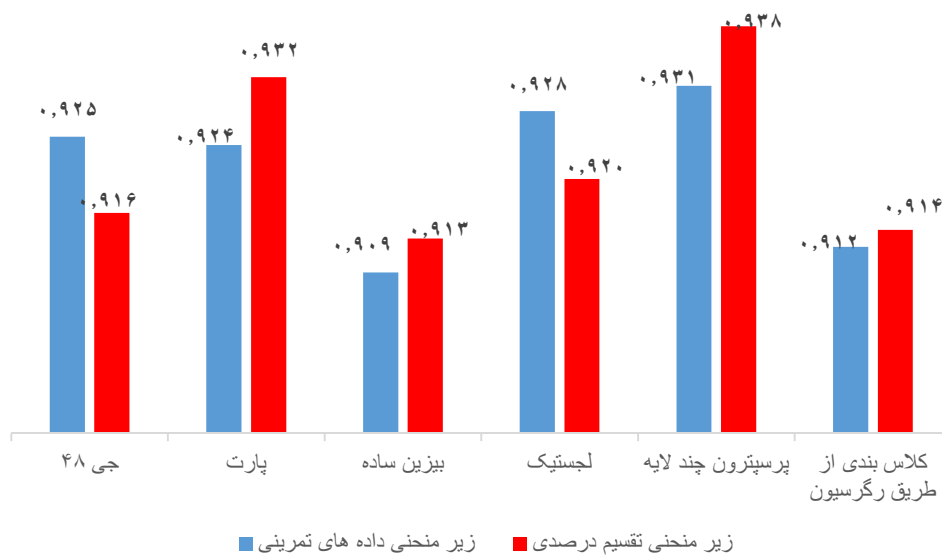
بهزاد مسلم، فرزاد موحدی سبحانی، عباس سفایی

بیشترین دقت را در میان مدل‌های آزمون دارد. پس از آن الگوریتم پارت نیز از دقت خوبی برخوردار است. الگوریتم‌های بیزین ساده و همچنین کلاس بندی از طریق رگرسیون دارای کمترین سطح زیر نمودار بوده و از دقت پایین تری نسبت به سایر مدل‌های آزمون برخوردارند.

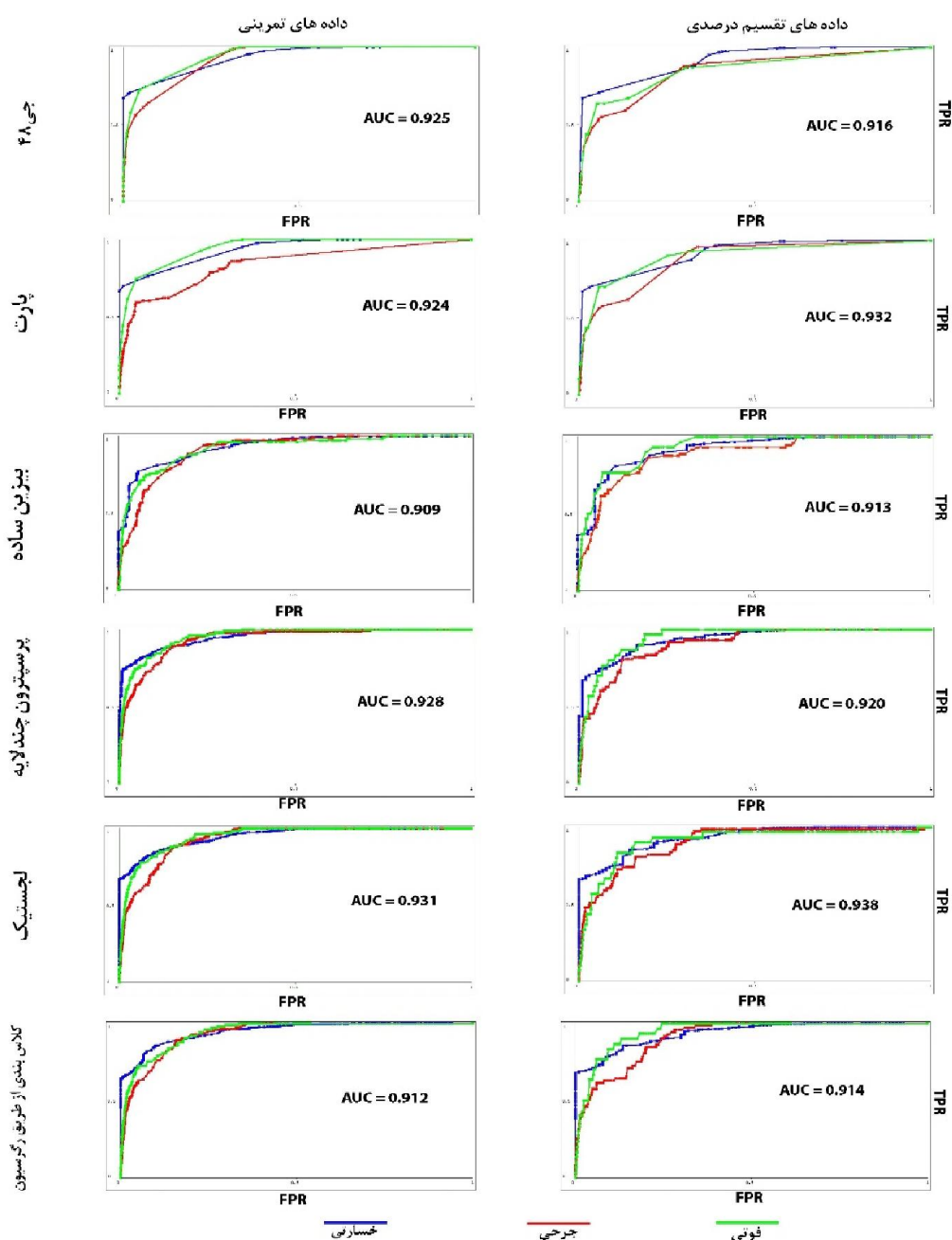
همان طور که در بخش پیشین گفته شده، از آنجا که کلیه الگوریتم‌های مورد آزمون سطح زیر نمودار بیشتر از ۰/۹ دارند، بنابراین جزو مدل‌های بسیار خوب به شمار می‌آیند. همان‌گونه که در شکل ۲ مشخص است، الگوریتم پرسپترون چندلایه برای هر دو روش داده‌های تمرینی و تقسیم درصدی دارای بیشترین پوشش در سطح زیر نمودار منحنی عملکرد سیستم را دارد و



شکل ۱. دقت مجموع پیش بینی مدل‌های مورد بررسی



شکل ۲. مجموع سطح زیر منحنی عملکرد سیستم برای مدل‌های مورد بررسی



شکل ۳. منحنی عملکرد سیستم برای مدل‌های داده‌کاوی مورد بررسی

برخوردار است. مدل‌های جی ۴۸ و پارت که مورد استفاده قرار گرفتند، توانایی تولید قوانین را دارا هستند. برای این منظور از الگوریتم اختصاصی اپروری نیز در کنار این دو مدل استفاده کردیم و بر اساس داده‌های موجود اقدام به تولید قوانین کردیم. ۲۰ قانون مهم استخراج شده از مجموعه قوانین در جدول ۹ ارائه گردیده است.

در شکل ۳ در زیر نمودارهای مشخصه عملکرد سیستم محاسبه شده برای هر سه متغیر تابع هدف و برای هر ۶ مدل داده‌کاوی نشان داده شده است. در شکل مجموع سطح زیر منحنی^{۳۳} برای هر مدل به صورت عددی مشخص شده است.

۳-۴ استخراج قوانین تصمیم‌گیری

استخراج قوانینی که بر اساس آنها بتوان تصمیمات مهمی جهت کاهش تصادفات گرفت، از درجه اهمیت بالایی

جدول ۹. قوانین استخراج شده

کل نمونه - خطای نمونه	قوانین تولید شده انتخاب شده	خصوصیت کلاس
۲۶-۲۵۳	مانع دید = بدون مانع دید عامل تصادف = در صحنه تصادف حضور دارد نوع برخورد = وسیله نقلیه با وسیله نقلیه هوا = صاف نحوه برخورد = از جلو به عقب کاربری محل = کشاورزی علل مستقیم بیان شده توسط پلیس = نقض قوانین رانندگی تاریخ = ۰۸/۹۲	
۴۲	عامل تصادف = در صحنه تصادف حضور دارد نوع برخورد = وسیله نقلیه با وسیله نقلیه نوع خط کشی جاده = متقاطع هوا = صاف موقعیت تصادف = باندا تندرو هندسه محل = مسیر مستقیم و مسطح علل قبلی بیان شده توسط پلیس = شتاب بی مورد علل مستقیم بیان شده توسط پلیس = نقض قوانین رانندگی	
۶	عامل تصادف = در صحنه تصادف حضور دارد نوع برخورد = وسیله نقلیه با موتورسیکلت نحوه برخورد = از جلو به عقب هندسه محل = مسیر مستقیم و مسطح تاریخ = ۹۱/۱۲ روشنایی = روز نوع راه = آزادراه	
۱۱	عامل تصادف = در صحنه تصادف حضور دارد نوع برخورد = وسیله نقلیه با وسیله نقلیه هوا = صاف نحوه برخورد = از جلو به جلو هندسه محل = مسیر مستقیم و مسطح علل قبلی بیان شده توسط پلیس = عدم توجه به رانندگی تاریخ = ۹۱/۰۷ نوع شانه راه = آسفالت	خسارتی
۲-۱۶	نوع برخورد = وسیله نقلیه با وسیله نقلیه هوا = صاف هندسه محل = مسیر مستقیم و مسطح کاربری محل = تجاری و اداری علل اولیه بیان شده توسط پلیس = نیاز به آموزش بیشتر راننده	
۱-۹	عامل تصادف = در صحنه تصادف حضور دارد نوع برخورد = واژگونی هوا = صاف کاربری محل = کشاورزی شرایط سطح راه = خشک عامل انسانی = عجله و شتاب نوع راه = آزادراه	
۱۲	عامل تصادف = در صحنه تصادف حضور دارد نوع برخورد = وسیله نقلیه با وسیله نقلیه نوع خط کشی جاده = متقاطع هوا = صاف نحوه برخورد = از سمت چپ به سمت راست شرایط سطح راه = خشک	
۱-۳	عامل تصادف = در صحنه تصادف حضور دارد نوع برخورد = وسیله نقلیه با موتورسیکلت نحوه برخورد = از جلو به عقب هندسه محل = مسیر مستقیم و مسطح روشنایی = شب نوع راه = آزادراه	
۱۴	مانع دید = بدون مانع دید عامل تصادف = در صحنه تصادف حضور دارد نوع برخورد = وسیله نقلیه با شی ثابت عامل انسانی = عجله و شتاب روشنایی = روز	
۳	نوع برخورد = وسیله نقلیه با شی ثابت علل قبلی بیان شده توسط پلیس = عدم توجه به رانندگی علل مستقیم بیان شده توسط پلیس = نقض قوانین رانندگی تاریخ = ۹۱/۰۷ روشنایی = شب	
۲-۷	عامل تصادف = در صحنه تصادف حضور دارد نوع برخورد = واژگونی هوا = صاف کاربری محل = کشاورزی علل مستقیم بیان شده توسط پلیس = نقض قوانین رانندگی شرایط سطح راه = خشک تاریخ = ۹۱/۱۱ عامل انسانی = خستگی و خواب آلودگی نوع راه = آزادراه	جرحی
۶	عامل تصادف = در صحنه تصادف حضور دارد نحوه برخورد = وسیله نقلیه با عابر پیاده کاربری محل = کشاورزی علل اولیه بیان شده توسط پلیس = نیاز به آموزش بیشتر راننده علل مستقیم بیان شده توسط پلیس = نقض قوانین رانندگی شرایط سطح راه = خشک نوع راه = جاده اصلی	

شناسایی عوامل موثر و بررسی تصادف‌های ترافیکی با استفاده از رویکردهای داده‌کاوی ...

۳	عامل تصادف = در صحنه تصادف حضور دارد نحوه برخورد = وسیله نقلیه با عابر پیاده کاربری محل = آموزشی علل مستقیم بیان شده توسط پلیس = نقض قوانین رانندگی شرایط سطح راه = خشک
۱-۹	نوع برخورد = وسیله نقلیه با عابر پیاده علل اولیه بیان شده توسط پلیس = نیاز به آموزش بیشتر راننده علل قبلی بیان شده توسط پلیس = عدم توجه به رانندگی نوع شانه راه = شانه راه ندارد
۱-۳	نوع برخورد = وسیله نقلیه با عابر پیاده موقعیت تصادف = باند تندرو علل اولیه بیان شده توسط پلیس = نیاز به آموزش بیشتر راننده علل قبلی بیان شده توسط پلیس = شتاب بی مورد نوع شانه راه = آسفالت
۱-۳	عامل تصادف = در صحنه تصادف حضور دارد نوع برخورد = وسیله نقلیه با وسیله نقلیه نحوه برخورد = از جلو به جلو هندسه محل = مسیر مستقیم و مسطح تاریخ = ۹۱/۱۲ نوع شانه راه = شانه راه ندارد
۶-۱۵	نوع برخورد = وسیله نقلیه با عابر پیاده موقعیت تصادف = باند تندرو کاربری محل = کشاورزی نوع راه = آزادراه
۱-۳	کاربری محل = غیرمسکونی علل اولیه بیان شده توسط پلیس = نیاز به آموزش بیشتر راننده علل قبلی بیان شده توسط پلیس = عدم توجه به رانندگی تاریخ = ۹۱/۱۱ روشنایی = شب
۲	عامل تصادف = در صحنه تصادف حضور دارد نوع برخورد = وسیله نقلیه با عابر پیاده کاربری محل = کشاورزی علل اولیه بیان شده توسط پلیس = نقص علائم جاده علل مستقیم بیان شده توسط پلیس = نقض قوانین رانندگی شرایط سطح راه = خشک
۲-۹	هوا = صاف کاربری محل = غیرمسکونی عامل انسانی = خستگی و خواب آلودگی

از جنبه کسب مهارت کافی رانندگی را می‌طلبد. آنچه که در بسیاری از موارد تصادفات به چشم می‌خورد، حضور عامل اصلی تصادف در صحنه تصادف است که دلیل آن را می‌توان به آموزش‌ها و پیام‌های پلیس در خصوص حضور در محل تصادف بجای گریختن از آن برای کمک و بررسی موضوع توسط پلیس، جستجو کرد. اما جنبه سوم مورد که از مجموعه قوانین به خوبی قابل درک است، توجه مسئولان به شرایط راه است. شرایط محیطی جاده‌ها، شانه‌های راه، وجود علائم مناسب هشدار دهنده در راه‌ها، وضعیت و شرایط سطح راه‌ها از عواملی هستند که نیازمند توجه برای کاهش تصادفات جاده هستند. در این میان توجه به مساله روش‌شنایی در معابر و نحوه ساخت معابر برون شهری به نحوی که نور خورشید در زمان‌های مختلف روز تاثیر کمتری در دید رانندگان داشته باشد از نکات مهم این عامل است. عامل چهارمی که جلب توجه می‌کند، وجود تاریخ‌هایی است که نشان از آن دارد در برخی ماه

مجموعه قوانین حاصل شده از سه روش بیان شده مورد بررسی قرار گرفتند و در نهایت مواردی که بیشتر می‌توانستند مفید باشند استخراج گردیدند. با تحلیل این مجموعه قوانین می‌توان دریافت، در مواردی که عوامل به درستی ثبت نشده‌اند، لزوم آموزش و توجه ویژه مسئولان راهنمایی و رانندگی و مسئولان راهداری به ثبت و نگهداری این داده‌ها اهمیت ویژه و خاصی را می‌طلبد. درخصوص عامل مانع دید می‌توان در موارد خاص مانع دید در تصادفات یافت ولی به دلیل وجود این تصادفات در مسیرهای برون شهری عموماً مانع دید قابل توجهی وجود ندارد. از سوی دیگر در بسیاری از قوانین به دلیل عدم مهارت راننده، تحت آموزش کافی قرار نگرفتن رانندگان، عدم توجه به مسئولیت‌های فردی و اجتماعی و عدم توجه به قوانین راهنمایی و رانندگی موجود شاهد تصادفات ترافیکی بودیم. این جنبه از تصادفات لزوم آموزش بیشتر و دقیق‌تر به رانندگان چه از جنبه آموزش قوانین و چه

رگرسیون و پرسپترون چند لایه مورد استفاده قرار گرفت. برای استخراج مجموعه قوانین از الگوریتم‌های اپریوری، جی ۴۸ و پارت استفاده گردید.

نتایج حاصل از استخراج قوانین نشان داد که، حضور عامل تصادف در صحنه تصادف، نوع برخورد، مانع دید، موقعیت تصادف، شرایط سطح راه، هندسه محل تصادف و علل مستقیم بیان شده توسط پلیس برای تصادف از مهم‌ترین عواملی بودند که در مجموعه قوانین به آن‌ها اشاره شده و بیشترین تعداد تکرار را داشتند. پس با در نظر گرفتن تدابیری در جهت کاهش و یا اصلاح این موارد می‌توان شاهد کاهش تصادفات ترافیکی در این جاده بوده و سالانه صدمات مالی و جانی کمتری به شهروندان و جامعه وارد آید.

نتایج حاصل از پژوهش در بحث مدل‌های داده کاوی، در مجموع نشان دهنده پیش‌بینی بهتر توسط الگوریتم پرسپترون چند لایه بود و پس از آن الگوریتم پارت نتایج را بهتر پیش‌بینی کرد. دقت مجموع برای مدل پرسپترون چند لایه داده‌های تمرینی ۰/۹۵۳ و برای تقسیم درصدی ۰/۹۵ بدست آمد. روش پارت نیز دارای دقت مدل ۰/۹۵۴ برای داده‌های تمرینی و ۰/۹۴۹ برای تقسیم درصدی بود. سطح زیر نمودار منحنی عملکرد سیستم برای پرسپترون در نهایت ۰/۹۳۸ و برای پارت ۰/۹۳۲ به دست آمد. الگوریتم بیزین ساده هم در دقت مدل و هم در میزان پوشش داده، علاقم و وجود در بازه‌های قابل قبول عملکرد ضعیف‌تری نسبت به سایر الگوریتم‌ها داشت.

پیشنهاد می‌گردد در پژوهش‌های آتی از روش‌های ترکیبی جهت تجزیه و تحلیل داده‌های تصادفات و پیش‌بینی تصادفات استفاده گردد. و یا از روش‌های فازی استفاده گردد و نتایج با روش حاضر مورد بررسی قرار گیرد. همچنین پیشنهاد می‌گردد بردار ورودی را با مقایسه چندین روش بررسی کرده و سپس بهترین روش انتخاب ورودی بر اساس نیاز انتخاب گردد. می‌توان پژوهش حاضر را برای چندین مسیر برون شهری مورد بررسی قرار داده و سپس به تجزیه و تحلیل آن‌ها

های خاص تصادفات به دلیل شرایط جاده، شرایط فصلی و شرایط زمانی افزایش پیدا می‌کنند. برای رفع این مشکل با بررسی دقیق‌تر این زمان‌ها با کمک یک سری زمانی می‌توان بیشتر پراکندگی تصادفات را به دست آورد و با توجه به آنها اقداماتی را مدنظر قرار داد. عامل بعدی که می‌توان آن را بررسی کرد نوع و نحوه برخورد خودرو هاست که نیاز به بررسی‌های بیشتری دارد. و در نهایت عامل تصادفات است که برای عابران پیاده رخ داده است. این نوع تصادفات می‌تواند دلایل مختلفی داشته باشد که در مجموعه قوانین نیز یاد شده است. اما توجه به این نکته که لزوم آموزش به عموم مردم و گسترش فرهنگ رانندگی، تشریح محیط جاده و نقش عابران پیاده در میان اقشار مختلف جامعه اهمیت زیادی دارد، بر کسی پوشیده نیست. هرچند متأسفانه در این زمینه غفلت‌هایی صورت پذیرفته که معضل عابران پیاده در شهرها و عدم کنترل آن به محیط برون شهری و جاده‌های برون شهری نیز کشیده شده است.

۵. نتیجه‌گیری و پیشنهادها

این پژوهش دو زمینه کلی را مورد بررسی قرار داد. نخست پیش‌بینی شدت تصادفات با استفاده از مدل‌های کلاس‌بندی و بررسی دقت مدل‌های مورد استفاده بود و دوم به ایجاد مجموعه قوانینی که می‌توان با تجزیه و تحلیل آن‌ها، زمینه‌های کاهش تصادف را برنامه‌ریزی نمود. داده‌ها شامل ۵۰۹۹ رکورد تصادفات ترافیکی صورت گرفته در محور تهران-قم در استان تهران طی یک دوره ۴ ساله بود. هر رکورد تصادف شامل ۵۷ خصوصیت درباره تصادفات بود. که پس از پیش پردازش رکوردهای مورد استفاده کاهش یافت. ۸ خصوصیت، عامل تصادف، نوع برخورد، مانع دید، موقعیت تصادف، شرایط سطح راه، هندسه محل، علل مستقیم و نوع تصادف در نهایت به عنوان بردار ورودی مورد بررسی قرار گرفتند. در پژوهش حاضر نرم افزار داده کاوی وکا با بهره‌گیری از الگوریتم‌های جی ۴۸، پارت، بیزین ساده، لجستیک، کلاس‌بندی از طریق

International Journal of Crashworthiness, Vol. 21, No. 2, pp. 104-111

-Chang, L. Y. and Wang, H. W. (2006) "Analysis of traffic injury severity: An application of non-parametric classification tree techniques", Accident Analysis and Prevention, Vol. 38, No. 5, pp. 1019-1027

-de Oña, J., López, G. and Abellán, J. (2013) "Extracting decision rules from police accident reports through decision trees", Accident Analysis & Prevention, Vol. 50, No., pp. 1151-1160

-Deb, R. and Liew, A. W. C. (2016) "Missing value imputation for the analysis of incomplete traffic accident data", Information Sciences, Vol. 339, No., pp. 274-289

-Geurts, K., Thomas, I., and Wets, G. (2005) "Understanding spatial concentrations of road accidents using frequent item sets", Accident Analysis and Prevention, Vol. 37, No. 4, pp. 787-799

-Haghighi, F. R. and GholamNejad, R. (2016) "Modeling the risk and safety of passage of students in roadside schools", Journal of Transportation Engineering Research, Vol. 7, No. 4, pp. 605-614

-Jung, S., Qin, X. and Oh, C. (2016) "Improving strategic policies for pedestrian safety enhancement using classification tree modeling", Transportation Research Part A: Policy and Practice, Vol. 85, No., pp. 53-64

-Kashani, A. T. and Besharati, M. M. (2017) "Fatality rate of pedestrians and fatal crash involvement rate of drivers in pedestrian crashes: a case study of Iran", International Journal of Injury Control and Safety Promotion, Vol. 24, No. 2, pp. 222-231

پرداخت تا بانک اطلاعاتی جامع از تصادفات برون شهری در ایران حاصل شود. این پژوهش به تصادفات جاده‌ای معطوف بود، می‌توان از زوایای دیگر به این نوع تصادفات توجه ویژه‌ای نشان داد.

۶. پی‌نوشت‌ها

1. Correlation-based Feature Selection (CFS)
2. Naïve Bayesian
3. Logistic
4. Multilayer Perceptron
5. Classification via Regression
6. Rule Induction(PART)
7. Decision Tree J48
8. Apriori
9. Greedy Search Algorithm
10. Forward
11. Backward
12. Training set
13. Split
14. Confusion Matrix
15. True Positive (TP)
16. True Negative (TN)
17. False Positive (FP)
18. False Negative (FN)
19. Model Accuracy
20. True Positive Rate (TPR)
21. False Positive Rate (FPR)
22. Receiver Operating Curve (ROC)
23. Area Under the Curve(AUC)

۷. مراجع

-Anvari, M. B., Tavakoli Kashani, A. and Rabieyan, R. (2017) "Identifying the most important factors in the at-fault probability of motorcyclists by data mining, based on classification tree models", International Journal of Civil Engineering, Vol. 15, No. 4, pp. 653-662

-Castro, Y. and Kim, Y. J. (2016) "Data mining on road safety: Factor assessment on vehicle accidents using classification models",

- Pakgozar, A., Tabrizi, R. S., Khalili, M. and Esmaeili, A. (2010) "The role of human factor in incidence and severity of road crashes based on the CART and LR regression: A data mining approach," 1st World Conference on Information Technology, WCIT-2010, Istanbul, 2011, pp. 764-769.
- Prati, G., Pietrantoni, L. and Fraboni, F. (2017) "Using data mining techniques to predict the severity of bicycle crashes", Accident Analysis and Prevention, Vol. 101, No., pp. 44-54
- Taamneh, M., Alkheder, S. and Taamneh, S. (2017) "Data-mining techniques for traffic accident modeling and prediction in the United Arab Emirates", Journal of Transportation Safety and Security, Vol. 9, No. 2, pp. 146-166
- Tao, G., Song, H., Liu, J., Zou, J. and Chen, Y. (2016) "A traffic accident morphology diagnostic model based on a rough set decision tree", Transportation Planning and Technology, Vol. 39, No. 8, pp. 751-758
- Tavakoli Kashani, A., Rabieyan, R. and Besharati, M. M. (2014) "A data mining approach to investigate the factors influencing the crash severity of motorcycle pillion passengers", Journal of Safety Research, Vol. 51, No., pp. 93-98
- Yau, K. K. W., Lo, H. P. and Fung, S. H. H. (2006) "Multiple-vehicle traffic accidents in Hong Kong", Accident Analysis and Prevention, Vol. 38, No. 6, pp. 1157-1161
- Kashani, A. T. and Mohaymany, A. S. (2011) "Analysis of the traffic injury severity on two-lane, two-way rural roads based on classification tree models", Safety Science, Vol. 49, No. 10, pp. 1314-1320
- Kumar, S. and Toshniwal, D. (2017) "Severity analysis of powered two wheeler traffic accidents in Uttarakhand, India", European Transport Research Review, Vol. 9, No. 2, pp.
- Kwon, O. H., Rhee, W. and Yoon, Y. (2015) "Application of classification algorithms for analysis of road safety risk factor dependencies", Accident Analysis and Prevention, Vol. 75, No., pp. 1-15
- Mohaymany, A. S., Kashani, A. T. and Ranjbari, A. (2010) "Identifying driver characteristics influencing overtaking crashes", Traffic Injury Prevention, Vol. 11, No. 4, pp. 411-416
- Montella, A. (2011) "Identifying crash contributory factors at urban roundabouts and using association rules to explore their relationships to different crash types", Accident Analysis and Prevention, Vol. 43, No. 4, pp. 1451-1463
- Montella, A., Aria, M., D'Ambrosio, A. and Mauriello, F. (2011) "Data-mining techniques for exploratory analysis of pedestrian crashes", Transportation Research Record, Vol. 2237, pp. 107-116. DOI: 10.3141/2237-12

شناسایی عوامل موثر و بررسی تصادف‌های ترافیکی با استفاده از رویکردهای داده‌کاوی ...

بهزاد مسلم، درجه کارشناسی در رشته علوم کامپیوتر را در سال ۱۳۹۴ از دانشگاه کاشان و درجه کارشناسی ارشد در رشته مهندسی صنایع را در سال ۱۳۹۶ از دانشگاه آزاد اسلامی واحد علوم و تحقیقات تهران اخذ نمود. زمینه‌های پژوهشی مورد علاقه ایشان، علوم داده، داده کاوی، کشف دانش و تجزیه و تحلیل داده‌ها، کیفیت، کنترل کیفیت، کنترل کیفیت آماری، بهبود فرآیندها و مدیریت و کنترل پروژه است.



دکتر فرزاد موحدی سبحانی، درجه کارشناسی خویش در رشته مهندسی صنایع را از دانشگاه صنعتی اصفهان و درجه کارشناسی ارشد در رشته مهندسی صنایع را از دانشگاه تربیت مدرس اخذ نمود. همچنین موفق به کسب درجه دکتری در رشته مهندسی صنایع از دانشگاه تربیت مدرس تهران گردید. زمینه‌های پژوهشی مورد علاقه ایشان، تجزیه و تحلیل چند متغیره، سیستم‌های دینامیکی، استراتژی‌های مدیریت دانش، مدیریت پروژه‌های فناوری اطلاعات، سازمان‌دهی و رهبری، مدیریت تغییر، بازمهندسی فرآیندهای کسب و کار و داده کاوی است. ایشان در حال حاضر عضو هیات علمی با مرتبه استادیار و مدیر گروه مهندسی صنایع در دانشگاه آزاد اسلامی واحد علوم و تحقیقات تهران است.



دکتر عباس سقایی، درجه کارشناسی و کارشناسی ارشد در رشته مهندسی صنایع را در سال‌های ۱۳۷۴ و ۱۳۷۶ از دانشگاه علم و صنعت ایران اخذ نمود. همچنین در سال ۱۳۸۴ موفق به کسب درجه دکتری در رشته مهندسی صنایع از دانشگاه علم و صنعت ایران گردید. زمینه‌های پژوهشی مورد علاقه ایشان، کنترل کیفیت آماری، مدیریت کیفیت، طراحی و تحلیل آزمایش‌ها، آمار و احتمال مهندسی، کنترل کیفیت پیشرفته، آمار و احتمال مهندسی، روش تحقیق، قابلیت اطمینان، داده کاوی و تحلیل سری‌های زمانی بوده و در حال حاضر عضو هیات علمی با مرتبه دانشیار در دانشگاه آزاد اسلامی واحد علوم و تحقیقات تهران است.

