تحلیل شدت تصادفات دوخودرویی درون‌شهری با الگوریتم جنگل تصادفی (RF) و تحلیل SHAP

مقاله علمی – پژوهشی

*حامد سیفی (نویسنده مسئول)، گروه مهندسی عمران، دانشگاه پیام نور، تهران، ایران
محمد کوهی، دانش آموخته کارشناسی ارشد، گروه مهندسی عمران، دانشگاه پیام نور، تهران، ایران
شاهین شعبانی، گروه مهندسی عمران، دانشگاه پیام نور، تهران، ایران
*پست الکترونیکی نویسنده مسئول: hamedsaify@pnu.ac.ir

دریافت:۱۴۰۴/۰۲/۲۴- پذیرش: ۱۴۰۵/۰۵/۰۵
صفحه ۲۴۰-۲۲۱

چکیده

تصادفات دوخودرویی درون‌شهری یکی از مهم‌ترین انواع تصادفات درون‌شهری محسوب می‌شوند که سهم قابل‌توجهی در تلفات انسانی و خسارات اقتصادی دارند. شناسایی عوامل مؤثر بر شدت این نوع تصادفات می‌تواند نقش مهمی در برنامه‌ریزی‌های ایمنی و مدیریت ترافیک شهری ایفا نماید. هدف این پژوهش، تحلیل عوامل مؤثر بر شدت تصادفات دوخودرویی درون‌شهری با استفاده از مدل‌های یادگیری ماشین و روش‌های تفسیرپذیر است. داده‌های مورد استفاده شامل ۵۸۶ فقره تصادف دوخودرویی ثبت‌شده در شهرستان شهرضا، طی سال‌های ۱۴۰۱ تا ۱۴۰۴ بود که از پایگاه اطلاعات پلیس راهور استخراج گردید. در مرحله پیش‌پردازش، داده‌ها کدگذاری و متعادل‌سازی شد، سپس مدل جنگل تصادفی (RF) به‌عنوان مدل اصلی توسعه یافت. همچنین عملکرد این مدل با دو مدل پایه شامل رگرسیون لجستیک (LR) و درخت تصمیم (DT) مقایسه شد. نتایج ارزیابی مدل‌ها نشان داد که مدل RF با دقت ۸۹ درصد، مقدار AUC برابر با ۰/۹۲ و مقادیر بالاتر Precision، Recall و F1-score نسبت به سایر مدل‌ها، بهترین عملکرد را در پیش‌بینی شدت تصادفات ارائه می‌دهد. به‌منظور تفسیرپذیری نتایج، از روش SHAP استفاده شد. یافته‌ها نشان داد که متغیرهای روز هفته، نوع خودرو، زمان وقوع تصادف، مانور خودرو، هندسه راه، وضعیت روشنایی و کنترل ترافیک از مهم‌ترین عوامل مؤثر بر شدت تصادفات دوخودرویی هستند. همچنین مشخص گردید وقوع تصادفات در شب، نبود کنترل ترافیک، راه‌های دوطرفه و حضور خودروهای سنگین احتمال بروز تصادفات فوتی را افزایش می‌دهد. نتایج این پژوهش نشان می‌دهد که ترکیب مدل‌های یادگیری ماشین و روش‌های تفسیرپذیر می‌تواند ابزار مؤثری برای تحلیل شدت تصادفات و تدوین راهکارهای ارتقای ایمنی ترافیک شهری باشد.

واژه‌های کلیدی: تصادفات دوخودرویی، تحلیل شدت، جنگل تصادفی (RF)، تحلیل SHAP

۱- مقدمه

نرخ‌های فوتی ناشی از تصادفات جاده‌ای در کشورهای کم‌درآمد و با درآمد متوسط، بالاتر از کشورهای توسعه‌یافته است (WHO, 2018). این کشورها با وجود داشتن تنها ۶۰ درصد از خودروهای دنیا، ۹۳ درصد از مرگ‌ومیر ناشی از تصادفات جاده‌ای را به خود اختصاص داده‌اند (WHO, 2022). خسارات جهانی ناشی از تصادفات جاده‌ای حدود ۵۱۸ میلیارد دلار برآورد می‌شود که هزینه‌ای معادل ۱ تا ۳ درصد از تولید ناخالص ملی کشورهای مختلف را در بر دارد. اگرچه نرخ تلفات تصادفات در بسیاری از کشورهای با درآمد بالا در دهه‌های اخیر کاهش یافته است، با این حال داده‌ها نشان می‌دهند که همه‌گیری

جهانی تصادفات جاده‌ای همچنان در بیشتر مناطق جهان در حال رشد است. انتظار می‌رود مرگ‌ومیر ناشی از تصادفات جاده‌ای تا سال ۲۰۳۰ به پنجمین علت اصلی مرگ‌ومیر تبدیل شود که منجر به حدود ۲/٤ میلیون فوت در سال خواهد شد، مگر اینکه اقدامات فوری صورت گیرد (WHO, 2009). تصادفات جاده‌ای را می توان به سه دسته عمده تصادفات خودرویی (تک‌خودرویی، خودرو با خودرو و چندخودرویی)، تصادفات عابرین پیاده و تصادفات موتورسیکلت تقسیم نمود. در این بین تصادفات خودرویی، مهمترین نوع تصادفات هستند زیرا علاوه بر داشتن سهم بیشتری در آمار کلی تصادفات، پتانسیل ایجاد آسیب‌های شدید و مرگ‌ومیر را نیز دارند. شدت تصادفات خودرویی در سراسر جهان، بسته به طیفی از عوامل مانند زیرساخت راه، رفتار راننده، اجرای اقدامات ایمنی موثر و غیره، متفاوت است. با وجود استفاده از خودروهای ایمن‌تر، بهبود طراحی راه و اجرای بهتر قوانین ترافیکی، هنوز هم تصادفات جاده‌ای خودرویی حتی در کشورهای توسعه‌یافته می‌تواند شدید باشد و منجر به مرگ‌ومیر و آسیب‌های شدید شود. برای مثال، در ایالات متحده، تصادفات خودرویی بیش از ٦۰ درصد از کل مرگ‌ومیر ترافیکی را به خود اختصاص داده است (NHTSA, 2020). یکی از ابزارهای اصلی پژوهشگران برای درک بهتر تصادفات به منظور مدیریت بهتر ایمنی ترافیک، استفاده از مدلسازی‌های مختلف تصادفات است. پژوهش‌ها نشان داده‌اند که تحلیل کلی تصادفات بدون تعریف زیرگروه‌های محتمل می‌تواند ارتباطات بین زیرگروه‌ها را از دست بدهد و منجر به نتایج نادرست هنگام توسعه مدل‌ها شود (Geedipally & Lord, 2010, Ma & Kockelman, 2006). به طور مناسب، پژوهشگران تلاش کرده‌اند تا مدل‌های چندگانه تصادفات را همزمان با تقسیم کل تصادفات به دسته‌های مختلف بر اساس شدت آسیب، انواع تصادف و تعداد خودروهای درگیر در یک تصادف توسعه دهند (Geedipally & Lord, 2010, Martensen & Dupont, 2013, Kitali & Sando, 2017). مدلسازی برخوردها با خوشه‌های ممکن در داده‌های تصادف می‌تواند به درک بهتر تأثیر عوامل متعدد بر هر دسته تصادف کمک کند و امکان توسعه اقدامات ایمنسازی مؤثر را فراهم کند. پژوهشگران اغلب داده‌های تصادف خودرویی (جدای از تصادفات عابرین پیاده و موتورسیکلت‌سواران) را هنگام مدلسازی تصادفات بر اساس تعداد کل خودروهای درگیر به دو گروه تقسیم می‌کنند: تصادفات تک‌خودرویی و تصادفات دو یا چند خودرویی (Geedipally & Lord, 2010, Geedipally & Lord, 2010, Martensen & Dupont, 2013, Chen & Chen, 2011, Pasupathy et al., 2000, Ma et al., 2016). پژوهش‌های قبلی نشان داده‌اند که تصادفات شامل دو یا چند خودرو به طور قابل توجهی با تصادفات شامل تنها یک خودرو متفاوت هستند. در نتیجه، این دو نوع تصادف باید به صورت جداگانه مدلسازی شوند (Geedipally & Lord, 2010; Ma et al., 2016; Qin et al., 2004; Lord et al., 2005; Griffith, 1999).

بر اساس یافته‌های این مطالعات، توسعه مدل‌های جداگانه برای تصادفات تک‌خودرویی و تصادفات دو یا چند خودرویی پیش‌بینی‌های دقیق‌تری نسبت به ایجاد مدل‌هایی ارائه می‌دهد که دو دسته تصادف را ترکیب می‌کنند (Geedipally & Lord, 2010). بنابراین توسعه مدل‌هایی که بتواند متغیرهای موثر در هر یک از این نوع تصادفات را نشان دهد و بتواند پیشبینی بهتری از انواع تصادفات در آینده را نمایان سازد، برای مدیریت بهتر ایمنی شبکه راه لازم و ضروری است. در این راستا برآورد و استفاده از مدل‌های شدت تصادف در سطح تفکیک‌شده، جزء مهمی از اقدامات پیشگیرانه در شناسایی و درک کامل عواملی است که منجر به شدت تصادفات می‌شوند. علاوه بر این، مدلسازی مستقل شدت انواع تصادفات خودرویی مورد نیاز است، زیرا مدلسازی تصادفات تجمعی (انواع تصادفات در ترکیب با هم) بدون تعیین زیرگروه‌های مرتبط ممکن است به کشف ارتباطات بین زیرگروه‌ها منجر نشود و در نتیجه برآوردهای پارامتری نادرستی ایجاد کند (Kitali et al. 2021). به همین منظور، هدف این پژوهش، بررسی شدت تصادفات دوخودرویی درون‌شهری با استفاده از روش‌های یادگیری ماشین و مدل‌های آماری بر پایه داده‌های تصادفات ثبت‌شده در شهرستان شهرضا از توابع استان اصفهان است. هدف اصلی این پژوهش، شناسایی عوامل موثر بر شدت تصادفات و

ارزیابی توان مدل‌های یادگیری ماشین در پیشبینی شدت آسیب در تصادفات دوخودرویی می‌باشد.

۲- پیشینه تحقیق

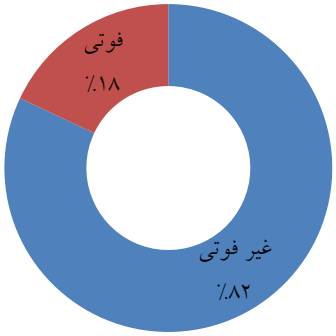
مطالعات متعددی برای بررسی مکانیسم تصادفات دوخودرویی توسط محققان مختلف با انواع مدل‌های آماری و یادگیری ماشین انجام شده است. یوان و همکاران (Yuan et al., 2022) در پنسیلوانیا، از مدل‌های لوجیت ترکیبی برای شناسایی عوامل تعیین‌کننده شدت آسیب در تصادفات دو خودرویی استفاده کردند و ویژگی‌های مختلف خودروهای درگیر در تصادف را در نظر گرفتند. نتایج نشان داد که نوع و حرکت خودروها تأثیر قابل توجهی بر شدت تصادف دارد. یوان و همکاران (Yuan et al., 2017) و لی و لی (Lee & Li, 2014) نشان دادند که تصادفات دو خودرویی در شرایط ترافیکی مشابه می‌توانند از نظر شدت بسیار متفاوت باشند، زیرا عملکرد، نوع، وزن و حرکت خودروها (خودروی برخوردکننده و خودروی برخوردشده) متفاوت است. شائو و همکاران (Shao et al., 2020) دریافتند که شدت آسیب‌های ناشی از تصادفات کامیون در برابر خودرو با آنهایی که در تصادفات صرفاً کامیونی رخ می‌دهد، تفاوت قابل توجهی دارد. زنگ و همکاران (Zeng et al., 2016) تأثیر تعامل بر شدت آسیب واحد خودرویی را در تصادفات دوخودرویی مطالعه کردند و دریافتند که در مقایسه با خودروهای سواری، سایر انواع خودروها به طور قابل توجهی شدت بیشتری دارند. چیو و همکاران (Chiou et al., 2020) در تایوان، شدت تصادف را توسط دو طرف (که به عنوان «طرف مسئول» و «طرف غیرمسئول» شناخته می‌شوند) با استفاده از معادلات برآورد عمومی در نظر گرفتند؛ نتیجه نشان داد که مهم‌ترین متغیری که به شدت تصادف کمک می‌کند، نوع خودرو است و پس از آن سرعت، زاویه برخورد و مصرف الکل قرار دارد. لومباردی و همکاران (Lombardi et al., 2017) نابرابری‌های مرتبط با سن را در تصادفات کشنده بین دو خودرو بررسی کردند. یافته‌ها نشان داد که رانندگان مسن‌تر و جوان‌تر برای درگیر شدن در تصادفات جاده‌ای مستعدتر از رانندگان با سن متوسط هستند. یانگ و همکاران (Yang et al., 2019) در ژاپن، متغیرهای کلیدی تعیین‌کننده شدت آسیب‌های رانندگان در تصادفات دوخودرویی شامل خودروی سواری و کامیون را با استفاده از مدل پروبیت ترتیبی بررسی کردند و دریافتند که هنگامه روز، مکان، شرایط ترافیکی، انواع تصادفات و انواع راه‌ها تأثیرات متفاوتی بر تصادفات دوخودرویی دارند، در حالی که شرایط جوی و سن رانندگان تأثیرات قابل توجهی برای هر دو نوع تصادف دارند. تحقیقات پیشین مبتنی بر مدلهای آماری کشف کرده‌اند که نوع تصادف یکی از عوامل تعیین‌کننده شدت تصادفات دو خودرویی است (Yang et al., 2019; Duncan et al., 1998; Chiou et al., 2020; Lee & Li, 2014, Ji & Levinson, 2020). عوامل مربوط به خودرو مانند نوع خودرو و حرکت خودرو نیز تأثیر قابل توجهی بر شدت تصادفات دو خودرویی نشان داده‌اند (Yuan et al., 2022; Zeng et al., 2016). ویژگی‌های راه و محیط اطراف (مانند شرایط ترافیکی، انواع راه و شرایط روشنایی) می‌توانند سطح شدت تصادف را تبیین کنند (Yang et al., 2019). طبق برخی تحقیقات، ویژگی‌های زمانی تصادفات (یعنی روز هفته و هنگامه روز) و میزان مصرف الکل توسط رانندگان نیز با سطح شدت در تصادفات دو خودرویی مرتبط هستند (Yang et al., 2019; Chiou et al., 2020). در سال‌های اخیر مدل‌های یادگیری ماشین به دلیل انعطاف‌پذیری بیشتر در پردازش داده‌های پرت، نویزی یا ناقص و همچنین با داشتن فرضیات پسین کمتر یا بدون فرضیه (که برای متغیرهای ورودی سازگارتر هستند) به عنوان فناوری‌های ارزشمند در مطالعات ایمنی راه‌ها ظهور کرده‌اند و همین سبب شده است که محققان برای غلبه بر نقاط ضعف رویکردهای آماری، مدل‌های مختلف یادگیری ماشین[1](ML) را برای مدل‌سازی روابط غیرخطی بین عناصر مؤثر در تصادف و پیامدهای شدت آسیب به کار گیرند (Abdel-Aty & Abdelwahab, 2004; Iranitalab & Khattak, 2017; Li et al. 2012; Pradhan & Sameen, 2020; Sameen & Pradhan, 2017; Sarkar et al., 2020; Tang et al., 2019). با این حال، مسائل مربوط به داده‌ها و روش‌شناسی در یادگیری ماشین هنوز به طور کامل حل نشده‌اند. اولاً، داده‌های مربوط به شدت تصادفات ذاتاً نامتوازن هستند و در برخی موارد نیز گزارش‌دهی ناقصی دارند. بسیاری از مطالعات نشان داده‌اند که اگرچه روش‌های یادگیری ماشین اغلب دقت کلی پیش‌بینی بالایی تولید می‌کنند، اما برای دسته‌های شدید و مرگبار که مشاهدات کمتری دارند، دقت پایینی دارند (Abdel-Aty & Abdelwahab, 2004; Chang & Wang, 2006; Chene et al., 2016a; Chene et al., 2016b; Lie et al., 2012). ثانیاً، اکثر رویکردهای یادگیری ماشین از مشکل «جعبه سیاه» رنج

می‌برند، به این معنی که تفسیر نتایج مدل‌سازی و استخراج روابط زیربنایی بین متغیرهای مستقل/توضیحی و پیامدهای شدت تصادف، نامشخص است. برای غلبه بر این مسئله، محققان تحلیل حساسیت[۲](SA) و به طور خاص تحلیل حساسیت محلی[۳] (LSA) را توسعه داده‌اند (Anvari et al., 2017; Yu & Abdel-Aty, 2014; Das et al., 2009; Zhang et al., 2018; Jiang et al., 2019; Li et al., 2008; Li et al. 2012; Zhang & Xie, 2007). یکی از کاربردهای تحلیل حساسیت، استخراج ویژگی‌ها و رتبه‌بندی اهمیت نسبی متغیرهای مختلف در رابطه با یک متغیر هدف مشخص است. این امر پیاده‌سازی مدل‌های یادگیری ماشین را در پژوهش‌های ایمنی وسایل نقلیه بسیار آسان‌تر کرده است. با این حال، برای ثبت اثرات مشترک چندین عامل خطر، تحلیل حساسیت باید فرضیات احتمالی نادرستی مانند خطی بودن، نرمال بودن و تغییرات محلی را در نظر بگیرد. هنگام به کارگیری رویکردهای یادگیری ماشین در تحلیل شدت تصادفات، مسائل اضافی مرتبط با داده‌ها و روش‌شناسی، مانند معیارهای عملکرد مدل، همبستگی‌های مکانی-زمانی تصادف، علیت، قابلیت انتقال و ناهمگونی، اغلب پدیدار می‌شوند. برخی از الگوریتم‌های رایج یادگیری ماشین که در حوزه مدل‌سازی شدت تصادف استفاده می‌شوند عبارتند از: شبکه عصبی مصنوعی[۴] (ANN) (Abdelwahab & Abdel-Aty, 2001; Amiri et al., 2020; Zeng & Huang, 2014)، ماشین بردار پشتیبان[۵] (SVM) (Dong et al., 2015; Mokhtarimousavi et al., 2019; Zhibin Li et al. 2012)، درخت تصمیم[۶] (DT) (Abellán et al., 2013; Oña et al., 2013; P. Lu et al., 2020)، خوشه‌بندی k-MEANS (Anderson, 2009; Fiorentini & Losa, 2020; Iranitalab & Khattak, 2017; Mondal et al., 2020; J. Zhang et al., 2018)، جنگل تصادفی[۷] (RF) (Mauro et al., 2013)، و بیز ساده (Arhin & Gatiba, 2020; Budiawan et al., 2019; C. Chen et al., 2016).

در تنها پیشینه یافت شده در خصوص استفاده از مدل‌های یادگیری ماشین در تحلیل شدت تصادفات دوخودرویی، جی و لوینسون (Ji & Levinson, 2020) در ایالات متحده، برخی از مدل‌های محبوب یادگیری ماشین و برخی از تکنیک‌های ترکیبی مانند K نزدیکترین همسایگی[۸] (KNN)، SVM، مدل خطی تعمیم‌یافته[۹] (GLM)، ماشین تقویت گرادیان[۱۰](GBM)، RF و AdaBoost را برای پیش‌بینی آسیب‌های سرنشینان در تصادفات بین دو خودرو تحلیل کردند. نتایج نشان داد که مدل‌های یادگیری ماشین می‌تواند عملکرد بهتری نسبت به مدل‌های آماری داشته باشد. مرور مطالعات پیشین نشان می‌دهد که بخش عمده پژوهش‌های مرتبط با شدت تصادفات دوخودرویی بر پایه مدل‌های آماری کلاسیک انجام شده و مطالعات محدودی از روش‌های یادگیری ماشین همراه با تکنیک‌های تفسیرپذیر استفاده کرده‌اند. همچنین در بسیاری از پژوهش‌ها، اثر هم‌زمان ویژگی‌های خودرو، شرایط محیطی، عوامل زمانی و خصوصیات راه به‌صورت جامع بررسی نشده است. از این‌رو، پژوهش حاضر تلاش می‌کند با بهره‌گیری از مدل جنگل تصادفی و روش SHAP، ضمن افزایش دقت پیش‌بینی، تفسیر روشن‌تری از نقش متغیرهای اثرگذار بر شدت تصادفات ارائه نماید.

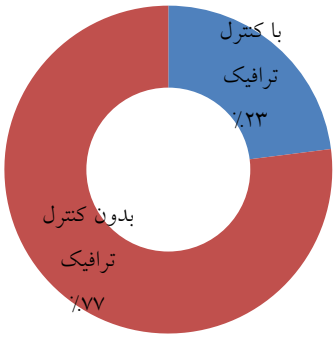
۳- داده‌ها

در این پژوهش، از داده‌های تصادفات ثبت‌شده توسط پلیس راهور شهرستان شهرضا طی سال‌های ۱۴۰۱ تا ۱۴۰۴ استفاده شد. داده‌ها شامل اطلاعات مربوط به ویژگی‌های تصادف، شرایط راه و محیط، مشخصات خودروها، ویژگی‌های زمانی و خصوصیات رانندگان بود که در قالب فایل‌های خام اکسل استخراج گردید. در مرحله پیش‌پردازش، تصادفات غیرمرتبط شامل تصادفات تک‌خودرویی، چندخودرویی، عابرپیاده و موتورسیکلت حذف شدند تا تنها تصادفات دوخودرویی مورد بررسی قرار گیرند. در نهایت، ۵۸۶ فقره تصادف دوخودرویی برای تحلیل نهایی انتخاب شد که توصیف آماری آن در شکل (۱) ارائه شده است.

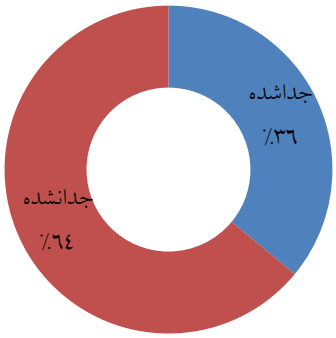


توزیع تصادفات بر اساس شدت

توزیع تصادفات بر اساس نوع برخورد

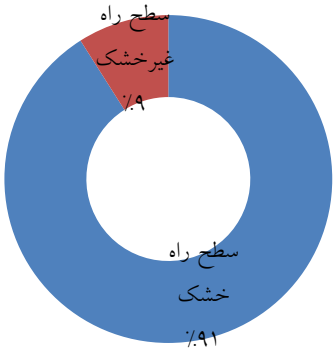


توزیع تصادفات بر اساس نوع کنترل ترافیک تقطع

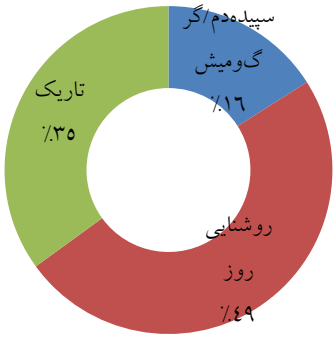


توزیع تصادفات بر اساس هندسه مسیر

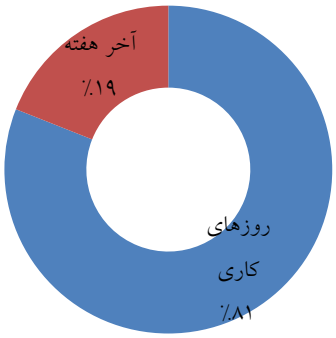
شکل ۱. توصیف آماری متغیرهای مورد استفاده در مدل



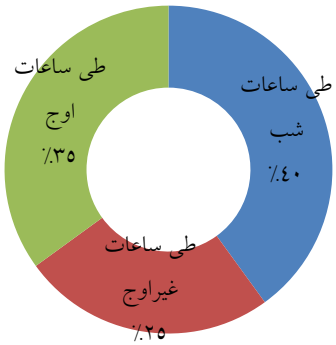
توزیع تصادفات بر اساس وضعیت سطح راه



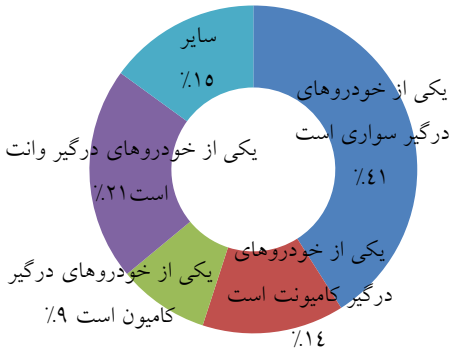
توزیع تصادفات بر اساس وضعیت روشنایی



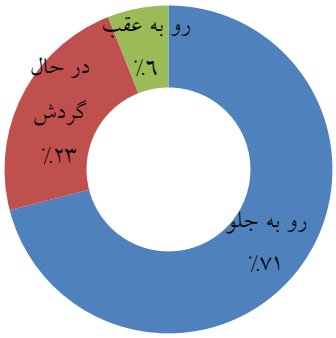
توزیع تصادفات بر اساس روزهای هفته



توزیع تصادفات بر اساس هنگامه روز

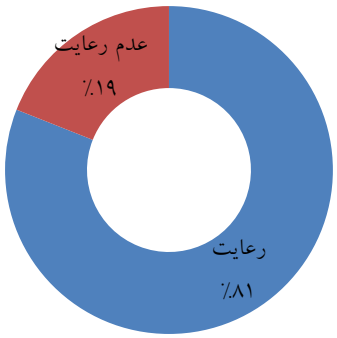


توزیع تصادفات بر اساس نوع خودرو



توزیع تصادفات بر اساس مانور خودرو

شکل ۱ (ادامه). توصیف آماری متغیرهای مورد استفاده در مدل



توزیع تصادفات بر اساس استفاده از کمربند/کلاه ایمنی

شکل ۱ (ادامه). توصیف آماری متغیرهای مورد استفاده در مدل

۳- روش تحقیق

در این پژوهش از مدل جنگل تصادفی (RF) برای تحلیل داده‌های شدت تصادفات دوخودرویی استفاده می‌شود. در پژوهش‌های حوزه ایمنی راه، این الگوریتم به عنوان ابزاری کارآمد برای پردازش کلان داده‌ها و چندمتغیره مورد توجه قرار گرفته است. یافته‌های موجود در ادبیات (yan & shen, 2022; Ji & Levinson, 2020; Hagenauer & Helbich, 2017; chen & chen, 2020; Yassin & Pooja, 2020) نشان می‌دهد که این روش از دقت و قابلیت اطمینان بالایی برخوردار بوده و گزینه‌ای مناسب برای انجام مطالعات ایمنی راه‌ها محسوب می‌شود.

۳-۱- جنگل تصادفی (RF)

مدل RF توسط بریمن (Breiman, 2001) بر اساس روش برچسب‌زدن (bagging) ساخته شد و یک روش یادگیری گروهی محبوب است که شامل مدل‌های مختلف DT با ویژگی‌های گوناگون است و نتایج مدل‌های آن‌ها را برای بهبود دقت پیش‌بینی ادغام می‌کند. بر اساس گفته‌های ون و همکاران (Wen et al., 2021)، در RF، مجموعه‌ای از طبقه‌بندهای درخت تصمیم آموزش داده شده و هر طبقه‌بند با استفاده از نمونه‌هایی که از طریق bagging به دست آمده‌اند، تشکیل می‌شود. برای هر طبقه‌بند درخت تصمیم، تنها یک زیرمجموعه تصادفی از متغیرهای مستقل برای جداسازی گره‌ها استفاده می‌شود. هر طبقه‌بند درخت تصمیم آموزش‌دیده، براساس ورودی‌ها به نتایج شدت رأی می‌دهد و طبقه‌بندی نهایی با رأی اکثریت تعیین می‌شود (Wen et al., 2021). قبل از شروع جستجو برای ویژگی‌ها و نقاط شکست بهینه، روش RF نیاز به تکمیل دو فرآیند دارد. ابتدا، تعداد مشخصی از مجموعه داده‌های آموزشی به صورت تصادفی انتخاب می‌شود. سپس هر بار یک زیرمجموعه تصادفی از درخت‌های در حال رشد توسط RF انتخاب می‌شود. نتیجه عملکرد مدل در RF با ترکیب نتایج مربوطه تمام یادگیرنده‌ها به دست می‌آید (Cai et al., 2022).

برای تعیین طبقه نهایی هر نمونه داده (i)، از توزیع احتمالی کلاس‌های شدت (k) استفاده می‌شود. در این فرمول‌بندی، p_{ik} نشان‌دهنده احتمال تعلق نمونه i به کلاس k است. طبقه انتخاب‌شده (C_i) با به حداکثر رساندن این احتمال تعیین می‌گردد (معادله ۱). به بیان دیگر، کلاسی به عنوان پیش‌بینی نهایی در نظر گرفته می‌شود که بیشترین وزن

احتمانی را در خروجی مدل RF تصادفی دارا باشد (Champahom et al., 2024).

$$C_i = argmax_k p_{ik} \quad (۱)$$

در فرآیند توسعه مدل RF، تعداد درخت‌ها برابر با ۲۰۰ در نظر گرفته شد و فرآیند آموزش با استفاده از معیار Gini برای تقسیم گره‌ها انجام گرفت. همچنین به منظور جلوگیری از بیش‌برازش و افزایش قابلیت تعمیم مدل، داده‌ها به دو بخش آموزش (۸۰ درصد) و آزمون (۲۰ درصد) تقسیم شدند. تنظیم پارامترهای مدل نیز با استفاده از جستجوی شبکه‌ای انجام شد تا بهترین ترکیب پارامترها انتخاب گردد.

۳-۲- انتخاب ویژگی

انتخاب ویژگی به‌عنوان یک استراتژی کلیدی در پیش‌پردازش داده‌ها، به‌ویژه در مجموعه‌داده‌های با ابعاد بالا، نقش مؤثری در بهبود فرآیند مدلسازی یادگیری ماشین ایفا می‌کند. اهداف اصلی این فرآیند شامل ساده‌سازی و افزایش قابلیت‌تفسیر مدل‌ها، ارتقای عملکرد الگوریتم‌ها و تهیه داده‌های تمیز و ساختاریافته است.

یکی از چالش‌های اصلی در داده‌های با ابعاد بالا، پدیده‌ای به نام «Curse of Dimensionality» یا به عبارتی «پیچیدگی ناشی از تعداد زیاد ویژگی‌ها» است. این پدیده منجر به پراکندگی داده‌ها در فضای چندبعدی شده و کارایی الگوریتم‌های طراحی‌شده برای فضاهای کم‌بعد را کاهش می‌دهد (Hastie et al., 2005). علاوه بر این، وجود تعداد زیادی ویژگی می‌تواند منجر به بیش‌برازش مدل شود که عملکرد آن را بر روی داده‌های مشاهده‌نشده تضعیف می‌کند. همچنین، داده‌های با ابعاد بالا هزینه‌های محاسباتی و نیازهای ذخیره‌سازی را به‌طور قابل‌توجهی افزایش می‌دهند. برای حل این مشکلات، کاهش ابعاد از طریق دو روش «استخراج ویژگی» و «انتخاب ویژگی» انجام می‌شود. در استخراج ویژگی، داده‌های اصلی به یک فضای جدید با ابعاد پایین‌تر نگاشت می‌شوند که معمولاً ترکیبی خطی یا غیرخطی از ویژگی‌های اولیه است. در مقابل، در انتخاب ویژگی، مستقیماً زیرمجموعه‌ای از ویژگی‌های مرتبط و مؤثر برای ساخت مدل انتخاب می‌شود (Guyon & Elisseeff, 2003; Liu & Motoda, 2007).

اگرچه هر دو روش مزایایی نظیر بهبود عملکرد، کاهش هزینه‌های محاسباتی و افزایش تعمیم‌پذیری مدل را دارند، با این حال معمولاً انتخاب ویژگی به دلیل حفظ معنای فیزیکی ویژگی‌های اصلی و افزایش قابلیت‌تفسیر، ترجیح داده می‌شود.

داده‌های واقعی اغلب حاوی ویژگی‌های غیرمرتبط، تکراری و نویزی هستند. حذف این ویژگی‌ها از طریق انتخاب ویژگی، بدون ایجاد افت معنادار در اطلاعات یا کاهش عملکرد یادگیری، هزینه‌های ذخیره‌سازی و محاسباتی را کاهش می‌دهد و دقت برآورد را بهبود می‌بخشد (Li et al., 2020).

در پژوهش حاضر، پیش از اعمال الگوریتم RF به‌عنوان طبقه‌بند، از RF به دلیل توانایی بالای آن در مدیریت داده‌های با ابعاد بالا و جلوگیری از بیش‌برازش، برای رتبه‌بندی ویژگی‌ها بر اساس اهمیت آن‌ها استفاده شد. این الگوریتم با ایجاد مجموعه‌ای از درختان تصمیم و ترکیب نتایج آن‌ها از طریق رأی‌گیری اکثریت، دقت و پایداری مدل را افزایش می‌دهد (Hossain, 2011). بدین ترتیب، ویژگی‌های منتخب با بالاترین امتیاز برای ساخت مدل نهایی انتخاب می‌شوند.

در این پژوهش، مرحله انتخاب ویژگی با استفاده از RF بوسیله بسته scikit-learn در برنامه پایتون انجام می‌شود.

۳-۳- معیارهای ارزیابی مدل

به منظور ارزیابی جامع عملکرد مدل‌ها، از معیارهای Recall، Precision، Accuracy و F1-score استفاده شد (جدول ۱). این معیارها توانایی مدل در تشخیص صحیح تصادفات فوتی و غیرفوتی را از جنبه‌های مختلف مورد سنجش قرار می‌دهند.

جدول ۱. معیارهای ارزیابی عملکرد مدل (Sunkpho et al., 2025)

$Accuracy = \frac{\text{تعداد دسته صحیح}}{\text{تعداد کل دسته ها}}$	(۱)
$Precision = \frac{\text{مثبت صحیح}}{\text{مثبت غلط} + \text{مثبت صحیح}}$	(۲)
$Recall = \frac{\text{مثبت صحیح}}{\text{منفی غلط} + \text{مثبت صحیح}}$	(۳)
$F1\ score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$	(٤)

٤-۳- تفسیر SHAP

با وجود اینکه یادگیری ماشین برای تولید تخمین‌های بسیار دقیق طراحی شده است، ارزیابی اثر متغیرهای توضیح‌دهنده بر خروجی آن چالش‌برانگیز است. روش SHAP که توسط لندبرگ و لی (Lundberg & Lee, 2017) توسعه یافته است، برای مشخص کردن اهمیت عوامل و تعیین نحوه تأثیر آن‌ها مورد استفاده قرار می‌گیرد. روش SHAP برای تشریح خروجی مدل پیش‌بینی، از نظریه بازی‌ها استفاده می‌کند. مقدار شاپلی (معادله ۹) را می‌توان با فرمول زیر تعیین کرد.

$$\varphi_t = \sum_{S \subseteq F-i} \frac{|S|!\,(|F|-|S|-1)!}{|F|!}\left[f_{SU\{i\}}\left(x_{SU\{i\}}\right) - f_s(x_s)\right] \quad (۹)$$

که در آن $|F|$ تعداد کل متغیرهای توضیح‌دهنده است، S نشان‌دهنده هر زیرمجموعه‌ای از متغیرهای توضیح‌دهنده است که شامل متغیر iام نمی‌شود و $|S|$ اندازه آن زیرمجموعه است. $f_{SU\{i\}}\left(x_{SU\{i\}}\right)$ نشان‌دهنده مدل آموزش‌دیده با i است و $f_S(x_s)$ مدل آموزش‌دیده بدون i است.

٤- نتایج

فرآیند پردازش داده‌ها و استخراج نتایج نهایی، طی مراحل زیر در نرم‌افزار پایتون انجام پذیرفت.

٤-۱- پیش‌پردازش داده‌ها

در گام نخست، برای آماده‌سازی داده‌ها، تمامی متغیرهای متنی با بهره‌گیری از الگوریتم LabelEncoder به مقادیر عددی کدگذاری شدند. پس از آن، ستون‌های فاقد کاربرد تحلیلی، از جمله «کد منحصربه‌فرد تصادف»، حذف گردیدند. متغیر هدف، یعنی «شدت تصادف»، نیز به فرمت عددی تبدیل شد و در نهایت، مجموعه داده به دو بخش مستقل شامل ویژگی‌ها (X) و برچسب‌ها (y) تفکیک گردید. با توجه به عدم تعادل در توزیع کلاس‌ها، از تکنیک SMOTE جهت ایجاد تعادل استفاده شد. بدین منظور، با تولید نمونه‌های مصنوعی، فراوانی کلاس‌های کم‌حجم افزایش یافت تا تعادل کلاس‌ها برقرار شود. داده‌های حاصل، با نسبت ۸۰ به ۲۰ به ترتیب به مجموعه‌های آموزش و آزمون تقسیم شدند. همچنین، در فرآیند آموزش

مدل‌ها، از وزن‌دهی متوازن برای کلاس‌ها استفاده گردید تا تأثیر نامتعادل بودن داده‌ها کاهش یابد.

٤-٢- انتخاب ویژگی با استفاده از RF

برای دستیابی به اهداف مشخص یا توسعه مدلی با عملکرد پیش‌بینی عالی، باید ویژگی‌های بهینه‌ای از داده‌های خام انتخاب شوند. در این پژوهش، از مدل RF برای کشف ویژگی‌های مؤثر استفاده شد، زیرا این مدل می‌تواند ارتباط ویژگی‌ها را ارزیابی کرده و گروهی از شاخص‌های مرتبط را با تفسیرپذیری بهبودیافته شناسایی کند (Li et al., 2020). مدل RF انواع مختلفی از ویژگی‌ها شامل عوامل مرتبط با خودرو، ویژگی‌های راننده، شرایط راه و محیط، ویژگی‌های تصادف و ویژگی‌های زمانی را از مجموعه ویژگی‌های خام استخراج نمود. اگرچه مدل RF برای غربال اولیه ویژگی‌ها استفاده شد، اما به دلیل ناسازگاری‌های شناخته‌شده در معیارهای سنتی با داده‌های نامتوازن یا دارای همبستگی مانند Gini Importance، تفسیر نهایی و شناسایی ویژگی‌های تاثیرگذار در این بخش، صرفا براساس مقادیر SHAP در ادامه و در زیربخش (٤-٤) ارائه می‌شود که قابلیت اطمینان بالاتری در تبیین مشارکت واقعی هر ویژگی دارد.

٤-٣- ارزیابی مدل

برای اعتبارسنجی بیشتر قدرت مدل طبقه‌بندی RF، دو مدل پایه‌ای شامل رگرسیون لجستیک[11] (LR) و درخت تصمیم (DT) مورد ارزیابی و مقایسه قرار گرفتند.

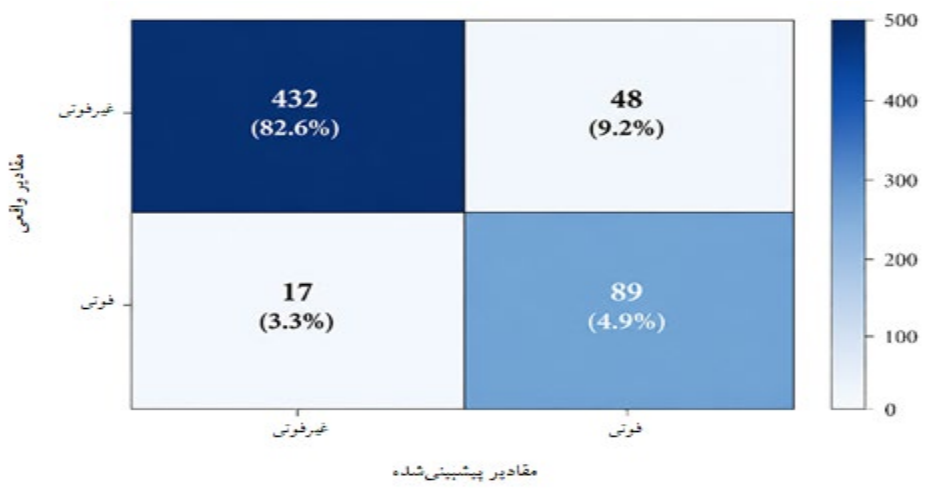
جدول (۲) معیارهای عملکرد این مدل‌ها را ارائه نموده است. نتایج نشان می‌دهد که RF در تمام شرایط، عملکرد بهتری نسبت به مدل‌های پایه داشته است که انتخاب آن به عنوان مدل اصلی در این پژوهش را تأیید می‌کند.

علاوه بر عملکرد قوی، این مدل برای یکپارچه‌سازی با روش SHAP بسیار مناسب است؛ چرا که این روش بینش‌های تفسیرپذیر و خاص هر سناریو را درباره اهمیت ویژگی‌ها در شرایط محیطی و زمانی متفاوت ارائه می‌دهد.

اگرچه مقدار دقت کلی مدل‌ها نسبت به سایر معیارها پایین‌تر است، اما این موضوع ناشی از نامتوازن‌بودن داده‌های شدت تصادف و دشواری پیشبینی دقیق کلاس‌های اقلیت است. با این وجود، مقادیر بالای recall و f1-score نشان می‌دهد که مدل RF توانایی مناسبی در شناسایی تصادفات فوتی دارد. همچنین مقدار AUC برابر با ۰/۹۲ مدل RF نشان می‌دهد که این مدل دارای توان تفکیک بسیار مطلوبی در تشخیص تصادفات فوتی و غیرفوتی است. این مقدار بیانگر عملکرد برتر مدل RF نسبت به مدل‌های LR و DT در پیشبینی شدت تصادفات دوخودرویی می‌باشد. همچنین شکل (۲) ماتریس سردرگمی عملکرد مدل جنگل تصادفی را در طبقه‌بندی شدت تصادفات دوخودرویی به دو کلاس «فوتی» و «غیر فوتی» نشان می‌دهد. نتایج بیانگر آن است که مدل RF توانسته است بخش عمده‌ای از نمونه‌ها را با دقت مناسبی طبقه‌بندی نماید. مطابق نتایج، از مجموع تصادفات غیر فوتی واقعی، تعداد ٤۳۲ نمونه به‌درستی به‌عنوان غیر فوتی پیش‌بینی شده‌اند، در حالی که تنها ٤۸ نمونه غیر فوتی به اشتباه در کلاس فوتی قرار گرفته‌اند. همچنین، از میان تصادفات فوتی واقعی، تعداد ۸۹ نمونه به‌درستی شناسایی شده و تنها ۱۷ نمونه به اشتباه به‌عنوان غیر فوتی طبقه‌بندی شده‌اند. تعداد پایین خطاهای طبقه‌بندی، به‌ویژه در شناسایی تصادفات فوتی، نشان‌دهنده توانایی مناسب مدل در تفکیک شدت تصادفات است. این موضوع از اهمیت بالایی برخوردار است؛ زیرا در مطالعات ایمنی ترافیک، شناسایی صحیح تصادفات شدید و فوتی نقش مهمی در تصمیم‌گیری‌های مدیریتی و برنامه‌ریزی اقدامات پیشگیرانه دارد. در مجموع، نتایج ماتریس سردرگمی و شاخص‌های ارزیابی تأیید می‌کند که مدل جنگل تصادفی از قابلیت بالایی در تحلیل شدت تصادفات دوخودرویی برخوردار بوده و می‌تواند به‌عنوان ابزاری کارآمد در مطالعات ایمنی ترافیک شهری مورد استفاده قرار گیرد.

جدول ۲. مقایسه عملکرد مدل RF با دو مدل پایه LR و DT در پیشبینی شدت تصادف

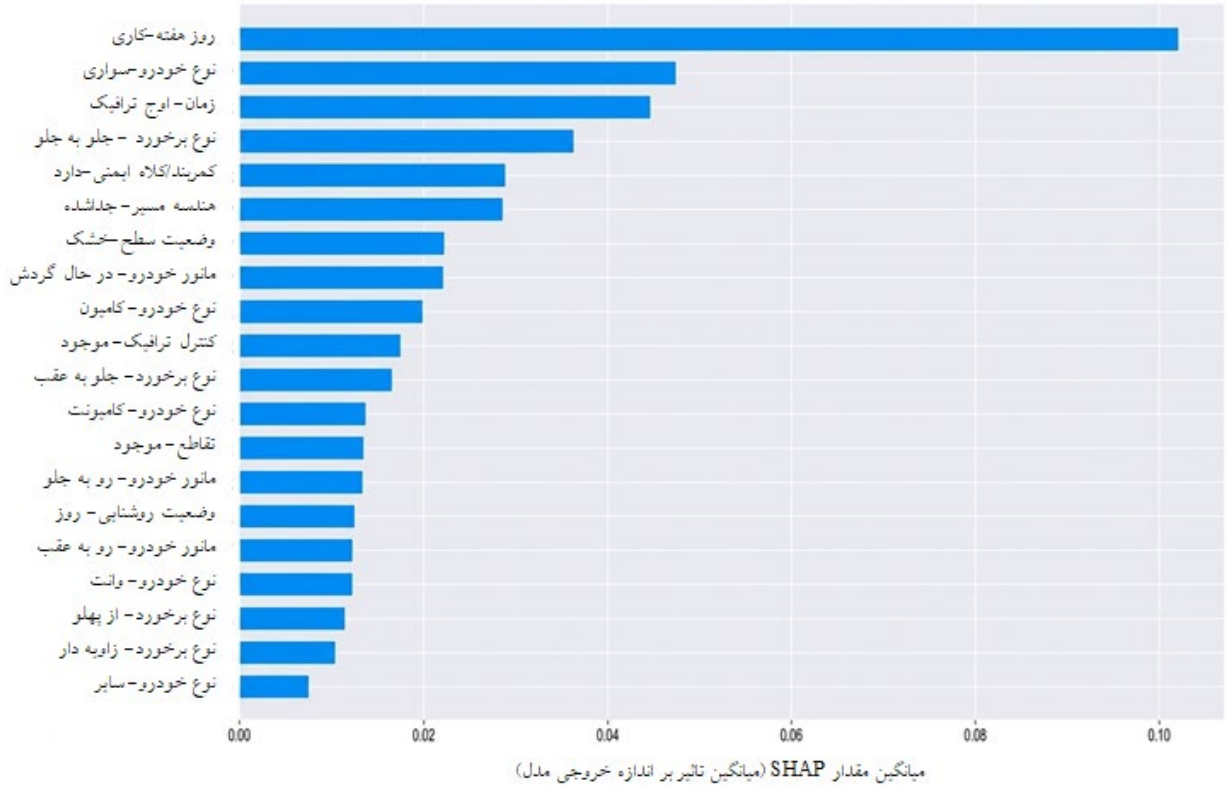
مدل	accuracy	precision	recall	f1-score	AUC
جنگل تصادفی (RF)	۰/۸۹	۰/۹۱	۰/۹۰	۰/۹۰	۰/۹۲
رگرسیون لجستیک (LR)	۰/۸۲	۰/۸٤	۰/۸۳	۰/۸۳	۰/۸٤
درخت تصمیم (DT)	۰/۷۹	۰/۸۱	۰/۸۰	۰/۸۰	۰/۸۱



شکل ۲. ماتریس سردرگمی مدل RF

٤-٤- تفسیر مدل با استفاده از روش‌شناسی SHAP

این پژوهش از روش SHAP برای تعیین میزان مشارکت ویژگی‌ها در پیش‌بینی شدت تصادفات دوخودرویی استفاده نموده و به این ترتیب عوامل تأثیرگذار را شناسایی کرده است. روش SHAP علاوه بر تعیین میزان اهمیت متغیرها، امکان تفسیر جهت و شدت اثر هر ویژگی بر خروجی مدل را نیز فراهم می‌سازد. به همین دلیل، این روش در سال‌های اخیر به عنوان یکی از مهمترین ابزارهای تفسیرپذیری مدل‌های یادگیری ماشین در مطالعات ایمنی راه شناخته شده است. شکل (۳) اهمیت جهانی ویژگی‌های استخراج‌شده است که با میانگین‌گیری از مقادیر شپلی مطلق برای هر ویژگی به دست آمده است. این امر نشان‌دهنده مشارکت حاشیه‌ای هر ویژگی در پیش‌بینی است. به عنوان مثال، مقدار شپلی برای اولین ویژگی «روز هفته» با نمونه‌برداری از یک همبستگی که شامل ویژگی اول «روز هفته» است و یک فرم همبستگی با حذف آن ویژگی، تعیین می‌شود. تفاوت بین مقادیر مربوطه این دو همبستگی، به عنوان مشارکت حاشیه‌ای اولین ویژگی «روز هفته» شناخته می‌شود. این بدان معناست که اولین ویژگی «روز هفته» چقدر به همبستگی متشکل از نوزده ویژگی دیگر کمک می‌کند. براساس روش SHAP، مشخص شد که «روز هفته»، «نوع خودرو»، «هنگامه روز»، «نوع برخورد» و «بستن کمربند/داشتن کلاه ایمنی» مهم‌ترین متغیرهای توضیحی هستند و مشارکت مهمی در پیش‌بینی شدت تصادفات دارند که با مطالعات قبلی (Chiou et al., 2020; Yuan et al., 2022; Lee & Li, 2014; Ji & Levinson, 2020) همخوانی دارد.



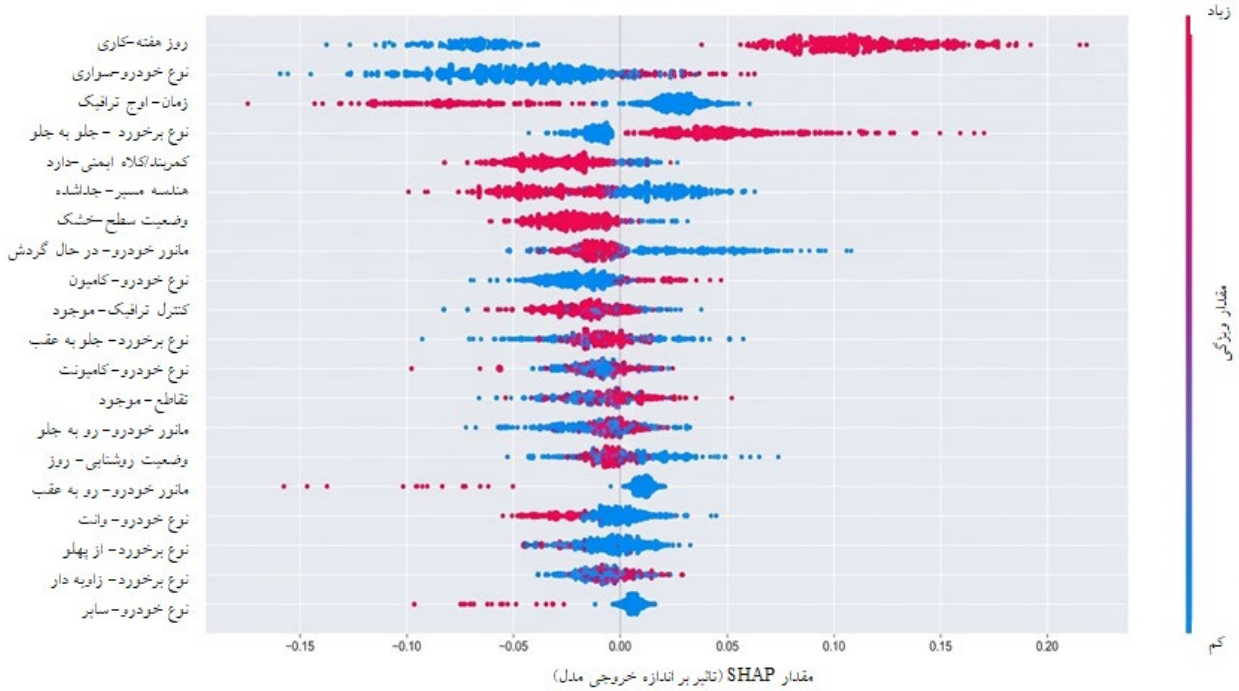
شکل ۳. عوامل مشارکت‌کننده در شدت تصادف دو خودرویی بر اساس اهمیت جهانی ویژگی‌ها بر اساس SHAP

برای شناسایی نحوه تأثیر عوامل مشارکت‌کننده بر شدت تصادفات دوخودرویی، شکل (٤) مقادیر SHAP عناصر تعیین‌کننده شدت تصادف دوخودرویی را نشان می‌دهد. باید توجه داشت که مقادیر SHAP بیشتر از صفر نشان‌دهنده اثرات مثبت بر خطر وقوع تصادف فوتی هستند، در حالی که مقادیر SHAP کمتر از صفر نشان‌دهنده پیامدهای جرحی است. این نمودار اندازه محلی، توزیع و جهت عناصر مشارکت‌کننده در تعیین اینکه آیا یک تصادف دوخودرویی فوتی خواهد بود یا جرحی را به تصویر می‌کشد. در این نمودار، هر نقطه نشان‌دهنده یک مورد تصادف منفرد است. موقعیت افقی هر نقطه مقدار SHAP (تأثیر آن بر پیش‌بینی) و رنگ آن نشان‌دهنده مقدار ویژگی است (مقادیر بالا به رنگ قرمز و مقادیر پایین به رنگ آبی). الگوهای خوشه‌بندی شده اطلاعات مهمی درباره رابطه بین ویژگی‌ها و پیش‌بینی‌ها آشکار می‌سازند:

- همبستگی مثبت: وجود نقاط قرمز (مقادیر بالا) در سمت راست و نقاط آبی (مقادیر پایین) در سمت چپ نشان می‌دهد که مقادیر بالاتر ویژگی با شدت بیشتر تصادفات مرتبط است.
- همبستگی منفی: قرارگیری نقاط قرمز در سمت چپ و نقاط آبی در سمت راست حاکی از آن است که مقادیر بالاتر، شدت تصادف را کاهش می‌دهند.
- توزیع‌های ترکیبی: پراکندگی نقاط قرمز و آبی در هر دو طرف، نشان‌دهنده اثرات پیچیده و وابسته به زمینه است که تحت تأثیر تعامل با سایر عوامل شکل می‌گیرد.
- عرض توزیع: پهنای توزیع در محور افقی نشان‌دهنده میزان تأثیر بالقوه یک ویژگی است؛ به‌طوری که پراکندگی گسترده‌تر، نشان‌دهنده تأثیر بیشتر آن ویژگی بر پیش‌بینی شدت تصادف در شرایط خاص است.

با توجه به توضیحات فوق مهمترین نتایجی که می‌توان از شکل (٤) استنباط نمود شامل موارد زیر است:

- اگر یک تصادف دوخودرویی در طول روزهای هفته رخ دهد، احتمال فوتی بودن تصادف بیشتر است. علاوه بر این، تصادفات آخر هفته به اندازه تصادفات روزهای کاری در شدت تصادفات دوخودرویی مهم نیستند.
- از متغیر نوع خودرو می‌توان استنباط کرد که اگر حداقل یکی از خودروهای درگیر در تصادف کامیون یا کامیونت باشد، احتمال وقوع تصادف فوتی افزایش می‌یابد زیرا وسایل نقلیه بزرگ‌تر انرژی برخورد بیشتری به خودروی مقابل وارد می‌کنند که منجر به تصادفات فوتی برای خودروهای مقابل می‌شود، که با تحقیقات قبلی (Chiou et al., 2020; Yuan et al., 2022; Lee & Li, 2014 and Ji & Levinson, 2020) همخوانی دارد.
- به نظر می‌رسد تصادفات دوخودرویی در ساعات غیراوج بیشتر منجر به مرگ و میر شوند. دلیل این امر می‌تواند ترافیک روان‌تر و سرعت بیشتر خودروها و متعاقبا بی‌احتیاطی بیشتر رانندگان باشد.
- احتمال فوتی بودن تصادفات دوخودرویی در مسیرهای دوطرفه بیشتر و برای مسیرهای یک‌طرفه کمتر است.
- وقتی سطح خیس/برفی یا یخ‌زده است، احتمال فوتی بودن تصادفات دوخودرویی بیشتر است و وقتی سطح خشک است، احتمال فوتی بودن کمتر است.
- وقتی کنترل ترافیک وجود ندارد، احتمال فوتی بودن تصادفات دوخودرویی بیشتر است. زیرا رانندگان احتیاط کمتری دارند و احتمال تداخلات بیشتر و مانورهای خطرناک بیشتر است.
- به نظر می‌رسد که احتمال فوتی بودن تصادفات دوخودرویی در طول شب بیشتر است.
- احتمال فوتی بودن تصادفات دوخودرویی در تقاطع‌ها بیشتر از مکان‌های غیرتقاطعی است. باید توجه داشت که در شهرستان‌های غیرکلانشهری، کنترل ترافیک تعداد زیادی از تقاطع‌ها با تابلوهای کنترل توقف است و یا حتی بدون کنترل ترافیک هستند که این به نوبه خود مانورهای پرخطر را ممکن ساخته و احتمال تصادفات با شدت بیشتر را افزایش می‌دهد.
- احتمال فوتی بودن تصادفات دوخورویی جلو به جلو و زاویه‌دار بیشتر از سایر انواع برخورد است. دلیل آن می‌تواند ضربه شدیدتر و نزدیکتر به رانندگان و سرنشینان خودروها در چنین برخوردهایی باشد.



شکل ٤. عوامل مشارکت‌کننده در شدت تصادف دو خودرو بر اساس توضیح محلی SHAP

محدودیت‌ها

باوجود نتایج ارزشمند پژوهش حاضر، برخی محدودیت‌ها نیز وجود دارد. اولا، داده‌های مورد استفاده محدود به تصادفات دوخودرویی ثبت‌شده در شهرستان شهرضا بود و ممکن است قابلیت تعمیم نتایج به سایر شهرها محدود باشد. ثانیا، برخی متغیرهای مهم مانند سرعت واقعی خودرو، میزان تخلفات رانندگی، وضعیت روحی-روانی رانندگان و حجم لحظه‌ای ترافیک در پایگاه داده موجود نبود. همچنین حجم نسبتا محدود نمونه‌ها می‌تواند بر عملکرد مدل‌های یادگیری ماشین تاثیرگذار باشد. پیشنهاد می‌شود در مطالعات آینده، از داده‌های گسترده‌تر و مدل‌های ترکیبی پیشرفته‌تر برای تحلیل شدت تصادفات دوخودرویی استفاده شود.

۵- نتیجه‌گیری

پژوهش حاضر با هدف تحلیل شدت تصادفات دوخودرویی درون‌شهری و شناسایی عوامل مؤثر بر آن، با استفاده از مدل جنگل تصادفی و روش تفسیرپذیر SHAP انجام شد. نتایج نشان داد که مدل‌های یادگیری ماشین، به‌ویژه مدل جنگل تصادفی، توانایی بالایی در پیش‌بینی شدت تصادفات دارند و می‌توانند به‌عنوان ابزارهای مؤثر در مطالعات ایمنی ترافیک مورد استفاده قرار گیرند. مقایسه عملکرد مدل‌ها نشان داد که مدل RF نسبت به مدل‌های رگرسیون لجستیک و درخت تصمیم دارای عملکرد برتر بوده و توانسته است با دقت و قابلیت تفکیک بالاتر، تصادفات فوتی و غیر فوتی را پیش‌بینی نماید. بر اساس تحلیل SHAP، متغیرهای روز هفته، نوع خودرو، زمان وقوع تصادف، مانور خودرو، هندسه راه، شرایط روشنایی، کنترل ترافیک و وضعیت سطح راه از مهم‌ترین عوامل مؤثر بر شدت تصادفات دوخودرویی شناخته شدند. نتایج بیانگر آن بود که وقوع تصادفات در ساعات شب، نبود کنترل ترافیک، راه‌های دوطرفه و حضور خودروهای سنگین با افزایش احتمال وقوع تصادفات فوتی همراه است. همچنین مشخص شد که شرایط محیطی و هندسی راه نقش مهمی در تشدید پیامدهای ناشی از تصادفات دارند. یافته‌های این پژوهش نشان می‌دهد که شدت تصادفات پدیده‌ای چندبعدی است و تنها به رفتار راننده یا ویژگی وسیله نقلیه محدود نمی‌شود، بلکه عوامل زمانی، محیطی، هندسی و مدیریتی نیز در شکل‌گیری پیامدهای شدید نقش اساسی دارند. بنابراین، ارتقای ایمنی ترافیک شهری مستلزم اتخاذ رویکردی جامع در مدیریت شبکه معابر، بهبود روشنایی راه‌ها، توسعه تجهیزات کنترل ترافیک، مدیریت سرعت در ساعات شبانه و افزایش ایمنی وسایل نقلیه آسیب‌پذیر است. یکی از مهم‌ترین نقاط قوت این پژوهش، استفاده هم‌زمان از مدل یادگیری ماشین و روش تفسیرپذیر SHAP بود که علاوه بر افزایش دقت پیش‌بینی، امکان تفسیر روشن‌تری از نحوه تأثیر متغیرها بر شدت تصادفات را فراهم ساخت. این موضوع می‌تواند به تصمیم‌گیرندگان و مدیران ایمنی کمک کند تا سیاست‌های هدفمندتر و مؤثرتری برای کاهش شدت تصادفات تدوین نمایند. با وجود نتایج ارزشمند این مطالعه، برخی محدودیت‌ها نیز وجود داشت. محدودبودن داده‌ها به یک محدوده جغرافیایی مشخص، نبود اطلاعاتی مانند سرعت واقعی خودروها، حجم لحظه‌ای ترافیک و برخی ویژگی‌های رفتاری رانندگان از جمله محدودیت‌های پژوهش محسوب می‌شود. پیشنهاد می‌شود در مطالعات آینده، از داده‌های گسترده‌تر و مدل‌های پیشرفته‌تر یادگیری ماشین و یادگیری عمیق برای تحلیل شدت تصادفات استفاده گردد.

٦- پانویس‌ها

1. Machine Learning (ML)
2. Sensitivity Analysis (LA)
3. Local Sensitivity Analysis (LSA)
4. Artificial Neural Network (ANN)
5. Support Vector Machine (SVM)
6. Decision Tree (DT)
7. Random Forest (RF)
8. K-Nearest Neighbors (KNN)
9. Generalized Linear Model (GLM)
10. Gradient Boosting Machine (GBM)
11. Logistic Regression (LR)

۷- مراجع

-Abdel-Aty, M. A., and Abdelwahab, H. T., (2004). Predicting injury severity levels in traffic crashes: A modeling comparison. *Journal of Transportation Engineering,* Vol. 130, No. 2, 204–210.

-Abellán, J., López, G., & de Oña, J. (2013). Analysis of traffic accident severity using decision rules via decision trees. *Expert Systems with Applications,* 40(15), 6047–6054.

-Amiri, A. M., Sadri, A., Nadimi, N., & Shams, M. (2020). A comparison between artificial neural network and hybrid intelligent genetic algorithm in predicting the severity of fixed object crashes among elderly drivers. *Accident Analysis and Prevention,* 138, 105468.

-Anvari, M. B., Kashani, A. T., 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, 653–662.

-Arhin, S. A., & Gatiba, A. (2020). Predicting crash injury severity at unsignalized intersections using support vector machines and naïve bayes classifiers. *Transportation Safety and Environment,* 2(2), 120– 132.

-Breiman, L. Random Forests (2001). *Machine Learning,* Vol. 45, No. 1, 5–32.

-Budiawan, W., Saptadi, S., Tjioe, C., & Phommachak, T. (2019). Traffic accident severity prediction using naive Bayes algorithm-a case study of Semarang toll road. *IOP Conference Series: Materials Science and Engineering,* 598, 012089.

-Cai, Q., M. Abdel-Aty, O. Zheng, and Y. Wu. (2022). Applying Machine Learning and Google Street View to Explore Effects of Drivers' Visual Environment on Traffic Safety. *Transportation Research Part C: Emerging Technologies,* Vol. 135.

-Champahom, T., Se, CH., Watcharamaisakul, F., Jomnonkwao, S., Karoonsoontawong, A., Ratanavaraha, V., (2024). Tree-based approaches to understanding factors influencing crash severity across roadway classes: A Thailand case study, *IATSS Research 48,* 464–476.

-Champahom, T., Se, CH., Watcharamaisakul, F., Jomnonkwao, S., Karoonsoontawong, A., Ratanavaraha, V., (2024). Tree-based approaches to understanding factors influencing crash severity across roadway classes: A Thailand case study, *IATSS Research 48,* 464–476.

-Chen, C., Zhang, G. H., Yang, J. F., Milton, C. J., & Alcántara, A. D. (2016). An explanatory analysis of driver injury severity in rear-end crashes using a decision table/Naïve Bayes (DTNB) hybrid classifier. *Accident Analysis & Prevention,* 90, 95–107.

-Chen, F., and Chen, S. (2011). Injury Severities of Truck Drivers in Single- and Multi-Vehicle Accidents on Rural Highways. *Accident*

-Hagenauer, J., and Helbich, M. (2017). A Comparative Study of Machine Learning Classifiers for Modeling Travel Mode Choice. *Expert Systems with Applications,* Vol. 78, 273–282.

-Hossain, M. (2011). Development of a Real-Time Proactive Road Safety Management System for Urban Expressways. Ph. D. Thesis, Department of Built Environment, *Tokyo Institute of Technology.*

-Iranitalab, A., and Khattak, A. (2017). Comparison of four statistical and machine learning methods for crash severity prediction. *Accident Analysis & Prevention,* Vol. 108, 27–36.

-Ji, A., and Levinson, D. (2020). Injury Severity Prediction fom Two-Vehicle Crash Mechanisms With Machine Learning and Ensemble Models. *IEEE Open Journal of Intelligent Transportation Systems,* Vol. 1, 217–226.

-Jiang, L., Xie, Y., and Ren, T. (2019). Modelling highly unbalanced crash injury severity data by ensemble methods and global sensitivity analysis. *In Procceding Transportation Research Board 98th Annual Meeting,* Washington, DC, USA, 13–17.

-Kitali, A. E., and Sando, T. (2017). A Full Bayesian Approach to Appraise the Safety Effects of Pedestrian Countdown Signals to Drivers. *Accident Analysis & Prevention,* Vol. 106, 327–335.

-Lee, C., and Li, X. (2014). Analysis of Injury Severity of Drivers Involved in Single- and Two-Vehicle Crashes on Highways in Ontario. *Accident Analysis and Prevention,* Vol. 71, 286–295.

-Li, Z., Liu, P., Wang, W., & Xu, C. (2012). Using support vector machine models for crash injury severity analysis. *Accident Analysis & Prevention,* 45, 478 486.

-Li, Z., Liu, P., Wang, W., & Xu, C. (2012). Using support vector machine models for crash injury severity analysis. *Accident Analysis & Prevention,* 45, 478 486.

Analysis & Prevention, Vol. 43, No. 5, 1677–1688.

-Chen, M.M. Chen, M.C., (2020). Modeling road accident severity with comparisons of logistic regression, decision tree and random forest, *Inf. 11*.

-Chen, M.M. Chen, M.C., (2020). Modeling road accident severity with comparisons of logistic regression, decision tree and random forest, *Inf. 11*.

-Chiou, Y. C., Fu, C. and Ke, C. Y. (2020). Modelling Two-Vehicle Crash Severity by Generalized Estimating Equations. *Accident Analysis and Prevention,* Vol. 148.

-Dong, N., Huang, H., & Zheng, L. (2015). Support vector machine in crash prediction at the level of traffic analysis zones: Assessing the spatial proximity effects. *Accident; Analysis and Prevention,* 82, 192–198.

-Duncan, C. S., Khattak, A. J. and Council, F. M. (1998). Applying the Ordered Probit Model to Injury Severity in Truck-Passenger Car Rear-End Collisions. Transportation Research Record: *Journal of the Transportation Research Board,* Vol. 1635, No. 1, 63–71.

-Fiorentini, N., & Losa, M. (2020). Handling imbalanced data in road crash severity prediction by machine learning algorithms. *Infrastructures,* 5(7), 61.

-Geedipally, S. R., and Lord, D. (2010). Investigating the Effect of Modeling Single Vehicle and Multi-Vehicle Crashes Separately on Confidence Intervals of Poisson–Gamma Models. *Accident Analysis & Prevention,* Vol. 42, No. 4, 1273–1282.

-Griffith, M. S. (1999). Safety Evaluation of Rolled-In Continuous Shoulder Rumble Strips Installed on Freeways. Transportation Research Record: *Journal of the Transportation Research Board,* 1665(1): 28–34.

-Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research,* 3:1157–82.

-Mokhtarimousavi, S., Anderson, J. C., Azizinamini, A., & Hadi, M. (2019). Improved support vector machine models for work zone crash injury severity prediction and analysis. Transportation Research Record: *Journal of the Transportation Research Board,* 2673(11), 680–692.

-Mondal, A. R., Bhuiyan, M. A. E., & Yang, F. (2020). Advancement of weather related crash prediction model using nonparametric machine learning algorithms. *SN Applied Sciences,* 2(8), 1–11.

-NHTSA, (2020), Traffic Safety Facts, A Compilation of Motor Vehicle Crash Data, Report No. DOT HS 813375, National Highway Traffic Safety Administration, U.S, *Department of Transportation.*

-NHTSA, (2020), Traffic Safety Facts, A Compilation of Motor Vehicle Crash Data, Report No. DOT HS 813375, National Highway Traffic Safety Administration, U.S, *Department of Transportation.*

-Pasupathy, Ivan, R. J. N., and Ossen, P. J. (2000). Single and Multi-Vehicle Crash Prediction Models for Two-Lane Roadways. NEUTC UCNR9-8, *Final Report,* Durham, NH.

-Pradhan, B., and Sameen, M. I., (2020). Predicting Injury Severity of Road Traffic Accidents Using a Hybrid Extreme Gradient Boosting and Deep Neural Network Approach. In Advances in Science, Technology and Innovation, *Springer Nature,* 119–127.

-Qin, X., Ivan J. N., & Ravishanker, N. (2004). Selecting Exposure Measures in Crash Rate Prediction for Two-Lane Highway Segments. *Accident Analysis & Prevention,* Vol. 36, No. 2, 183–191.

-Sameen, M. I., & Pradhan, B. (2017). Severity Prediction of Traffic Accidents with Recurrent Neural Networks. *Applied Sciences,* 7(6), 476.

-Sarkar, S., Pramanik, A., Maiti, J., & Reniers, G. (2020). Predicting and analyzing injury severity: A machine learning-based approach using class-imbalanced proactive and reactive data. *Safety Science,* 125, 104616.

-Li, Z., Wang, H. X., Zhang, Y. W., and Zhao, X. H. (2020). Random Forest–Based Feature Selection and Detection Method for Drunk Driving Recognition. *International Journal of Distributed Sensor Networks,* Vol. 16, No. 2.

-Lombardi, D. A., Horrey, W. J., & Courtney, T. K. (2017). Age-related differences in fatal intersection crashes in the United States. *Accident Analysis & Prevention,* 99, 20–29.

-Lord, D., Washington, S. P., and Ivan, J. N. (2005). Poisson, Poisson-Gamma and Zero-Inflated Regression Models of Motor Vehicle Crashes: Balancing Statistical Fit and Theory. *Accident Analysis & Prevention,* Vol. 37, No. 1, 35–46.

-Lu, P., Zheng, Z., Ren, Y., Zhou, X., Keramati, A., Tolliver, D., & Huang, Y. (2020). A gradient boosting crash prediction approach for highway-rail grade crossing crash analysis. *Journal of Advanced Transportation,* 1–10.

-Lundberg, S. M., and Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions.

-Ma, J., and Kockelman, K. M. (2006). Bayesian Multivariate Poisson Regression for Models of Injury Count, by Severity. Transportation Research Record: *Journal of the Transportation Research Board,* 1950(1): 24–34.

-Ma, X., Chen, S., & Chen, F. (2016). Correlated Random-Effects Bivariate Poisson Lognormal Model to Study Single-Vehicle and Multivehicle Crashes. *Journal of Transportation Engineering,* 04016049.

-Martensen, H., and Dupont, E. (2013). Comparing Single Vehicle and Multivehicle Fatal Road Crashes: A Joint Analysis of Road Conditions, Time Variables and Driver Characteristics. *Accident Analysis & Prevention,* Vol. 60, 466–471.

-Mauro, R., De Luca, M., & Dell'Acqua, G. (2013). Using a K-means clustering algorithm to examine patterns of vehicle crashes in before-after analysis. *Modern Applied Science,* 7(10), 11.

Trucks Considering Vehicle Types. *Asian Transport Studies,* Vol. 5, Issue. 4, 720-733.

-Yassin, S. S., and Pooja (2020). Road Accident Prediction and Model Interpretation Using a Hybrid K-Means and Random Forest Algorithm Approach. *SN Applied Sciences,* Vol. 2, No. 9, 1–13.

-Yu, R., and Abdel-Aty, M. (2014). Analyzing crash injury severity for a mountainous freeway incorporating real-time traffic and weather data. *Safety Science,* Vol. 63, 50–56.

-Yuan, Q., Lu, M., Theofilatos, A., & Li, Y. B. (2017). Investigation on occupant injury severity in rear-end crashes involving trucks as the front vehicle in Beijing area, China. *Chinese Journal of Traumatology,* 20(1), 20–26.

-Yuan, R., Gan, J., Peng, Z., and Xiang, Q. (2022). Injury Severity Analysis of Two Vehicle Crashes at Unsignalized Intersections Using Mixed Logit Models. *International Journal of Injury Control and Safety Promotion.*

-Zeng, Q., & Huang, H. (2014). A stable and optimized neural network model for crash injury severity prediction. *Accident Analysis and Prevention,* 73, 351 358.

-Zeng, Q., Wen, H., and Huang, H. (2016). The Interactive Effect on Injury Severity of Driver-Vehicle Units in Two-Vehicle Crashes. *Journal of Safety Research,* Vol. 59, 105–111.

-Zhang, J., Li, Z., Pu, Z., & Xu, C. (2018). Comparing prediction performance for crash injury severity among various machine learning and statistical methods. *IEEE Access,* 6, 60079–60087.

-Zhang, J., Li, Z., Pu, Z., & Xu, C. (2018). Comparing prediction performance for crash injury severity among various machine learning and statistical methods. *IEEE Access,* 6, 60079–60087.

-Shao, X., Ma, X., Chen, F., Song, M., Pan, X., & You, K. (2020). A Random Parameters Ordered Probit Analysis of Injury Severity in Truck Involved Rear-End Collisions. *International Journal of Environmental Research and Public Health,* 17(2), 395.

-Sum, S., Se, Ch., Champahom, T., Jomnonkwao, S., Sinha, S., Ratanavaraha, V., (2025). A random forest and SHAP-based analysis of motorcycle crash severity in Thailand: Urban-rural and day-night perspectives, *Transportation Engineering,* 21, 100369.

-Tang, J., Liang, J., Han, C., Li, Z., & Huang, H. (2019). Crash injury severity analysis using a two-layer Stacking framework. *Accident Analysis & Prevention,* Vol. 122, 226-238.

-Wen, X., Xie, Y., Jiang, L., Pu, Z., & Ge, T. (2021). Applications of Machine Learning Methods in Traffic Crash Severity Modelling: Current Status and Future Directions. *Transport Reviews,* Vol. 41, No. 6, 855–879.

-World Health Organization (WHO), (2009). Global status report on road safety: Time for action.

-World Health Organization (WHO), (2018). World report on road traffic injury prevention.

-World Health Organization (WHO), (2022). Road Traffic Injuries.

-Yan, M., Shen, Y., (2022). Traffic accident severity prediction based on random forest, *Sustain 14.*

-Yan, M., Shen, Y., (2022). Traffic accident severity prediction based on random forest, *Sustain 14.*

-Yang, J., Ren, P., and Ando, R. (2019). Examining Drivers' Injury Severity of Two Vehicle Crashes between Passenger Cars and

Analysis of the Severity of Urban Two-Vehicle Crashes Using the Random Forest (RF) Algorithm and SHAP Analysis

***Hamed Sefi**, Department of Civil Engineering, Payame Noor University (PNU), Tehran, Iran.*

Mohammad Koohi, M.Sc., Grad., Department of Civil Engineering, Payame Noor University (PNU), Tehran, Iran.

Shahin Shabani, Department of Civil Engineering, Payame Noor University (PNU), Tehran, Iran.

E-mail: hamedsaify@pnu.ac.ir

Received: March 2026- Accepted: August 2026

ABSTRACT

Urban two-vehicle crashes constitute one of the most significant types of traffic crashes, accounting for a substantial proportion of human casualties and economic losses. Identifying the factors influencing the severity of these crashes can play a crucial role in urban traffic management and safety planning. This study aims to analyze the factors affecting the severity of urban two-vehicle crashes using machine learning models and interpretable methods. The dataset comprised 586 recorded two-vehicle crashes in Shahrud County between 2022 and 2025, extracted from the Traffic Police database. During the preprocessing stage, the data were encoded and balanced, followed by the development of the Random Forest (RF) model as the primary classifier. The performance of the RF model was compared with two baseline models: Logistic Regression (LR) and Decision Tree (DT). Evaluation results indicated that the RF model achieved the best performance in predicting collision severity, with an accuracy of 89%, an AUC of 0.92, and higher Precision, Recall, and F1-score values compared to the other models. To enhance the interpretability of the results, the SHAP method was employed. The findings revealed that the day of the week, vehicle type, time of occurrence, vehicle maneuver, road geometry, lighting conditions, and traffic control are the most significant factors affecting collision severity. Furthermore, the analysis indicated that crashes occurring at night, the absence of traffic control, two-way roads, and the presence of heavy vehicles increase the likelihood of fatal accidents. The results demonstrate that combining machine learning models with interpretable methods can serve as an effective tool for analyzing collision severity and formulating strategies to enhance urban traffic safety.

Keywords: Two-Vehicle Crashes, Severity Analysis, Random Forest (RF), SHAP Analysis