مقایسه الگوریتم‌های یادگیری ماشین در انتخاب ویژگی و طبقه‌بندی شدت تصادفات عابر پیاده (مطالعه موردی: شهر قم)

مقاله علمی – پژوهشی

صادق مسلمی مهنی، دانشجوی دکتری، گروه عمران، واحد علوم و تحقیقات، دانشگاه آزاد اسلامی، تهران، ایران
علی نادران *، استادیار، گروه عمران، واحد علوم و تحقیقات، دانشگاه آزاد اسلامی، تهران، ایران
سید صابر ناصر علوی، استادیار، گروه عمران، واحد کرمان، دانشگاه باهنر، کرمان، ایران
*پست الکترونیکی نویسنده مسئول: naderan@iau.ac.ir
دریافت: ۱۴۰۴/۰۱/۲۹– پذیرش: ۱۴۰۴/۰۵/۰۱
صفحه ۲۳۲-۲۱۷

چکیده

این مقاله به بررسی کارایی الگوریتم‌های یادگیری ماشین در طبقه‌بندی شدت تصادفات عابر پیاده باتوجه‌به عوامل مختلفی مانند وضعیت آب‌وهوا، دوره‌های زمانی و نوع تصادف پرداخته است. در ابتدا، تکنیک‌های مختلفی برای انتخاب ویژگی مورد بررسی و ارزیابی قرار گرفته و با استفاده از داده‌های واقعی شهر قم، بهینه‌سازی ترکیب ویژگی‌ها برای بهبود دقت طبقه‌بندی انجام شده است. سپس، کارایی الگوریتم‌های k نزدیک‌ترین همسایه، درخت تصمیم و جنگل تصادفی با استفاده از چندین معیار، بررسی و مقایسه شده است. نتایج نشان داده‌اند که درخت تصمیم بهترین کارایی را در طبقه‌بندی شدت تصادفات دارد. ضمن اینکه برای این مدل، انجام انتخاب ویژگی توصیه نمی‌شود. دقت طبقه بندی هر مدل یادگیری به ابعاد داده‌ها و ویژگی‌های انتخاب شده بستگی دارد. بنابراین، پیشنهاد می‌شود که تأثیر عوامل دیگری مانند شرایط ترافیک و ویژگی‌های جاده به همراه وضعیت آب‌وهوایی بررسی شود تا بهبود کارایی الگوریتم‌های طبقه‌بندی در این زمینه انجام شود.

واژه‌های کلیدی: انتخاب ویژگی، طبقه‌بندی، ماتریس درهم‌ریختگی، یادگیری ماشین

۱-مقدمه

شهر قم با جمعیتی بالغ بر ۱٫۴۳ میلیون نفر، یکی از شهرهای پرجمعیت و پرتردد ایران محسوب می‌شود. باتوجه‌به موقعیت جغرافیایی و افزایش تردد روزانه خودروها، تعداد تصادفات رانندگی در این شهر بسیار زیاد است. طبق گزارش سازمان اورژانس ایران، در سال ۲۰۱۸، تعداد ٤٤٦٧ حادثه رانندگی با عابران پیاده در شهر قم ثبت شده که از این تعداد ۲۲٦۲ تصادف با شدت متوسط و ۲۲۰٤ تصادف با شدت شدید بوده است. تأثیرات این حوادث قابل‌توجه بوده و باعث ایجاد خسارات جبران‌ناپذیری از جمله هزینه‌های درمانی و اقتصادی شده است. به‌منظور کاهش این تعداد تصادفات، استفاده از الگوریتم‌های هوشمند و یادگیری ماشین به‌عنوان رویکردی نوین، می‌تواند اثرات مثبتی داشته باشد. این روش‌ها با استفاده از مجموعه‌داده‌هایی که عوامل تصادف را مشخص می‌کنند، مدل‌هایی برای تحلیل و پیش‌بینی تصادفات ایجاد می‌کنند (Nguyen, Wang, & Nguyen, 2013). در این راستا، هدف این مطالعه، استفاده از الگوریتم‌های انتخاب ویژگی برای انتخاب یک مجموعه از عوامل مؤثر و سپس ساخت مدل‌هایی برای طبقه‌بندی شدت آسیب در تصادفات رانندگی است. علاوه

بر این الگوریتم‌های مذکور مورد مقایسه و بررسی قرار گرفته‌اند. اطلاعاتی که از این تحلیل‌ها به دست می‌آید، به سازمان‌ها امکان می‌دهد تا برای مقابله با خطرات احتمالی، اقدامات لازم را با بیشترین اثربخشی انجام دهند.

۲-پیشینه تحقیق

همانطور که در مقدمه ذکر شده است، تصادفات جاده‌ای، یکی از شایع‌ترین علل مرگ و صدمات جسمی در جوامع امروزی هستند. آمارها نشان می‌دهند که هر ساله میلیون‌ها نفر در سراسر جهان در اثر تصادفات جاده‌ای کشته یا مصدوم می‌شوند. علاوه بر این، تصادفات جاده‌ای به شدت بر اقتصاد جوامع تأثیر می‌گذارند. در این راستا، الگوریتم‌های یادگیری ماشین برای پیشگیری از تصادفات جاده‌ای، توانسته‌اند نتایج بسیار قابل توجهی را به دنبال داشته باشند. در این زمینه، الگوریتم‌های مبتنی بر درخت تصمیم، جزو موثرترین الگوریتم‌ها هستند. این مدل ها با تحلیل داده‌های تصادف، عوامل موثر در کارایی هر الگوریتم مورد استفاده را شناسایی می‌کنند.ضمن اینکه، این الگوریتم‌ها می‌توانند به صورت هوشمندانه، طبقه بندی تصادفات را انجام دهند. نتایج پژوهش‌های انجام شده نیز نشان می‌دهد که استفاده از مدل های یادگیری ماشین می‌تواند به طور قابل‌توجهی تعداد تصادفات را کاهش دهد. در تحقیق (Yang, Huang, Guo, & Sun, 2022) روشی با نام ادغام چندین طبقه‌بندی به منظور بهبود دقت ارائه شده است. این روش شامل دو بخش اساسی انتخاب ویژگی و طبقه‌بندی می باشد. برای انتخاب ویژگی، از روش اطلاعات مشترک و انتخاب تک تک ویژگی‌ها به صورت پیاپی استفاده شده است. سپس مدل‌هایی با استفاده از ماشین بردار پشتیبان، نزدیکترین همسایه و جنگل تصادفی برای طبقه‌بندی در نظر گرفته شده است.نتایج نشان داد که استفاده از ترکیب طبقه‌بندهای مختلف، باعث بهبود دقت و افزایش قابلیت اطمینان مدل می‌شود.

در تحقیق (Singh & Haider, 2022)، بر روی تکنیک انتخاب ویژگی که یک روش کاهش ابعاد است و باعث بهبود دقت و کاهش زمان اجرا می‌شود تمرکز شده است. در این تحقیق، برای پیش‌بینی از چهار روش طبقه‌بندی ماشینی k نزدیکترین همسایه، رگرسیون لجستیک، درخت تصمیم و ماشین بردار پشتیبان استفاده شده است. همچنین، سه روش فیلتر، پوششی و روش‌های ترکیبی برای انتخاب ویژگی مورد استفاده قرار گرفته است. در کل، این تحقیق نشان می‌دهد که استفاده از روش‌های ترکیبی برای انتخاب ویژگی‌ها به همراه یک الگوریتم طبقه‌بندی باعث بهبود دقت در پیش‌بینی می‌شود. در مقاله (Vandana & Chikkamannur, 2021) تکنیک‌های مختلفی برای انتخاب ویژگی مورد بررسی و ارزیابی قرار گرفته‌اند. این تکنیک‌ها شامل روش‌های فیلتر و روش‌های وابسته به مدل هستند. در روش فیلتر، ویژگی‌ها بر اساس معیارهای خاصی انتخاب می‌شوند و در روش وابسته به مدل، ویژگی‌ها بر اساس تأثیر آنها در عملکرد الگوریتم‌های یادگیری مورد بررسی قرار می‌گیرند. همچنین، برای ارزیابی عملکرد الگوریتم‌های انتخاب ویژگی، از دو الگوریتم یادگیری نظارت شده بیز ساده و درخت تصمیم استفاده شده است. نتایج نشان می‌دهد که الگوریتم پیشنهادی با استفاده از تعداد کمتری ویژگی، کارایی بهتری نسبت به الگوریتم‌های دیگر انتخاب ویژگی دارد. در مقاله (Got, Moussaoui, & Zouache, 2021)، یک روش جدید هیبریدی با ادغام دو روش فیلتر و پوششی برای انتخاب ویژگی با استفاده از الگوریتم بهینه‌سازی نهنگ پیشنهاد شده است. در این روش، ابتدا با استفاده از یک روش فیلتر، ویژگی‌هایی که دارای ارتباط بیشتری با کلاس‌های مختلف هستند، انتخاب می‌شوند. سپس با استفاده از یک روش پوششی، زیرمجموعه‌ای از ویژگی‌های انتخاب شده باعث افزایش دقت طبقه‌بندی می‌شود. در این روش، الگوریتم بهینه‌سازی نهنگ برای بهینه‌سازی استفاده می‌شود. نتایج آزمایش‌ها نشان داده‌اند که الگوریتم پیشنهادی قادر به بدست آوردن زیرمجموعه‌هایی با تعداد کمتری از ویژگی‌ها با دقت طبقه‌بندی بالا است. در مطالعه(Danaei Mehr & Polat, 2022)، عملکرد چندین طبقه‌بندی جنگل تصادفی، درخت اضافی، آدابوست و شبکه عصبی چند لایه با استفاده از تمام ویژگی‌ها و زیرمجموعه کاهش یافته از ویژگی‌هایی که توسط روش‌های فیلتر، تعبیه شده[10] و دسته بند[11] تولید شده‌اند، بررسی شد. نتایج آزمایش‌ها نشان داد که روش‌های انتخاب ویژگی اثرات مفیدی بر بهبود عملکرد تمامی طبقه‌بندی‌ها دارند. علاوه بر این، طبقه‌بندی جنگل تصادفی با استفاده از زیرمجموعه کاهش یافته ویژگی‌ها از سایر

طبقه‌بندها پیشی گرفت. در تحقیق (Ali et al., 2022)، از روش‌های همبستگی پیرسون و اطلاعات متقابل برای بررسی رابطه بین ویژگی‌های ورودی و خروجی استفاده شده است. با استفاده از ویژگی‌های ورودی بهینه، مدلهای مختلفی با استفاده از الگوریتم‌های شبکه عصبی مصنوعی، درخت تصمیم و بیز ناخالص توسعه داده شده‌اند. نتایج نشان داده‌اند که با استفاده از الگوریتم‌های انتخاب ویژگی، می‌توان به دقت بیشتری دست یافت. در مقاله (Zhou, Zhang, Zhou, Guo, & Ma, 2021)، روشی برای بهبود دقت دسته‌بندی درخت تصمیم ارائه شده است. در این راستا، یک الگوریتم جدید انتخاب ویژگی به نام درخت تصمیم بازگشتی شبکه‌ای[۱۲] پیشنهاد شده است. نتایج تجربی نشان می‌دهد که این الگوریتم بهبود چشمگیری در دقت، بازخوانی و امتیاز F1 دسته‌بندی درخت تصمیم دارد. همچنین، استفاده از الگوریتم مذکور می‌تواند بهبود قابل توجهی در دقت و سرعت دسته‌بندی درخت تصمیم داشته باشد. در پژوهش (Tsai & Chen, 2019)، برای انتخاب ویژگی، سه روش مختلف فیلتر، پوششی و تعبیه شده استفاده شده‌اند که هر کدام از آن‌ها از الگوریتم‌های تجزیه و تحلیل مؤلفه‌های اصلی، ژنتیک و درخت تصمیم چندگانه استفاده کرده‌اند. نتایج نشان داده‌اند که برای طبقه‌بند ماشین بردار پشتیبان، انجام انتخاب ویژگی با استفاده از درخت تصمیم چندگانه ، بهترین عملکرد را در مقایسه با دیگر ترکیب‌ها داشته‌اند. برای طبقه‌بند درخت تصمیم نیز، انجام انتخاب ویژگی با استفاده از درخت تصمیم چندگانه توصیه شده است. در کل، این پژوهش به دنبال ارائه یک ترکیب بهینه از انتخاب ویژگی برای بهبود دقت طبقه‌بندی و کاهش زمان محاسبات می باشد. پژوهش (Kunhare, Tiwari, & Dhar, 2020) بررسی کرده است که چگونه با استفاده از الگوریتم‌های یادگیری ماشین و الگوریتم‌های بهینه‌سازی، می‌توان از تعداد زیاد ویژگی‌ها کاسته و عملکرد سیستم را بهبود بخشید. در این پژوهش، ابتدا با استفاده از الگوریتم درخت تصادفی، ویژگی‌های بی‌مفهوم و نویزی حذف شده‌اند. سپس با استفاده از الگوریتم بهینه‌سازی ذرات، ویژگی‌های انتخاب شده بهینه سازی گردیده است. در نهایت، با استفاده از روش‌های مختلف طبقه بندی، عملکرد سیستم با ویژگی‌های انتخاب شده ارزیابی شده است. نتایج نشان داده‌اند که استفاده از الگوریتم درخت تصادفی برای حذف ویژگی‌هایی با تاثیر کمتر، باعث بهبود عملکرد سیستم شده است. هم‌چنین، استفاده از الگوریتم بهینه‌سازی ذرات برای بهینه‌سازی ویژگی‌های انتخاب شده، باعث کاهش نرخ خطای نادرست و افزایش دقت شده است. نتایج نشان داده‌اند که استفاده از طبقه بندهای مختلف، از جمله نزدیکترین همسایه، ماشین‌بردار پشتیبان، رگرسیون لجستیک، درخت تصمیم و بیز ساده، می‌تواند با استفاده از ویژگی‌های انتخاب شده، بهبود قابل توجهی در عملکرد ایجاد کنند. در مطالعه (AlMamlook, Kwayu, Alkasisbeh, & Frefer, 2019)، پژوهشگران هدف خود را برای توسعه مدل‌های دقیق جهت پیش‌بینی شدت تصادفات جاده‌ای با استفاده از تکنیک‌های یادگیری ماشین مطرح کردند. آن‌ها یک مجموعه از عوامل مؤثر را که می‌توانند بر شدت آسیب‌ها تأثیر بگذارند، انتخاب کردند و سپس از الگوریتم‌های یادگیری ماشین نظارت‌شده از جمله آدابوست، رگرسیون لجستیک، بیز ساده و جنگل تصادفی برای ساخت مدل‌ها استفاده کردند. نتایج مطالعه نشان داد که الگوریتم جنگل تصادفی نسبت به سایر الگوریتم‌ها، عملکرد بهتری داشت. این بدان معنی است که مدل ساخته شده با استفاده از الگوریتم مذکور می‌تواند یک ابزار قابل اعتماد برای پیش‌بینی شدت تصادفات جاده‌ای باشد. در مطالعه (Komol et al., 2021)، داده‌های تصادف جاده‌ای در استرالیا برای سال‌های ۲۰۱۳ تا ۲۰۱۹ مورد بررسی قرار گرفته‌است و ویژگی های مختلف مرتبط با خصوصیات کاربران جاده، آب و هوا، محیط زیست، وضعیت وسیله نقلیه، وضعیت راننده، دوره زمانی، مشخصات جاده، مشخصات ترافیک و محدوده سرعت جمع‌آوری شده‌اند. سپس الگوریتم‌های یادگیری ماشین مانند نزدیکترین همسایه، ماشین بردار پشتیبان و جنگل تصادفی برای پیش‌بینی آموزش داده شده‌اند. نتایج نشان می‌دهد که الگوریتم جنگل تصادفی برای همه حالت‌ها دقت آزمون نسبتاً بالایی داشته است. همان‌طور که ملاحظه می‌شود، استفاده از الگوریتم‌های یادگیری ماشین در پیشگیری از تصادفات جاده‌ای و بهبود ایمنی جاده‌ها، در سال‌های اخیر به یکی از موضوعات پژوهشی پرطرف‌دار تبدیل شده است و می‌تواند بهبود قابل‌توجهی در ایمنی جاده‌ها ایجاد کند.

۳-روش تحقیق

هدف این مطالعه، استفاده از الگوریتم‌های یادگیری ماشین برای شناسایی ویژگی‌های مؤثر و ساخت مدل‌های با بهترین عملکرد است. در شکل ۱، طرح کلی مطالعه نمایش‌داده‌شده است که در آن مراحل مختلف از جمله جمع‌آوری داده‌ها، پیش‌پردازش، استخراج ویژگی‌ها، ساخت مدل‌های طبقه‌بندی و ارزیابی عملکرد مدل‌ها به شکل واضحی مشخص شده است. در این مطالعه، از مدل‌های متنوعی برای پیداکردن بهترین ویژگی‌ها استفاده شده است. سپس برای آموزش و آزمون هر مدل طبقه‌بندی متناسب با مدل انتخاب ویژگی استفاده می‌شود. به‌این‌ترتیب، می‌توان اطمینان حاصل کنیم که مدل بهترین عملکرد را دارد و قابلیت تعمیم‌پذیری و استفاده در داده‌های جدید را دارا است. این بخش به روش‌های استفاده شده در این مطالعه از جمله مشخصات داده‌ها، شناسایی ویژگی‌های مؤثر، ساخت مدل‌های طبقه‌بندی، ارزیابی و استخراج دانش موردنیاز می‌پردازد.

۳-۱- منابع و جزئیات داده

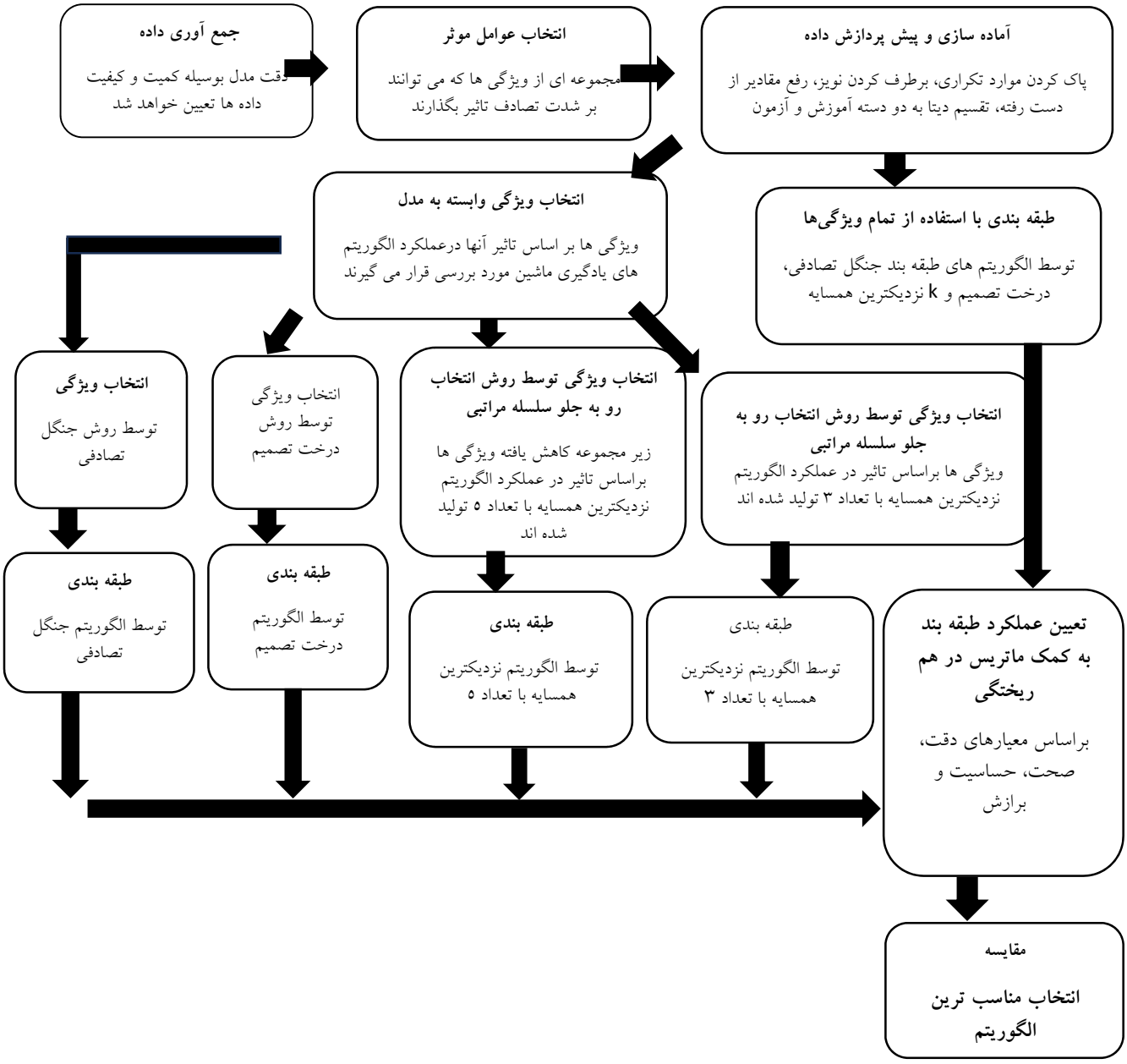
در این مقاله، از داده‌های جمع‌آوری شده توسط سازمان‌های هواشناسی، اورژانس و پزشکی قانونی برای بررسی تصادفات در شهر قم استفاده شده است. در این مجموعه‌داده، یازده ویژگی شامل شدت تصادف و عوامل مختلف محیطی مرتبط با آن بررسی شده‌اند. باتوجه‌به احتمال وجود شرایط متغیر محیطی و اهمیت آن در تعیین شدت تصادف، این مجموعه‌داده، در بازه‌های زمانی سه ساعتی مورد تجزیه و تحلیل قرار گرفته است. اطلاعات مربوط به تعداد کل حوادث رانندگی شهر قم در شکل ۲ نمایش‌داده شده‌اند. در این نمودار، هر میله نمایانگر تعداد تصادفات درون‌شهری در یک نوع خاص از برخورد در سال ۲۰۱۸ است. با استفاده از این نمودار، تعداد تصادفات عابر پیاده برای ارزیابی سطح ایمنی مورد استفاده قرار گرفته است.

۳-۲- پیش‌پردازش داده‌ها

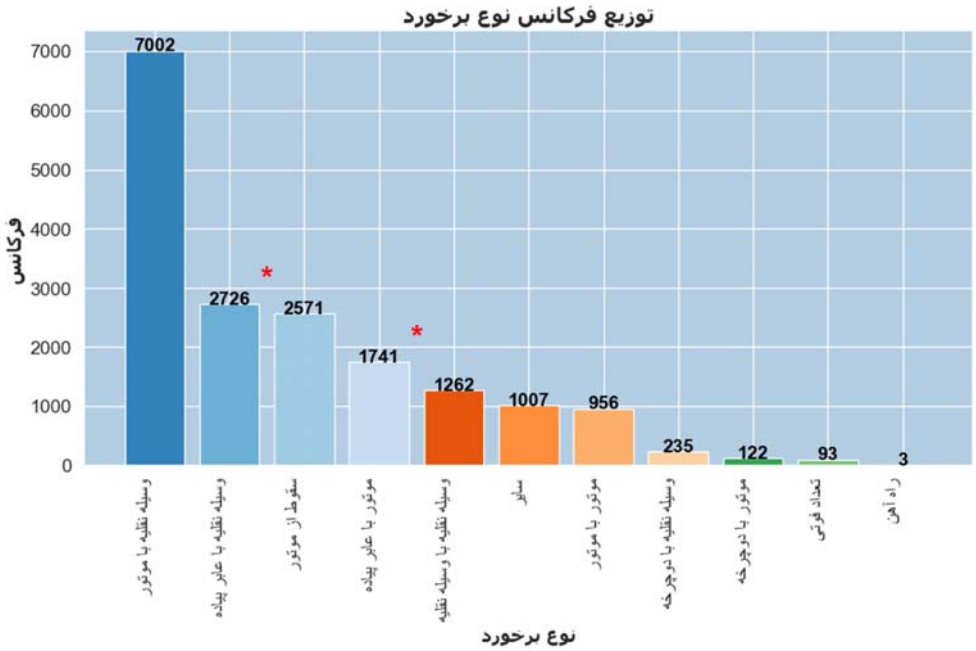
یک مرحله مهم در فرایندهای یادگیری ماشین و استخراج داده است. در این مرحله، داده‌های خام به یک قالب قابل‌استفاده تبدیل می‌شوند که می‌تواند توسط الگوریتم‌های یادگیری ماشین تحلیل شود. این فرایند شامل چند مرحله است که از جمله پاک‌سازی داده‌ها، نرمال‌سازی آن‌ها و تبدیل آن‌ها به شکلی قابل‌استفاده برای الگوریتم‌های یادگیری ماشین است.

۳-۳- مدل‌های یادگیری ماشین

در ادامه این تحقیق، تکنیک‌های معروف یادگیری ماشین که برای توسعه مدل‌های انتخاب ویژگی و طبقه‌بندی مورد استفاده قرار گرفته‌اند، به‌صورت خلاصه تعریف شده‌اند.



شکل ۱. روند انجام مطالعه



شکل ۲. نمودار میله‌ای تعداد تصادفات درون‌شهری

۳-۳-۱- مدل انتخاب روبه‌جلو سلسله‌مراتبی

این الگوریتم یک ابزار قدرتمند برای انتخاب بهینه ویژگی در مدل‌های یادگیری ماشین است و با موفقیت در بسیاری از برنامه‌های پردازش و طبقه‌بندی به کار گرفته شده است. این مدل، با یک مجموعه خالی از ویژگی‌ها شروع می‌کند و در هر مرحله، یک ویژگی به زیرمجموعه انتخاب شده اضافه می‌شود تا مقدار تابع معیار را به حداکثر برساند. در هر تکرار، الگوریتم ویژگی را که منجر به بیشترین افزایش در مقدار تابع معیار می‌شود، به زیرمجموعه اضافه می‌کند تا تعداد مورد نظر ویژگی‌ها انتخاب شود (Chotchantarakun & Sornil, ۲۰۲۱)

به عبارت بهتر این روش با استفاده از طبقه‌بند سعی می‌کند ویژگی‌هایی را انتخاب کند که اطلاعات جدیدی به مدل اضافه کنند و عملکرد بهتری را برای مدل به ارمغان آورند.

۳-۳-۲- الگوریتم K نزدیک‌ترین همسایه

در این الگوریتم، فاصله بین هدف و تمام نقاط در مجموعه آموزشی محاسبه می‌شود و K نزدیک‌ترین نقاط به هدف انتخاب می‌شوند. با استفاده از برچسب‌های نزدیک‌ترین K نقطه، پیش‌بینی خروجی برای هدف محاسبه می‌شود. برای ساخت مدل با استفاده از k نزدیک‌ترین همسایه، عامل K و تابع فاصله مورداستفاده نیاز است. برای محاسبه فاصله بین مشاهدات، از فاصله اقلیدسی(Niyogisubizo et al., 2023) استفاده می‌شود. فاصله اقلیدسی براساس قضیه‌ی فیثاغورس با معادله ۱ نشان داده می‌شود.

$$dij = \sqrt[2]{(x_i - x_j)^2 + (y_i + y_j)^2} \qquad (۱)$$

فاصله دونقطه i و j اندازه پاره‌خطی است که آنها را به هم متصل می‌کند.

۳-۳-۳- مدل درخت تصمیم

تکنیک درخت تصمیم یکی از پرکاربردترین الگوریتم‌های یادگیری ماشین است که در زمینه‌های مختلفی از جمله طبقه‌بندی و انتخاب ویژگی استفاده می‌شود. در این الگوریتم، داده‌ها به‌صورت یک درخت تصمیم ساختارمند نمایش داده می‌شوند و هر گره داخلی به‌عنوان یک آزمون از یک ویژگی عمل می‌کند و هر شاخه نتیجه آزمون را نشان می‌دهد. با پیمایش مسیر درخت از ریشه تا برگ، اطلاعات لازم برای پیش‌بینی رده جدید به دست می‌آید.
(Shiran, Imaninasab, & Khayamim, 2021).

به طور خلاصه، فرایند کلی درخت تصمیم شامل ساخت مدل با استفاده از داده‌های آموزشی و پیش‌بینی برچسب داده‌های آزمون با استفاده از این مدل است.

۳-۳-٤- الگوریتم جنگل تصادفی

یک الگوریتم یادگیری جمعی است که برای پیش‌بینی، از چندین درخت تصمیم‌گیری استفاده می‌کند. در این الگوریتم، هر درخت تصمیم، بر روی یک زیرمجموعه مختلف از داده‌های آموزشی آموزش داده می‌شود و پیش‌بینی نهایی بر اساس میانگین پیش‌بینی تمامی درخت‌ها صورت می‌گیرد.

با استفاده از الگوریتم جنگل تصادفی، می‌توان ویژگی‌های مهم را شناسایی کرد و از آن‌ها برای ساخت مدل استفاده نمود. این مدل باعث بهبود عملکرد و کاهش زمان آموزش می‌شود. (Liu, Tang, Gao, & Ding, 2023) علاوه بر این، الگوریتم جنگل تصادفی به دلیل کاربرد تصادفی متغیرها برای تولید درخت‌های طبقه‌بندی، توانایی خوبی برای مقاومت در برابر نویز دارد. این مدل نه‌تنها داده‌های گسسته، بلکه داده‌های پیوسته را نیز می‌تواند پردازش کند.

۳-٤- ارزیابی مدل

ماتریس درهم‌ریختگی یک ابزار ارزشمند برای ارزیابی عملکرد مدل‌های یادگیری ماشین است که با استفاده از آن، می‌توان اطلاعاتی درباره دقت و سایر عامل‌های عملکرد مدل بر روی داده‌های آزمون به دست آورد. این ماتریس به‌صورت یک جدول ۲x۲ نمایش داده می‌شود (جدول ۱) که در آن تعداد نمونه‌های واقعی و پیش‌بینی‌شده برای دودسته مثبت و منفی مشخص می‌شود. این جدول شامل چهار عامل است که عبارت‌اند از مثبت صحیح (TP)، منفی صحیح(TN)، مثبت کاذب[18] (FP) و منفی کاذب[19] (FN) . مثبت صحیح شامل تعداد نمونه‌هایی است که به‌درستی به‌عنوان مثبت شناخته شده‌اند، منفی صحیح شامل تعداد نمونه‌هایی است که به‌درستی به‌عنوان منفی شناخته شده‌اند. مثبت کاذب یا اتهام غلط شامل تعداد نمونه‌هایی است که به‌اشتباه به‌عنوان مثبت شناخته شده‌اند و منفی کاذب یا اعتماد غلط شامل تعداد نمونه‌هایی است که به‌اشتباه به‌عنوان منفی شناخته شده‌اند.

باتوجه‌به مطالعه (Fiorentini & Losa, 2020) معیار دقت تنها می‌تواند حد آموزش یافتن مدل را نشان دهد و اطلاعات کافی درباره کارایی و کیفیت کلی مدل را ارائه نمی‌دهد؛ بنابراین پس از هر تحلیل، ارزیابی مدل نباید به یک معیار محدود شود. برای ارزیابی دقیق عملکرد یک الگوریتم، علاوه بر معیار دقت، باید از دیگر معیارهای مندرج در جدول ۲، همچون صحت، حساسیت و برازش نیز استفاده کرد. این معیارها باتوجه‌به تعداد واقعی از نتایج درست و نادرست، عملکرد مدل را بهتر ارزیابی می‌کنند. معیار صحت نسبت تعداد نتایج مثبت واقعی به تعداد کل نتایج مثبت پیش‌بینی شده توسط مدل را محاسبه می‌کند و نشان می‌دهد که وقتی مدل یک نتیجه مثبت را پیش‌بینی می‌کند، این نتیجه تا چه اندازه درست است. معیار حساسیت یا فراخوانی نشان می‌دهد که الگوریتم چه تعداد از موارد مثبت کلاس را به‌درستی شناسایی کرده است. معیار برازش F۱ از ترکیب دو معیار دقت و حساسیت برای ارزیابی عملکرد الگوریتم‌های طبقه‌بندی استفاده می‌کند. با بهبود هر دو معیار دقت و حساسیت، مقدار معیار F۱ نیز بهبود می‌یابد. علاوه بر این معیار نشان می‌دهد که چقدر الگوریتم در تعیین نمونه‌های مثبت و منفی کلاس دقیق است.

در جدول (۳)، ماتریس الگوریتم‌های استفاده شده در مطالعه حاضر نمایش‌داده‌شده‌اند. عامل‌های موجود در این جدول شامل TP، TN، FP و FN با استفاده از نرم‌افزارهای متلب و پایتون محاسبه شده است.

جدول ۱. ماتریس درهم‌ریختگی و عامل‌های آن

		رده پیش‌بینی‌شده: مثبت	رده پیش‌بینی‌شده: منفی
رده واقعی	مثبت	مثبت صحیح (TP)	منفی کاذب (FN) خطای نوع دو (۲)
رده واقعی	منفی	مثبت کاذب (FP) خطای نوع یک (۱)	منفی صحیح (TN)

جدول ۲. معیارهای ارزیابی الگوریتم‌های یادگیری ماشین

معیار	معادله ریاضی
دقت (Accuracy)	$\frac{TP+TN}{TP+FP+TN+FN}$
صحت (Precision)	$\frac{TP}{TP+FP}$
حساسیت یا فراخوانی (Recall)	$\frac{TP}{TP+FN}$
برازش F_1 (F1-Score)	$\frac{2*Recall*Precision}{Recall+Precision}$

جدول ۳. ماتریس درهم‌ریختگی الگوریتم‌های طبقه‌بندی برای شدت تصادفات

عامل‌های ماتریس درهم‌ریختگی	الگوریتم k نزدیک‌ترین همسایه				الگوریتم درخت تصمیم		الگوریتم جنگل تصادفی	
	ویژگی‌های منتخب		تمام ویژگی‌ها					
	k=۵	k=۳	k=۵	k=۳	تمام ویژگی‌ها	ویژگی‌های منتخب	تمام ویژگی‌ها	ویژگی‌های منتخب
مثبت صحیح (TP)	۳۷۰	٤۳٤	۳٦٤	۳٦۲	٤۳٦	۳۷٤	۳٦۰	۳۷۵
منفی صحیح (TN)	۳٤۳	۲۸۳	۳٤۱	۳۰۸	۲۹۰	۳۲٦	۳۵۰	۳٤۷
مثبت کاذب (FP)	۳۱۱	۲٤۷	۳۱۷	۳۱۹	۲۵۵	۳۱۷	۲۹۹	۳۰۲
منفی کاذب (FN)	۳۱٦	۳۷٦	۳۱۸	۳۵۱	۳۵۹	۳۲۳	۳۳۱	۳۱٦

با استفاده از جداول ۲ و ۳، می‌توان به‌راحتی میزان کارایی الگوریتم‌های یادگیری ماشین را در برابر داده‌های آموزش و آزمون اندازه‌گیری کرد. از این جداول برای محاسبه معیارهای ارزیابی در جدول ۵ بخش بحث و نتایج استفاده می‌شود.

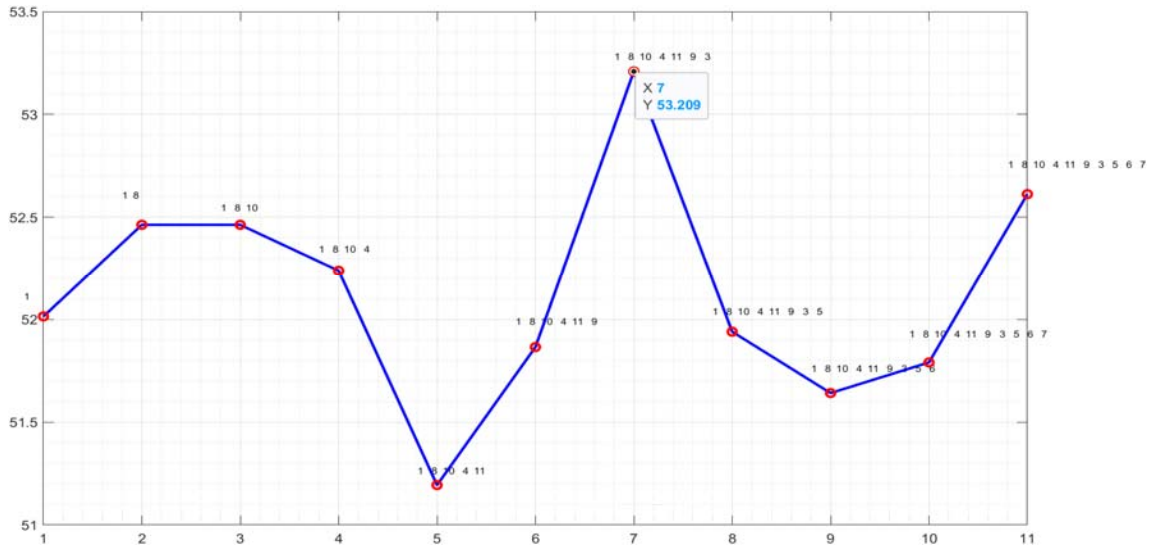
٤- بحث و نتایج

در این پژوهش، یک رویکرد یکپارچه برای مقایسه عملکرد مدل‌های مختلف ارائه شده است. در این رویکرد، ابتدا با استفاده از تمامی ویژگی‌ها، عملکرد سیستم ارزیابی می‌شود. سپس متغیرهایی که بیشترین تأثیر را دارند با استفاده از روش انتخاب ویژگی تعیین می‌شوند. در مرحله بعد، پروژه با استفاده از

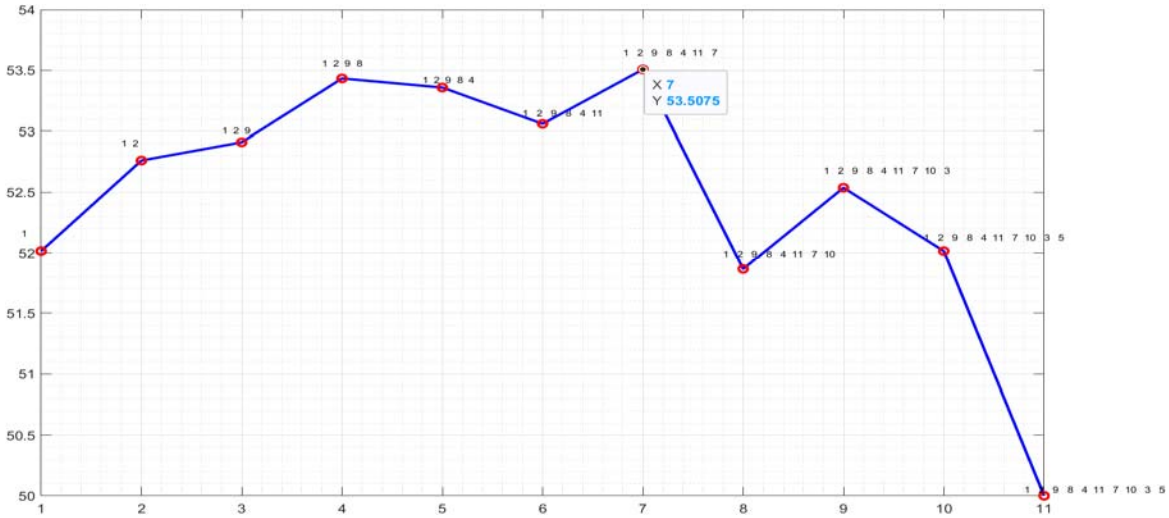
ویژگی‌های انتخاب شده مجدداً اجرا شده و در نهایت عملکردها مقایسه می‌شوند. به‌منظور دستیابی به اهداف تعیین شده دسته‌بندی‌های زیر انجام شده است.

٤-۱- روش‌های انتخاب ویژگی

در این مطالعه، از روش‌های یادگیری ماشین مختلفی مانند انتخاب روبه‌جلو سلسله‌مراتبی (SFFS)، جنگل تصادفی و درخت تصمیم برای انتخاب ویژگی استفاده شده است. بهترین ترکیب ویژگی‌ها، با استفاده از الگوریتم SFFS در شکل ۳ و ٤ مشخص شده است. محور عمودی نمودارها به معنای عملکرد مدل و محور افقی به معنای تعداد ویژگی‌های ترکیبی است. نتایج نشان داد که استفاده از ۷ ویژگی سبب بهبود حداکثری عملکرد در هر دو مدل می‌شود؛ اما با افزایش بیشتر تعداد ویژگی‌ها، عملکرد کاهش می‌یابد. ویژگی‌هایی مانند زمان، فصل، رطوبت، بارش، دید، روز – هفته و دما نسبت به سایر عامل‌های تعیین‌کننده، عملکرد بهتری در الگوریتم نزدیک‌ترین همسایه با k=۳ داشتند. همچنین، نتایج الگوریتم نزدیک‌ترین همسایه با k=۵ نشان داد ویژگی‌های، بارش، روشنایی، دید، روز–هفته، رطوبت و نوع تصادف نسبت به سایر متغیرها، عملکرد بهتری دارند



شکل ۳. نمودار انتخاب ویژگی SFFS وابسته به مدل نزدیک‌ترین همسایه با k=۳

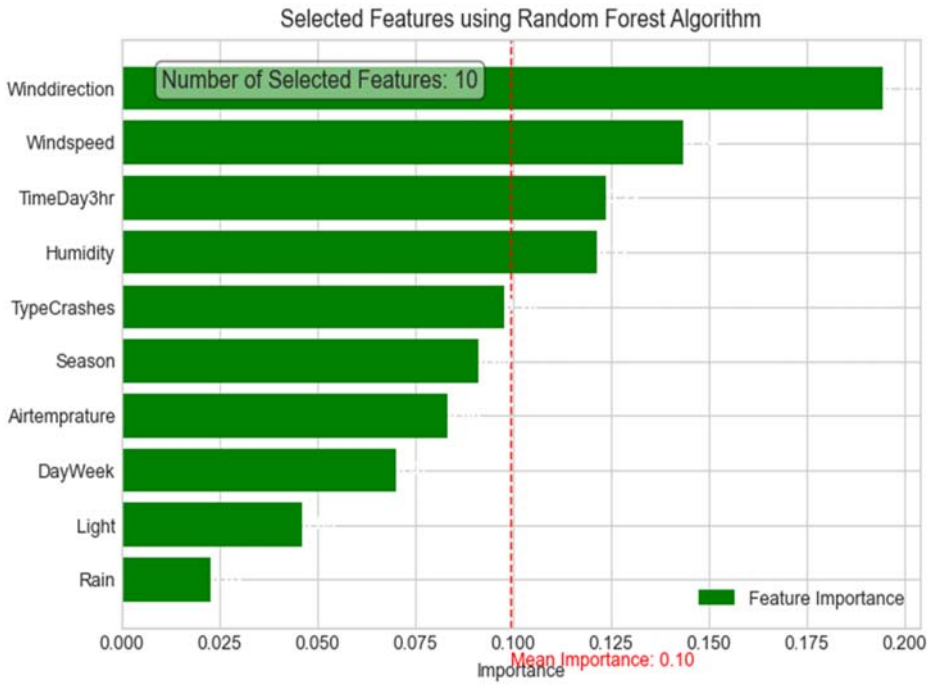


شکل ٤. نمودار انتخاب ویژگی SFFS وابسته به مدل نزدیک‌ترین همسایه با k=۵

جدول ٤. راهنمای نتایج نمودار

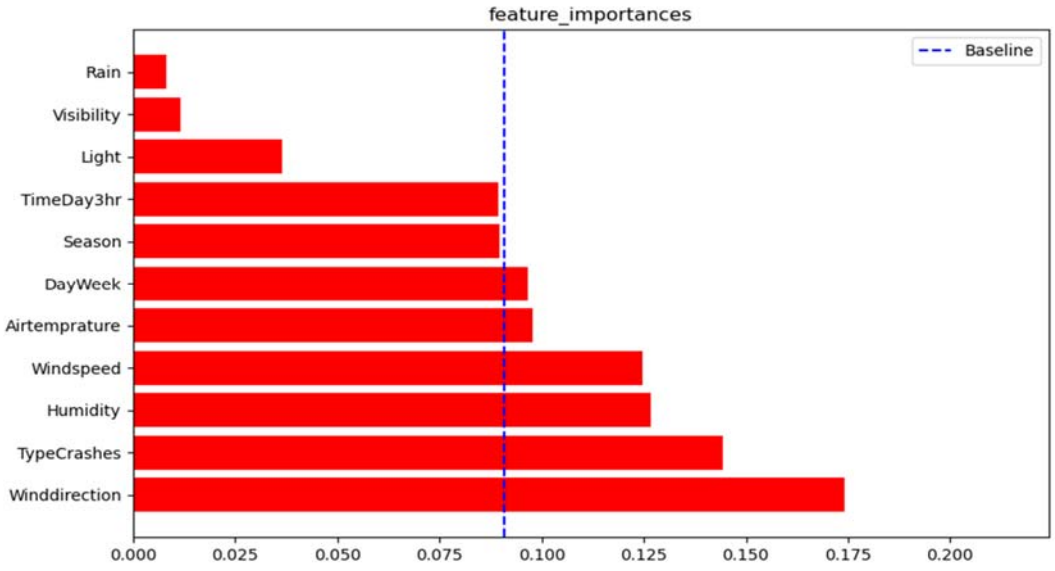
نتایج نمودار	توضیح	نتایج نمودار	توضیح
۱	زمان (Time Day)	۷	دما (Air Temperature)
۲	فصل (Season)	۸	بارش (Rain)
۳	نوع تصادف (Type Crashes)	۹	رطوبت (Humidity)
٤	دید (Visibility)	۱۰	روشنایی (Light)
٥	سرعت باد (Wind Speed)	۱۱	روز- هفته (Day Week)
٦	جهت باد (Wind Direction)	۱۲	شدت آسیب (Injury Severity)

استفاده از جنگل تصادفی برای انتخاب ویژگی، از این جهت مزیت دارد که علاوه بر این که می‌تواند ویژگی‌های مهم را شناسایی کند، قابلیت تعیین اهمیت ویژگی‌ها را نیز ارائه می‌دهد. جنگل تصادفی برای محاسبه اهمیت ویژگی‌ها از دو روش استفاده می‌کند: میزان کاهش در دقت و میزان کاهش در خطای جینی. در روش اول، با حذف هر ویژگی، دقت مدل پیش‌بینی محاسبه می‌شود و ویژگی‌هایی که با حذف آن‌ها دقت مدل بیشتر کاهش پیدا می‌کند، به‌عنوان ویژگی‌های مهم شناسایی می‌شوند. در روش دوم، با حذف هر ویژگی، خطای جینی محاسبه می‌شود و ویژگی‌هایی که با حذف آن‌ها خطای جینی مدل بیشتر کاهش پیدا می‌کند، به‌عنوان ویژگی‌های مهم شناسایی می‌شوند. در این مطالعه، از دقت مدل برای انتخاب ویژگی استفاده شده است و باتوجه‌به بهترین عملکرد مدل، تعداد ویژگی‌های انتخابی به ۱۰ مورد محدود شده است. (شکل ٥)



شکل ٥. اهمیت ویژگی‌های انتخاب شده توسط الگوریتم جنگل تصادفی

در بسیاری از مقالات، نمودار اهمیت ویژگی‌ها قبل از انتخاب ویژگی‌های مهم مورداستفاده قرار می‌گیرد. نتایج اهمیت هر ویژگی توسط الگوریتم درخت تصمیم در شکل ٦ نمایش‌داده‌شده است. این نمودار نشان می‌دهد که ویژگی‌هایی مانند جهت باد، نوع تصادف، میزان رطوبت و سرعت باد برای پیش‌بینی متغیر هدف از اهمیت بالاتری برخوردار هستند



شکل ٦. اهمیت ویژگی با مدل درخت تصمیم

در ادامه و بعد از انتخاب ویژگی توسط مدل درخت تصمیم، ویژگی‌های زمان و نوع تصادف به‌عنوان عوامل مهم در پیش‌بینی شدت جراحت شناسایی شدند. این نشان می‌دهد که ویژگی جدیدی مانند زمان همچنان می‌تواند برای طبقه‌بندی مفید باشد، حتی اگر در نمودار اهمیت قبلی جزء بااهمیت‌ترین ویژگی‌ها شناخته نشده باشد. دلیل این امر، قابلیت مدل در شناسایی ویژگی‌های جدید با الگوریتم خود و استفاده از آن‌ها در پیش‌بینی است.

نتیجهٔ دیگری که پس از انتخاب ویژگی توسط مدل درخت تصمیم در اینجا مطرح می‌شود این است که استفاده از دو ویژگی زمان و نوع تصادف، می‌تواند به سازمان‌ها کمک کند تا هزینه و زمان زیادی را برای جمع‌آوری و بررسی ویژگی‌های دیگر صرف نکنند. بااین‌حال، ممکن است استفاده از این دو ویژگی به‌تنهایی، باعث ازدست‌رفتن اطلاعات مهم و افزایش خطا در طبقه‌بندی شود. جهت روشن‌شدن مطلب در بخش ارزیابی الگوریتم‌های یادگیری، به این پرسش پاسخ خواهیم داد که آیا استفاده از ویژگی‌های زمان و نوع تصادف به‌تنهایی باعث افزایش دقت در طبقه‌بندی شدت تصادفات می‌شود یا خیر؟

٤-۲- مقایسه و ارزیابی الگوریتم‌های یادگیری ماشین برای طبقه‌بندی داده‌ها

بهترین روش برای ارزیابی الگوریتم‌های یادگیری ماشین، استفاده از داده‌های آزمون و ارزیابی کارایی الگوریتم بر روی این داده‌هاست. برای این منظور، معیارهایی مانند دقت، صحت، فراخوانی و برازش F_1 روی داده‌های آزمون محاسبه می‌شوند. این معیارها به همراه نوع هر الگوریتم و تعداد ویژگی‌های مورداستفاده آن مدل در جدول ٥ نشان‌داده‌شده است.

برای نمایش رتبه‌بندی‌های جدول، از رنگ‌های سبز، قرمز و آبی استفاده شده است. رتبه‌بندی‌هایی که با رنگ سبز مشخص شده‌اند، بهترین رتبه‌بندی‌ها هستند و رتبه‌بندی‌هایی که با رنگ قرمز مشخص شده‌اند، بدترین رتبه‌بندی‌ها هستند. رتبه‌بندی‌هایی که با رنگ آبی مشخص شده‌اند، متوسط رتبه‌بندی‌هایی هستند که در مقایسه با سایر الگوریتم‌ها، کمتر موفق بوده‌اند. این جدول کمک می‌کند تا بر اساس نیاز، بهترین مدل انتخاب شود.

جدول ۵. تحلیل مقایسه‌ای نتایج الگوریتم‌های یادگیری ماشین بر اساس معیارهای فراخوانی، دقت، برازش F_1 و صحت

الگوریتم های طبقه‌بندی یادگیری ماشین	تعداد ویژگی‌ه ا	صحت (درصد)	رتبه‌بندی صحت	برازش F_1	رتبه‌بندی برازش F_1	دقت (درصد)	رتبه‌بند ی دقت	فراخوانی (درصد)	رتبه‌بندی فراخوانی
K نزدیک ترین همسایه با k=۳	تمام ویژگی‌ها	۱۵۷۱/۵۳	۷	۹۳۶۹/۵۱	۸	۵۰	۸	۷۷۱۴/۵۰	۸
	۷ ویژگی منتخب	۷۲۹۸/۶۳	۱	۲۱۵۹/۵۸	۲	۵۰۷۵/۵۳	۳	۵۸۰۲/۵۳	۶
K نزدیک ترین همسایه با k=۵	تمام ویژگی‌ها	۴۵۰۸/۵۳	۶	۴۱۱۶/۵۳	۶	۶۱۱۹/۵۲	۶	۳۷۲۴/۵۳	۷
	۷ ویژگی منتخب	۳۳۱۹/۵۴	۳	۱۳۳۲/۵۴	۴	۲۰۸۹/۵۳	۴	۹۳۵۹/۵۳	۴
درخت تصمیم	تمام ویژگی‌ها	۰۹۶۹/۶۳	۲	۶۸۱۰/۵۸	۱	۱۷۹۱/۵۴	۱	۸۴۲۸/۵۴	۲
	۲ ویژگی منتخب	۱۲۴۵/۵۴	۵	۸۹۰۵/۵۳	۵	۲۳۸۸/۵۲	۷	۶۵۸۵/۵۳	۵
جنگل تصادفی	تمام ویژگی‌ها	۰۹۸۴/۵۲	۸	۳۳۳۳/۵۳	۷	۹۸۵۱/۵۲	۵	۶۲۸۲/۵۴	۳
	۱۰ ویژگی منتخب	۲۶۹۲/۵۴	۴	۸۲۴۵/۵۴	۳	۸۸۰۶/۵۳	۲	۳۹۱۴/۵۵	۱

باتوجه‌به جدول ارائه شده، مشاهده می‌شود که انتخاب ویژگی می‌تواند بهبود دقت مدل را به دنبال داشته باشد. در واقع، اگر برخی از ویژگی‌ها برای مدل طبقه‌بندی اهمیت کمتری داشته باشند و یا با ویژگی‌های دیگر همبستگی زیادی داشته باشند، حذف آن‌ها می‌تواند به بهبود دقت مدل منجر شود. تنها در مورد الگوریتم درخت تصمیم، نتایج نشان داده‌اند که استفاده از تمام ویژگی‌های موجود، بهترین عملکرد را برای این الگوریتم به دنبال دارد. دلیل اینکه با انتخاب ویژگی‌های مؤثر بهبودی در دقت این روش حاصل نشده است، ممکن است به دلایل کم‌بودن تعداد نمونه‌های داده، نوع داده‌ها و نوع الگوریتم مورداستفاده باشد. البته برای توسعه این مطالعه می‌توان با استفاده از تنظیمات مختلف روش درخت تصمیم، دقت مدل را بهبود بخشید. یکی از راه‌های افزایش دقت، تغییر عامل‌های روش انتخاب ویژگی است. می‌توان با تغییر مقدار آستانه بهترین حد را برای انتخاب ویژگی‌های مهم پیدا کرد. همچنین، می‌توان عامل‌های درخت تصمیم را تنظیم نمود تا بهترین عملکرد را برای داده‌ها داشته باشد. برای مثال، می‌توان محدودیت عمق درخت، تعداد نمونه در هر گره و حداقل تعداد نمونه در یک برگ را تغییر داد. این عامل‌ها را باید بر اساس داده‌ها تنظیم نمود تا بتوان دقت روش انتخاب ویژگی را بهبود بخشید.

٤-۳- نتایج مقایسه معیارهای ارزیابی

باتوجه‌به جدول ۳، معیارهای ارزیابی مدل‌های مختلف با یکدیگر مقایسه شده و مدل‌ها بر اساس عملکرد خود در هر معیار، از بالاترین به پایین‌ترین رتبه قرار گرفته‌اند. بر اساس نتایج به‌دست‌آمده، مدل‌های نزدیک‌ترین همسایه با $K=۳$ و درخت تصمیم با تمام ویژگی‌ها صحت آزمون بسیار بهتری با نمرات ۷۲۹۸/٦۳ و ۰۹٦۹۰/٦۳ به دست آوردند. علاوه بر این، در معیار دقت، درخت تصمیم با تمام ویژگی‌ها بهترین عملکرد را داشت و مدل‌های جنگل تصادفی و نزدیک‌ترین همسایه با $k=۳$ همراه با ویژگی‌های منتخب خود به ترتیب در جایگاه دوم و سوم قرار گرفتند. در معیار فراخوانی، روش جنگل تصادفی با ویژگی‌های انتخاب شده خود بهترین عملکرد را داشت و درخت تصمیم در جایگاه دوم قرار گرفت. باتوجه‌به معیار برازش $F۱$، الگوریتم درخت تصمیم بهترین نتیجه را به دست آورد و پس از آن به ترتیب مدل‌های نزدیک‌ترین همسایه با $k=۳$ و جنگل تصادفی با ویژگی‌های منتخب خود قرار گرفتند.

با استفاده از نتایج فوق، اگر یک سازمان تصمیم بگیرد که از بودجه ایمنی خود به بهترین شکل ممکن استفاده کند، دقت و امتیاز $F۱$ دارای اهمیت است. باتوجه‌به این معیار، الگوریتم درخت تصمیم احتمالاً یک انتخاب مناسب برای اهداف سازمان خواهد بود.

٥-نتیجه‌گیری

این مطالعه به بررسی کارایی الگوریتم‌های یادگیری ماشین در ساخت مدل طبقه‌بندی می‌پردازد. در این تحقیق، عملکرد چندین طبقه‌بند از جمله الگوریتم‌های k نزدیک‌ترین همسایه، درخت تصمیم و جنگل تصادفی بررسی شده‌اند. این الگوریتم‌ها با استفاده از مجموعه‌داده با تمام ویژگی‌ها و با استفاده از زیرمجموعه‌هایی از ویژگی‌ها که توسط روش‌های انتخاب ویژگی کاهش یافته‌اند، تولید و مقایسه شده‌اند.با درنظرگرفتن محدودیت بودجه، نتایج این مطالعه نشان می‌دهد که الگوریتم درخت تصمیم با استفاده از تمام ویژگی‌ها، در طبقه‌بندی شدت تصادفات نسبت به سایر الگوریتم‌ها بهترین کارایی را ارائه می‌دهد. همچنین در اکثر موارد، با استفاده از روش‌های انتخاب ویژگی دقت مدل افزایش می‌یابد و اثرات مفیدی برای بهبود عملکرد طبقه‌بندها ایجاد می‌شود. بااین‌حال، برای مدل طبقه‌بندی درخت تصمیم، انجام انتخاب ویژگی توصیه نمی‌شود؛ زیرا مشاهده گردید که باعث ازدست‌رفتن اطلاعات مهم و کاهش دقت آن می‌شود.

در تحقیقات آینده، بهتر است توجه بیشتری به بررسی تأثیر عوامل مختلفی مانند شرایط ترافیک، ویژگی‌های جاده، وضعیت آب‌وهوا، و ویژگی‌های کاربران جاده مانند سن، جنسیت، سطح تحصیلات صورت پذیرد. علاوه بر این، بررسی میزان تأثیر هر یک از این عوامل بر شدت و فراوانی تصادفات جاده‌ای نیز باید موردتوجه قرار گیرد تا بر اساس آن راهکارهای مناسبی برای کاهش آنها ارائه شود. این اقدامات می‌توانند به دقیق‌تر شدن نتایج و بهبود کارایی الگوریتم‌های طبقه‌بندی در مسائل مرتبط با تصادفات کمک کنند.

٦-مراجع

-Ali, N. M., Farouk, A. I. B., Haruna, S. I., Alanazi, H., Adamu, M., & Ibrahim, Y. E. (2022). Feature selection approach for failure mode detection of reinforced concrete bridge columns. *Case Studies in Construction Materials, 17*, e01383.

-AlMamlook, R. E., Kwayu, K. M., Alkasisbeh, M. R., & Frefer, A. A. (2019). Comparison of machine learning algorithms for predicting traffic accident severity.

-Chotchantarakun, K., & Sornil, O. (2021). An Adaptive Multilevels Sequential Feature Selection. *International Journal of Computer Information Systems and Industrial Management Applications, 13*, 010-019.

-Niyogisubizo, J., Liao, L., Sun, Q., Nziyumva, E., Wang, Y., Luo, L., . . . Murwanashyaka, E. (2023). Predicting Crash Injury Severity in Smart Cities: a Novel Computational Approach with Wide and Deep Learning Model. *International Journal of Intelligent Transportation Systems Research, 21*(1), 240-258.

-Shiran, G., Imaninasab, R., & Khayamim, R. (2021). Crash severity analysis of highways based on multinomial logistic regression model, decision tree techniques, and artificial neural network: a Modeling comparison. *Sustainability, 13*(10), 5670.

-Singh, S., & Haider, T. U. (2022, 2022). Selection of best feature reduction method for module-based software defect prediction.

-Tsai, C.-F., & Chen, Y.-C. (2019). The optimal combination of feature selection and data discretization: An empirical study. *Information Sciences, 505*, 282-293.

-Vandana, C. P., & Chikkamannur, A. A. (2021). Feature selection: An empirical study. *International Journal of Engineering Trends and Technology, 69*(2), 165-170.

-Yang, H., Huang, S., Guo, S., & Sun, G. (2022). Multi-Classifier Fusion Based on MI–SFFS for Cross-Subject Emotion Recognition. *Entropy, 24*(5), 705.

-Zhou, H., Zhang, J., Zhou, Y., Guo, X., & Ma, Y. (2021). A feature selection algorithm of decision tree based on feature weight. *Expert Systems with Applications, 164*, 113842.

-Danaei Mehr, H., & Polat, H. (2022). Diagnosis of polycystic ovary syndrome through different machine learning and feature selection techniques. *Health and Technology, 12*(1), 137-150.

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

-Got, A., Moussaoui, A., & Zouache, D. (2021). Hybrid filter-wrapper feature selection using whale optimization algorithm. A multi-objective approach. *Expert Systems with Applications, 183*, 115312.

-Komol, M. M. R., Hasan, M. M., Elhenawy, M., Yasmin, S., Masoud, M., & Rakotonirainy, A. (2021). Crash severity analysis of vulnerable road users using machine learning. *PLoS one, 16*(8), e0255828.

-Kunhare, N., Tiwari, R., & Dhar, J. (2020). Particle swarm optimization and feature selection for intrusion detection system. *Sādhanā, 45*, 1-14.

-Liu, X., Tang, J., Gao, F., & Ding, X. (2023). Time and Distance Gaps of Primary-Secondary Crashes Prediction and Analysis Using Random Forests and SHAP Model. *Journal of Advanced Transportation.*

-Nguyen, C., Wang, Y., & Nguyen, H. N. (2013). Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic.

Comparison of Machine Learning Algorithms for Feature Selection and Severity Classification of Pedestrian Accidents (Case Study: Qom City)

Seyed Mehdi Moslemi Mahani, Ph.D., Candidate, Department of Civil Engineering, SR.C., Islamic Azad University, Tehran, Iran.

Ali Naderan, Assistant Professor, Department of Civil Engineering, SR.C., Islamic Azad University, Tehran, Iran.

Seyed Saber Naser Alavi, Assistant Professor, Civil Engineering Department, Shahid Bahonar University of Kerman, Kerman, Iran.

E-mail: naderan@iau.ac.ir

Received: May 2025- Accepted: August 2025

ABSTRACT

This article investigates the performance of machine learning algorithms in classifying the severity of pedestrian accidents, taking into account various factors such as weather conditions, environment, time periods, and accident types. Initially, feature combinations were optimized using real-world data from Qom City to improve classification accuracy. The algorithms were trained on 70% of the data and tested on the remaining 30% to evaluate their effectiveness. The performance of three popular classification algorithms, including k-Nearest Neighbor, Decision Tree, and Random Forest, was compared using all features and subsets of features. The results revealed that the Decision Tree algorithm with all features achieved the highest performance in classifying the severity of pedestrian accidents. Future studies should examine the impact of various factors, such as traffic conditions and road features, along with weather conditions, on pedestrian accidents to enhance the performance of classification algorithms in this area.

Keywords: Feature Selection, Classification, Confusion Matrix, Machine Learning