تحلیل شدت مصدومیت ناشی از تصادفها در راههای دوخطه برون‌شهری با استفاده از مدل‌های داده‌کاوی

افشین شریعت‌مهیمنی*، استادیار، دانشکده عمران، دانشگاه علم و صنعت ایران، تهران، ایران
علی توکلی‌کاشانی، دانشجوی دکتری، دانشکده عمران، دانشگاه علم و صنعت ایران، تهران، ایران
E-mail: shariat@iust.ac.ir
دریافت: ۱۳۸۸/۰۹/۲۱– پذیرش:۱۳۸۹/۰۳/۱۲

چکیده

راههاي دوخطه دوطرفه، بخش عظيمي از شبكه برون‌شهري اكثر كشورهاي دنيا را تشكيل مي‌دهند كه كشور ايران نيز از اين امر مستثني نيست. در اين مطالعه شناسايي عوامل مؤثر بر افزايش شدت مصدوميت رانندگان و سرنشينان ناشي از تصادفها در اين راهها، مورد تجزيه و تحليل قرار گرفته است. استفاده از مدل‌های آماری، یکی از متداول‌ترین روش‌هایی است که تاکنون در مدل‌های شدت تصادفها بکارگرفته شده است. در این مطالعه، ضمن بررسی این مدل‌ها و بیان نقاط ضعف آنها، روش کارت معرفی می‌شود. روش كارت كه يكي از پركاربردترين روش‌هاي داده‌كاوي است، براي تحليل آمار تصادفهاي سه سال (۱۳۸۵-۱۳۸۷) مورد استفاده واقع شد. از خروجي‌های بسیار مفید این روش، ارایه مدل‌ها در قالب درخت‌های تصمیم است که می‌تواند نتایج را به صورت قوانین واضحی بیان کند. در روند مدل‌سازی با تبدیل متغیر پاسخ سه‌تایی به مدل‌هایی با متغیر پاسخ دوتایی، هشت مدل مورد تجزیه و تحلیل قرار گرفتند. تقریباً در همه مدل‌ها وضعیت استفاده از کمربند ایمنی و علت تامه تصادف به عنوان مهم‌ترین متغیرها در افزایش شدت مصدومیت شناسایی شدند.

واژه‌های کلیدی: داده‌کاوی، درخت دسته‌بندی و رگرسیون (CART)، شدت مصدومیت، راههای دوخطه برون‌شهری

۱. مقدمه

هزینه بسیار زیاد تصادفهای ترافیکی، باعث شده که بهبود وضع ایمنی راهها، به یکی از مهم‌ترین اهداف مهندسی حمل و نقل در اکثر کشورها تبدیل شود. در کشور ایران، طی ۳ سال اخیر به طور متوسط، سالیانه حدود ۲٤۰۰۰ نفر در تصادفهای ترافیکی جان خود را از دست داده‌اند که از این مقدار بیش از ۷۰٪ مربوط به راههای برون‌شهری و بقیه مربوط به نواحی درون شهری بوده است [سازمان پزشکی قانونی کشور، ۱۳۸۸]. بررسی تصادفهای برون‌شهری نشان می‌دهد که ۵۰ تا ٦۰ درصد آنها در راههای دوخطه[1] اتفاق می‌افتند [صفارزاده، ۱۳۸٦]. در ایران، بخش عظیمی از شبکه راههای برون‌شهری از این نوع راهها تشکیل می‌شده که شناسایی عوامل مؤثر بر افزایش شدت تصادفها در این راهها، نه تنها می‌تواند باعث جلوگیری از فوت افراد درگیر در تصادفها شود، بلکه از وقوع مصدومیت‌های شدید نیز جلوگیری می‌کند. در ایران بیش از ۹۰ درصد جابجایی مسافر، در مد حمل و نقل جاده‌ای انجام می شود که این موضوع در کنار سهم بالای تصادفها، لزوم بررسی وضعیت ایمنی سرنشینان وسایل نقلیه اعم از رانندگان و مسافران را دوچندان می‌کند.

هدف اصلی این مطالعه، شناسایی مهم‌ترین عوامل تعیین‌کننده در شدت مصدومیت[2] رانندگان و سایر سرنشینان وسایل نقلیه درگیر در تصادفها برای راههای اصلی دوخطه برون‌شهری ایران با استفاده از داده‌کاوی[3] است.

داده‌کاوی، اکتشاف و تحلیل حجم زیادی از داده‌ها برای کشف الگوها و قواعد معنی‌دار است [Han (et. al.), 2006]. با توجه به حجم زیاد داده‌های تصادفهای راههای برون‌شهری کشور، در این مطالعه که برای ۱٦۹٦٤۸ داده مربوط به راننده و ۷۲٤۱ داده مربوط به سرنشین انجام شد، استفاده از داده‌کاوی، ابزاری مناسب تشخیص داده شد. این مطالعه اولین تحقیق از نوع خود در کشور است که در آن داده‌کاوی داده‌های تصادفها، انجام شده است.

*نویسنده مسئول

تاکنون اکثر مطالعاتی که در این حوزه انجام شده‌اند، از مدل‌های رگرسیون خطی تعمیم یافته[4] استفاده کرده‌اند که رابطه بین شدت مصدومیت و عوامل تصادفها را خطی فرض می‌کنند. این پیش‌فرض می‌تواند باعث خطا در تخمین احتمال شدت مصدومیت شود [Mussone (et. al.), 1999]. از سوی دیگر هنگامی که تعداد متغیرهای مستقل و ارتباط بین آنها زیاد باشد، باعث افزایش خطا در مدل آماری می‌شود که این مهم برای داده‌های تصادفها بیشتر اتفاق می‌افتد، چرا که فرم ثبت اطلاعات تصادفها، حاوی متغیرهای متعدد ترافیکی، هندسی و انسانی است که در بسیاری مواقع با یکدیگر رابطه تنگاتنگی دارند، برای مثال وضعیت هوا و شرایط روسازی و یا نوع وسیله نقلیه و نوع گواهینامه. برای توسعه مدل‌های آماری لازم است که قبل از آن، فرم تابع[5] مدل را فرض کرد. اگر فرم تابع مدل درست فرض نشود، باعث ایجاد خطا در بیان رابطه متغیرهای مستقل با متغیر پاسخ خواهد شد، ضمن آنکه پیش‌بینی‌های مدل، آمیخته با مقدار زیادی خطا خواهد شد. وجود داده‌های پرت از دیگر مواردی است که مدل‌های آماری شدت تصادفها را با چالش مواجه می‌کند.

مدل کارت[6] که یک مدل ناپارامتری و بدون هرگونه پیش فرض در خصوص رابطه بین متغیرهای مستقل[7] و متغیر هدف[8] است و از روش‌های مهم داده‌کاوی است، به طور گسترده در تجارت، صنعت، مهندسی و سایر علوم استفاده شده است. مدل کارت، ابزاری قدرتمند در تعیین مهم‌ترین متغیرهای مستقل و حل مسایل دسته‌بندی و پیش‌بینی[9] است.

در این تحقیق با استفاده از روش کارت مهم‌ترین متغیرهای مستقل اثرگذار بر شدت مصدومیت سرنشینان وسایل نقلیه، شناسایی شدند. در مدل‌های به دست آمده، علت تامه تصادف و وضعیت استفاده از کمربند ایمنی، از عوامل تعیین‌کننده در شدت مصدومیت بودند، به نحوی که سرنشینان درگیر در تصادفهایی که علت‌تامه آنها انحراف به چپ ناشی از سبقت بوده و از کمربند ایمنی استفاده نکرده بودند، به احتمال بیشتر، دچار مصدومیت شدیدتر شده و یا فوت گردیده‌اند.

در ادامه این بخش، به مطالعات پیشین در این زمینه پرداخته شده است. بخش سوم به روش تحقیق و معرفی روش کارت و نحوه محاسبه اهمیت متغیرها می‌پردازد. بخش چهارم، داده‌ها را توصیف کرده و بخش پنجم به تجزیه و تحلیل آنها می‌پردازد و نهایتاً بخش ششم نتایج این تحقیق را بیان می‌کند.

۲. مطالعات پیشین

تاکنون مطالعات زیادی در خصوص پیش‌بینی و مدل‌سازی شدت تصادفها و نیز شدت مصدومیت وارده بر افراد درگیر در تصادفها، انجام شده است. اکثر این مطالعات با هدف شناسایی عوامل مهم و اثرگذار بر افزایش شدت تصادف صورت گرفته است تا با حذف یا کنترل آنها، از وقوع تصادفهای شدید و به تبع آن مصدومیت‌های شدید و منجر به فوت جلوگیری شود.

از دیدگاه نظری، مدل‌های آماری لوجیت[10] یکی از پرکاربردترین مدل‌های بکاررفته در تحلیل شدت تصادفها و یا مصدومیت هستند [(Al-Ghamdi, 2002), Bedard (et. al.), 2002 Valent (et. al.), 2002, Wood (et. al.), 2002, (Yau, 2004)]. به عنوان مثال مطالعه‌ای در عربستان نشان داد که موقعیت محل وقوع و علت تامه تصادف، از عوامل مهم در افزایش شدت تصادف هستند (Al-Ghamdi, 2002). وود و سیمز اثر ابعاد وسیله نقلیه [Wood (et. al.), 2002] و بدارد و ولنت، کمربند ایمنی و کلاه ایمنی را، عوامل تعیین‌کننده در شدت مصدومیت تشخیص دادند [Bedard (et. al.), 2002, Valent (et. al.), 2002]. مدل‌های پروبیت ترتیبی[11]، دسته‌ای دیگر از مدل‌های آماری هستند که در مدل‌سازی شدت تصادف عمومیت یافته‌اند [Renski (et. al.), 1999, Khattak (et. al.), 2002, Kockelman (et. al.), 2002, Kweon (et. al.), 2003, Zajac (et. al.), 2003, Abdel-Aty (et. al.), 2005]. به عنوان مثال؛ وِان، نشان داد که استفاده از کمربند ایمنی، ریسک مصدومیت در تصادفها را کاهش می‌دهد [Kweon (et. al.), 2003]. رِنسکی، افزایش سرعت مجاز را باعث افزایش شدت مصدومیت دانست [Renski (et. al.), 1999]. زاجاک، به بررسی شرایط محل تصادف و شدت تصادفهای عابرین پیاده در راههای برون‌شهری پرداخت [Zajac (et. al.), 2003].

در سال‌های اخیر پژوهشگران مختلف سعی کرده‌اند با استفاده از روش‌های ناپارامتری و ابزارهای داده‌کاوی، شدت تصادفها را پیش‌بینی کنند. این پیش‌بینی که براساس تکنیک‌های دسته‌بندی داده‌ها انجام می‌شود، از میان حجم عظیم داده‌ها، الگوهایی را استخراج می‌کند که شدت تصادفها برحسب غیرجرحی بودن، جرحی یا فوتی بودن، هر یک در دسته‌های مربوط به خود قرار می‌گیرند و به این ترتیب شرایطی که باعث می‌شود تا یک تصادف برحسب متغیرهای مستقل مربوط به خود در یک دسته مشخص از شدت تصادفها قرارگیرد، معلوم شود.

سان و شین، با استفاده از روش کارت، شبکه‌های عصبی و تحلیل رگرسیون، به دسته‌بندی تصادفها بر حسب شدت آن در کشور کره پرداختند و نشان دادند که کمربند و کلاه ایمنی، مهم‌ترین عوامل تعیین کننده شدت تصادف هستند [Sohn (et. al.), 2001] . شبکه‌های عصبی مصنوعی[12] از دیگر ابزارهای داده‌کاوی و روش‌های ناپارامتری است که برخی از پژوهشگران به کمک آن به تحلیل شدت تصادفها و مصدومیت افراد درگیر در تصادفها پرداخته‌اند [Abdelwahab (et. al.), 2001], [Abdelwahab (et. al.), 2002], [Delen (et. al.), 2006] . به عنوان مثال دِلِن و همکاران، با تحلیل شدت مصدومیت ۳۰۳۵۳ راننده آمریکایی درگیر در تصادفها، نشان دادند که وضعیت استفاده از کمربند ایمنی و مصرف مشروبات الکلی، سن و جنسیت راننده، از عوامل مهم در شدت مصدومیت رانندگان است [Delen (et. al.), 2006].

روش کارت که از متداول‌ترین روش‌های دسته‌بندی و پیش‌بینی بوده و نتایج آن در قالب درخت تصمیم[13] قابل ارایه است، توسط برخی پژوهشگران، در مطالعات ایمنی ترافیک، مورد استفاده قرار گرفته است [Stewart, (1996), Sohn (et. al.), 2001, Chong (et. al.), 2005, Tesema (et. al.), 2005, Chang (et. al.), 2006] .

چانگ و همکاران، با تحلیل ۱۲٦۰٤ داده تصادف در تایوان و با استفاده از روش کارت، نشان دادند که کاربران آسیب‌پذیر (عابرین‌پیاده، دوچرخه‌سواران، موتورسیکلت سواران و سرنشینان وسایل نقلیه سواری) با احتمال بیشتری دچار مصدومیت‌های شدید در تصادفها می‌شوند و به این نتیجه رسیدند که مسیرهای مشترک دوچرخه و موتورسیکلت با سایر وسایل‌نقلیه، موقعیت‌های بسیار خطرناکی را در افزایش شدت تصادفها ایجاد می‌کند[Chong (et. al.), 2005]. مطالعه دیگری با ٤٦٥۸ داده تصادف که مربوط به پایتخت کشور اتیوپی بود و با همین روش انجام شد، نشان داد که عدم رعایت حق‌تقدم عابر پیاده و نیز سرعت غیرمجاز، از مهم‌ترین عوامل افزایش شدت مصدومیت هستند. آنچه که این مطالعه را از سایر مطالعات پیشین در سایر کشورها نیز متمایز می‌کند، گستردگی جغرافیایی و حجم زیاد داده‌هایی است که مورد تحلیل واقع شده‌اند. از طرفی در ایران نیز استفاده از داده‌کاوی برای تحلیل تصادفهای سرنشینان برای اولین بار انجام شده است که نشان می‌دهد در صورت داشتن اطلاعات کافی می‌توان تحلیل‌های مناسبی برای بررسی وضعیت تصادفها توسط مدل‌هایی مانند کارت انجام داد.

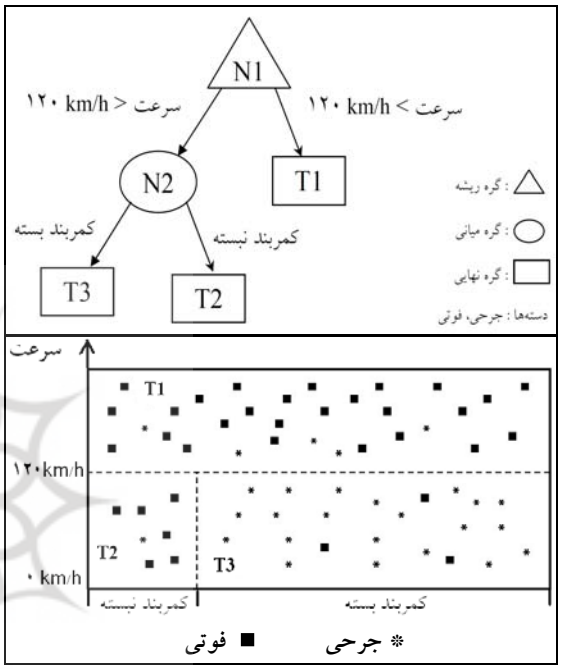
۳. روش تحقیق

برای رسیدن به اهداف این مطالعه لازم بود تا از مفاهیمی چون درخت تصمیم و روش‌هایی برای یافتن متغیرهای مهم‌تر استفاده شود، بنابراین در ادامه و در شرح چگونگی روش تحقیق، به معرفی درخت تصمیم و روش کارت که برای دسته‌بندی متغیر هدف لازم است، پرداخته می‌شود. شناسایی متغیرهایی که نقش مهم‌تری در دسته‌بندی متغیر هدف دارند، براساس روابطی به نام اهمیت متغیر[14] انجام می‌گیرد. توضیحات مربوط به درخت تصمیم و اهمیت متغیر در پی آمده است.

۳-۱ درخت تصمیم

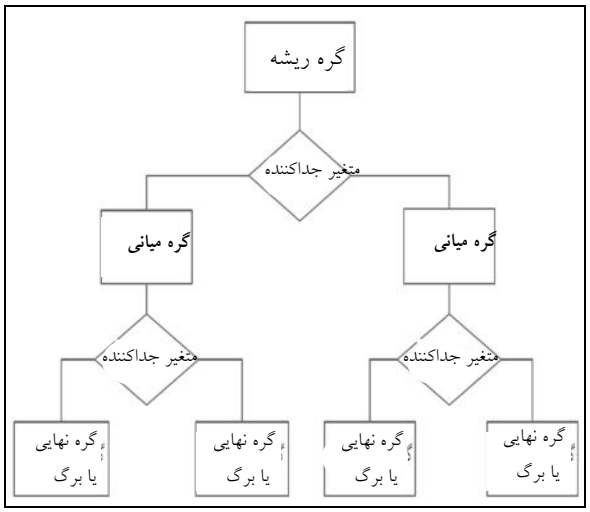
دسته‌بندی، عملیاتی برای تحلیل داده‌ها و استخراج مدل به منظور توصیف و تعریف دسته‌های مهم، قابل فهم و پیش‌بینی رفتار آینده آنهاست. مدل دسته‌بندی براساس داده‌های آموزشی که متغیرهای مستقل و هدف (وابسته) آن مشخص هستند تولید شده و برای داده‌های جدیدی که هدف، پیش‌بینی متغیر هدف است، بکار گرفته می‌شود. روش‌های مختلفی برای دسته‌بندی وجود دارد که از جمله آنها می‌توان به شبکه‌های عصبی، درخت تصمیم و رگرسیون اشاره کرد. زمانی که از درخت تصمیم برای دسته‌بندیِ متغیر هدف اسمی[15] استفاده می‌شود، به آن درخت دسته‌بندی، و برای پیش‌بینیِ متغییر هدف پیوسته، به آن، درخت رگرسیون گفته می‌شود [Han (et. al.), 2006]. روش کارت، برای هر دو نوع متغیر هدف، قابل استفاده است. از آنجا که در این مطالعه، متغیر هدف، از نوع اسمی (شدت مصدومیت) است، از درخت دسته‌بندی استفاده شده است. درخت ایجاد شده به‌گونه‌ای است که در ابتدا همگی داده‌ها در گره ریشه (اولین گره) که در بالا قرار می‌گیرد، وجود دارند. سپس براساس متغیری که می‌تواند بیشترین همگنی را برای هر شاخه ایجاد کند، در ریشه، انشعاب ایجاد می‌شود. این عمل آنقدر ادامه می‌یابد تا داده‌های موجود در هر گره، بیشترین همگنی را داشته و به یک دسته خاص تعلق گیرند. چنین گرهی که در انتها قرار می‌گیرد و از آن انشعابی ایجاد نمی‌شود، گره نهایی یا برگ[16] نامیده می‌شود[Berry (et. al.), 2004] .

شکل ۱ با مثالی ساده و با فرض اینکه مثلاً متغیر هدف، دوتایی و در دو دسته جرحی و فوتی باشد، چگونگی رشد و ایجاد درخت و تفسیر نتایج آنرا نشان می‌دهد. در واقع درخت دسته‌بندی، فضای پاسخ را به نواحی مستطیلی تقسیم می‌کند. این درخت فرضی نشان می‌دهد که مثلاً اگر سرعت قبل از تصادف بالای ۱۲۰km/h باشد، سرنشینان کشته می‌شوند و اگر زیر ۱۲۰km/h بوده ولی کمربند بسته باشند، مجروح شده و در غیراین صورت کشته می‌شوند.



شکل ۱. نمونه‌ای از اصول عملکرد مدل درختی

اساس کار روش کارت، در رشد درخت این است که هر گرهی براساس متغیر جداکننده‌ای[17] که بهترین خلوص[18] را ایجاد کند (به عنوان نمونه در مثال قبلی، برای گره N_1، سرعت، بهترین متغیر جداکننده است و یا وضعیت کمربند بهترین متغیر جداکننده برای گره N_2 است)، به دو زیرگره[19] تقسیم می‌شود. اگر گره‌های ایجاد شده به اندازه کافی یکدست و خالص باشند انشعاب آنها متوقف شده و در حکم گره نهایی یا برگ خواهند بود، اما اگر خلوص لازم را نداشته باشند، مجدداً به گره‌های دیگری منشعب خواهند شد. در واقع داده‌های هر زیر گره، همگن‌تر از گره قبلی[20] خود هستند. به گره‌هایی که مابین گره ریشه و گره‌های نهایی هستند، گره میانی گفته می‌شود. گره‌های میانی لایه بالاتر، والدین گره‌های لایه پایین‌تر از خود هستند. نمودار شماتیکی از ساختار کلی یک درخت در شکل ۲ نمایش داده شده است.



شکل ۲. شکل شماتیکی از ساختار کلی یک درخت تصمیم

برای جداسازی هر گره به دو زیرگره، شاخص‌های مختلفی وجود دارد که معروف‌ترین آن برای داده‌های اسمی، شاخص جینی[21] است که به شکل زیر تعریف می‌شود:

$$p(j \mid m) = \frac{p(j,m)}{p(m)} \ , \ p(j,m) = \frac{\pi(j) N_j(m)}{N_j} \quad (۱)$$

$$p(m) = \sum_{j=1}^{J} p(j,m)$$

$$Gini(m) = 1 - \sum_{j=1}^{J} p^2(j \mid m) \quad (۲)$$

که در آن:

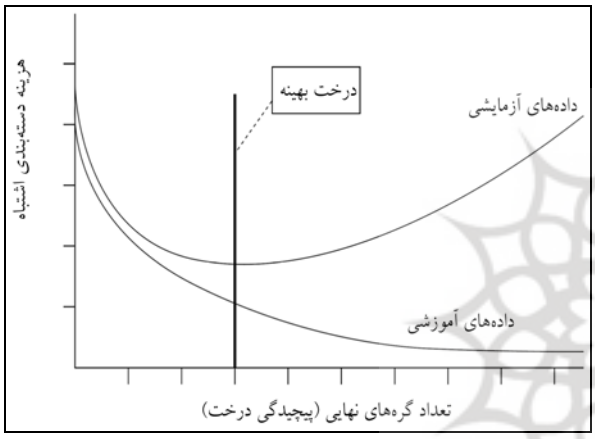
J تعداد دسته‌ها یا همان متغیرهای هدف، $\pi(j)$، احتمال اولیه[22] مربوط به دسته j و توسط تصمیم‌گیرنده مشخص می‌شود (توضیحات بیشتر در بخش ۵-۲ آمده است)، $N_j(m)$ تعداد مشاهدات مربوط به دسته j در گره m، N_j تعداد کل مشاهدات مربوط به کلاس j در گره ریشه، p(j|m)، احتمال قرارگیری مشاهدات مربوط به دسته j در گره m، و Gini(m)، که همان شاخص جینی است، معرف عدم خلوص یا ناهمگنی در گره m است. به این معنی که مثلاً اگر همه مشاهدات در یک گره از یک دسته باشند، Gini(m)، برابر صفر شده و مبین کمترین ناخالصی و به عبارتی بیشترین خلوص در گره است و بر عکس، بیشترین مقدار Gini(m)، زمانی حاصل می‌شود که از همه مشاهدات به یک نسبت، در گره وجود داشته باشند. شاخص جینی در هر گره برای تمام متغیرها محاسبه شده و متغیری به عنوان متغیر جداکننده انتخاب می‌شود که کمترین مقدار برای جینی از آن به دست آید. احتمال اولیه، مبین سهم هر یک از دسته‌ها در جامعه مرجع است.

رشد درخت براساس شاخص جینی از همان گره ریشه[۲۳] (گره N1 در شکل ۱) که اولین گره بوده و در برگیرنده تمام مشاهدات است، آغاز شده و برای هر درختی که ایجاد می‌شود، هزینه دسته‌بندی اشتباه[۲۴] آن – که می‌توان از آن به عنوان شاخص خوبی برازش[۲۵] یاد کرد– طبق رابطه زیر محاسبه می‌شود:

$$misclassification\cos t = \sum_{t=1}^{T} p(t)\left[1-\sum_{j=1}^{J} p^2(j\mid t)\right] \quad (۳)$$

که در آن:

$p(t)$ ، سهم مشاهدات موجود در گره نهایی t از کل مشاهدات بوده و T، تعداد گره‌های نهایی است. رابطه فوق مبین آن دسته از داده‌هایی است که به اشتباه در دسته‌های غیرمرتبط با خود، دسته‌بندی شده‌اند (مثلاً مصدومیت‌هایی که جرحی بوده ولی در گره نهایی فوتی قرار گرفته‌اند). در روش کارت، رشد درخت تا جایی ادامه می‌یابد که در هر گره نهایی، فقط مشاهدات یکسان وجود داشته باشد. در چنین حالتی، بزرگ‌ترین درخت[۲۶]، ایجاد شده است که بر روی داده‌های آموزش، برازش بیش از حد[۲۷] دارد. در این مرحله برای این که از پیچیدگی درخت نهایی کاسته شده و درخت‌های ساده‌تری ایجاد شوند، عملیات هرس‌کردن[۲۸] آن براساس الگوریتم هزینه- پیچیدگی[۲۹]، انجام می‌شود. هر چقدر یک درخت ساده‌تر (با گره‌های نهایی کمتر) باشد، هزینه دسته‌بندی اشتباه آن بیشتر است. اگر با قطع یک زیر درخت، افزایش هزینه دسته‌بندی اشتباه، به اندازه کافی کمتر از کاهش هزینه پیچیدگی آن باشد، شاخه مربوطه هرس شده و درخت جدیدی به وجود می‌آید. شکل ۳ نشان می‌دهد که چگونه از بین درخت‌های (مدل‌های) ایجاد شده، درخت بهینه انتخاب می‌شود. با افزایش پیچیدگی (تعداد گره‌های نهایی بیشتر)، هزینه دسته‌بندی اشتباه برای داده‌های آموزشی مرتباً کاهش می‌یابد، به این معنی که در برگ‌های ایجاد شده نهایی، فقط داده‌های مربوط به یک دسته قرار دارند و مدل ایجاد شده به گونه‌ای است که مثلاً اگر یک برگ را با برچسب فوتی پیش‌بینی کرده است، تمام داده‌های موجود در آن فوتی هستند و هیچ داده جرحی در آن قرار ندارد، یعنی هیچ داده‌ای را به اشتباه در دسته‌ای غیر از دسته خودش قرار نداده است. اما در عمل و برای یک سری داده جدید اگر بخواهد براساس مدلی که قبلاً تهیه کرده است عمل کند، چنین امکانی وجود نخواهد داشت و داده‌های زیادی را به اشتباه در دسته‌های غیر مربوط با خودش قرار خواهد داد، به همین دلیل است که وقتی با بزرگ‌ترین درخت با تعداد زیادی گره نهایی (برگ) مواجه هستیم و داده‌های جدیدی مثل داده‌های آزمایشی به آن معرفی می‌شود، هزینه دسته‌بندی اشتباه آن زیاد است و نمی‌تواند طبق قوانینی که براساس داده‌های آموزشی به دست آورده بود، داده‌های آزمایشی را هم درست دسته‌بندی کند، بنابراین هزینه دسته‌بندی اشتباه برای داده‌های آزمایشی ابتدا کاهش و سپس افزایش می‌یابد. درخت بهینه درختی است که کمترین هزینه دسته‌بندی اشتباه را برای داده‌های آزمایشی داشته باشد[Giudici, 2003]. برای ایجاد درخت بهینه در این تحقیق از نرم‌افزار [SPSS, Clementine, 2008] که یکی از قوی‌ترین نرم‌افزارهای داده‌کاوی برای توسعه مدل درختی با روش کارت است، استفاده شد.



شکل۳. رابطه پیچیدگی درخت و هزینه دسته‌بندی اشتباه برای داده‌های آموزشی و آزمایشی

۳-۲ اهمیت متغیر

در مدل‌سازی، شناسایی متغیرهایی که نقش اساسی‌تری در پیش‌بینی متغیر هدف دارند یکی از مهم‌ترین مراحل مدل‌سازی است. یکی از خروجی‌های روش کارت، شاخص اهمیت متغیر است که براساس آن پس از ساخت درخت بهینه، اهمیت متغیر x_j (مثلاً وضعیت استفاده از کمربند، جنسیت، علت تامه و...) به صورت نسبی طبق رابطه زیر به دست می‌آید:

$$VIM(x_j) = \sum_{m=1}^{M} \frac{n_m}{N} \Delta Gini(S(x_j, m)) \quad (٤)$$

که در آن :

$S(x_j, m)$ جداکننده متغیر x_j در محل گره m و $\Delta Gini(S(x_j, m))$، کاهش شاخص جینی در گره m براساس متغیر x_j، $\frac{n_m}{N}$ نسبت مشاهدات در گره m، M تعداد کل گره‌ها و

N تعداد کل مشاهدات است. ممکن است در یک گره، پس از جداسازی توسط متغیر x_j، خلوص داده‌هایی که در دو زیرگره آن قرار می‌گیرند بالا، اما نسبت این داده‌ها به کل داده‌ها، پایین باشد. بنابراین میانگین وزنی آنها از طریق وارد کردن نسبت $\frac{n_m}{N}$ در رابطه در نظر گرفته می‌شوند. این مقدار برای همه متغیرهای مستقل x_j محاسبه شده و به نحوی مقیاس می‌شود که مجموع آن برای همه متغیرها، ۱ شود. متغیری که بیشترین سهم را دارد، نسبت به بقیه، به صورت نسبی بزرگ‌ترین عدد را اختیار می‌کند.

٤. توصیف داده‌ها

داده‌های تصادفهای برون‌شهری ایران مربوط به ۳ سال اخیر (۱۳۸۵-۱۳۸۷) از معاونت راهور ناجا اخذ شد. از آنجا که محدوده این مطالعه مربوط به کل راههای اصلی دوخطه دوطرفه (۲۱۵۷۹ km) کشور بود، بنابراین تصادفهای مربوط به آزادراهها، بزرگراهها و راههای فرعی در نظر گرفته نشدند. فرم ثبت اطلاعات تصادفهای معروف به کام ۱۱٤، حاوی داده‌های مهمی مربوط به هر تصادف است. در این فرم اطلاعات کلی مربوط به هر تصادف از قبیل زمان وقوع تصادف (سال، ماه، روز و ساعت)، نوع راه (آزادراه، بزرگراه، ...)، نوع برخورد (چند وسیله‌ای، برخورد با شیئ ثابت، ...)، علت تامه تصادف (تجاوز به چپ ناشی از سبقت، عدم رعایت فاصله طولی، ...)، شرایط سطح معبر (خشک، یخبندان وبرفی، ...)، وضعیت شانه راه (آسفالتی، شنی) و نیز اطلاعات مربوط به کلیه رانندگان درگیر در تصادفها، از قبیل سن، جنسیت، نوع گواهینامه و شدت مصدومیت (اعم از آسیب ندیده، جرحی و فوتی) و اطلاعات عابرین پیاده آسیب‌دیده در تصادف و نیز اطلاعات مربوط به سرنشینان صدمه‌دیده درگیر در تصادف برداشت می‌شود.

از آنجا که در این مطالعه، هدف، بررسی عوامل مؤثر در شدت مصدومیت دو گروه افراد (رانندگان و سرنشینان) درگیر در تصادف بوده است، بنابراین داده‌های مربوط به عابرین پیاده حذف شده است. شدت مصدومیت رانندگان درگیر در هر تصادف در سه سطحِ آسیب‌ندیده، جرحی و فوتی و برای سرنشینان در سه سطحِ جراحت سطحی، جراحت شدید و فوتی ثبت می‌شود.

با تلفیق بانک اطلاعات مربوط به مشخصات کلی تصادف با بانک اطلاعات مربوط به رانندگان، نهایتاً یک بانک اطلاعاتی که هر رکورد آن معرف یک راننده درگیر در تصادف بود، با ۱٦۹٦٤۸ رکورد ایجاد شد. همین روند برای داده‌های مربوط به سرنشینان نیز انجام شد و به تبع آن یک بانک اطلاعاتی با ۷۲٤۱ رکورد ایجاد شد. لازم به بیان این مطلب است که اطلاعات مربوط به سرنشینان آسیب ندیده، در فرم ثبت تصادفها، ثبت نمی‌شود.

جدول ۱، توضیحات مربوط به متغیرهای وارد شده در مدل‌سازی را نشان می‌دهد. نهایتاً برای رانندگان، ۱۳ متغیر مستقل و۱ متغیر هدف شدت مصدومیت (آسیب ندیده، جرحی و فوتی) و برای سرنشینان ۱۱ متغیر مستقل (برای سرنشینان، سن سرنشین و کاربری وسیله نقلیه‌ای که در آن بوده است ثبت نمی‌شود) و ۱ متغیر هدف شدت مصدومیت (سطحی، شدید و فوتی) تعریف شده و وارد مدل کارت شدند. همان‌طور که مرجع [Stewart, 1996] هم پیشنهاد داده است بایستی بخشی از داده‌ها به آموزش[۳۰] و بخشی دیگر به آزمایش[۳۱] تعلق گیرد که در این مطالعه، ۷۰٪ داده‌ها به صورت تصادفی انتخاب و به آموزش مدل اختصاص داده شد و ۳۰٪ دیگر برای آزمایش مدل استفاده شد. روش دیگری که برای ارزیابی این مدلها بکار می‌رود، ارزیابی متقابل k قسمتی[۳۲] نام دارد. از این روش زمانی استفاده می‌شود که حجم داده‌ها کم باشد [Stewart, 1996]. چون در این مطالعه حجم داده‌ها به مقدار کافی زیاد بود، بنابراین فقط با همان ۳۰٪ داده‌ها این کار انجام شد.

۵. تجزیه و تحلیل

۵-۱ تبدیل متغیر هدف چندتایی به دوتایی

همان‌طورکه قبلاً اشاره شد، متغیر هدف (شدت مصدومیت) در این مطالعه در سه دسته قرار دارد (یعنی مدل‌سازی شدت مصدومیت رانندگان دارای متغیر پاسخی در سه سطح آسیب ندیده، جرحی و فوتی است و مدل‌سازی شدت مصدومیت سرنشینان دارای متغیر پاسخی در سه سطح جرحی خفیف، جرحی شدید و فوتی است). در چنین مواردی که متغیر هدف، چندتایی (در اینجا سه تایی) است، ممکن است دقت پیش‌بینی مدل دسته‌بندی ایجادشده – در اینجا مدل دسته‌بندی ایجاد شده توسط روش کارت- خواه برای کل مدل و خواه برای هر یک از دسته‌ها کاهش یابد (این مسئله به عنوان نمونه برای رانندگان در جدول۲ آمده است). به‌عنوان مثال فقط ۲۷/۹٪ از رانندگانی که در تصادفها مجروح شده‌اند، توسط مدل دسته‌بندی ایجاد شده با

روش کارت، در دسته مربوط به جرحی، دسته‌بندی شده‌اند. برای افزایش دقت پیش‌بینی مسایلی که متغیر هدف آنها چنددسته‌ای است، تبدیل آنها به یک سری از مسایلی که متغیر هدف آنها دوتایی است، پیشنهاد شده است [Dissanayake (et. al.), 2002 and Delen (et. al.), 2006] . برای انجام این تبدیل برخی پژوهشگران پیشنهاد کرده‌اند که این کار به نحوی صورت گیرد که سطوح پایین‌تر شدت مصدومیت را در مقابل سطوح بالاتر قرار دهد. شکل ٤ این موضوع را به صورت گرافیکی نشان می‌دهد. مثلاً در حالت ب-۲-۱، داده‌های فوتی اصلاً وارد مدل دسته‌بندی کارت نشده‌اند و فقط مصدومیت‌های سطحی و شدید، با یکدیگر مقایسه شده‌اند و یا در حالت ب-۱-۲، داده‌هایی که متغیر هدف آنها جراحت شدید و فوتی بوده، با هم در یک دسته با برچسب یک قرار گرفته‌اند و جرحی‌های سطحی در دسته‌ای با برچسب صفر قرار گرفته‌اند. در واقع به جای اینکه برای پیش‌بینی شدت مصدومیت رانندگان یک مدل ارایه شود که متغیر پاسخ (هدف) آن در سه سطحِ صدمه ندیده، جرحی و فوتی باشد، چهار مدل ارایه شده است که متغیر پاسخِ آن در دو سطحِ صفر و یک است. همین وضعیت برای ساخت مدل پیش‌بینی شدت مصدومیت سرنشینان هم وجود دارد. یعنی به جای ساخت فقط یک مدل، با متغیر پاسخی در سه سطحِ جرحی خفیف، جرحی شدید و فوتی، چهار مدل ساخته شده است که متغیر پاسخ آن در دو سطحِ صفر و یک است. از سوی دیگر، از آنجا که ممکن است در صحنه تصادف، فردی به عنوان جرحی شدید ثبت شود ولی در حین انتقال به بیمارستان و یا بعد از آن فوت کند، در این نوع تبدیلِ دوتایی، چون مدل‌های مختلفی تولید و با یکدیگر مقایسه می‌شوند، بنابراین، این امکان فراهم می‌شود که مواردی از این نوع نیز به نحوی دیده شود و این، یکی دیگر از مزایای شیوه ایجاد متغیر هدف دوتایی است.

۵-۲ دقت پیش‌بینی مدل‌ها

دقت پیش‌بینی مدل، عبارت است از: آن دسته از داده‌هایی که درست پیش‌بینی شده‌اند، تقسیم بر تعداد کل داده‌ها. این دقت برای هر دسته نیز قابل تعریف است، به این معنی که داده‌هایی که در هر دسته، درست پیش‌بینی شده‌اند تقسیم بر تعداد کل داده‌های همان دسته می‌شوند. روش کارت برای تحلیل هر ۸ مدل دوتایی، بکارگرفته شد. برای رشد درخت از شاخص جینی استفاده شد و احتمال اولیه یا همان $\pi(j)$ (که در رابطه ۱ به آن اشاره شد) به دلیل اهمیت یکسان پیش‌بینی هر دو (دسته با برچسب یک و دسته با برچسب صفر) حالت متغیر هدف، یکی در نظرگرفته شد (یعنی $\pi(j)=0.5$). جدول ۳ دقت پیش‌بینی مدل را برای سرنشینان و رانندگان نشان می‌دهد. این مقادیر برای داده‌های آزمایشی است.

احتمال اولیه ($\pi(j)$)، مبین سهم هر یک از دسته‌ها در جامعه است، اما در مواردی که سهم یک دسته به مراتب بیشتر از یک دسته دیگر باشد (مثلاً در مدل الف- ۱-۲، سهم دسته صفر حدود ۱۲ برابر سهم دسته یک است)، چنان که احتمال اولیه آنها هم براساس همان نسبت موجود در داده‌ها، تنظیم شود (یعنی مثلاً برای همین حالت الف- ۱-۲، ۰/۹۲ =(۱۷۹۱+۱۱۱۳۸+۱۵٦۷۱۹)/۱۵٦۷۱۹ =$\pi(0)$ و ۰/۰۸=(۱۷۹۱+۱۱۱۳۸+۱۵٦۷۱۹)/(۱۱۱۳۸+۱۷۹۱) =$\pi(1)$)، در نتیجه، برای اینکه دقت کل مدل افزایش یابد، مدل حاصل، تمام داده‌ها را در دسته غالب، پیش‌بینی می‌کند. مثلاً برای حالت الف- ۱-۲، اگر π، طبق مقادیر فوق در مدل اعمال شود، مدل حاصل به نحوی خواهد بود که هر داده جدیدی به آن معرفی شود، آنرا در دسته صفر (در اینجا آسیب ندیده) پیش‌بینی خواهد کرد. اتفاقاً از آنجا که در مطالعات شدت مصدومیت، غالباً سهم داده‌های فوتی در برابر سایر داده‌ها از قبیل خسارتی و یا جرحی کمتر است، مشاهده می‌شود که دقت پیش‌بینی آنها کاهش می‌یابد، به نحوی که در تعداد قابل توجهی از مطالعات پیشین، دقت پیش‌بینی داده‌های فوتی برابر صفر و یا نزدیک به صفر به دست آمده‌اند. برای برون‌رفت از این مشکل پیشنهاد شده است در مواردی که نسبت متغیرهای هدف، نامتعادل بوده و به عبارتی سهم یک دسته به مراتب بیشتر از دسته یا دسته‌های دیگر است، ولی اهمیت دقت پیش‌بینی هر یک از آنها با هم برابر است، احتمال اولیه یکسان در نظرگرفته شود، تا آنهایی که سهم کمتری هم دارند در پیش‌بینی در نظر گرفته شوند [Steinberg (et. al.), 2007] . در این صورت اگرچه دقت کل مدل ممکن است کاهش یابد، ولی در عوض دقت پیش‌بینی داده‌های با سهم کمتر افزایش می‌یابد که این موضوع در اکثر موارد برای تصمیم‌گیر مهم‌تر است. این نکته یکی از ویژگی‌های بسیار خوب روش کارت است که این امکان را ایجاد می‌کند که داده‌های با نسبت کمتر، مانند بقیه درنظر گرفته شوند. در جدول ۳، دقت پیش‌بینی دسته یک تقریباً برای همه مدل‌ها حدود ٦۰٪ و بالای آن به دست آمده که در مقایسه با اکثر مطالعات پیشین که دقت پیش‌بینی دسته فوتی‌ها یا جراحت‌های شدید در حد صفر بوده، بسیار قابل قبول است.

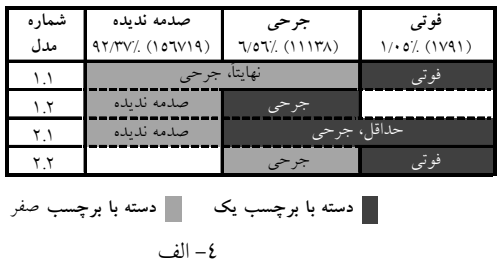
جدول ۱. نام و کد حالت‌های مختلف هر متغیر

نام متغیر	کد حالتهای مختلف
شدت مصدومیت (برای رانندگان)	متغییر وابسته: ۱.صدمه ندیده، ۲.جرحی، ۳.فوتی
شدت مصدومیت (برای سرنشینان)	متغییر وابسته: ۱.سطحی، ۲.شدید، ۳.فوتی
جنسیت	۱.مرد، ۲.زن
سن (فقط برای رانندگان)	کمّی (بر حسب سال)
وضعیت استفاده از کمربند	۱.استفاده کرده، ۲.استفاده نکرده، ۳.نامشخص
علت تامه تصادف	۱.عدم رعایت فاصله طولی، ۲.عدم رعایت فاصله عرضی، ۳.عدم رعایت حق تقدم، ٤.عدم توجه به جلو، ۵.عدم مهارت در رانندگی، ٦.عدم توانایی در کنترل وسیله نقلیه، ۷.تخطی از سرعت مطمئنه، ۸.تجاوز از سرعت مقرره، ۹.تجاوز به چپ ناشی از سبقت، ۱۰.انحراف به چپ، ۱۱.انحراف به راست، ۱۳.گردش به طرز غلط، ١٤.عبور از محل ممنوع، ۱۵.حرکت در خلاف جهت، ١٦.حرکت با دنده عقب، ۱۷.نقص فنی حادث وسیله نقلیه، ۱۸.نقص فنی مستمر در وسیله نقلیه، ۱۹.تغییر مسیر ناگهانی، ۲۰.خطای عابر، ۲۱.نقص ماده ٤ قانون ایمنی راهها ۲۲.نقص مقررات حمل بار، ۲۳.یدک کشی بطرز غلط، ٢٤.عبور از چراغ قرمز، ۲۵.دورزدن در محل ممنوع، ۲٦.سایر علل
نوع برخورد	۱.برخورد وسیله نقلیه با موتورسیکلت یا دوچرخه، ۲.برخورد وسیله نقلیه با یک وسیله نقلیه، ۳.برخورد وسیله نقلیه با چند وسیله نقلیه، ٤.برخورد وسیله نقلیه با عابر، ۵.برخورد وسیله نقلیه با حیوان، ٦.برخورد وسیله نقلیه با شیئ ثابت، ۷.واژگونی و سقوط، ۸.ایجاد حریق، ۹.برخورد موتورسیکلت با دوچرخه یا عابر، ۱۰.برخورد موتورسیکلت با موتورسیکلت، ۱۱.سایرموارد
مشخصه محل تصادف	۱.عادی، ۲.تقاطع، ۳.پل، ٤.تونل، ۵.میدان، ٦.سایر
وضع روشنایی	۱.روز، ۲.شب، ۳.هنگام غروب یا طلوع
وضع هوا	۱.صاف، ۲.مه آلود، ۳.ریزش باران، ٤.ریزش برف، ۵.طوفانی، ٦.ابری، ۷.غبارآلود
شرایط سطح معبر	۱.خشک، ۲.تر، ۳.یخبندان و برفی، ٤.شنی و خاکی، ۵.گل آلود، ٦.روغنی، نفتی، کثیف، ۷.سایر
موقعیت تصادف	۱.باند سواره رو، ۲.شانه، ۳.میانه(رفوژ میانه راه)، ٤.کنار جاده(حریم جاده)، ۵.خارج از حریم جاده، ٦.سایر
نوع شانه راه	۱.ندارد، ۲.خاکی، ۳.آسفالته
عرض شانه	کمّی (بر حسب متر)
کاربری وسیله نقلیه (فقط برای رانندگان)	۱.سواری، ۲.مینی بوس، ۳.اتوبوس، ٤.وانت بار، ۵.کامیونت، ٦.کامیون، ۷.آمبولانس، ۸.تریلر، ۹.موتورسیکلت، ۱۰.دوچرخه، ۱۱.ادوات کشاورزی، ۱۲.ادوات راهسازی، ۱۳.خودرو آتش نشانی، ١٤.خودرو پلیس، ۱۵.سایر وسایل

جدول ۲. دقت پیش‌بینی مدل با متغیر هدف سه‌تایی برای رانندگان

	داده های آموزشی		داده های آزمایشی	
	داده های مشاهده شده	داده های درست پیش بینی شده	داده های مشاهده شده	داده های درست پیش بینی شده
صدمه ندیده	۱۰۹٦۵٦	۸۳٤٦۳ (۷٦/۱۱٪)	٤۷۰٦۳	۳۵۷۳٦ (۷۵/۹۳٪)
جرحی	۷۷٦۵	۲۳۲٤ (۲۹/۹۳٪)	۳۳۷۳	۹٤۱ (۲۷/۹۰٪)
فوتی	۱۲۹۷	٦٦۵ (۵۱/۲۷٪)	٤۹٤	۲٤٤ (٤۹/۳۹٪)
کل مدل	۱۱۸۷۱۸	۸٦٤۵۲(۷۲/۸۲٪)	۵۰۹۳۰	۳٦۹۲۱ (۷۲/٤۹٪)

٤- ب



٤- الف

شکل ٤. نمایش گرافیکی مدل‌های با متغیر هدف دوتایی و سهم هر یک از دسته‌ها از کل داده‌ها؛ الف: برای رانندگان، ب: برای سرنشینان

جدول ۳. دقت پیش‌بینی مدل‌ها

	مدل ۱-۱		مدل ۲-۱		مدل ۱-۲		مدل۲-۲	
	سرنشینان	رانندگان	سرنشینان	رانندگان	سرنشینان	رانندگان	سرنشینان	رانندگان
متغیر پاسخ صفر	٥٦/٨٠٪	٥٥/٠٠٪	٥٨/١٢٪	٧١/١٠٪	٥٧/٥٥٪	٧٣/٣٦٪	٤٠/٧٠٪	٤٧/٤٠٪
متغیر پاسخ یک	٥٨/٣٣٪	٨٨/٠٥٪	٥٩/٨٨٪	٦٧/٣٣٪	٥٧/٦٢٪	٦٣/١٧٪	٦٩/٥١٪	٧٠/٠٠٪
کل مدل	٥٦/٩١٪	٥٥/٣٢٪	٥٩/٠٦٪	٧٠/٨٦٪	٥٧/٥٩٪	٧٢/٥٩٪	٤٤/٤٣٪	٥٠/٣٥٪

۵-۳ اهمیت متغیرها

جدول ٤ به صورت نسبی اهمیت متغیرهای مستقل را نسبت به هم نشان می‌دهد. در اکثر مدل‌ها، وضعیت استفاده از کمربند جزء اولین متغیر تعیین‌کننده افزایش شدت مصدومیت است، به این معنی که زمانی که سرنشین یا راننده از کمربند استفاده نکرده است، با احتمال بیشتری در یک وضعیت مصدومیت شدیدتر قرار می‌گیرد. این نتیجه در بسیاری از تحقیقات پیشین نیز ثابت شده بود [Sohn (et. al.), 2001 and Bedard (et. al.), 2002, Valent (et. al.), 2002, Kweon (et. al.), 2003, Delen (et. al.), 2006] . با توجه به اینکه در کشورمان، پلیس، وضعیت مصدومین هر تصادف را فقط در صحنه تصادف برداشت می‌کند، بنابراین همواره بسیاری از مصدومیت‌هایی که در صحنه تصادف به عنوان جراحت و یا جراحت شدید ثبت می‌شوند، پس از انتقال به بیمارستان و یا در حین انتقال، جان می‌سپارند. یادآوری می‌شود بانک اطلاعاتی موجود، از پلیس اخذ شده بود. بنابراین می‌توان گفت که در بین٤ مدل مربوط به سرنشینان، مدل‌های ۲-۱و ۱-۲ ، اهمیت بیشتری دارند، چرا که به نوعی مرز بین حیات و فوت یک سرنشین مصدوم را مشخص می‌کنند.

همان‌طور که در جدول ٤- ب هم آمده است، ترتیب اهمیت دو متغیر اول برای این دو مدل یکی است که این موضوع حاکی از شباهت مفهومی این دو مدل است. وضعیت مشابهی برای رانندگان وجود دارد، به نحوی که در جدول ٤- الف، و در مدل‌های ۲-۱ و ۱-۲ ، اهمیت متغیر کمربند بیش از ۹۰٪ است و این به آن معنی است که اگر راننده‌ای از کمربند استفاده نکند، از وضعیت صدمه‌ندیده وارد وضعیت جراحت و یا وضعیت حداقل جراحت و فوت می‌شود.

اهمیت بسیار زیاد متغیر وضعیت استفاده از کمربند در سه مدل اول جدول ٤- الف، نشان می‌دهد که متأسفانه همچنان در کشور ما فرهنگ استفاده از کمربند ایمنی حتی در بین رانندگان، فراگیر نشده است و این عامل، همچنان عاملی تعیین‌کننده در افزایش شدت مصدومیت است. از طرف دیگر متأسفانه در ایران بستن کمربند ایمنی برای سرنشینان ردیف عقب وسایل نقلیه نیز اجباری نیست. این مطالعه اهمیت استفاده از آن را برای این دسته از سرنشینان نیز نشان می‌دهد.

متغیر دیگری که در بین این هشت مدل، جایگاه بعدی را به خود اختصاص داده است، علت تامه تصادف است. این نتیجه، تأییدی بر مطالعه مرجع (Al-Ghamdi, 2002) در عربستان است که در آنجا هم علت تامه یکی از عوامل مهم افزایش شدت تصادف بود. با بررسی درخت‌های تصمیم به دست آمده از مدل‌ها معلوم شد، در اکثر موارد وقتی که علت تامه، انحراف به چپ یا تجاوز به چپ ناشی از سبقت باشد (که در مورد راههای دوخطه دوطرفه این هر دو مانور، به سبقت خطرناک برمی‌گردد)، شاخه مربوطه آن درخت، مصدومیت‌های شدیدتر (متغیر هدف یک) را پیش‌بینی می‌کند.

دلیل اصلی آن هم این است که مانور سبقت در این نوع راهها با استفاده از باند جهت مقابل انجام می‌شود که در پی آن تصادفات شاخ به شاخ که از شدیدترین نوع تصادفها هستند رخ می‌دهد. از اقدامات کم‌هزینه‌ای که برای رفع این مشکل می‌توان داشت، احداث خطوط کمکی باندهای سبقت در برخی از قسمت‌های راه است تا در آن نواحی، رانندگان بتوانند سبقت بگیرند. سرعت غیر مجاز نیز از دیگر حالت‌های مربوط به علت تامه است که باعث افزایش شدت تصادفها می‌شود. حضور گشت‌های ثابت و سیار پلیس در این راهها و برخورد جدی با این دو نوع تخلف، از راهکارهایی است که تا حد زیادی می‌تواند از بروز تصادفهای ناشی از این تخلفات بکاهد.

سهم بقیه متغیرها در قیاس با این دو متغیر بسیار ناچیز است. در واقع جداول ٤ کمک می‌کند که این شناخت به دست آید با سرمایه‌گذاری روی کدام عامل بیشترین بهره‌دهی را داشته باشیم.

۵-٤ درخت تصمیم

یکی از ویژگی‌های بسیار خوب درخت‌های تصمیم نسبت به سایر روش‌های مدل‌سازی، این است که با در اختیار گذاشتن یک سری قواعد[۳۳]، به تصمیم‌گیر کمک می‌کند تا به خوبی به مسایل «اگر – آنگاه» پاسخ دهد.

در اینجا به عنوان نمونه فقط درخت تصمیم حاصل از مدل ۲- ۱ سرنشینان را که در شکل ۵ معرفی شده است، آورده‌ایم. تفسیر درخت به دست آمده در زیر می‌آید:

جدول ٤- الف. اهمیت نسبی متغیرها برای رانندگان

۱-۱		۱-۲		۲-۱		۲-۲	
۰/٦٥	وضعیت استفاده از کمربند	۰/۹٤	وضعیت استفاده از کمربند	۰/۹٦	وضعیت استفاده از کمربند	۰/٤	علت تامه تصادف
۰/۰۷	موقعیت تصادف	۰/۰۱	موقعیت تصادف	۰/۰۱	موقعیت تصادف	۰/۱۳	نوع برخورد
۰/۰۷	عرض شانه	۰/۰۱	عرض شانه	۰/۰۱	عرض شانه	۰/۰۹	وضعیت استفاده از کمربند
۰/۰۷	مشخصه محل تصادف	۰/۰۱	نوع شانه راه	۰/۰۱	نوع شانه راه	۰/۰۸	کاربری وسیله نقلیه
۰/۰۷	وضع هوا	۰/۰۱	مشخصه محل تصادف	۰/۰۱	مشخصه محل تصادف	۰/۰٤	عرض شانه
۰/۰۷	کاربری وسیله نقلیه	۰/۰۱	شرایط سطح معبر	۰/۰۱	وضع هوا	۰/۰٤	موقعیت تصادف
۰	علت تامه تصادف	۰/۰۱	سن	۰/۰۱	کاربری وسیله نقلیه	۰/۰٤	مشخصه محل تصادف
۰	نوع برخورد	۰/۰۱	کاربری وسیله نقلیه	۰/۰۱	سن	۰/۰٤	شرایط سطح معبر
۰	وضع روشنایی	۰	علت تامه تصادف	۰	علت تامه تصادف	۰/۰٤	سن
۰	شرایط سطح معبر	۰	نوع برخورد	۰	شرایط سطح معبر	۰/۰٤	نوع شانه راه
۰	نوع شانه راه	۰	وضع روشنایی	۰	نوع برخورد	۰/۰٤	وضع هوا
۰	سن	۰	وضع هوا	۰	وضع روشنایی	۰	وضع روشنایی
۰	جنسیت	۰	جنسیت	۰	جنسیت	۰	جنسیت

جدول ٤- ب. اهمیت نسبی متغیرها برای سرنشینان

۱-۱		۱-۲		۲-۱		۲-۲	
۰/٦۱	علت تامه تصادف	۰/٥۹	وضعیت استفاده از کمربند	۰/٥۳	وضعیت استفاده از کمربند	۰/۳۱	جنسیت
۰/۱۱	وضعیت استفاده از کمربند	۰/۲	علت تامه تصادف	۰/۲۳	علت تامه تصادف	۰/۲٦	علت تامه تصادف
۰/۰۷	شرایط سطح معبر	۰/۰٦	وضع هوا	۰/۰٦	شرایط سطح معبر	۰/۱٤	موقعیت تصادف
۰/۰۷	مشخصه محل تصادف	۰/۰٥	شرایط سطح معبر	۰/۰٦	نوع برخورد	۰/۱٤	مشخصه محل تصادف
۰/۰۷	نوع برخورد	۰/۰٤	نوع شانه راه	۰/۰٤	وضع هوا	۰/۱٤	نوع برخورد
۰/۰۷	وضع هوا	۰/۰۳	نوع برخورد	۰/۰۳	وضع روشنایی	۰	وضع هوا
۰	وضع روشنایی	۰/۰۲	موقعیت تصادف	۰/۰۲	نوع شانه راه	۰	شرایط سطح معبر
۰	موقعیت تصادف	۰/۰۲	مشخصه محل تصادف	۰/۰۲	موقعیت تصادف	۰	وضعیت استفاده از کمربند
۰	جنسیت	۰	جنسیت	۰/۰۲	مشخصه محل تصادف	۰	نوع شانه راه
۰	نوع شانه راه	۰	وضع روشنایی	۰	جنسیت	۰	وضع روشنایی
۰	عرض شانه	۰	عرض شانه	۰	عرض شانه	۰	عرض شانه

گره صفر که همان ریشه و اولین گره است، توسط متغیر کمربند، به دو شاخه تقسیم شده است. این نشان می‌دهد که مهم‌ترین متغیر در دسته‌بندی و پیش‌بینی شدت مصدومیت، کمربند است. گره ۲ واقع در شاخه سمت راست، داده‌های مربوط به سرنشینانی را نشان می‌دهد که از کمربند ایمنی استفاده کرده‌اند و یا نامشخص است. در سمت راست درخت، ٤ گره پایانی داریم (٥،۱۱،۱۷،۱۸) که در همه آنها به غیر از ۱۸، جراحت سطحی بالاتر از جراحت شدید است (به نوار توسی رنگی که روی عدد صفر قرار دارد دقت شود). این نشان می‌دهد که چنانچه سرنشینی از تجهیزات ایمنی استفاده کرده باشد، فارغ از در نظر گرفتن همه متغیرها، با احتمال بیشتر، آسیب کمتری خواهد دید.

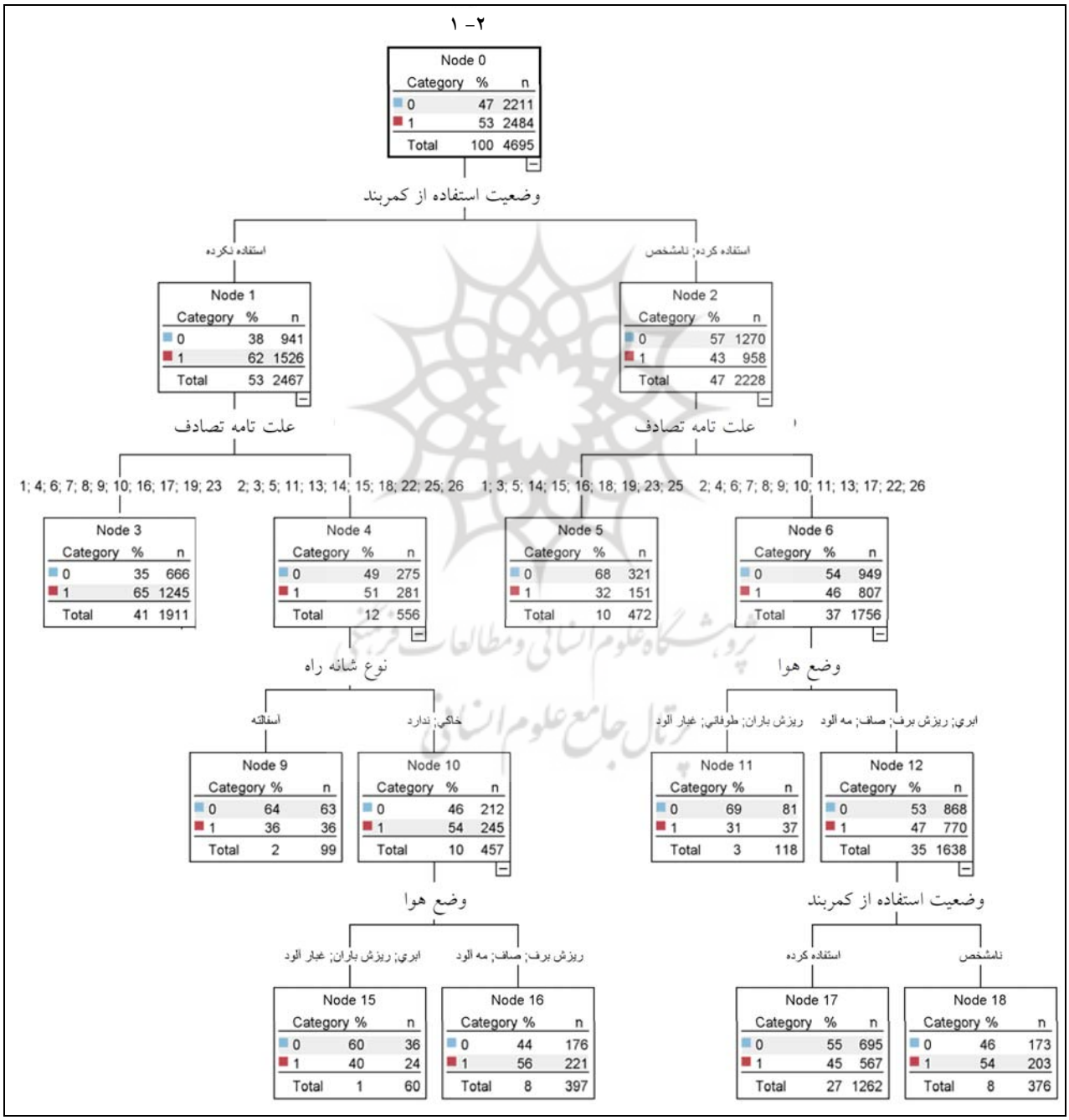
گره ۲، توسط متغیر علت تامه به گره پایانی ٥ و گره میانی ٦ رسیده است. گره پایانی ٥ نشان می‌دهد که اگر سرنشینی از کمربند ایمنی استفاده کرده باشد و علت تامه تصادفی که در آن درگیر بوده است (یعنی علت تامه تصادفی که این سرنشین در آن حضور داشته و درگیر شده بوده است)، موارد ۱، ۳، ٥، ۱٤، ۱٥، ۱٦، ۱۸، ۱۹، ۲۳ و ۲٥ باشد، با احتمال ٦۸٪ جراحت سطحی خواهد داشت. گره میانی ٦ نیز بر اساس متغیر وضع هوا به گره میانی ۱۲ و گره پایانی ۱۱ ختم می‌شود که برای وضعیت هوای بارانی، طوفانی و غبار آلود، جراحت سطحی محتمل‌تر از جراحت شدید به دست آمده است. در دو گره پایانی ۱۷ و ۱۸ نیز ملاحظه می‌شود که برای حالتی که به طور قطع از کمربند استفاده شده است، احتمال جراحت شدید کمتر است.

گره میانی ۱ در شاخه سمت چپ براساس متغیر علت تامه، به گره میانی ٤ و گره پایانی ۳ رسیده است. گره پایانی ۳ نشان می‌دهد که چنانچه سرنشینی از کمربند ایمنی استفاده نکرده باشد و علت تامه تصادفی که در آن بوده است مواردی چون رعایت نکردن فاصله طولی، توجه نکردن به جلو، ناتوانی درکنترل وسیله نقلیه، تخطی از سرعت مطمئنه، تجاوز از سرعت مقرره، تجاوز به چپ ناشی از سبقت، انحراف به چپ، حرکت با دنده عقب، نقص فنی حادث وسیله نقلیه، تغییر مسیر ناگهانی و یدک کشی به گونه‌ای که نادرست باشد، با احتمال بیشتری دچار جراحت و مصدومیت شدیدتر می‌شود. یادآوری می‌شود که چون محدوده این مطالعه راههای دوخطه دوطرفه برون‌شهری بود،

مانور سبقت غیر ایمن و انحراف بـه چـپ (بانـد حرکـت جهـت مقابل) می‌تواند منجر به تصادفهای شدید و مرگباری شود (همـان چیزی که گره پایانی ۳ آن را تأیید می‌کند).

گره میانی ٤ براساس متغیر نوع شانه به گـره پایـانی ۹ رسـیده و اگـر شـانه آسـفالتی باشـد، جراحـت سـطحی بـه دنبـال دارد. نهایتـاً گـره‌هـای پایـانی ۱۵ و ۱٦ براسـاس متغیـر وضـع هـوا تولید شده‌اند و در هوای برفی، صاف و مـه‌آلـود، جراحـت‌هـای شدیدتری دیده شده و در شرایط جوی دیگر، جراحت سطحی مشاهده می‌شود.

در مجموع، آنچه که از دیدگاه ایمنی تـردد، بیشـتر حـائز اهمیـت است، توجه بـه علـل و عوامـل وضـعیت‌هـایی اسـت کـه در آن جراحت‌هـا و مصـدومیت‌هـای شـدید، بیشـتر رخ مـی‌دهنـد. در درخت شکل ۵، گره ۳ جای توجه بیشتری دارد و قاعـده‌ای هـم که از آن استخراج می‌شود، کاملاً منطقی است و بـر لـزوم توجـه ویژه به بستن کمربند و اجتناب از مانورهای سـبقت خطرنـاک در این راهها تأکید دارد.



شکل ۵. درخت دسته‌بندی ایجاد شده از روش کارت برای سرنشینان مدل ۱.۲

٦. جمع‌بندی و نتیجه‌گیری

مدل کارت به دلیل نمایش گرافیکی از نتایجش، به سادگی قابل فهم و تفسیر است. ساختار درختی مدل این امکان را می‌دهد که به سادگی به سوالات "if-then" پاسخ داده شود. ضمناً این ساختار می‌تواند تعاملات پیچیده بین متغیرها را به سادگی نمایش دهد.

از طرف دیگر برای تحلیل حجم زیادی از داده‌ها با متغیرهای مستقل زیاد، مدل کارت بسیار کارآمد بوده و به راحتی می‌تواند هم متغیرهای مهم و هم متغیرهای کم اهمیت را شناسایی کرده و عملاً متغیرهای کم اهمیت را حذف کند.

این مطالعه که با مدل کارت انجام شد، نشان داد که عدم استفاده از کمربند ایمنی مهم‌ترین عامل در افزایش شدت مصدومیت سرنشینان و رانندگان وسایل نقلیه است. این مهم نشان می‌دهد که متأسفانه هنوز میزان استفاده از کمربند در بین سرنشینان و حتی رانندگان پایین است و این متغیر همچنان در بین بقیه متغیرها بیشتر خودنمایی می‌کند. لازم است که در کشور، برخوردهای سخت‌گیرانه‌تری در اعمال مقررات استفاده از کمربند اعمال شود. این در حالی است که حتی از نظر قانونی، هنوز استفاده از کمربند ایمنی برای سرنشینان ردیف عقب اجباری نشده است. علت تامه تصادف به عنوان دومین متغیر مهم شناسایی شد. سبقت خطرناک، به عنوان یکی از مهم‌ترین عوامل تشدیدکننده مصدومیت شناسایی شد (با توجه به اینکه این مطالعه مربوط به راههای دوخطه برون‌شهری بود، این نتیجه دور از ذهن نبود). لزوم توجه به احداث باندهای سبقت و تشدید اعمال مقررات برای کاهش مانورهای سبقت خطرناک در این راهها، ضروری به نظر می‌رسد.

۷. سپاسگزاری

شایسته است که نویسندگان این مقاله از جناب آقای سرهنگ میشانی و جناب آقای صباغ (از کارکنان محترم معاونت راهور ناجا)، برای در اختیار گذاشتن آمار و اطلاعات لازم سپاسگزاری کنند.

۸ . پانویس‌ها

1. Two-lane Two-way rural roads
2. Injury severity
3. Data mining
4. Regression type generalized linear models
5. Functional form
6. Classification and Regression Tree (CART)
7. Independent variable
8. Target variable
9. Classification and Prediction
10. Logit models
11. Ordered probit
12. Artificial Neural Networks
13. Decision tree
14. Variable importance
15. Nominal
16. Leaf or terminal node
17. Splitter
18. Purity
19. Child node
20. Parent node
21. Gini index
22. Prior probability
23. Root node
24. Misclassification cost
25. Goodness of fit
26. Maximal tree
27. Overfitting
28. Pruning
29. Cost-complexity
30. Train
31. Test
32. K-fold cross validation
33. Rules

۹. مراجع

- صفارزاده، محمود، شعبانی، شاهین و آذرمی، اکبر (۱۳۸٦) "مدل پیش‌بینی تصادفات در قوس‌های واقع در راههای دوخطه برون‌شهری،" پژوهشنامه حمل و نقل، سال چهارم، شماره ۳، ص. ۲۱۳-۲۲۱.

- سازمان پزشکی قانونی کشور (۱۳۸۸) "آمار مقایسه‌ای متوفیات ناشی از تصادفهای ارجاعی به مراکز پزشکی قانونی"

- Abdel-Aty, M. and Keller, J (2005) "Exploring the overall and specific crash severity levels at signalized intersections", Accident Analysis and Prevention 37(3), pp. 417-425.

- Abdelwahab, H. T. and Abdel-Aty, M. A. (2001) "Development of artificial neural network models to predict driver injury severity in traffic accidents at signalized intersections.", Transportation Research Record 1746(1), pp. 6-13.

- Abdelwahab, H. T. and Abdel-Aty, M. A. (2002) "Artificial neural networks and logit models for

- Kweon, Y. and Kockelman, K. (2003) "Driver attitudes and choices: Seatbelt use, speed limits, alcohol consumption and crash histories". 82nd. Annual Meeting of Transportation Research Board, Washington D. C.

- Mussone, L. and Ferrari, A.(1999) "An analysis of urban collisions using an artificial intelligence model", Accident Analysis and Prevention 31(6), pp. 705-718.

- Renski, H. (1999) "Effect of speed limit increases on crash injury severity: Analysis of single-vehicle crashes on North Carolina interstate highways", Transportation Research Record, 1665(1), pp. 100-108.

- Sohn, S. Y. and Shin, H. (2001) "Pattern recognition for road traffic accident severity in Korea", Ergonomics, 44(1), pp. 107-117.

- SPSS Clementine Software (2008)

- Steinberg, D. and Golovnya, M. (2007) "CART 6.0, User`s Guide".

- Stewart, J. (1996) "Applications of classification and regression tree methods in roadway safety studies", Transportation Research Record: Journal of Transportation Research Board, 1542(1), pp.1-5.

- Tesema, T. B. and Abraham, A. (2005) "Rule mining and classification of road traffic accidents using adaptive regression trees", International Journal of Simulation Systems, 6, pp. 80-94.

- Valent, F. and Schiava, F. (2002) "Risk factors for fatal road traffic accidents in Udine, Italy", Accident Analysis and Prevention, 34(1): pp.71-84.

- Wood, D. and Simms, C. (2002) "Car size and injury risk: a model for injury risk in frontal collisions", Accident Analysis & Prevention, 34(1), pp. 93-99.

- Yau, K. (2004) "Risk factors affecting the severity of single vehicle traffic accidents in Hong Kong", Accident Analysis and Prevention, 36(3), pp. 333-340.

- Zajac, S. and Ivan, J. (2003) "Factors influencing injury severity of motor vehicle-crossing pedestrian crashes in rural Connecticut", Accident Analysis & Prevention, 35(3), pp. 369-379.

traffic safety analysis of toll plazas" , Transportation Research Record 1784(1), pp. 115-125.

- Al-Ghamdi, A. (2002) "Using logistic regression to estimate the influence of accident factors on accident severity.", Accident Analysis and Prevention 34(6), pp.729-741.

- Bedard, M. and Guyatt, G. (2002) "The independent contribution of driver, crash and vehicle characteristics to driver fatalities", Accident Analysis and Prevention, 34(6), p. 717.

- Berry, M. J. A. and Linoff, G. S. (2004) "Data mining techniques: for marketing, sales, and customer relationship management", Wiley Computer Publishing.

- Chang, L. Y. and Wang, H. W. (2006) "Analysis of traffic injury severity: An application of non-parametric classification tree techniques", Accident Analysis and Prevention , 38(5), pp. 1019-1027.

- Chong, M. and Abraham, A. (2005). "Traffic accident analysis using machine learning paradigms", Informatica, 29, pp.89-98.

- Delen, D. and Sharda, R. (2006) "Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks", Accident Analysis and Prevention, 38(3), pp. 434-444.

- Dissanayake, S. and Lu, J. J. (2002) "Factors influential in making an injury severity difference to older drivers involved in fixed object-passenger car crashes", Accident Analysis and Prevention 34(5), pp. 609-618.

- Giudici, P. (2003) "Applied data mining: Statistical methods for business and industry", Wiley.

- Han, J. and Kamber, M. (2006) "Data mining: concepts and techniques", Morgan Kaufmann.

- Khattak, A. [et. al.] (2002). "Risk factors in large truck rollovers and injury severity: Analysis of single-vehicle collisions"

- Kockelman, K. and Kweon, Y. (2002). "Driver injury severity: An application of ordered probit models", Accident Analysis and Prevention 34(3), pp. 313-322.