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@iau.ac.ir

دریافت: ۱۳۹٦/۰۸/۰٤ پذیرش: ۱۳۹٦/۱۲/۱۳

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 برآنها 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 جمع آوری 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 جی٤٨، برای پیش‌بینی دقت مدل‌های مورد ارزیابی با استفاده از نرم افزار داده‌کاوی وکا مورد استفاده قرار گرفتند؛ همچنین از الگوریتم اپریوری به همراه دو مدل جی٤٨ و پارت جهت استخراج قوانین استفاده شد. 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 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

با پیچیده‌تر شدن جوامع انسانی، نقش حمل و نقل برای جوامع بشـــری روز به روز اهمیت بیشـــتری پیدا می‌کند. یکی از مهم‌ترین وظایف در ایمنی جاده ها، تعیین دلایل تصـــادفات ترافیک جاده‌ای اســت. ســهم قابل توجهی از شــبکه جاده‌های برون شهری ایران شامل مسیرهای دو طرفه و دوبانده است. ۲۱۷۹۷ کیلومتر از جاده های اصلی برون شهری، ۳۰ درصد از شبکه‌های جاده‌ای ایران را تشکیل می‌دهد و بیش از ۹۰ درصد مســـافران ایرانی، حالت حمل و نقل جاده‌ای را با توجه به ســهولت و قیمت تمام شــده، برای ســفرهای خود انتخاب می کنند. طبق آمار منتشــرشــده توســط پژوهشــکده 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 تصــادفات جاده‌ای در کشـــور ما یک تهدید جدی به شـــمار می‌رود و نیازمند توجه جدی و سریع در این راستا است. بنابراین نیازمند راهکاری برای رفع این مشــکل و به دنبال پاســخی برای رفع این مســاله حیاتی که ســالانه صــدمات جبران ناپذیری به بدنه اجتماعی و اقتصادی کشور وارد می‌سازد هستیم.

این پژوهش شـــکاف تحقیقاتی در زمینه دقت و صـــحت روش‌های مورد اســتفاده و عدم مقایســه روش‌های مختلف به منظور انتخاب روشـــی مناســـب‌تر جهت انجام کلاس‌بندی و کشــف الگوهای مناســب از آن‌ها و همچنین شــکاف در ارائه قوانینی جامع در حوزه تصـــادفات که بتواند مفید باشـــد را شناسایی کرده و اهداف خود را برای پوشش این شکاف های علمی بر مبنای یافتن مهم‌ترین خصــیصــه های موثر و ســپس تحلیل و بررســـی دقت مدل های داده کاوی برای پیش بینی تصادفات برون شهری در ایران قرار داده است. به همین منظور از ٦ الگوریتم داده کاوی، بیزین ساده، لجیستیک، پرسپترون چندلایه، کلاس بندی از طریق رگرسیون، قوانین استنتاجی(پارت) و درخت تصمیم‌گیری جی۴۸ با استفاده از نرم افزار داده‌کاوی وکا استفاده شده است.

در بخش بعدی این مقاله، پیشـــینه پژوهشـــی درباره پیش‌بینی شدت تصادفات ترافیکی با روش‌های کلاس‌بندی بررسی شده اســت. بخش ســه، مقدمه‌ای بر مدل‌های کلاس‌بندی، الگوریتم های مورد استفاده و روش‌های ارزیابی و همچنین توضیحاتی در مورد داده‌های مورد مطالعه ارائه می‌دهد. ســـپس، نتایج و تحلیل‌های مربوطه در بخش چهار ارائه شده است. در نهایت، نتیجه‌گیری بر اساس نتایج مطالعه توصیف می‌شود.

۲. پیشینه پژوهش

۲-۱ پیشینهٔ 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 برآنها است. این مجموعه داده‌ها می‌تواند شـــامل ده‌ها پارامتر مختلف باشـــد که بر اســـاس رویکردها و تکنیک‌های مختلف می‌توان آنها را مورد بررسی قرار داد.

با بررســـی ادبیات و پژوهش‌های صـــورت گرفته در این زمینه مشـــاهده شـــد که تنها مقالات [Castro and Kim,2016, Taamneh, Alkheder and Taamneh,2017] بـــه بررســـی الگوریتم جی۴۸ پرداختند. البته [de Oña, López and Abellán,2013, Deb and Liew,2016, Tao, et al.,2016] نیز با الگوریتم سی ٤/٥، که مشابه با الگوریتم جی ٤٨ اســـت اقدام به بررســـی کردند. در این میان مدل درخت کلاس‌بندی و رگرســـیون(کارت) با بیشـــترین اقبال از ســـوی نویســـندگان روبرو بوده اســـت [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, Pakgohar, et al.,2011, Tavakoli Kashani, Rabieyan and Besharati,2014]. در کنار

این‌ها [de Oña, López and Abellán,2013, Kwon, Rhee and Yoon,2015, Montella, et al.,2011] نیز با مدل‌های دیگری از درخت تصمیم گیری 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 شده اند [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 الگوریتم ها کاربرد کمتری دارد ولی همچنان در [Kwon, Rhee and Yoon,2015, Pakgohar, et al.,2011, Yau, Lo and Fung,2006] مورد استفاده قرار گرفته است. با وجود آنکه در بیشتر پژوهش‌ها با استفاده از الگوریتم های درخت تصمیم‌گیری اقدام به ارائه قوانین کردند اما تنها در [Geurts, Thomas and Wets,2005, Montella,2011, Montella, et al.,2011] از الگوریتم های 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, Pakgohar, 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 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 جی۴۸، 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] جدید ترین الگوریتم های مورد استفاده در تحقیقات اخیر هستند.

۲-۲ پیشینهٔ تجربی

تامنه و همکارانش(۲۰۱۷) به بررسی تصادفات ترافیکی و یافتن دلایل آن با توجه با مشخصات رانندگان، جاده و تصادفات پرداخته اند. آنها با بررسی ۵۹۷۳ مورد اطلاعات تصادفات در شهر ابوظبی امارات در سال های ۲۰۰۸ تا ۲۰۱۳و با استفاده از روش‌های مختلف داده‌کاوی و استفاده از الگوریتم های مختلف کلاس‌بندی همچون درخت تصمیم‌گیری، قوانین استنتاجی، شبکه های 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 کرده اند. 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 مدل های درخت کلاس‌بندی پرداختند. آنها داده هایی 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

پرداختند. نتایج نشـان داد که نوع تصـادف و سـن راننده در احتمال مقصـر بودن موثر بود. همچنین تصـادف جلو به عقب بیشـترین میزان تصـادف در این زمینه را داشـت. در نهایت پیشنهاد آموزش بیشتر به رانندگان و نصب علائم هشدار دهنده ارائـه گردیـده بود[Anvari, Tavakoli Kashani and Rabieyan,2017].

کومار و توشـنیوال، به تجزیه و تحلیل تصـادفات ترافیکی وسـایل نقلیه با دو چرخ، در هند پرداختند. آنها با تجزیه و تحلیل عوامل موثر بر شدت تصادفات به وسیله سه الگوریتم درخت تصـمیم‌گیری رگرسـیون و کلاس‌بندی، بیزین سـاده و بردار حمایت ماشینی دریافتند که برای ۱۳ منطقه مورد بررسی عوامل متفاوتی وجود دارد. همچنین مدل درخت رگرسـیون و کلاس بنـدی دقـت بهتری نسـبت بـه دو الگوریتم دیگر داشت[Kumar and Toshniwal,2017].

پراتی و همکارنش، با اسـتفاده از تکنیک‌های داده کاوی اقدام به پیش بینی شدت تصادفات دوچرخه سواران کردند. آنها با بهره گیری از دو الگوریتم درخت تصمیم گیری چاید و شبکه بیزین دریافتند برای درخت تصـمیم گیری چاید، نوع جاده، نوع تصادف و سن دوچرخه سوار مهم‌ترین عوامل بودند. و برای شـبکه بیزین نوع تصـادف، نوع جاده و نوع سـرنشـینان وسـیله به عنوان مهم‌ترین عامل شـدت تصـادفات دوچرخه سـواران بـه دسـت آمـد[Prati, Pietrantoni and Fraboni,2017].

جانگ و همکارانش به بررسـی بهبود اسـتراتژی‌های پلیس جهت بهبود ایمنی عابران پیاده با اسـتفاده از مدل‌های درخت کلاس‌بندی پرداختند. با اسـتفاده از داده‌ها و تحلیل آنها با اسـتفاده از روش‌های درخت کلاس‌بندی به دنبال راهکاری برای کمک به پلیس کره جنوبی بودند. نتایج این تحقیق نشـان داد که سـن عابران پیاده و نوع حرکت آنها دو عامل اولیه در شـدت جراحات بودند. همچنین در ادامه تجزیه و تحلیل کلی و الگوهای خاص جهت اسـتفاده پلیس تهیه گردید که می‌توان از این الگوها در سـایر کشـورها نیز بهره برد. این پژوهش اسـتفاده از چندین روش کاوش داده‌ها را در انجام کارهای آینده پیشنهاد می‌دهد[Jung, Qin and Oh,2016].

توکلی‌کا شانی و همکارانش به بررسی رویکردهای داده کاوی در عوامل موثر بر شدت تصادفات سرنشینان موتور سیکلت پرداخته‌اند. آنها با اسـتفاده از روش درخت کلاس‌بندی و رگرسیون، داده‌های ایران را در یک دوره ٤ سال مورد بررسی قرار دادند و دریافتند نوع منطقه، سـطح مورد اسـتفاده و ناحیه آسـیب دیده بدن بیشـترین تاثیر را بر جراحات موتورسـواران دارند. همچنین دقت پیش‌بینی مدل سـاخته شـده بهبود زیادی نسبت به تحقیقات پیشین خود داشت و استفاده از کلاه ایمنی تاثیر بسیار زیادی در کاهش آسیب‌ها داشته است[Tavakoli Kashani, Rabieyan and Besharati,2014].

توکلی‌کا شانی و شریعت‌مهیمنی، شدت جراحات ترافیکی در جاده‌های دوبانده و دوطرفه برون شـهری ایران را بر اسـاس مدل‌های درخت کلاس بندی و طی مدت سـه سـال تجزیه و تحلیل کردند. آنها دریافتند که حرکات مارپیچ و نبستن کمربند ایمنی مهم‌ترین عوامل موثر بر شـدت آسـیب اسـت. همچنین نتایج به دسـت آمده نسـبت به تحقیقات پیشـین بهبود یافته است[Kashani and Mohaymany,2011].

حقیقی و غلام نژاد در پژوهشـی ابتدا شـرایط عبور از عرض خیابان در یک مدر سه داخل شهری درکناره خیابان به و سیله ویدیوگرافی مورد ارزیابی و تجزیه و تحلیل قرارگرفته، سـپس با اسـتفاده از تکنیک‌های ریاضـی با اسـتفاده از متغیرهای اسـتخراج شـده مدل‌سـازی تعارض و ریسـک ایمنی عبور پرداخته شـده اسـت. در همین راسـتا ارزیابی‌های آماری عبور دانش‌آموزان از عرض خیابان نیز مورد مطالعه قرار گرفته است. نتایج این تحقیق نشـان می‌دهد که ۱۰ متغیر نظیر اندازه گروه عبور، سـرعت نزدیک شـدن، زمان انتظار قبل از عبور، زمان توقف و انتظار روی خط، متوسط زمان عبور، سرعت متوسط عبور، ، توجه به وسایل نقلیه نزدیک شونده، عقب برگشتن از

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 تصمیم گیری‌های عبور از عرض خیابان داشته اند[Haghighi 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 مدل های پیش بینی از ٦ الگوریتم داده‌کاوی بیزین ساده[۲]، لجیستیک[۳]، پرسپترون چندلایه[٤]، کلاس‌بندی از طریق رگرسیون[٥]، قوانین استنتاجی(پارت)[٦] و درخت تصمیم گیری جی٤٨[۷] با استفاده از نرم افزار داده‌کاوی وکا استفاده شد. همچنین از الگوریتم اپریوری[۸] به همراه الگوریتم جی٤٨ 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 ٤ بخش تعریف کرد: جمع آوری داده‌ها، پیش پردازش 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.

۳-۱ تعریف داده ها

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 ٤ ساله بود.

۳-۲ تعریف مدل‌ها

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

- درخت تصمیم گیری جی ٤٨ : درخت‌های تصمیم گیری 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 داده های 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 طبقه بندی استفاده می‌شود.
- شبکه عصبی پرسپترون چندلایه : این روش یک تکنیک یادگیری تحت نظارت است که برای طبقه بندی و رگرسیون استفاده می‌شود. پرسپترون چند لایه، یک شبکه عصبی مصنوعی ایجاد می‌کند که متشکل از گره‌های چندگانه در سه

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 رگر سیون برای متغیرهای وابسته دو سویی است. این مدل را می‌توان به عنوان مدل خطی تعمیم‌یافته‌ای که از تابع لجیت به عنوان تابع پیوند استفاده می‌کند و خطایش از توزیع چندجمله‌ای پیروی می‌کند، به‌حساب‌آورد.
- کلاس بندی از طریق رگرسیون : درختان طبقه بندی و رگر سیون، 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 رگر سیون برای متغیرهای وابسته هستند که مقادیر مستقل یا دستورالعمل را با خطای پیش بینی، معمولا با تفاوت مربع بین مقادیر مشاهده شده و پیش بینی شده اندازه گیری می‌کنند.
- اپریوری : اپریوری یک الگوریتم کلاسیک برای یادگیری قوانین وابستگی است. اپریوری روی پایگاه‌های داده شامل تراکنش‌ها ساخته شده است. ورودی این الگوریتم مجموعه‌ای از مجموعه آیتم ها است. این الگوریتم تلاش می‌کند تا زیرمجموعه‌هایی از آیتم‌ها را که حداقل بین چند مجموعه آیتم مشترک است بیابد. اپریوری یک الگوریتم پایین به بالا است، آن‌گونه که در هر مرحله یک آیتم به زیرمجموعه های مکرر اضافه می‌شود.
- انتخاب ویژگی مبتنی بر همبستگی : این الگوریتم با در نظر گرفتن توانایی پیش بینی فردی هر یک از ویژگی ها و همچنین میزان همبستگی بین آن‌ها، ارزش یک زیر مجموعه از ویژگی ها را ارزیابی می‌کند. زیرمجموعه ویژگی هایی که با کلاس همخوانی زیادی دارند، در مقایسه با آن‌ها که همخوانی کمتری دارند، ترجیح داده می‌شوند.

۳-۳ پیش پردازش داده ها

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 داده های کد شده در کنار داده های کیفی بود. پس از پیش پردازش ستون های مربوط به کد، ستون های خالی و نامرتبط از مجموعه حذف گردیدند. در نهایت مشخص گردید تنها ۲٤ خصیصه قابلیت اطمینان و اتکا جهت بررسی انتخاب خصیصه های موثر بر تصادفات را داشتند. بر اساس متغیر هدف، نوع تصادفات به ۳ دسته فوتی، جرحی و خسارتی تقسیم بندی شده است.

۳-٤ 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[9] استفاده شد. این الگوریتم یک جستجوی پیشرانه[10] یا پسرانه[11] را از طریق خصوصیات زیر مجموعه‌ها انجام می‌دهد.

در مرحله بعد ٦ الگوریتم، بیزین ساده، لجیستیک، پرسپترون چندلایه، کلاس بندی از طریق رگرسیون، قوانین استنتاجی(پارت) و درخت تصمیم‌گیری جی٤٨ که هر یک از خانواده متفاوتی هستند انتخاب گردیدند تا علاوه بر انتخاب روش های مختلف انواع مختلف مدل های آماری، درخت ها، 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

مجموعه داده تمرینی[12] مورد ارزیابی اولیه و تمرین قرار گرفتند که ارزیابی و ارزش هر مدل مورد بررسی مشخص گردد. سپس تو سط روش تق سیم در صدی[13]، 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 ا ستخراج قوانین از الگوریتم اپریوری به همراه الگوریتم جی۴۸ و پارت استفاده می‌شود و قانون‌های برتر مشخص می‌گردند.

۳-۵ روش ارزیابی مدل

عملکرد مدل کلاس بندی شده توسط یک ماتریس اشتباهات[14] تعریف می‌شود که نمونه‌های درست و نادرست کلاس‌بندی را برای هر کلاس نشان می‌دهد. اندازه‌گیری‌هایی که برای ارزیابی عملکرد یک کلاس‌بندی نیاز ا ست تو سط ماتریس ا شتباهات محا سبه می شوند. جدول ۱ این ماتریس را ن شان می‌دهد. این ماتریس دارای ٤ مقدار زیر است:

- درست-مثبت(TP)[15]: نمونه‌هایی که درست هستند و به عنوان درست کلاس‌بندی شده‌اند. این مقدار بیانگر میزان موفقیت پیش بینی است.
- درست-منفی(TN)[16]: نمونه‌هایی که درست هستند و به عنوان اشتباه کلاس‌بندی شده‌اند. این مقدار بیانگر به طور صحیح مردود شده است.
- اشتباه-مثبت(FP)[17]: نمونه‌هایی که اشتباه هستند و به عنوان در ست کلاس‌بندی شده‌اند. این مقدار بیانگر یک هشدار غلط است یا به بیان دیگر بیانگر خطای نوع ۱ است.
- ا شتباه-منفی(FN)[18] : نمونه‌هایی که ا شتباه ه ستند و به عنوان اشتباه کلاس‌بندی شده‌اند. این مقدار بیانگر مقادیر از د ست رفته یا به بیان دیگر بیانگر **خطای نوع ۲ است.**

جدول ۱. ماتریس اشتباهات

		پیش بینی	
		مثبت	منفی
مشاهدات	مثبت	درست-مثبت(**TP**)	اشتباه-منفی(**FN**)
	منفی	اشتباه-مثبت(**FP**)	درست-منفی(**TN**)

بر ا ساس ماتریس ا شتباهات شاخص دقت مدل[19]، را تعریف می‌کنیم که بیانگر تفاوت میان نتیجه واقعی و نتیجه پیش بینی شده ا ست. برای این منظور مواردی که به در ستی کلاس‌بندی و محاسبه شده‌اند را نسبت به کل موارد پیش بینی شده درصد بندی می‌کند. فرمول ۱ نحوه محاسبه آن را نشان می‌دهد:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \quad (۱)$$

شاخص های نرخ مثبت صحیح[20] و نرخ مثبت اشتباه[21] برای اساس ماتریس اشتباهات به ترتیب از طریق فرمول های ۲ و ۳ محا سبه شده و برای محا سبه شاخص نمودار منحنی عملکرد سیستم[22] در ادامه استفاده می‌شوند.

$$\text{True Positive Rate (TPR)} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (۲)$$

$$\text{False Positive Rate (FPR)} = \frac{\text{FP}}{\text{FP} + \text{TN}} \quad (۳)$$

شاخص دیگری که علاوه بر دقت مدل مورد ارزیابی قرار می‌گیرد منحنی عملکرد سیستم ا ست که محورهای آن را نرخ مثبت صحیح و نرخ مثبت ا شتباه ت شکیل می‌دهند. برای ر سم نمودار، مقادیر نرخ مثبت صحیح و نرخ مثبت اشتباه، نقاط برش بین ۰ و ۱را محا سبه 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 ب سیار عالی داریم. اگر ناحیه بین ۰/۸ تا ۰/۹ با شد با یک مدل خوب مواجه هستیم و اگر بین ۰/۷ و ۰/۸ باشد با یک مدل قابل قبول روبرو هستیم.

۳-٦ استخراج قوانین

برای درک عوامل اصلی که بر شدت تصادفات تاثیرگذارند، مجموعه قوانینی را توسط الگوریتم های اپریوری، جی٤٨ و پارت استخراج می‌کنیم. از مجموعه داده ها به عنوان یک مجموعه آموزشی برای الگوریتم فوق استفاده کرده و در نهایت قانون‌هایی که دارای درجه اهمیت بالاتری هستند را انتخاب می‌کنیم.

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

۱. **تعریف و جمع آوری داده ها**
۲. **پیش پردازش داده ها**
   ۲. ۱. **یکپارچه سازی داده ها**
   ۲. ۲. **حذف داده های نامناسب، غیرقابل اعتماد و دارای نویز از مجموعه داده ها**
۳. **روش شناسی پژوهش**
   ۳. ۱. **انتخاب خصیصه ها بر اساس مدل انتخاب ویژگی مبتنی بر همبستگی(CFS)**
   ۳. ۲. **انتخاب خصیصه هدف**
   ۳. ۳. **محاسبه مدل‌های داده‌کاوی(جی٤٨، پارت، بیزین ساده، پرسپترون، لجستیک، کلاس‌بندی از طریق رگرسیون) برای مجموعه داده تهران-قم**
      ۳. ۳. ۱. **روش داده های تمرینی**
      ۳. ۳. ۲. **روش تقسیم درصدی (۷۰ درصد آموزش، ۳۰ درصد آزمون)**
٤. **ارزیابی نتایج و شاخص ها**
   ٤. ۱. **شاخص ماتریس ماتریس اشتباهات**
   ٤. ۲. **شاخص نرخ مثبت‌های صحیح (TPR)**
   ٤. ۳. **شاخص نرخ مثبت‌های اشتباه (FPR)**
   ٤. ٤. **شاخص دقت مدل (Accuracy)**
   ٤. ٥. **شاخص منحنی مشخصه عملکرد سیستم (ROC)**
٥. **استخراج قوانین**
٦. **پایان**

٤- یافته‌های پژوهش

٤-۱ 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

٤-۲ پیش بینی دقت مدل ها

نتایج حاصل از بررسی داده‌ها توسط ٦ مدل مطرح شده به ترتیب در زیر آورده شده است. نتایج پیش‌بینی برای الگوریتم جی٤٨ در جدول۳ ارایه شده است. دقت مدل برای الگوریتم جی٤٨ بر اساس داده‌های تمرینی به ترتیب برای خسارتی، جرحی و فوتی، ۰/۹۹٥ ، ۰/٤۰۱ و ۰/۱۳۳ به دست آمد. همچنین بر اساس روش تقسیم درصدی به ترتیب ۰/۹۹٦ ،

۰/٤۳۲ و ۰/۰۳۷ به دست آمد. ارزیابی دقت مجموع با نمونه‌گیری توسط داده‌های تمرینی ۰/۹٤۹ 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 ۰/۹۱٦ رسید. دقت الگوریتم پارت بر اساس داده‌های تمرینی به ترتیب برای خسارتی، جرحی و فوتی به مقدار ۰/۹۹۵ ، ۰/٤۰۹ و ۰/۱٦۳ به دست آمد و برای تقسیم درصدی به ترتیب ۰/۹۹٦ ، ۰/٤۷۳ ، ۰/۰۳۷ بود. نتایج پیش بینی در جدول٤ آورده شده است. از نتایج مشخص است که دقت مجموع در روش نمونه‌گیری و در روش تقسیم درصدی ۰/۹٤۹ و ۰/۹۵٤ است. همچنین سطح زیر نمودار عملکرد سیستم به ترتیب برای داده‌های تمرینی و تقسیم درصدی ۰/۹۲٤ و ۰/۹۳۲ بود.

جدول ۳. دقت پیش بینی مدل جی٤۸

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 ٤۸	داده های تمرینی	خسارتی	٤۷۲۱	۲۳	۰/۹۹۵	۰/۹۲٦	ایجاد:
		جرحی	۱۰۳	۱۵٤	۰/٤۰۱	۰/۹۱۱	۰/۱٦
		فوتی	۱۳	۸۵	۰/۱۳۳	۰/۹۳۳	تمرین:
		مجموع	٤۸۳۷	۲٦۲	۰/۹٤۹	۰/۹۲۵	۰/۰۲
	تقسیم درصدی (تمرین ۷۰، آزمون ۳۰)	خسارتی	۱٤۲۳	٦	۰/۹۹٦	۰/۹۱۸	ایجاد:
		جرحی	۳۲	٤۲	۰/٤۳۲	۰/۹۰۸	۰/۱۷
		فوتی	۱	۲٦	۰/۰۳۷	۰/۸۰٦	آزمون:
		مجموع	۱٤۵٦	۷٤	۰/۹۵۲	۰/۹۱٦	۰/۰۱

جدول ٤. 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	داده های تمرینی	خسارتی	٤۷۱۸	۲٦	۰/۹۹۵	۰/۹۲٤	ایجاد:
		جرحی	۱۰۵	۱۵۲	۰/٤۰۹	۰/۹۰۷	۰/٤۱
		فوتی	۱٦	۸۲	۰/۱٦۳	۰/۹۳۲	تمرین:
		مجموع	٤۸۳۹	۲٦۰	۰/۹٤۹	۰/۹۲٤	۰/۰۲
	تقسیم درصدی (تمرین ۷۰، آزمون ۳۰)	خسارتی	۱٤۲۳	٦	۰/۹۹٦	۰/۹۳۳	ایجاد:
		جرحی	۳۵	۳۹	۰/٤۷۳	۰/۹۱۸	۰/۳۹
		فوتی	۱	۲٦	۰/۰۳۷	۰/۸۸٦	آزمون:
		مجموع	۱٤۵۹	۷۱	۰/۹۵٤	۰/۹۳۲	۰/۰

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	داده های تمرینی	خسارتی	٤٧٠٣	٤١	٠/٩٩١	٠/٩١٠	ایجاد:
		جرحی	٦٢	١٩٥	٠/٢٤١	٠/٨٨٩	٠/٠٢
		فوتی	٣٨	٦٠	٠/٣٨٨	٠/٩١١	تمرین:
		مجموع	٤٨٠٣	٢٩٦	٠/٩٤٢	٠/٩٠٩	٠/٠٦
	تقسیم درصدی (تمرین ۷۰، آزمون ۳۰)	خسارتی	١٤٢١	٨	٠/٩٩٤	٠/٩١٧	ایجاد:
		جرحی	١٧	٥٧	٠/٢٣٠	٠/٨٣٩	٠/٠٢
		فوتی	٨	١٩	٠/٢٩٦	٠/٩٠٥	آزمون:
		مجموع	١٤٤٦	٨٤	٠/٩٤٥	٠/٩١٣	٠/٠

جدول ٦. 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	داده های تمرینی	خسارتی	٤٧١٨	٢٦	٠/٩٩٥	٠/٩٢٩	ایجاد:
		جرحی	١١٢	١٤٥	٠/٤٣٦	٠/٩١٢	١/٠٨
		فوتی	١	٩٧	٠/٠١٠	٠/٩٤٤	تمرین:
		مجموع	٤٨٣١	٢٦٨	٠/٩٤٧	٠/٩٢٨	٠/٠٦
	تقسیم درصدی (تمرین ۷۰، آزمون ۳۰)	خسارتی	١٤٢٤	٥	٠/٩٩٧	٠/٩٢٢	ایجاد:
		جرحی	٢٥	٤٩	٠/٣٣٨	٠/٩٠٠	٠/٩٢
		فوتی	٤	٢٣	٠/١٤٨	٠/٨٣٧	آزمون:
		مجموع	١٤٥٣	٧٧	٠/٩٥٠	٠/٩٢٠	٠/٠

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	داده های تمرینی	خسارتی	٤٧٢١	٢٣	٠/٩٩٥	٠/٩٣٢	ایجاد:
		جرحی	٩١	١٦٦	٠/٣٥٤	٠/٩١٧	٥٣/١١
		فوتی	٣٢	٦٦	٠/٣٢٧	٠/٩٤٣	تمرین:
		مجموع	٤٨٤٤	٢٥٥	٠/٩٥٠	٠/٩٣١	٠/٠٦
	تقسیم درصدی	خسارتی	١٤٢٣	٦	٠/٩٩٦	٠/٩٤٠	ایجاد:
		جرحی	٣٠	٤٤	٠/٤٠٥	٠/٩٢٤	٥٣/٥٥

(تمرین ۷۰،	فوتی	۵	۲۲	۰/۱۸۵	۰/۹۱۹	آزمون:
آزمون ۳۰)	مجموع	۱٤۵۸	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	داده های تمرینی	خسارتی	٤۷۰۸	۳٦	۰/۹۹۲	۰/۹۱٤	ایجاد:
		جرحی	۱۰۵	۱۵۲	۰/٤۰۹	۰/۹۰۰	۱/۰۹
		فوتی	۰	۹۸	۰/۰۰۰	۰/۸۵۵	تمرین:
		مجموع	٤۸۱۳	۲۸٦	۰/۹٤٤	۰/۹۱۲	۰/۰۳
	تقسیم درصدی (تمرین ۷۰، آزمون ۳۰)	خسارتی	۱٤۲۱	۸	۰/۹۹٤	۰/۹۱٦	ایجاد:
		جرحی	۳۱	٤۳	۰/٤۱۹	۰/۹۰٤	۰/۷۲
		فوتی	۱	۲٦	۰/۰۳۷	۰/۸۱۰	آزمون:
		مجموع	۱٤۵۳	۷۷	۰/۹۵۰	۰/۹۱٤	۰/۰

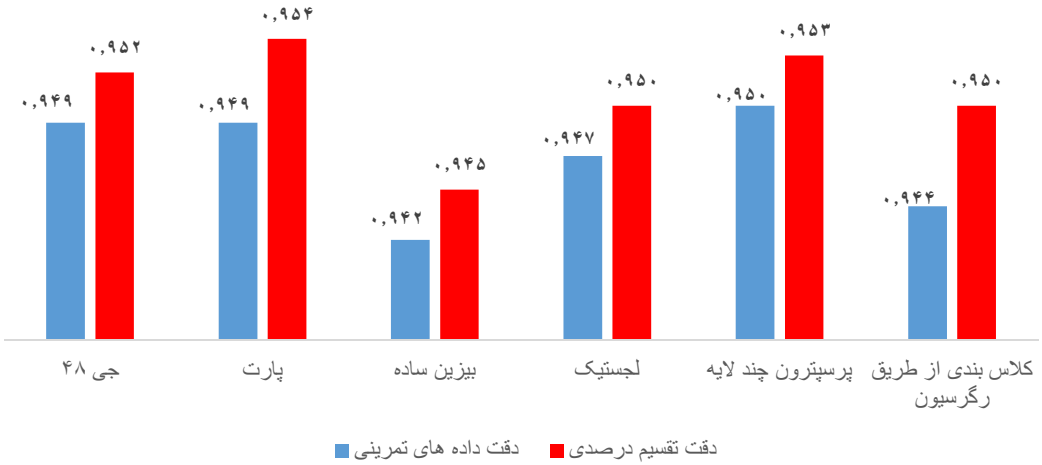
نتایج پیش‌بینی برای الگوریتم بیزین ساده در جدول۵ ارایه شده است. نتایج دقت مجموع برای این الگوریتم نشان می‌دهد که داده‌های تمرینی ۰/۹٤۲ بوده و برای تقسیم در صدی بهبود یافته و به ۰/۹٤۵ رسیده است. این بهبود در نتایج بررسی در سطح زیر نمودار هم مشاهده می‌شود و از ۰/۹۰۹ در داده‌های تمرینی به ۰/۹۱۳ در تصمیم در صدی رسیده است. نتایج به دست آمده برای الگوریتم لجستیک در جدول٦ آورده شده است. پیش‌بینی‌های این الگوریتم برای داده‌های تمرینی دارای دقت ۰/۹٤۷ و برای تقسیم درصدی دارای دقت ۰/۹۵۰ است. سطح زیر منحنی در داده‌های تمرینی ۰/۹۲۸ بود و در تقسیم درصدی ۰/۹۲۰ به دست آمد.

نتایج حاصل شده از الگوریتم پرسپترون چندلایه در جدول۷ آورده شده است. دقت مجموع در داده‌های تمرینی ۰/۹۵۰ و در تقسیم در صدی 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.

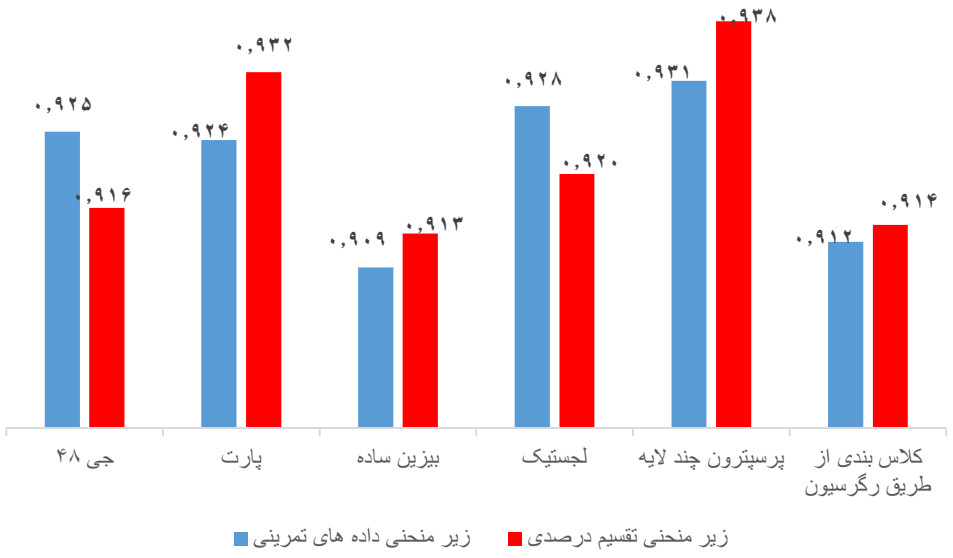
نتایج حاصل شده از الگوریتم کلاس‌بندی از طریق رگرسیون در جدول۸ ارائه شده است. علارغم عدم تشخیص مناسب در داده‌های تمرینی برای کلاس فوتی ولی در مجموع دقت مدل برای داده‌های تمرینی برابر با ۰/۹٤٤ و برای تقسیم درصدی ۰/۹۵۰ بود. همچنین سطح زیر نمودار عملکرد سیستم برای داده‌های تمرینی ۰/۹۱۲ و برای تقسیم درصدی ۰/۹۱٤ به دست آمد

شکل۱ نشان می‌دهد که دقت مجموع برای مدل پرسپترون چندلایه بر اساس روش داده‌های تمرینی عملکرد بهتری را نسبت به سایر مدل‌های مورد بررسی از خود نشان داد. اما در روش تقسیم درصدی، روش پارت بیشترین دقت پیش‌بینی 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 مدل‌ها دارای دقت پیش بینی بسیار خوبی هستند اما در میان همین روش‌ها الگوریتم بیزین ساده هم در داده‌های تمرینی و هم در تقسیم درصدی عملکرد ضعیف‌تری نسبت به سایرین از خود نشان داد هرچند 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

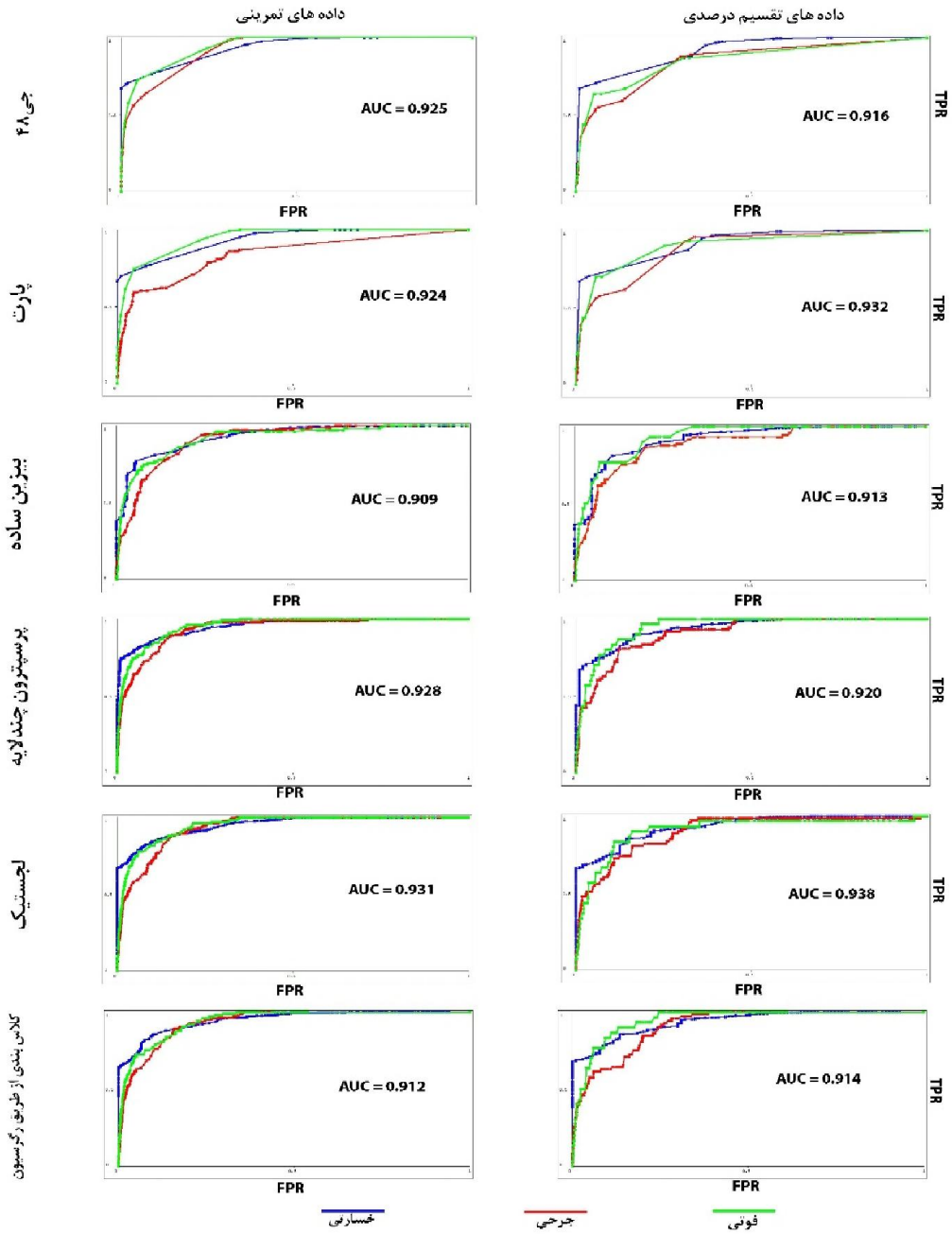
همان طور که در بخش پیشـــین گفته شـــده، از آنجا که کلیه الگوریتم‌های مورد آزمون سطح زیر نمودار بیشتر از ۰/۹ دارند، بنابراین جزو مدل‌های بسیارخوب به شمار می‌آیند. همان‌گونه که در شکل۲ مشخص است، الگوریتم پرسپترون چندلایه برای هر دو روش داده‌های تمرینی و تقسیم درصدی دارای بیشترین پوشش در سطح زیر نمودار منحنی عملکرد سیستم را دارد و بیشـــترین دقت را در میان مدل های آزمون دارد. پس از آن الگوریتم پارت نیز از دقت خوبی برخوردار است. الگوریتم‌های بیزین ساده و همچنین کلاس بندی از طریق رگر سیون دارای کمترین ســطح زیر نمودار بوده و از دقت پایین‌تری نســبت به سایر مدل‌های آزمون برخوردارند.



شکل ۱. دقت مجموع پیش بینی مدل‌های مورد بررسی



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



شکل ۳. منحنی عملکرد سیستم برای مدل های داده کاوی مورد بررسی

در شکل ۳ در زیر نمودارهای مشخصه عملکرد سیستم محاسبه شده برای هر سه متغیر تابع هدف و برای هر ٦ مدل داده کاوی نشان داده شده است. در شکل مجموع سطح زیر منحنی[۲۳] برای هر مدل به صورت عددی مشخص شده است.

٤-۳ استخراج قوانین تصمیم گیری

استخراج قوانینی که بر اساس آنها بتوان تصمیمات مهمی جهت کاهش تصادفات گرفت، از درجه اهمیت بالایی برخوردار است. مدل‌های جی۴۸ و پارت که مورد استفاده قرار گرفتند، توانایی تولید قوانین را دارا هستند. برای این منظور از الگوریتم اختصاصی اپریوری 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.

جدول ۹. قوانین استخراج شده

خصوصیت کلاس	قوانین تولید شده انتخاب شده	کل نمونه- خطای نمونه
خسارتی	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 |	۲۶-۲۵۳
	عامل تصادف = در صحنه تصادف حضور دارد | نوع برخورد = وسیله نقلیه با وسیله نقلیه | نوع خط کشی جاده = متقاطع | هوا = صاف | موقعیت تصادف = باند 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 |	۴۲
	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 |	۶
	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 |	۱۱
	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 |	۲-۱۶
	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 |	۱-۹
	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 |	۱۲
	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 |	۱-۳
	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 |	۱۴
جرحی	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 |	۳
	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 |	۲-۷
	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 |	۶

	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 |	۳
	نوع برخورد = وسیله نقلیه با عابرپیاده | 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 |	۱-۹
	نوع برخورد = وسیله نقلیه با عابرپیاده | 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 |	۱-۳
	نوع برخورد = وسیله نقلیه با عابرپیاده | موقعیت تصادف = باند تندرو | کاربری محل = کشاورزی | نوع راه = آزادراه |	۶-۱۵
فوتی	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 |	۱-۳
	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
	هوا = صاف | کاربری محل = غیرمسکونی | عامل انسانی = خستگی و خواب آلودگی |	۲-۹

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 نشده اند، لزوم آموزش و توجه ویژه مسئولان راهنمایی و رانندگی و مسئولان راهداری به ثبت و نگهداری این داده ها 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 عموما 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 مسئولیت های فردی و اجتماعی و عدم توجه به قوانین راهنمایی و رانندگی موجود شاهد تصادفات ترافیکی بودیم. این جنبه از تصادفات لزوم آموزش بیشتر و دقیق تر به رانندگان چه از جنبه آموزش قوانین و چه 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 آموزش ها و پیام های 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 جاده ها، شانه های راه، وجود علائم مناسب هشدار دهنده در راه ها، وضعیت و شرایط سطح راه ها از عواملی هستند که نیازمند توجه برای کاهش تصادفات جاده هستند. در این میان توجه به مساله روشنایی در معابر و نحوه ساخت معابر برون شهری به نحوی که نور خورشید در زمان های مختلف روز تاثیر کمتری در دید رانندگان داشته باشد از نکات مهم این عامل است. عامل چهارمی که جلب توجه می کند، وجود تاریخ هایی است که نشان از آن دارد در برخی ماه

های خاص تصادفات به دلیل شرایط جاده، شرایط فصلی و شرایط زمانی افزایش پیدا می‌کنند. برای رفع این مشکل با بررسی دقیق تر این زمان ها 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 بررسی های بیشتری دارد. و در نهایت عامل تصادفاتی 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 متاسفانه در این زمینه غفلت‌هایی صورت پذیرفته که معضل عابران پیاده در شهرها و عدم کنترل آن به محیط برون شهری و جاده های برون شهری نیز کشیده شده است.

۵. نتیجه‌گیری و پیشنهادها

این پژوهش دو زمینه کلی را مورد بررسی قرار داد. نخست پیش بینی شدت تصادفات با استفاده از مدل‌های کلاس‌بندی و بررسی دقت مدل های 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 ٤ 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 جی٤٨، پارت، بیزین ساده، لجستیک، کلاس‌بندی از طریق رگرسیون و پرسپترون چند لایه مورد استفاده قرار گرفت. برای استخراج مجموعه قوانین از الگوریتم‌های اپریوری، جی٤٨ 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 چندلایه داده های تمرینی ۰/۹۵۳ و برای تقسیم درصدی ۰/۹۵ بدست آمد. روش پارت نیز دارای دقت مدل ۰/۹٥٤ برای داده های تمرینی و ۰/۹٤۹ 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، علارقم وجود در بازه‌های قابل قبول عملکرد ضعیف‌تری نسبت به سایر الگوریتم‌ها داشت.

پیشنهاد می‌گردد در پژوهش های آتی از روش های 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 آن ها

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 تصـــادفات توجه ویژه‌ای نشان داد.

٦. پی نوشت‌ها

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)

٧. 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", 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

-Pakgohar, 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

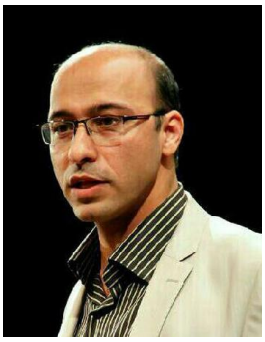
-Mohaymany, A. S., Kashani, A. T. and Ranjbari, A. (2010) "Identifying driver characteristics influencing overtaking crashes", Traffic Injury Prevention, Vol. 11, No. 4, pp. 411-416

-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

بهزاد مسلم، درجه کارشناسی در رشته علوم کامپیوتر را در سال ۱۳۹٤ از دانشگاه کاشان و درجه کارشناسی ارشد در رشته مهندسی صنایع را در سال ۱۳۹٦ از دانشگاه آزاد اسلامی واحد علوم و تحقیقات تهران اخذ نمود. زمینه های پژوهشی مورد علاقه ایشان، علوم داده، داده کاوی، کشف دانش و تجزیه و تحلیل داده ها، 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. زمینه های 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 ۱۳۷٤ و ۱۳۷٦ از دانشگاه علم و صنعت ایران اخذ نمود. همچنین در سال ۱۳۸٤ موفق به کسب درجه دکتری در رشته مهندسی صنایع از دانشگاه علم و صنعت ایران گردید. زمینه های پژوهشی مورد علاقه ایشان، کنترل کیفیت آماری، مدیریت کیفیت، طراحی و تحلیل آزمایش ها، آمار و احتمال مهندسی، کنترل کیفیت پیشرفته، آمار و احتمال مهندسی، روش تحقیق، قابلیت اطمینان، داده کاوی و تحلیل سری های زمانی بوده و در حال حاضر عضو هیات علمی با مرتبه دانشیار در دانشگاه آزاد اسلامی واحد علوم و تحقیقات تهران است.