استفاده از روش‌های یادگیری ماشین در پیش‌بینی شدت تصادفات جاده‌ای (مطالعه موردی: استان زنجان)

مقاله علمی – پژوهشی

علی توکلی کاشانی*، دانشیار، دانشکده مهندسی عمران، دانشگاه علم و صنعت ایران، تهران، ایران
علی مدقالچی، دانشکده عمران، واحد قزوین، دانشگاه آزاد اسلامی، قزوین، ایران
اعظم محمدی، دانش آموخته کارشناسی، دانشگاه زنجان، زنجان، ایران
محمد جزونقی، دانش آموخته کارشناسی ارشد، دانشگاه بین المللی امام خمینی (ره)، قزوین، ایران
*پست الکترونیکی نویسنده مسئول:alitavakoli@iust.ac.ir

دریافت: ۱۴۰۲/۰۸/۲۸- پذیرش: ۱۴۰۳/۰۱/۲۵
صفحه ۱۱۸-۱۰۷

چکیده

بیشترین سهم تصادفات در جهان مربوط به کشورهای با درآمد متوسط و پایین است. از طرفی، آمار مجروحین و فوتی‌ها در تصادفات ترافیکی ایران رو به افزایش است؛ که بیانگر لزوم توجه و تمرکز بیش از پیش بر تحلیل تصادفات ترافیکی و یافتن علل موثر بر شدت تصادفات به منظور ارتقاء ایمنی راه‌های کشور و کاهش پیامدهای ناشی از آن می‌باشد. در مطالعه‌ی حاضر سعی شده‌است مهم‌ترین عوامل موثر بر شدت تصادفات برون‌شهری استان زنجان با دو مدل ماشین بردار پشتیبان و درخت تصمیم شناسایی شوند. بدین منظور از ۲۵ هزار داده‌های تصادفات استان طی ۹ سال اخیر استفاده شده‌است. پس از فرآیند پاکسازی داده‌ها، مدل‌ها در محیط برنامه‌نویسی پایتون توسعه داده شدند. نتایج تحلیل‌ها نشان داد در مدل ماشین بردار پشتیبان، نحوه تصادف، نوع وسیله و کیلومتر وقوع تصادف، و در مدل درخت تصمیم نحوه تصادف، نوع وسیله مقصر و کیلومتر وقوع تصادف به ترتیب سه متغیر دارای اهمیت برای پیش‌بینی شدت این تصادفات هستند. هم‌چنین بطور کلی درخت تصمیم قدرت پیش‌بینی بیش‌تری دارد و دقت این مدل در جراحات شدیدتر بیش‌تر از ماشین بردار پشتیبان می‌باشد.

واژه‌های کلیدی: شدت تصادف، درخت تصمیم، ماشین بردار پشتیبان، تصادفات جاده‌ای، ایمنی ترافیک

۱-مقدمه

تصادفات ترافیکی هرساله جان ۱/۳ میلیون نفر را در جهان می‌گیرد که ۹۳ درصد از آن سهم کشورهای با درآمد متوسط و کم است و هزینه‌ای معادل با ۳ درصد تولید ناخالص داخلی کشورها را دربر دارد. هم‌چنین برآورد شده است که ۲۰ تا ۵۰ میلیون نفر در جهان از جراحت و نقص عضو ناشی از تصادفات رنج می‌برند (World Health Organization, 2021). در ایران طبق گزارش‌های سازمان پزشکی قانونی، طی ده سال اخیر آمار مصدومین و متوفیان تصادفات ترافیکی تا پیش از سال ۱۳۹۹ رشد منفی داشته است اما پس از آن تا کنون بطور میانگین تعداد متوفیان ۱۱/۷ درصد و مصدومین ۱۷/۸ درصد افزایش یافته‌است (Iranian Legal Medicine Organization, 1401). این آمار نشان‌دهنده‌ی لزوم توجه و تمرکز هرچه بیش‌تر بر تحلیل تصادفات ترافیکی و یافتن علل موثر بر شدت تصادفات برای ارتقاء ایمنی راه‌ها و کاهش پیامدهای ناشی از آن می‌باشد. مدل‌سازی شدت تصادفات صورت گرفته بصورت کلی در دو چارچوب مدل‌سازی آماری و مدل‌سازی یادگیری ماشین تقسیم می‌شوند. مدل‌های آماری بعلت سهولت در قابلیت تفسیر نتایج در بسیاری از مطالعات شدت

تصادفات استفاده شده‌است. برای مثال مدل‌های رگرسیونی رایج‌ترین شیوه برای تعیین عوامل موثر بر ریسک شدت تصادفات بوده‌است (Abrari Vajari et al., 2020) (Kaplan and Prato, 2012) و برخی نیز از سایر روش‌های آماری استفاده کرده‌اند (Yuan et al., 2021). اما مدل‌های آماری نیازمند برخی فرضیات درباره‌ی توزیع احتمال داده‌ها و روابط پیشین بین متغیرهای وابسته و مستقل است و درصورتیکه این فرضیات برآورده نشوند برآوردهای انجام شده و تفسیرهای مربوطه نادرست خواهند بود. از طرف دیگر، مدل‌های یادگیری ماشین بعلت عدم نیاز به فرضیات پیشین مورد توجه بسیاری از پژوهش‌های قرار گرفته‌است(Santos, Dias and Amado, 2022). ونگ و کیم به تحلیل عوامل موثر بر شدت تصادفات در ایالت مریلند امریکا پرداخت. بدین منظور دو مدل لوجیت چندگانه و جنگل تصادفی را مقایسه کردند. نتایج نشان داد الگوریتم جنگل تصادفی دقت پیش‌بینی بیش‌تری دارد (Wang and Kim, 2019). لابیب و همکاران از چهار مدل درخت تصمیم، نزدیک‌ترین همسایگی، بیز خالص و آدابوست برای پیش‌بینی شدت تصادفات بنگلادش استفاده کرد (Labib et al., 2019). المملوک و همکاران شدت تصادفات را با استفاده از روش‌های رگرسیون لوجستیک، بیز خالص، آدابوست، و جنگل تصادفی مدل‌سازی کردند. نتایج نشان داد دقت مدل جنگل تصادفی بیش از بقیه بود (AlMamlook et al., 2019). آرهین و گتیبا برای پیش‌بینی شدت تصادفات تقاطعات بدون چراغ از دو مدل ماشین بردار پشتیبان با درنظر گرفتن کرنل‌های مختلف و بیز خالص قوسی در واشنگتن امریکا استفاده نمودند. آنان نتیجه گرفتند ماشین بردار پشتیبان با کرنل شعاعی بیش‌ترین دقت را بدست می‌دهد و مدل بیز دقتی در حد نصف آن را دارد (Arhin and Gatiba, 2020). در ایران، حسین‌زاده و همکاران شدت جراحت رانندگان کامیون در هشت شهر ایران را بوسیله‌ی مدل‌های ماشین بردار پشتیبان و لوجیت مدل‌سازی کردند. مطابق نتایج، ماشین بردار پشتیبان دقت پیش‌بینی بیش‌تری را ارائه داد (Hosseinzadeh, Moeinaddini and Ghasemzadeh, 2021). المغری و همکاران از پنج مدل درخت تصمیم، بیز، جنگل تصادفی و ماشین بردار پشتیبان و یک مدل آماری برای پیش‌بینی شدت جراحت تصادفات ترافیکی استفاده کرد و نتیجه گرفت الگوریتم جنگل تصادفی بیش‌ترین دقت مدل‌سازی را دارد (Al-Moqri et al., 2020). توکلی کاشانی و شریعت مهیمنی شدت جراحت تصادفات را در راه‌های دوخطه‌ی دوطرفه بوسیله‌ی مدل درخت تصمیم بررسی نمود. نتایج درخت تصمیم دقت پیش‌بینی مناسبی را داشت (Tavakoli Kashani and Mohaymany, 2011). در مطالعه‌ی دیگری توکلی کاشانی و امیری‌فر (Tavakoli Kashani and Amirifar, 2020) از مدل درخت تصمیم برای پیش‌بینی شدت تصادفات عبور از چراغ قرمز استفاده کرد. بهبانی و همکاران نیز عوامل موثر بر شدت تصادفات رخ داده در قوس‌های افقی را در یک محور برون‌شهری توسط درخت تصمیم بررسی کردند (Behbahani, Effati and Mortezaei, 2020).

۲- پیشینه تحقیق

مرور ادبیات پیشین نشان می‌دهد که به دلیل مزیت‌های مدل‌های یادگیری ماشین، روند مطالعات صورت گرفته با این مدل‌ها در حال رشد است. هم‌چنین، در میان پژوهش‌های انجام شده در حوزه‌ی شدت تصادفات با رویکرد مدل‌های یادگیری ماشین، سه مدل جنگل تصادفی، ماشین بردار پشتیبان و درخت تصمیم بهترین عملکرد و دقت پیش‌بینی را داشته‌اند. از طرفی، این مدل‌ها در کم‌تر مطالعه‌ای برای پیش‌بینی شدت تصادفات راه‌های برون‌شهری استفاده شده‌است. لذا در پژوهش پیش‌رو سعی برآن است که از دو مدل درخت تصمیم و ماشین بردار پشتیبان برای تحلیل شدت تصادفات راه‌های برون‌شهری استفاده شود. بخش بعدی به بررسی داده‌های مورد استفاده و روش مطالعه خواهد پرداخت و سپس نتایج تشریح خواهند شد.

۲-۱-داده‌ها

در این مطالعه از حدود ۲۵ هزار داده‌ی مربوط به تصادفات برون‌شهری استان زنجان طی سال‌های ۱۳۹۲ تا ۱۴۰۰ استفاده شد. پس از پاکسازی داده‌ها و حذف رکوردهای با مقادیر نامشخص، نزدیک به ۲۰ هزار داده باقی ماند. خلاصه‌ای از متغیرها و دسته‌بندی‌های هریک در جدول ۱ آورده شده‌است. این داده‌ها مربوط به شرایط محیطی، وسیله نقلیه و ویژگی‌های راننده‌ی دخیل در وقوع تصادف می‌باشد. متغیر تصمیم به دو دسته‌ی تصادفات خسارتی و تصادفات جرحی و فوتی تقسیم شده‌است. لازم به ذکر است بعلت کم بودن تصادفات فوتی، در دسته‌ی مشترک با تصادفات جرحی درنظر گرفته شدند.

جدول ۱. توصیف متغیرها و دسته‌های آن‌ها

متغیر	دسته‌بندی
شدت تصادف (متغیر هدف)	(۱)خسارتی (۲) جرحی و فوتی
ماه وقوع تصادف	(۱)فروردین (۲)اردیبهشت (۳)خرداد (٤)تیر (٥)مرداد (٦)شهریور (۷)مهر (۸) آبان (۹)آذر (۱۰)دی (۱۱)بهمن (۱۲)اسفند
روز وقوع تصادف (ایام هفته)	(۱)شنبه (۲)یکشنبه (۳)دوشنبه (٤)سه‌شنبه (٥)چهارشنبه (٦)پنج‌شنبه (۷)جمعه
ساعت وقوع تصادف	متغیر پیوسته
نوع راه	(۱)آزادراه (۲)بزرگراه (۳)دوخطه دوطرفه (٤)فرعی
کیلومتر وقوع تصادف	متغیر پیوسته
نحوه برخورد	(۱)ایجاد حریق (۲)برخورد با چند وسیله (۳)برخورد با حیوان/احشام (٤)برخورد با دوچرخه (٥)برخورد با شیء ثابت (٦)برخورد با عابر (۷)برخورد با موتورسیکلت (۸) برخورد با وسیله پارک شده (۹)برخورد با یک وسیله (۱۰)خروج از جاده (۱۱)واژگونی و سقوط (۱۲)سایر موارد
روشنایی معبر	(۱)روز (۲) شب با روشنایی کافی (۳)شب بدون روشنایی (٤)طلوع (٥)غروب
سطح معبر	(۱)آب گرفتگی (۲)خشک (۳)شنی و خاکی (٤)قیرزدگی (٥)مرطوب و خیس (٦)یخبندان و برف (۷) گل آلود (۸)سایر
موقعیت تصادف	(۱)باندسواره (۲)خارج از حریم راه (۳)رفوژمیانه (٤)شانه (٥)کنار جاده (٦)سایر
علت تامه تصادف	(۱)انحراف به چپ (۲)انحراف به راست (۳)تجاوز از سرعت مجاز (٤)تجاوز به چپ (٥)تغییر مسیر ناگهانی (٦)حرکت با دنده عقب (۷)حرکت در خلاف جهت (۸)دور زدن در محل ممنوع (۹)عبور از محل ممنوع (۱۰)عدم توانایی در کنترل وسیله (۱۱)عدم توجه به جلو (۱۲)عدم رعایت حق تقدم (۱۳)عدم رعایت فاصله طولی و عرضی (۱٤)نقص راه (۱٥)نقص فنی وسیله نقلیه (۱٦)نقض ماده ۲۱۱ و ۲۱۲(رعایت مقررات ایمنی تعمیر و نگهداری راه) (۱۷)توقف وسیله نقلیه در خطوط عبور (۱۸)نقض مقررات حمل بار (۱۹)گردش به طرز غلط (۲۰)سایر
نوع وسیله مقصر	(۱)آمبولانس (۲)اتوبوس (۳)ادوات راهسازی (٤)ادوات کشاورزی (٥)تانکر حمل موارد خطرناک (٦)تریلر (۷)خودرو پلیس (۸)سواری (۹)کامیون (۱۰)کامیونت (۱۱)موتورسیکلت (۱۲)مینی‌بوس (۱۳)وانت‌بار (۱٤)سایر
سن راننده مقصر	متغیر پیوسته
جنسیت راننده مقصر	(۱)زن (۲)مرد

۳-روش تحقیق

به منظور دستیابی به هدف تحقیق، از دو روش درخت تصمیم و ماشین بردار پشتیبان استفاده شده تا نتایج و قدرت پیش‌بینی آن‌ها با یکدیگر مقایسه شوند. لازم به ذکر است که مدل‌سازی با زبان برنامه‌نویسی Python در محیط Anaconda3 (Spyder) انجام شده‌است. درخت طبقه‌بندی یکی از روش‌های قدرتمند یادگیری با نظارت در یادگیری ماشین برای طبقه‌بندی و شناخت الگوی میان داده‌ها است. استفاده از آن بعلت سهولت در بیان نتایج بصورت شکل و قابلیت اجرا بر داده‌های با حجم زیاد کاربرد بسیاری داشته‌است. هم‌چنین نتایج مطالعات نشان می‌دهد که این روش یکی از دقیق‌ترین مدل‌ها در پیش‌بینی شدت تصادفات است (Santos, Dias and Amado, 2022). ابتدا

داده‌های وارد شده به مدل در گره‌ی ریشه یا گره‌ی ابتدایی قرار می‌گیرند. سپس بر اساس بهترین متغیر ممکن، که توسط ضریب جینی یا ضریب آنتروپی مشخص می‌گردد، تقسیم‌بندی شروع می‌شود. هر متغیر پس از انتخاب به زیر مجموعه‌هایی شامل دسته‌های خود تقسیم می‌شود. در صورتیکه گره‌های انتهایی از نظر شاخص ارزیابی خلوص کافی را داشته باشند بعنوان گره‌ی نهایی یا برگ خواهند بود و در غیر این‌صورت انشعاب آن‌ها بر اساس سایر متغیرها ادامه خواهد یافت. یکی دیگر از مزایای این مدل آن است که برای جلوگیری از بیش‌برازش می‌توان آن را هرس کرد.

ضریب جینی معیاری از میزان ناخالصی در داده‌ها برای ایجاد دسته‌بندی است که بصورت رابطه ۱ محاسبه می‌شود.

برای تمام شکست‌های ممکن در دسته‌بندی‌ها ضریب جینی محاسبه می‌شود سپس شکستی که کم‌ترین مقدار Gini را دارد انتخاب می‌شود. الگوریتم CART نیز در دسته‌بندی‌های باینری خود از ضریب جینی استفاده می‌کند. P احتمال وقوع هریک از دسته‌ها می‌باشد.

$$Gini(t) = 1 - \sum_{i=1}^{c} p_i^2 \qquad (۱)$$

ماشین‌های بردار پشتیبان برای دسته‌بندی خطی و غیرخطی داده‌ها بکار می‌روند. یک ماشین بردار پشتیبان برای تعیین بهترین مرز میان مجموعه‌ها از خط استفاده می‌کند. اما هنگامیکه مرز میان دسته‌ها بصورت خطی قابل تفکیک نباشد از تابع کرنل استفاده می‌کند و آن‌ها را در فضای چند بعدی ترسیم می‌نماید. به بیان دیگر، این مدل به دنبال کمینه کردن خطای دسته‌بندی با یافتن بهترین اَبَرصفحه‌ی جداکننده‌ی میان دسته‌ها است. معادله‌ی ابرصفحه بصورت رابطه ۲ است. در میان معادلات ممکن برای ابرصفحه‌ها، ابرصفحه‌ی بهینه با کمینه کردن رابطه‌ی ۳ تحت شرایط رابطه ٤ بدست می‌آید. X مجموعه نقاط هستند، W بردار نرمال عمود بر ابرصفحه است، ε_i برابر است با خطای دسته‌بندی داده‌ها، C ضریب توازن برای دو جمله‌ی کمینه شونده است بطوریکه با افزایش آن خطای قابل قبول کم‌تر خواهد شد.

$$f(x, \alpha) = (W_\alpha . x) + b \qquad (۲)$$

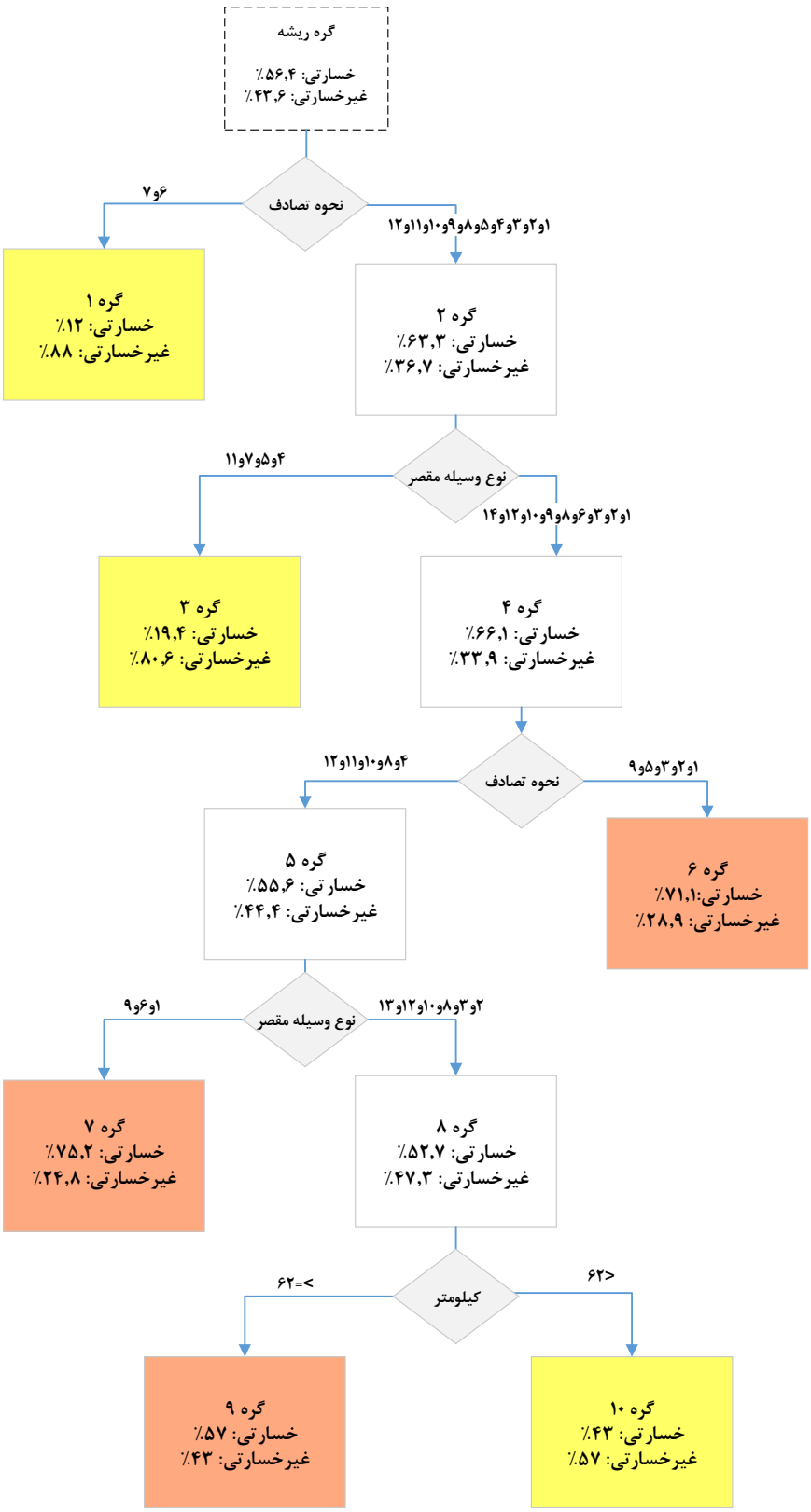
$$Min\left(\frac{1}{2}\|W\|^2 + C\sum_{i=1}^{m}\varepsilon_i\right) \qquad (۳)$$

$$y_i\left(\overrightarrow{W}\cdot\overrightarrow{X_i} + b\right) \geq 1 - \varepsilon_i \quad .\varepsilon_i \geq 0 \qquad (٤)$$

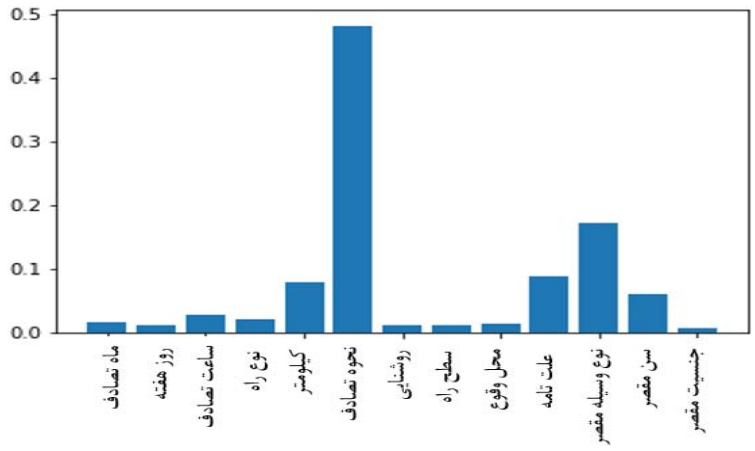
٤-نتایج

نتایج مدل درخت تصمیم در شکل ۱ آورده شده‌است. مطابق این شکل، مهم‌ترین متغیر برای تقسیم‌بندی داده‌ها نحوه تصادف است. در گره ۱ تصادفات شامل برخورد با عابر و موتورسیکلت در ۸۸ درصد مواقع منجر به تصادفات جرحی یا فوتی خواهد شد که بعلت آسب‌پذیری این کاربران راه است. سایر برخوردها بر اساس نوع وسیله نقلیه مقصر تقسیم‌بندی می‌شوند. هنگامی‌که وسیله نقلیه راننده مقصر از نوع ادوات کشاورزی، تانکر، خودروی پلیس و موتورسیکلت باشد در نزدیک ۸۱ درصد پیش‌بینی می‌شود که تصادف از نوع جرحی یا فوتی باشد. سایر انواع وسایل نقلیه بر اساس متغیر نحوه برخورد تقسیم‌بندی شده‌اند. براین اساس برخورد با چند وسیله نقلیه، برخورد با حیوانات، برخورد با شیء ثابت و برخورد با یک وسیله نقلیه در ۷۱ درصد خسارتی خواهد بود. برخوردهای از نوع برخورد با دوچرخه، برخورد با وسیله پارک شده، خروج از جاده و واژگونی در صورتیکه وسیله نقلیه مقصر در تصادفات از نوع آمبولانس، تریلر، و کامیون باشد با احتمال ۷۵ درصد خسارتی خواهند بود. وسایل نقلیه‌ی اتوبوس، ادوات راهسازی، مینی‌بوس و وانت بار در کیلومتراژ کم‌تر از ٦۲ کیلومتر عمدتا جرحی و فوتی هستند. در کیلومترهای بیش‌تر از ٦۲ کیلومتر عمدتاً تصادفات خسارتی خواهند بود.

شکل ۲ اهمیت نسبی متغیرها را در مدل درخت تصمیم نشان می‌دهد. مطابق این شکل، متغیرهای نحوه تصادف، نوع وسیله و کیلومتر وقوع تصادف به ترتیب بیش‌ترین اهمیت را در پیش‌بینی شدت تصادفات در این مدل دارند.

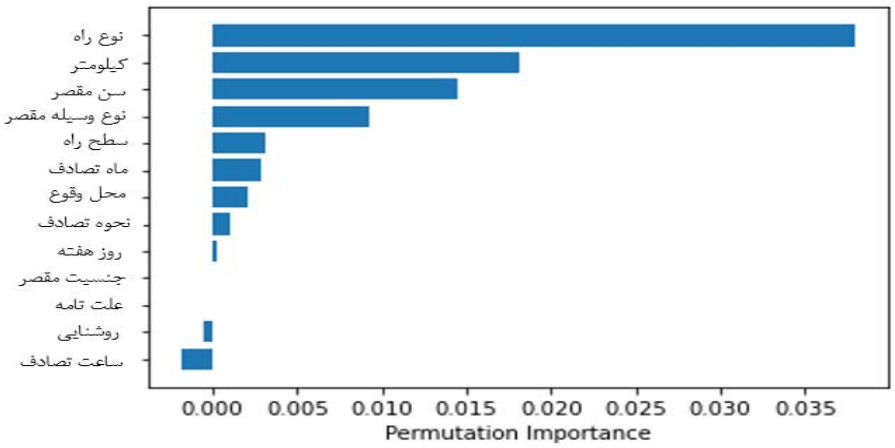


شکل ۱. مدل درخت تصمیم



شکل ۲. ضرایب اهمیت متغیرها در مدل درخت تصمیم

در مدل SVM به منظور یافتن بهترین مقادیر هایپرپارامترهای[۲] مدل از روش جست‌وجوی شبکه‌ای[۳] استفاده شد. در این مدل برای کرنل‌هایی بجز کرنل خطی (گاوسی و شعاعی) نمی‌توان مقادیر ضرایب اهمیت متغیرها را تعیین نمود چرا که کرنل‌ها در محیط n بعدی دسته‌بندی را انجام می‌دهند اما می‌توان از ضرایب اهمیت جایگشتی[۴] استفاده کرد. بدین صورت که متغیرها چندین دفعه به اصطلاح مخلوط[۵] می‌شوند (ترتیبشان بهم می‌ریزد) و پیش‌بینی انجام می‌شود. سپس مقدار خطای پیش‌بینی داده‌های مخلوط شده محاسبه می‌شود. تفاوت بین مقدار خطای پایه و خطای داده‌های آمیخته شده محاسبه می‌شود. این روند حداقل سه مرتبه انجام می‌شود و میانگین آن بدست آورده می‌شود. در آخر متغیرها بر اساس مقادیر میانگین تاثیرشان بر پیش‌بینی مدل رتبه‌بندی می‌شوند. متغیرهایی که تاثیر بیش‌تری دارند نمره‌ی بالاتری دارند. مقادیر منفی نیز نشان دهنده‌ی افزایش خطای مدل در صورت بکار بردن این متغیرها است. در شکل ۲ ملاحظه می‌شود که نوع راه، کیلومتر و سن وسیله بیش‌ترین اهمیت را در این مدل داشته‌اند.



شکل ۲. ضرایب اهمیت جایگشتی متغیرها در مدل ماشین بردار پشتیبان

در اعتبارسنجی مدل‌ها توسط شاخص مشخصه عملکرد ROC[۶] دو نوع شاخص micro-average ROC و macro-average ROC استفاده شده‌است. هر داده در واقع به یکی از کلاس‌های مثبت یا منفی (خسارتی یا غیرخسارتی) تعلق دارد و الگوریتم مورد استفاده در مدل‌سازی نیز در نهایت پیش‌بینی خواهد نمود که هر نمونه متعلق به کدام دسته‌ی مثبت یا منفی است. بر این اساس برای هریک از داده‌ها یکی از چهار حالت بوجود خواهد آمد: داده عضو دسته مثبت باشد و درست پیش‌بینی شود (مثبت درست[۷])، و اگر منفی پیش‌بینی شود منفی نادرست[۸]، داده عضو دسته منفی باشد و درست پیش‌بینی شود (منفی درست[۹]) و اگر مثبت پیش‌بینی شود مثبت نادرست[۱۰] خواهد بود. در نمودار ROC محور افقی برابر با FPR و محور عمودی برابر TPR می‌باشد. مطابق با معادلات ۵ و ٦ در نوع اول این شاخص (micro-average) استفاده می‌شود. این روش دقت بیش‌تری به هنگام نامتوازن بودن کلاس‌ها دارد. در نوع دوم (macro-average) مقادیر TPR و FPR در هر دسته بطور جداگانه محاسبه و سپس بین آن‌ها میانگین‌گیری انجام می‌شود. بنابراین وزن یکسانی برای هر دسته لحاظ می‌گردد. این روش نامتوازن بودن داده‌ها را درنظر نمی‌گیرد و دیدگاه کلی‌تری دارد.

$$Micro - Average\ TPR = \frac{\sum_{n=c} TP_n}{\sum_{n=c} TP_n + FN_n} \quad (۵)$$

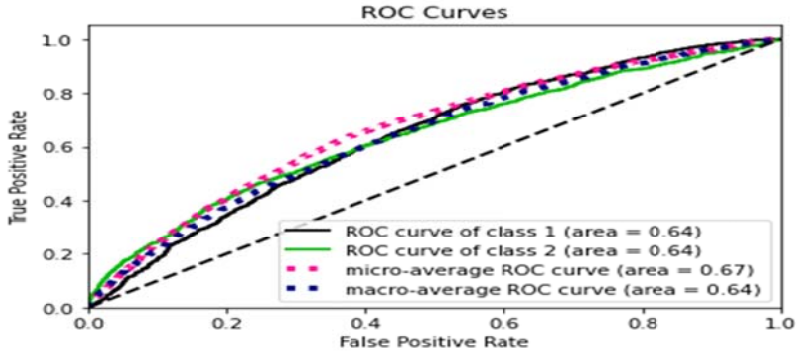
$$Micro - Average\ FPR = \frac{\sum_{n=c} FP_n}{\sum_{n=c} FP_n + TN_n} \quad (٦)$$

$$Macro - Average\ TPR = \frac{\sum_{n=c} TPR}{n} \quad (۷)$$

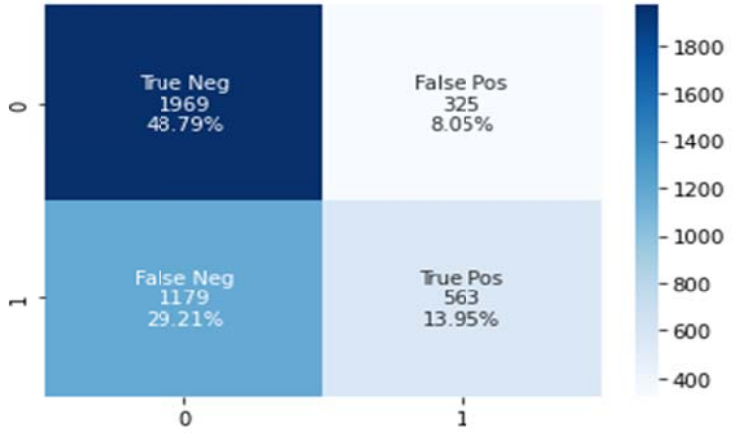
$$Macro - Average\ FPR = \frac{\sum_{n=c} FPR}{n} \quad (۸)$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (۹)$$

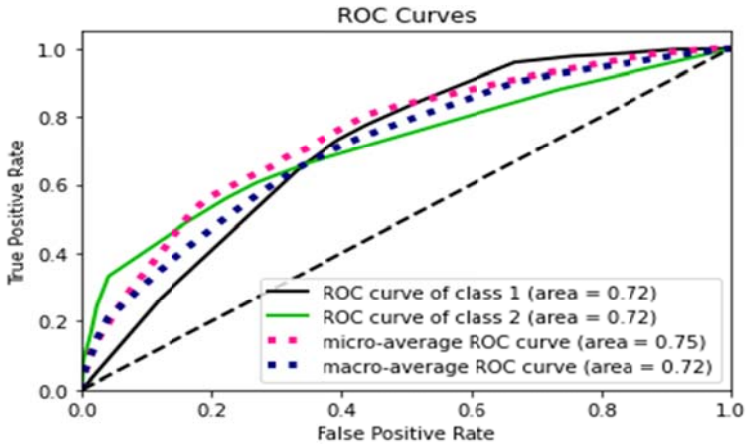
همانطور که در شکل۳ مشاهده می‌شود، مدل SVM در بهینه‌ترین حالت با کرنل RBF[۱۱] مقدار دقت برابر با ٦۳ درصد را بدست می‌دهد و مساحت زیر نمودار ROC آن برابر با ۰/٦۷ شده است. شکل٤ مربوط به ماتریس درهم‌ریختگی این مدل است. همانطور که ملاحظه می‌شود این مدل در پیش‌بینی تصادفات جرحی و فوتی نمی‌تواند عملکرد مطلوبی داشته باشد و تنها درصد پیش‌بینی تصادفات خسارتی در آن بالا است. لازم به ذکر است دقت مدل از رابطه ۹ بدست می‌آید. نتایج اعتبارسنجی مدل درخت تصمیم در شکل ۵ نشان داده شده‌است. مطابق این شکل، مساحت زیر نمودار ROC درخت تصمیم برابر با ۰/۷۵ است و دقت پیش‌بینی این مدل برای شدت تصادفات برابر با ٦۸ درصد می‌باشد. ماتریس درهم ریختگی درخت تصمیم نیز در شکل ٦ نشان می‌دهد که این مدل عملکرد مطلوب‌تری نسبت به مدل SVM دارد و توانسته در هر دو دسته پیش‌بینی‌های درست داشته باشد.



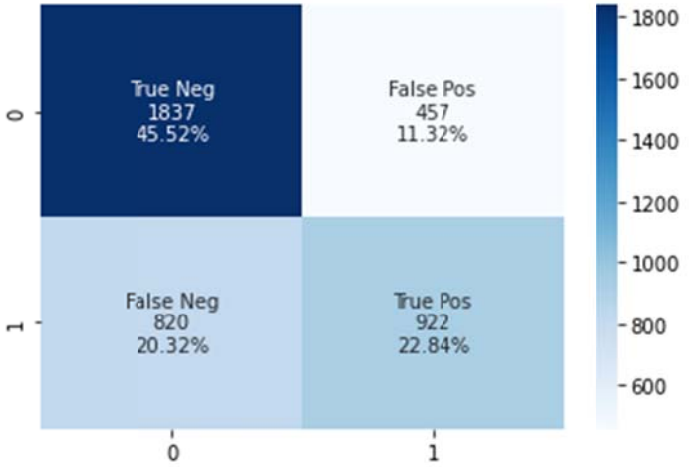
شکل ۳. منحنی ROC در مدل ماشین بردار پشتیبان



شکل ٤. ماتریس درهم‌ریختگی در مدل ماشین بردار پشتیبان



شکل ٥. منحنی ROC در مدل درخت تصمیم



شکل ٦. ماتریس درهم‌ریختگی در مدل درخت تصمیم

۵-نتیجه‌گیری

در پژوهش حاضر به پیش‌بینی و شناسایی عوامل موثر بر شدت تصادفات راه‌های برون شهری استان زنجان پرداخته شد. بدین منظور از دو مدل یادگیری نظارت شده ماشین بردار پشتیبان و درخت تصمیم استفاده شد و نتایج هردو مدل با یکدیگر مقایسه شدند. همانطور که ملاحظه شد، در مدل‌های ساخته شده برای داده‌های مطالعه، بطور کلی درخت تصمیم قدرت پیش‌بینی بیش‌تری دارد. هم‌چنین دقت درخت تصمیم در جراحات شدیدتر بیش‌تر از ماشین‌بردار پشتیبان می‌باشد. مقایسه‌ی ضرایب اهمیت متغیرها نیز نشان داد مهم‌ترین متغیر از نظر درخت تصمیم نحوه تصادف و از نظر ماشین بردار پشتیبان نوع راه است. از طرفی ملاحظه می‌شود که نتایج پیش‌بینی درخت تصمیم به راحتی قابل مشاهده است. اما ماشین بردار پشتیبان مانند یک جعبه سیاه عمل می‌نماید و تنها می‌توان دقت عملکرد آن را بررسی نمود. مهم‌ترین نتایج مدل درخت تصمیم در ادامه اشاره می‌شوند. تصادفاتی که با عابران پیاده و موتورسیکلت‌ها می‌شود بعلت عدم وجود حفاظ کافی در آن‌ها آسیب‌پذیری بیش‌تری دارند لذا شدت تصادفات آنان نیز بیش‌تر خواهد بود و لزوم تمرکز بر افزایش ایمنی این کاربران راه بویژه در مناطق برون‌شهری که سرعت وسایل نقلیه بیش‌تر است وجود دارد.

هم‌چنین مدل نشان داد در فواصل کمتر از ۶۲ کیلومتر نسبت به شهر شدت تصادفات بیش‌تر است.

۶- سپاسگزاری

مقاله حاضر با همکاری مشترک اداره کل راهداری و حمل‌ونقل جاده‌ای استان زنجان در قالب مطالعات بازدید ایمنی راه‌های استان زنجان انجام شده‌است. بدینوسیله از همکاری صمیمانه‌ی جناب آقای علی مدقالچی، معاون حمل‌ونقل اداره کل راهداری و حمل‌ونقل جاده‌ای استان زنجان، سرکار خانم اعظم محمدی، مدیر اداره ایمنی ترافیک اداره کل راهداری و حمل‌ونقل جاده‌ای استان زنجان، و جناب آقای محمد جزونقی، کارشناس ایمنی ترافیک اداره کل راهداری و حمل‌ونقل جاده‌ای استان زنجان کمال تشکر و قدردانی می‌شود.

۷-پی‌نوشت‌ها

1. Classification Tree
2. Hyper Parameter Tuning
3. Grid Search
4. Permutation Importance
5. Shuffle
6. Receiver Operating Characteristic
7. True Positive (Tp)
8. False Negative (Fn)
9. True Negative (Tn)
10. False Positive (Fp)
11. Radial Basis Function

۸-مراجع

-Abrari Vajari, M. *et al.* (2020). A multinomial logit model of motorcycle crash severity at Australian intersections. *Journal of Safety Research*, 73, 17–24. **doi.org/10.1016/j.jsr.2020.02.008**

-Al-Moqri, T. *et al.* (2020). Exploiting Machine Learning Algorithms for Predicting Crash Injury Severity in Yemen: Hospital Case Study. *Appl. Comput. Math*, 9(5), 55–164.

-AlMamlook, R.E. *et al.* (2019). Comparison of machine learning algorithms for predicting traffic accident severity. in *2019 IEEE Jordan International Joint Conference On Electrical Engineering And Information Technology (JEEIT)*. IEEE, 272–276.

-Arhin, S.A. and 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.

Behbahani, H., Effati, M. and Mortezaei, S. (2020). roviding a Method for Accident Severity Analysis Using Geospatial Clustering Functions and Decision Tree, Case Study: Qazvin-Loshan Freeway (in persian). *Amirkabir J. Civil Eng.*, 52(6), 1419–1438.

-Hosseinzadeh, A., Moeinaddini, A. and

-Tavakoli Kashani, A. 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*, 49(10), 1314–1320.
doi.org/10.1016/j.ssci.2011.04.019
-Wang, X. and Kim, S.H. (2019). Prediction and factor identification for crash severity: comparison of discrete choice and tree-based models. *Transportation Research Record*, 2673(9), 640–653.
-World Health Organization (2021). extranet.who.int/roadsafety/death-on-the-roads/#country_or_area/IRN.
-Yuan, Y. et al. (2021). Risk factors associated with truck-involved fatal crash severity: Analyzing their impact for different groups of truck drivers. *Journal of Safety Research*, 76, 154–165.
doi.org/10.1016/j.jsr.2020.12.012
Ghasemzadeh, A. (2021). Investigating factors affecting severity of large truck-involved crashes: Comparison of the SVM and random parameter logit model. *Journal of safety Research*, 77, 151–160.
Iranian Legal Medicine Organization (1401). Available at: https://www.lmo.ir/.
-Kaplan, S. and Prato, C.G. (2012). Risk factors associated with bus accident severity in the United States: A generalized ordered logit model. *Journal of Safety Research*, 43(3), 171–180.
doi.org/10.1016/j.jsr.2012.05.003
-Labib, M.F. *et al.* (2019). Road accident analysis and prediction of accident severity by using machine learning in Bangladesh. in *2019 7th International Conference On Smart Computing & Communications (ICSCC)*. IEEE. 1–5.
-Santos, K., Dias, J.P. and Amado, C. (2022). A literature review of machine learning algorithms for crash injury severity prediction. *Journal of Safety Research*, 80, 254–269.
-Tavakoli Kashani, A. and Amirifar, S. (2020). Analyzing the effect of drivers' characteristics on red-light running crash severity, case study: Isfahan (In persian). *In The 18th International Conference on Traffic & Transportation*. The 18th International Conference on Traffic & Transportation.

Applying Machine Learning Methods to Predict Crash Severity at Rural Roads- Case Study of Zanjan Province

Ali Tavakoli Kashani, Associate Professor, School of Civil Engineering, Iran University of Science and Technology, Tehran, Iran.

Ali Medghalchi, School of Civil Engineering, Azad University of Qazvin, Qazvin, Iran.

Azam Mohammadi, School of Civil Engineering, Zanjan University, Zanjan, Iran.

Mohammad Jazvanaqi, M.Sc., Grad., Engineering, Imam Khomeini International University, Qzavin, Iran.

E-mail: alitavakoli@iust.ac.ir

Received: February 2024- Accepted: June 2024

ABSTRACT

Traffic crashes are a significant problem in low and middle-income countries, while there is a worrying trend of increasing fatal and injury crashes Iran. This highlights the urgent need to analyze the causes of such accidents to improve road safety and reduce their negative consequences. To address this issue, a study was conducted to investigate the factors that contribute to the severity of rural crashes in Zanjan province, using advanced machine learning models such as Support Vector Machine and Decision Tree. The study utilized a crash database of 25,000 incidents over a 9-year period, and after cleaning the data, the models were developed in Python. The findings suggest that “type of crash”, “at-fault driver's vehicle type”, and “kilometer occurrence of the crash” are key variables for predicting the severity of these crashes. The Decision Tree model was also found to be more accurate than the Support Vector Machine model, particularly in predicting severe crashes. This study provides valuable insights for improving road safety and reducing the harmful effects of traffic crashes in rural areas.

Keywords: Crash Severity, Decision Tree, Support Vector Machine, Rural Crashes, Traffic Safety