فصلنامه مهندسی حمل و نقل، دوره ۱۳، شماره ۱، پاییز ۱۴۰۰، ص. ۱۱۴۸-۱۱۳۵
مقاله علمی-پژوهشی

پیش‌بینی تأخیر قطارهای مسافری با استفاده از تکنیک‌های داده‌کاوی

مسعود یقینی، دانشیار، دانشکده‌ی مهندسی راه آهن، دانشگاه علم و صنعت ایران، تهران، ایران
E-mail: yaghini@iust.ac.ir
حامد زارعی، دانش آموخته کارشناسی ارشد، دانشکده‌ی مهندسی راه آهن، دانشگاه علم و صنعت ایران، تهران، ایران

دریافت: ۱۳۹۸/۰۳/۱۱ پذیرش: ۱۳۹۹/۰۱/۱۶

چکیده

هدف از این مقاله، پیش‌بینی تأخیرات قطارهای مسافری در راه‌آهن جمهوری اسلامی ایران با استفاده از تکنیک‌های داده‌کاوی است. پیش‌بینی تأخیرات می‌تواند برای تعیین زمان‌های حائل در جداول زمانی قطارهای مسافری مورد استفاده قرار گیرد. داده‌های مورد استفاده در این مطالعه شامل پایگاه داده تأخیر قطارهای مسافری از سال ۹۲ تا ۹۷ است که دربردارنده ۳۸۰,۷۴۸ رکورد می‌باشد. متغیرهای مستقل جهت پیش‌بینی شامل سال، ماه، روز ماه، روز هفته، ساعت حرکت، محورهای حرکت، نوع قطار، نوع سالن، مبدا و مقصد قطار و همچنین نام مالک قطار هستند. مدلسازی پیش‌بینی به دو صورت عددی و طبقه‌ای انجام شده است. جهت پیش‌بینی طبقه‌ای، داده‌های تأخیر با استفاده از روش خوشه‌بندی دو مرحله‌ای گسسته‌سازی شده‌اند. از دو روش شبکه عصبی و C5.0 جهت پیش‌بینی طبقه‌ای و سه روش رگرسیون، CHAID و شبکه عصبی برای پیش‌بینی عددی استفاده شده است. برای ارزیابی نتایج پیش‌بینی، مدل ساخته شده ابتدا بر اساس داده‌های سال‌های ۹۲ تا ۹۵ آموزش می‌بینند، سپس به پیش‌بینی تأخیر برای سال ۹۶ می‌پردازد. نتایج نشان می‌دهد که در پیش بینی عددی، روش شبکه عصبی و در پیش‌بینی به صورت طبقه‌ای، روش C5.0 از دقت بالاتری نسبت به سایر روش‌ها برخوردار هستند، لذا از این دو تکنیک برای پیش‌بینی تأخیر قطارهای سال ۹۷ استفاده شده است. همچنین جهت ارزیابی تکرارپذیری نتایج، تاخیرات پیش بینی شده سال ۹۷ با سال ۹۶ مقایسه گردیده اند. در انتها، پیش‌بینی عددی به صورت گروه‌بندی بر روی رکوردهای پایگاه داده نیز انجام شده است. نتایج نشان می‌دهد که دقت پیش‌بینی به صورت گروه‌بندی، بالاتر است.

واژه‌های کلیدی: پیش‌بینی، تأخیر قطارهای مسافری، تکنیک‌های داده‌کاوی، راه‌آهن ایران

۱. مقدمه

تأخیر قطارهای مسافری از مهمترین چالش‌های سیستم‌های ریلی درتمام دنیا به شمار می‌آید که هزینه‌های زیادی را برای مسافران و اپراتورها اعمال می‌کند و موجب ناکارآمدی عملیات قطارها می‌شود [Van Oort, 2011]. سیستم راه‌آهن شامل چندین زیر سیستم از جمله زیرساخت شبکه، آلات ناقله، کنترل و ارتباطات، سیاست‌ها و قوانین مختلف عملیاتی است که هدف آن ارائه خدمات قابل اطمینان برای حمل مسافر یا کالا می‌باشد. با این حال، عدم اطمینان‌های متعددی ممکن است از این زیر سیستم‌ها ایجاد شود که می‌تواند فعالیت‌های برنامه‌ریزی شده را مختل کند و موجب تأخیر غیر منتظره گردد. خطوط راه‌آهن به علت حجم زیاد، قطارهای بزرگ و زیرساخت‌های محدود، بسیار متراکم هستند. گاهی اوقات یک رویداد نسبتا جزئی مانند تأخیر خدمه یا نقص فنی در واگن یا لکوموتیو می‌تواند اثرات موجدار بزرگی را در سراسر شبکه راه‌آهن ایجاد کند و باعث تأخیر در عملیات تعداد زیادی از قطارها شود [Wen et al, 2017]. اعتبار تمامی سطوح برنامه‌ریزی عملیات راه‌آهن، از جمله ایجاد جداول زمانی قابل اجرا، پیش‌بینی ترافیک در زمان واقعی، پیش‌بینی ناسازگاری‌ها و ارائه اطلاعات قابل اطمینان برای مسافر، به شدت بستگی به تخمین دقیق از زمان‌های عملیات قطار دارد که این زمان‌ها وابسته به تأخیرها هستند [Zhang and Yang, 2018].

با استفاده از پیش‌بینی تأخیر قطارهای مسافری می‌توان قطارهای با تأخیر زیاد را شناسایی کرد و نسبت به برنامه‌ریزی مناسب جهت کاهش تأخیرها و اثرات ضربه‌ای[1] آنها اقدام نمود. در هنگام طراحی جداول زمانی، می‌توان با محاسبه میانگین تأخیرات پیش‌بینی شده برای هر قطار، یک بافر زمانی در نظر گرفت تا در صورت بروز تأخیر در یک قطار، از انتشار آن تأخیر در قطارهای بعدی جلوگیری شود که نتیجه آن بهینه‌سازی زمان‌های سیر قطارها و استفاده بهتر از ظرفیت‌های منابع ریلی و طراحی بهتر جداول زمانی است.

در این مطالعه، از ویژگی‌های جدیدی شامل ساعت حرکت قطار، نام مالک قطار و نوع سالن‌های مسافری جهت پیش‌بینی تأخیر قطار استفاده شده است که در هیچ یک از مطالعات قبلی انجام شده، این ویژگی‌ها به مدل پیش‌بینی اضافه نشده‌اند. مدلسازی پیش‌بینی تأخیر قطار به دو صورت عددی و طبقه‌ای انجام شده است که در آن تأخیر قطارهای سال ۹۷ که داده‌های آن موجود نبود نیز پیش‌بینی شده‌اند. همچنین میانگین تأخیرات پیش‌بینی شده قطارهای سال ۹۷ جهت تعیین بافر در جداول زمانی قطارها محاسبه شده است. بمنظور بالاتر بردن دقت مدل، پیش‌بینی به صورت گروه‌بندی بر روی رکوردهای تأخیر نیز انجام شده است.

ساختار این مقاله به این صورت سازماندهی شده است که ابتدا در بخش دوم، مرور ادبیات موضوع بررسی می‌شود. در بخش سوم مطابق متدولوژی CRISP-DM، داده‌های جمع‌آوری شده تأخیر قطار تشریح شده‌اند، سپس داده‌ها پاکسازی و تحت فرآیند پیش‌پردازش قرار می‌گیرند. در مرحله بعد، با استفاده از تکنیک‌های داده‌کاوی، مدلسازی پیش‌بینی تأخیر قطار صورت گرفته و در انتها نتایج حاصل شده مورد ارزیابی قرار می‌گیرند. در بخش چهارم به جمع‌بندی و نتیجه‌گیری پرداخته می‌شود.

۲. مروری بر ادبیات موضوع

در ادامه به بررسی برخی از مقالاتی که در بازه زمانی سال‌های ۲۰۰٤ تا ۲۰۱۸ در رابطه با پیش‌بینی تأخیر منتشر شده است، پرداخته می‌شود.

چن و همکاران [Chen et al, 2004] یک مدل پویا جهت پیش‌بینی زمان‌های رسیدن اتوبوس‌ها ارائه نموده‌اند. مدل آنها از دو قسمت تشکیل شده است، قسمت اول بر مبنای شبکه عصبی بوده و جهت پیش‌بینی زمان‌های رسیدن اتوبوس‌ها مورد استفاده قرار می‌گیرد. قسمت دوم از الگوریتم پویا بر مبنای فیلتر کالمن استفاده

می‌کند تا پیش‌بینی زمان‌های رسیدن را با استفاده از اطلاعات بر مبنای دقیقه محل اتوبوس تنظیم کند.

یوان [Yuan, 2006] به مدل سازی تأخیرهای قطار و انتشار تأخیر در ایستگاه‌ها پرداخته است تا بتواند بهره‌برداری از ظرفیت و طراحی جدول زمانی را در سطح قابلیت اطمینان مورد نظر بهبود دهد. بر اساس مدل انتشار تأخیر به پیش‌بینی تأخیرهای ضربه‌ای قطار پرداخته شده است.

میر و همکارانش [Meer, Goverde and Hansen, 2010] مدلی جهت پیش‌بینی زمان سیر قطارها ارائه کرده‌اند. هدف اصلی این مدل پیش‌بینی انتشار تأخیر در شبکه ریلی است. در این پژوهش جهت یافتن تخمین دقیق وزن کمان‌ها (زمان‌های سیر و سرفاصله‌ها) از روش‌های داده‌کاوی بر روی داده‌های اشغالی خطوط استفاده شده است.

هنسن و همکاران [Hansen et al, 2010] به پیش‌بینی تأخیر و زمان حرکت قطار پرداختند. در این مقاله از داده‌کاوی برای محاسبه دقیق تأخیر قطار در ایستگاه‌ها استفاده شده است. هدف این مقاله پیش‌بینی دقیق زمان حرکت واقعی قطار تا ایستگاه بعدی است. داده‌های مورد استفاده در این تحقیق متعلق به کشور هلند می‌باشد.

کلو و همکاران [Clue et al, 2011] به آشکارسازی الگوهای تأخیرهای ضربه‌ای قطار پرداخته‌اند. هدف این مقاله شناسایی الگوهای قطارهایی است که تأخیر را به یکدیگر انتقال می‌دهند که به آن تأخیر ضربه ای گفته می‌شود. در این تحقیق به منظور آشکارسازی الگوهای تأخیرهای ضربه ای از روش اپیزودهای مکرر استفاده شده است تا مجموعه هایی از تأخیرهای قطار که به طور مکرر اتفاق می افتند، مشخص شوند.

یقینی و خوشرفتار [Yaghini and Khoshraftar, 2013] به پیش‌بینی تأخیر قطارهای مسافری با استفاده از شبکه‌های عصبی پرداختند. در این مقاله، مدلی مبتنی بر شبکه‌های عصبی پیشخور با دقت بالا برای پیش‌بینی تأخیر قطارهای مسافری درراه آهن جمهوری اسلامی ایران ارائه شده است.

یک مدل برای پیش‌بینی تأخیر قطار در زمان واقعی با استفاده از شبکه بیزی در مقاله کچمن و گاوردی [Kechman and Goverde, 2015] ارائه شده است. آنها یک ماه از داده‌های تحلیلی ترافیکی از مدیریت زیرساخت راه‌آهن سوئد در یک محیط شبیه سازی شده در زمان واقعی استفاده کردند. نتایج محاسباتی نشان داد که پیش‌بینی‌ها برای افق‌های ۳۰ دقیقه‌ای قابل اعتماد هستند. با این وجود فرض اصلی آنها این است که ترتیب و مسیرهای قطار در افق پیش‌بینی شناخته شده است، که اغلب در جهان واقعی وجود ندارد.

لسان و همکاران [Lessan, Liping and Chao, 2018] یک مدل پیش بینی تأخیر قطار بر پایه شبکه بیزی ارائه کرده‌اند. سه نوع ساختار ابتکاری، خطی و ترکیبی شبکه بیزی بر روی داده‌های تأخیر قطارهای پر سرعت به کار بردند. اعتبارسنجی حاصل از مدل نشان داده است که یک مدل مبتنی بر شبکه بیزی می‌تواند یک ابزار کارآمد برای رفع اثرات متقابل تأخیرهای قطار باشد. نتایج بدست آمده از این تحقیق نشان می‌دهد که مدل پیشنهاد شده می‌تواند دقت پیش‌بینی بیش از ۸۰٪ برای افق‌های ٦۰ دقیقه‌ای داشته باشد.

۳. متدولوژی تحقیق

برای انجام این تحقیق، از مراحل متدولوژی CRISP-DM[۲] که متدولوژی پروژه‌های داده‌کاوی در سازمان‌ها می‌باشد، الهام گرفته شده است. این متدولوژی دربردارنده شش فاز اصلی شامل: ۱) شناخت سازمان، ۲) شناخت داده‌ها، ۳) پیش‌پردازی داده‌ها، ٤) مدلسازی، ۵) ارزیابی و ٦) پیاده‌سازی است. مراحل بعدی این تحقیق بر اساس فازهای شماره ۲ تا ۵ این متدولوژی می‌باشد.

۳-۱ شناخت داده‌های تأخیر

این فاز با جمع آوری اولیه داده و آشنایی با آنها آغاز می‌گردد. در ادامه کیفیت داده‌ها مورد بررسی قرار می‌گیرند، شناخت دقیق‌تری روی داده‌ها حاصل می‌شود و مجموعه داده‌هایی که می‌توانند جهت پشتیبانی فرضیه‌های پروژه مورد استفاده قرار گیرند، شناسایی می‌گردند. داده‌های مورد استفاده در این تحقیق، بخشی از پایگاه داده تأخیر قطارهای مسافری در راه‌آهن ج.ا.ا. می‌باشد. این پایگاه داده مربوط به بازه زمانی سال‌های ۹۲ تا ۹۷ است که در آن برای هر سفر ریلی انجام شده با یک قطار مسافری در تاریخ مشخص، یک رکورد ثبت شده است. جدول ۱ شامل بعضی از فیلدهای مهم این پایگاه داده می‌باشد. در جدول ۲ نیز خلاصه‌ای از پایگاه داده تأخیر قطارهای مسافری ارائه شده است.

جدول ۱. فیلدهای پایگاه داده تأخیر قطار

عنوان فیلد	توضیح فیلد
مسیر حرکت	نشان دهنده مبدا- مقصد سفر است.
محور	محوری که سفر در آن صورت گرفته است.
تاریخ حرکت	تاریخ شمسی که سفر در آن آغاز شده است.
نام مالک	نام شرکتی که مالک قطار است.
علت تأخیر	علتی که باعث تأخیر قطار شده است.
تأخیر در مبدا	مدت تأخیر قطار هنگام حرکت از ایستگاه مبدا.
کل تأخیر	مدت تأخیر قطار هنگام ورود به ایستگاه مقصد.

جدول ۲. خلاصه داده‌های جمع آوری شده از پایگاه داده تأخیر

سال	تعداد قطارهای اعزامی	مجموع تأخیر (دقیقه)	میانگین تأخیر قطار(دقیقه)
۹۲	٦٢,٠٥٢	٢,٠١٢,٤٠١	٣٢/٤٣
۹۳	٦٤,٥٦٨	٢,٦٤٨,٩٩٦	٤١/٠٢
۹٤	٦٤,١٨٤	١,٩١٢,٢٦٧	٢٩/٧٩
۹٥	٦٤,٨٣٠	٨٧٧,١٣٢	١٣/٥٢
۹٦	٦٣,٤٤٧	٧٦٨,٥٤٧	١٢/١١
۹۷	٦١,٦٦٧	٧١١,٦٣٥	١١/٥٣
مجموع	٣٨٠,٧٤٨	٨,٩٣٠,٩٧٨	٢٣/٤ (میانگین)

۳-۲ آماده‌سازی داده‌ها

با توجه به این که ثبت داده‌ها توسط کاربران انسانی انجام می‌شوند، وجود خطا در داده‌ها امری طبیعی است. بنابراین داده‌ها باید قبل از مدلسازی، آماده سازی شوند. فاز آماده‌سازی و پیش‌پردازش داده‌ها شامل تمام فعالیت‌هایی می‌گردد که روی داده‌های خام صورت می‌پذیرد تا ورودی‌های مناسبی را برای ابزارهای داده‌کاوی مهیا سازند. فعالیت‌های صورت گرفته از انتخاب جداول، رکوردها و ویژگی‌ها تا تبدیل و پاکسازی داده‌های پرت را در بر می‌گیرد.

برای بالا بردن دقت پیش‌بینی در این تحقیق، فیلدهای ساعت حرکت، نوع سالن مسافری و نوع قطار از برنامه حرکت قطارهای مسافری بر اساس شماره قطار استخراج شده است و به پایگاه داده اولیه اضافه شدند. برای استفاده بهتر از ویژگی تاریخ حرکت، این ویژگی به سه فیلد مجزا شامل سال حرکت، ماه حرکت و روز حرکت تبدیل شده است. ویژگی روز هفته نیز از تقویم استخراج شده و به پایگاه داده اضافه شده است. ویژگی مسیر حرکت نیز به دو فیلد مجزا مبدا و مقصد تبدیل شده است. پس از اضافه کردن ویژگی‌های مهم، تبدیل بعضی از فیلدها به فیلدهای دیگر و همچنین حذف بعضی از ویژگی‌های غیر ضروری، ویژگی‌های نهایی برای استفاده در مدل پیش‌بینی مطابق جدول ۳ می‌باشد.

جدول ۳. ویژگی‌های نهایی پایگاه داده جهت پیش‌بینی

عنوان فیلد	توضیح فیلد
سال	سال شمسی که در آن سفر انجام شده است.
ماه	ماهی که در آن سفر انجام شده است.
روز	روز ماه که در آن سفر انجام شده است.
روز هفته	روزی از هفته که در آن سفر انجام شده است.

عنوان فیلد	توضیح فیلد
مبدا	نشان دهنده مبدا سفر است.
مقصد	نشان دهنده مقصد سفر است.
محور	محوری که سفر در آن صورت گرفته است.
ساعت حرکت	ساعتی که سفر آغاز شده است.
نام مالک	نام شرکتی که مالک قطار است.
نوع قطار	قطار از نوع خودکشش بوده است یا عادی.
نوع سالن	نوع سالن های مسافری که قطار دارا بوده است.

۳-۲-۱ پاکسازی داده‌های پرت

به منظور حذف اثر نامطلوب مقادیر پرت از یک روش آماری تحت عنوان روش دامنه بین چارکی[۳] استفاده شده است. در این شیوه اگر اعداد خارج از فاصله $[Q_1 - 1.5\times IQR, Q_3 + 1.5\times IQR]$ قرار گیرند به عنوان داده پرت محسوب شده و حذف می‌شوند. Q_i چارک i ام است و IQR اختلاف چارک اول و سوم می‌باشد. نخست با استفاده از فرمول (۱)، چارک اول و سوم محاسبه خواهد شد [Seo, 2006].

$$Q_p = L_p + \left(\frac{p\times n - g_p}{f_p} \right) \times w \qquad (۱)$$

در رابطه بالا، n تعداد داده‌ها، L_p مرز پایین رده، Q_p یعنی نخستین رده‌ای که فراوانی انباشته آن برابر یا بیشتر از $n\times p$ است، g_p فراوانی انباشته رده بلافاصله قبل از رده Q_p، f_p فراوانی رده Q_p و w طول رده می‌باشد. مقادیر چارک‌ها و دامنه میان چارکی در جدول ٤ ارائه شده است.

جدول ٤. مقادیر چارک ها و دامنه میان چارکی

Q_1	Q_3	IQR	$Q_1 - 1.5\times IQR$	$Q_3 + 1.5\times IQR$
۰	۳۶	۳۶	-۵۴	۹۰

با توجه به این جدول، بازه [۹۰، ٥٤-] به عنوان بازه‌ی مشخص کردن مقادیر پرت است. بر اساس این روش، تأخیر قطارهایی که بیش از ۹۰ دقیقه باشند، داده پرت محسوب شده و از پایگاه داده حذف می‌شوند. در این پایگاه داده تعداد ۱۲,٦٦٨ رکورد دارای مقادیر پرت بودند که حدود ۳/۳٪ از کل پایگاه داده را شامل می‌شدند، لذا از پایگاه داده حذف شدند. خلاصه داده‌های تأخیر قطار پس از فرآیند پاکسازی داده‌های پرت، مطابق جدول ۵ می‌باشد.

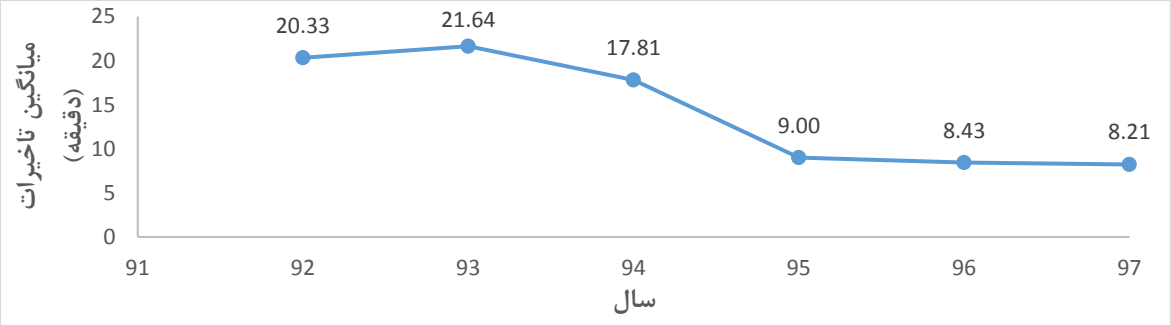
جدول ۵. خلاصه داده‌های تأخیر قطار پس از فرآیند پاکسازی

سال	تعدادکل قطارهای اعزامی	مجموع تأخیر (دقیقه)	میانگین تأخیر قطار(دقیقه)
۹۲	٦۱,۳۸۷	۱,۱۲٦,۰۱۳	۲۰/۳۳
۹۳	٦۰,٦٦۰	۱,۱۸۸,۲٦۷	۲۱/٦٤
۹٤	٦۱,۱۲۲	۱,۰۲۹,٦۲۰	۱۷/۸۱
۹٥	٦۳,۳٥۸	٥٦٤,۰۳٦	۹/۰۰
۹٦	٦۱,٦۷٥	٥۲۰,٦٤۲	۸/٤۳
۹۷	٥۹,۸۷۸	٤۹٤,۸٤۱	۸/۲۱
مجموع	۳٦۸,۰۸۰	٤,۹۲۳,٤۱۹	۱٤/۲ (میانگین)

۳-۲-۲ تحلیل‌های آمار توصیفی از پایگاه داده تأخیر قطار

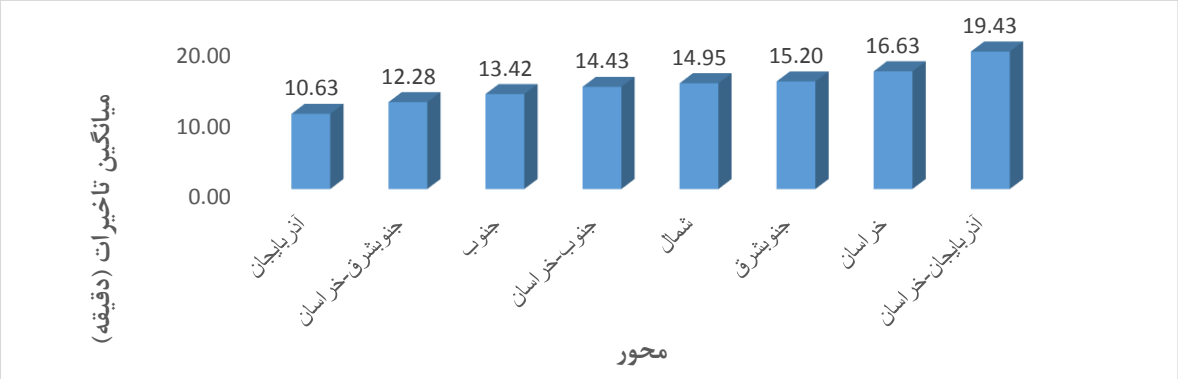
پس از پاکسازی داده‌های پرت، تحلیل‌های آمار توصیفی بر روی پایگاه داده تأخیر قطار صورت گرفته است. تحلیل آماری تأخیرات قطار، نه تنها یک دید در مورد ویژگی‌های تأخیر می‌دهد بلکه کمک به شناسایی منابع تأخیر می‌کند. شکل ۱ نشان دهنده میانگین تأخیر برای هر قطار در سال‌های مختلف است. سال ۹۳ بیشترین میانگین تأخیر را داشته است. از سال ۹۳ به بعد، میانگین تأخیر روند نزولی دارد. شکل ۲ نشان دهنده میانگین تأخیر ماهانه قطارهای مسافری می‌باشد. همانطور که مشاهده می‌شود، ماه شهریور با میانگین ۱۹/۸۱ دقیقه بیشترین میانگین تأخیر و ماه اسفند با میانگین ۱۱/٥٦ دقیقه کمترین میانگین تأخیر را برای هر قطار دارند. میانگین تأخیر هر یک از ۸ محور راه‌آهن ایران در شکل ۳ ارائه شده است. با توجه

به این نمودار، کمترین مقدار تأخیر مربوط به محور آذربایجان و بیشترین مقدار تأخیر مربوط به محور آذربایجان–خراسان است.



شکل ۱. میانگین تأخیر قطارها از سال ۹۲ تا ۹۷

شکل ۲. میانگین تأخیر ماهانه قطارهای مسافری از سال ۹۲ تا ۹۷



شکل ۳. میانگین تأخیر هر محور

۳-۲-۳ تحلیل همبستگی

پیش از ساختن مدل پیش‌بینی لازم است که متغیرهای مستقل ورودی برای ساختن مدل پیش‌بینی تعیین شوند. متغیرهایی که در

ساخت مدل پیش‌بینی از آن‌ها استفاده شده است، شامل سال، ماه، روز ماه، روز هفته، ساعتی که قطار اعزام شده است، محورهایی که قطار بر روی آنها حرکت می‌کند، نوع قطار (قطار از نوع خودکشش بوده است یا عادی)، نوع سالن، مبدا و مقصد قطار و نام مالک قطار می‌باشد. با استفاده از آزمون مربع‌کای پیرسون[4] که می‌تواند ارتباط بین دو متغیر رده‌ای را آزمایش کند، می‌توان وابستگی ۲*۲ را در بین ویژگی‌ها اندازه‌گیری کرد. اگر مقدار آزمون بین بازه $[0, \chi^2_{\alpha(r-1)\times(c-1)}]$ باشد، فرضیه استقلال بین دو ویژگی پذیرفته شده است، در غیر این صورت رد می‌شود. در این رابطه، α سطح معناداری است که در اینجا ۰/۰۵ در نظر گرفته می‌شود و همچنین r و c نیز به ترتیب تعداد ردیف‌ها و ستون‌ها می‌باشند. وابستگی بین دو ویژگی دسته‌ای بالا است درصورتی که احتمال عدم وابستگی کمتر از ۰/۰۵ باشد [Yaghini et al, 2010]. مقادیر آزمون مربع‌کای پیرسون برای متغیرهای پیش‌بینی مذکور، مطابق جدول ٦ می‌باشد.

جدول ٦. مقادیر آزمون مربع‌کای پیرسون

ویژگی‌ها	مقادیر آزمون مربع‌کای پیرسون	درجه آزادی	احتمال عدم وابستگی
تأخیر، سال	٨٤,٩٩٤	٩٧٢	۰
تأخیر، ماه	١٨,٦٩٤	٢,١٤٥	۰
تأخیر، روز ماه	٣,٦١٠	٥,٨٥٠	۰
تأخیر، روز هفته	٣,٠٠٨	١,١٧٠	۰
تأخیر، محور	١٠٠,٠٩٥	١,٧٥٥	۰
تأخیر، نوع قطار	٤,٥١٢	١٩٥	۰
تأخیر، ساعتحرکت	٨٥٤,٤٦٣	٦٤,١٥٢	۰
تأخیر، مبدا	١١٨,٣٥٦	٢,١٨٧	۰
تأخیر، نوع سالن	٣٨٦,٢٧٧	٣٠,١٣٢	۰
تأخیر، مالک	١٨٠,٤٠٨	٣,٦٤٥	۰
تأخیر، مقصد	٦١,٠٥٤	٢,٩١٦	۰

٣-٢-٤ داده های آموزشی و آزمایشی

داده‌ها به دو دسته شامل مجموعه داده‌های آموزشی[5] و داده‌های آزمایشی[6] تقسیم شده‌اند که ارزیابی مدل پیش‌بینی بر اساس داده‌های آزمایشی سنجیده می‌شود. از داده‌های آزمایشی برای مقایسه میزان واقعی آنها با میزان تخمین آنها استفاده می‌شود تا دقت مدل پیش‌بینی تعیین شود. ابتدا مدل بر اساس داده‌های آموزشی آموزش می‌بیند و سپس به پیش‌بینی تأخیر برای داده‌های آزمایشی می‌پردازد.

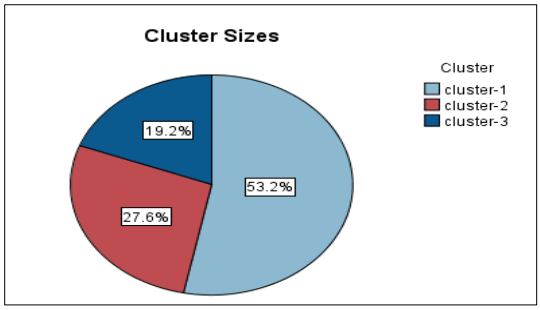
٣-٣ پیش بینی به صورت طبقه‌ای بر روی کل پایگاه داده

٣-٣-١ گسسته‌سازی داده‌های پیوسته

با استفاده از روش خوشه‌بندی دومرحله‌ای[8]، داده‌های تأخیر قطار به سه بازه با اندازه‌های مختلف تقسیم شده‌اند. جدول ٧ بازه تأخیر هر برچسب خوشه‌بندی را نشان می‌دهد. همچنین فراوانی هر خوشه در شکل ٤ ارائه شده است.

جدول ٧. بازه تأخیر هر خوشه

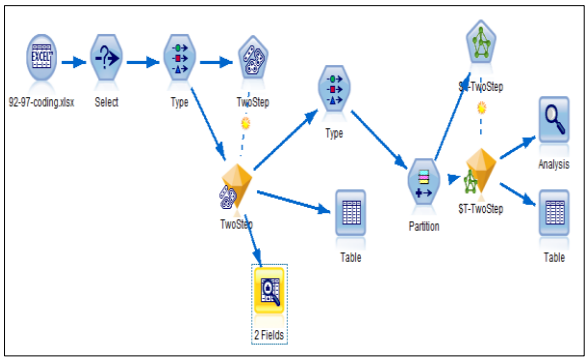
شماره خوشه	بازه تأخیر
١	[١٣-٠]
٢	[٤٢-١٤]
٣	[٩٠-٤٣]



شکل ٤. فراوانی هر خوشه

۳-۳-۲ مدلسازی

از دو روش درخت تصمیم و شبکه عصبی برای مدلسازی پیش‌بینی تأخیر قطار به صورت طبقه‌ای در محیط نرم افزار SPSS Modeler 18.0 استفاده شده است. در این تحقیق، یک نوع خاص از شبکه عصبی به نام پرسپترون چند لایه (MLP) [۹] مورد استفاده قرار می‌گیرد. پرسپترون چند لایه، یک شبکه‌ی آموزشی تحت نظارت است که دارای حداکثر دو لایه پنهان است. برای ساختن درخت تصمیم از الگوریتم C5.0 استفاده شده است. در شکل ۵ نمایی از نرم افزار مورد استفاده نشان داده شده است.



شکل ۵. مدلسازی پیش‌بینی با تکنیک شبکه عصبی

۳-۳-۳ ورودی‌های مدل

متغیرهای مورد استفاده در مدل پیش‌بینی شامل روز هفته، روز، ماه، مبدا، مقصد، نوع سالن، نوع قطار، نام مالک قطار و ساعت حرکت می‌باشند.

۳-۳-٤ خروجی‌های مدل

در مساله کلاس‌بندی تأخیر، خروجی مدل دقیقا مشخص نیست. هر یک از برچسب‌های بدست آمده با استفاده از روش خوشه‌بندی، یک واحد خروجی را ایجاد می‌کنند. در نهایت، مدل پیشنهادی می‌تواند یکی از این برچسب‌ها را پیش‌بینی کند؛ به این معنی است که زمان تقریبی تأخیر پیش‌بینی می‌شود.

۳-۳-۵ ارزیابی مدل پیش‌بینی

برای اینکه توانایی مدل آموزش دیده با روش‌های پیشنهادی در پیش‌بینی برای دوره‌های آتی مشخص گردد، می‌توان مدل‌هایی بر اساس دوره‌های زمانی (سال، ماه، هفته یا هر دوره زمانی دلخواه دیگر) قبلی ایجاد کرده تا تأخیر را در دوره زمانی بعدی پیش‌بینی کند. مدل بر اساس داده‌های سال‌های ۹۲ تا ۹۵ آموزش می‌بیند (درواقع داده‌های سال‌های ۹۲ تا ۹۵ مجموعه آموزشی را تشکیل می‌دهند)، سپس از مدل آموزش دیده برای پیش‌بینی تأخیر قطارهای مسافری در سال ۹۶ استفاده خواهد شد. تقسیمات و تعداد الگوها درجدول ۸ ارائه شده است.

جدول ۸. تعداد الگوها در مجموعه داده‌های آموزشی و آزمایشی

تعداد الگو	
مجموعه آموزشی	مجموعه آزمایشی
۲٤٦,٥۲۷	٦۱,٦۷٥

جدول ۹ مقایسه‌ای بین دقت پیش‌بینی حاصل از اجرای مستقل الگوریتم C5.0 و روش شبکه عصبی بر روی پایگاه داده تأخیر قطار را نشان می‌دهد. همانطور که مشاهده می‌شود، الگوریتم C5.0 دقت بالاتری نسبت به روش شبکه عصبی دارد.

جدول ۹. مقایسه دقت پیش‌بینی شبکه عصبی و C5.0

روش	مجموعه آموزشی	مجموعه آزمایشی
	دقت پیش‌بینی	دقت پیش‌بینی
C5.0	٤۸/۸٦٪	٥٥/۷٦٪
شبکه عصبی	٦٤/۷۸٪	۷۷/۸۲٪

در جدول ۱۰، مشخصات شبکه عصبی استفاده شده ارائه شده است.

جدول ۱۰. مشخصات شبکه عصبی

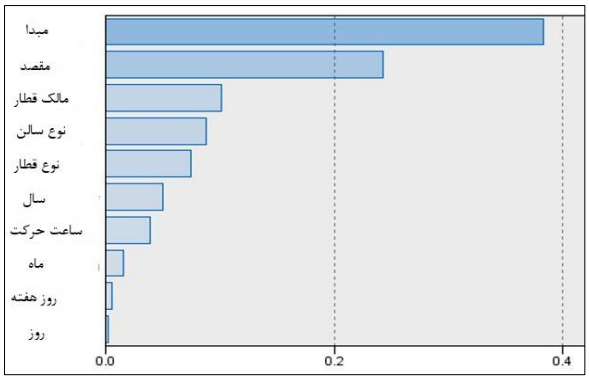
نوع شبکه	تعداد نرونهای لایه ورودی	تعداد نرونهای لایه مخفی اول	تعداد نرونهای لایه خروجی
MLP	۹	۱۱	۱

میزان دقت روش شبکه عصبی برای هر بازه پیش‌بینی بر روی مجموعه آموزشی مطابق شکل ٦ است. همانطور که مشاهده می‌شود، دقت برچسب اول که مربوط به بازه تأخیر صفر تا ۱۳ دقیقه می‌باشد، ۹۰/۷٪ است. دقت برچسب دوم که مربوط به بازه تأخیر ١٤ تا ٤٢ دقیقه است، ٦٣/٨٪ می‌باشد. دقت برچسب سوم نیز که مربوط به بازه ٤٣ تا ۹۰ دقیقه است، ۷۹/۷٪ می‌باشد.

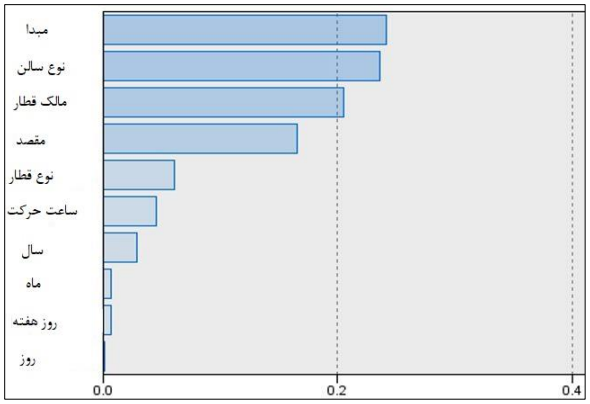
Observed	Predicted		
	cluster-1	cluster-2	cluster-3
cluster-1	90.7%	6.6%	2.7%
cluster-2	29.4%	63.8%	6.8%
cluster-3	13.9%	6.4%	79.7%

شکل ٦. میزان دقت پیش‌بینی با شبکه عصبی بر روی مجموعه آموزشی

اهمیت متغیرهای پیش‌بینی‌کننده برای روش شبکه عصبی و C5.0 به ترتیب در شکل‌های شماره ۷ و ۸ ارائه شده است. در هر دو روش، متغیرهای مبدا، مقصد، نوع سالن و مالک قطار بیشترین اهمیت را در بین متغیرهای پیش‌بینی‌کننده داشته‌اند.



شکل ۷ . اهمیت متغیرهای پیش‌بینی‌کننده با روش شبکه عصبی



شکل ۸ . اهمیت متغیرهای پیش‌بینی‌کننده با الگوریتم C5.0

۳-۳-٦ پیش‌بینی تأخیر قطارهای سال ۹۷

در این بخش، از الگوریتم C5.0 که دقت بالاتری نسبت به شبکه عصبی برخوردار بود، جهت پیش‌بینی تأخیر قطارهای مسافری سال ۹۷ استفاده شده است.

۳-۳-۷ ارزیابی مدل پیش‌بینی

مدل بر اساس داده‌های سال‌های ۹۲ تا ۹٦ آموزش می‌بیند و به پیش‌بینی تأخیر قطارهای سال ۹۷ می‌پردازد. تقسیمات و تعداد الگوها در جدول ۱۱ ارائه شده است.

جدول ۱۱. تعداد الگوها در مجموعه داده‌های آموزشی و آزمایشی

مجموعه آموزشی	*مجموعه آزمایشی*
۳۰۸,۲۰۲	۵۹,۸۷۸

میزان دقت پیش‌بینی با استفاده از الگوریتم C5.0 در جدول ۱۲ ارائه و همچنین جهت آزمون تکرارپذیری نتایج، با دقت پیش‌بینی سال ۹٦ مقایسه گردیده است.

جدول ۱۲. دقت پیش‌بینی الگوریتم C5.0

سال	مجموعه آموزشی	مجموعه آزمایشی
۹٦	٤٨/٨٦٪	٥٥/٧٦٪
۹۷	٤١/٧٦٪	٥٥/٧٣٪

۳-٤ پیش بینی به صورت عددی بر روی کل پایگاه داده

برای مدلسازی پیش‌بینی عددی تأخیر قطار، سه تکنیک رگرسیون خطی، درخت تصمیم CHAID و شبکه عصبی به کار گرفته شده است.

۳-٤-۱ ورودی‌ها و خروجی‌های مدل

در اینجا، ورودی‌های مدل دقیقا همانند ورودی‌های مدل پیش‌بینی به صورت طبقه‌ای می‌باشد. خروجی مدل برای پیش‌بینی عددی، یعنی مجموع تأخیرها مشخص است، در واقع مدل یک خروجی داشته که بر اساس آن مقدار عددی پیش‌بینی تأخیر مشخص می‌شود.

۳-٤-۲ ارزیابی مدل پیش‌بینی

برای ارزیابی مدل، همانند پیش‌بینی به صورت طبقه‌ای، داده‌های سال‌های ۹۲ تا ۹۵ مجموعه آموزشی و داده‌های سال ۹٦ مجموعه آزمایشی را تشکیل می‌دهند. مقایسه میانگین و انحراف استاندارد سه روش رگرسیون، شبکه عصبی و CHAID درجدول ۱۳ ارائه شده است. با توجه به جدول، کمترین میانگین خطای مطلق مربوط به روش شبکه عصبی با مقدار ۱۲/۳۰ دقیقه است.

جدول ۱۳. مقایسه میانگین و انحراف استاندارد سه روش رگرسیون، شبکه عصبی و CHAID

روش	مجموعه آموزشی		مجموعه آزمایشی	
	میانگین خطای مطلق(دقیقه)	انحراف استاندارد (دقیقه)	میانگین خطای مطلق (دقیقه)	انحراف استاندارد (دقیقه)
CHAID	۱۲/۵۲	۱۷/٦۰	۱٦/٤٤	۲۲/٤۹
رگرسیون	۱۹/٦۵	۲۳/۵۳	۱۹/۷٤	۲۰/۲٤
شبکه عصبی	۱۲/۵٦	۱۷/٦۱	۱۲/۳۰	۱۷/۰٦

در جدول ۱٤ مشخصات شبکه عصبی استفاده شده ارائه شده است.

جدول ۱٤. مشخصات شبکه عصبی

نوع شبکه	تعداد نرونهای لایه ورودی	تعداد نرونهای لایه مخفی اول	تعداد نرونهای لایه خروجی
MLP	۹	۹	۱

۳-٤-۳ پیش‌بینی عددی تأخیر قطارهای سال ۹۷

برای پیش‌بینی عددی تأخیر قطارهای مسافری سال ۹۷، از روش شبکه عصبی که دقت بالاتری نسبت به دو روش CHAID و رگرسیون برخوردار بود، استفاده شده است. مدل بر اساس داده‌های سال‌های ۹۲ تا ۹٦ آموزش می‌بیند و به پیش‌بینی تأخیر قطارهای سال ۹۷ می‌پردازد.

۳-٤-٤ ارزیابی مدل پیش‌بینی

مدل بر اساس داده‌های سال‌های ۹۲ تا ۹٦ آموزش می‌بیند و به پیش‌بینی تأخیر قطارهای سال ۹۷ می‌پردازد. میزان دقت پیش‌بینی با استفاده از روش شبکه عصبی در جدول ۱۵ ارائه شده و با دقت پیش‌بینی سال ۹٦ مقایسه گردیده است.

جدول ۱۵. دقت پیش‌بینی روش شبکه عصبی

سال	مجموعه آموزشی		مجموعه آزمایشی	
	میانگین خطای مطلق(دقیقه)	انحراف استاندارد (دقیقه)	میانگین خطای مطلق (دقیقه)	انحراف استاندارد (دقیقه)
۹٦	۱۲/۵٦	۱۷/٦۱	۱۲/۳۰	۱۷/۰٦
۹۷	۱٤/۵۳	۱۹/۸۸	۱٤/۵۱	۱۹/۸٦

۳-٤-۵ میانگین تأخیرات پیش‌بینی شده قطارهای سال ۹۷

در این بخش، میانگین تأخیرات پیش‌بینی شده برای هر قطار محاسبه می‌شود. از آنجایی که در راه‌آهن ایران، برنامه حرکت قطارهای مسافری برای چهار دوره (نوروز، بهار، تابستان، پاییز و زمستان) طراحی می‌شود، میانگین تأخیرات پیش‌بینی شده قطار برای هر دوره محاسبه خواهد شد. یک نمونه از نتایج در اینجا ارائه

شده است. میانگین تأخیرات پیش‌بینی شده برای قطار با مشخصات جدول ١٦، ١٣/٩٣دقیقه می‌باشد.

جدول ١٦. مشخصات قطار

عنوان فیلد	مقدار فیلد
ماه	٤،٥،٦
مبدا	تهران
مقصد	مشهد
ساعت حرکت	٢٢:٠٠
مالک قطار	رجا
نوع قطار	عادی
نوع سالن	پلور سبز رجا ٤٠

٣-٤-٦ پیش‌بینی عددی به صورت گروه‌بندی رکوردهای تأخیر

بمنظور بالاتر بردن دقت مدل و بهبود میزان برازش مدل با الگوهای تأخیر قطارهای مسافری و همچنین تحلیل ساده‌تر نتایج مدل می‌توان پایگاه داده را به گروه‌های مختلفی از رکوردهای تأخیر تقسیم نموده و سپس پیش‌بینی را بر اساس گروه‌های ایجاد شده انجام داد. در اینجا یک گروه بر اساس متغیرهای مبدا، مقصد، ماه حرکت، ساعت حرکت و نوع قطار ایجاد شده است. مقادیر این متغیرها مطابق جدول ١٧ می‌باشند.

جدول ١٧. مقادیر هر متغیر جهت گروه‌بندی

عنوان فیلد	مقدار فیلد
ماه	٤،٥،٦
مبدا	مشهد
مقصد	تهران
ساعت حرکت	از ٦ تا ١٢
نوع قطار	خودکشش

پس از اجرای مدل پیش‌بینی عددی به صورت گروه‌بندی با سه روش شبکه عصبی، رگرسیون و CHAID، نتایج بدست آمده در جدول ١٨ ارائه شده است.

جدول ١٨. مقایسه میانگین و انحراف استاندارد پیش‌بینی عددی به صورت گروه‌بندی رکوردهای تأخیر

روش	مجموعه آموزشی		مجموعه آزمایشی	
	میانگین خطای مطلق (دقیقه)	انحراف استاندارد (دقیقه)	میانگین خطای مطلق (دقیقه)	انحراف استاندارد (دقیقه)
CHAID	٤/٨١	٧/٩٢	٣/٩٤	٣/٧٧
رگرسیون	٥/٠٦	٨/٠٣	٣/٨٩	٣/٤٨
شبکه عصبی	٤/٧١	٧/٩٦	٣/٣٦	٣/٦٣

همانطور که از مقادیر این جدول مشخص است، میانگین خطای مطلق هر سه روش نسبت به پیش‌بینی بر روی کل پایگاه داده کاهش داشته است و دقت مدل پیش‌بینی بسیار بالاتر رفته است.

٤. جمع بندی، نتیجه‌گیری و پیشنهادات

در این تحقیق، به پیش‌بینی تاخیر قطارهای مسافری در خطوط راه‌آهن جمهوری اسلامی ایران با استفاده از تکنیک‌های داده‌کاوی پرداخته شد. برای این منظور مراحل متدولوژی داده‌کاوی CRISP-DM برای پایگاه داده تأخیر قطارهای مسافری به کار گرفته شد. داده‌هایی که در این تحقیق جهت پیش‌بینی تأخیر استفاده شد، شامل پایگاه داده تأخیر قطارهای مسافری راه‌آهن ایران از سال ٩٢ تا ٩٧ بود. مدلسازی پیش‌بینی تأخیر قطار به دو صورت عددی و طبقه‌ای بر روی کل پایگاه داده انجام شد. پیش از انجام پیش‌بینی به صورت طبقه‌ای، داده‌های تأخیر قطار با استفاده از روش خوشه‌بندی دومرحله‌ای در سه برچسب با بازه‌های مختلف خوشه‌بندی شدند.

نتایج نشان داد که در پیش‌بینی عددی، روش شبکه عصبی دقت بالاتری نسبت به روش‌های رگرسیون و CHAID داشت، لذا از روش شبکه عصبی برای پیش‌بینی عددی تأخیر قطارهای سال ۹۷ استفاده شد.

در پیش‌بینی به صورت طبقه‌ای نیز، روش C5.0 از دقت بالاتری نسبت به روش شبکه عصبی برخوردار بود، لذا از این تکنیک برای پیش‌بینی تأخیر قطارهای سال ۹۷ به صورت طبقه‌ای استفاده شد. در انتها، پیش‌بینی عددی به صورت گروه‌بندی بر روی رکوردهای پایگاه داده تأخیر نیز محاسبه شد، نتایج نشان داد که دقت پیش‌بینی بالاتر از زمانی است که پیش‌بینی بر روی کل پایگاه داده صورت گیرد.

شرکت‌های حمل و نقل ریلی با به کارگیری نتایج بدست آمده از این تحقیق می‌توانند برنامه‌ریزی‌های مناسبی را برای تأخیرات قطارها انجام دهند. می‌توان قطارهای با تأخیر زیاد را شناسایی کرد و نسبت به برنامه‌ریزی مناسب جهت کاهش تأخیرها و اثرات ضربه‌ای آنها اقدام نمود. در هنگام طراحی جداول زمانی، می‌توان با محاسبه میانگین تأخیرات پیش‌بینی شده برای هر قطار، یک بافر زمانی[10] در نظر گرفت تا در صورت بروز تأخیر در یک قطار، از انتشار آن تأخیر در قطارهای بعدی جلوگیری شود که نتیجه آن بهینه‌سازی زمان‌های سیر قطارها و استفاده بهتر از ظرفیت‌های منابع ریلی موجود و طراحی بهتر جداول زمانی است.

۵. پی‌نوشت‌ها

1. Knock on Delay
2. Cross Industry Standard Process for Data Mining
3. Interquartile range
4. Pearson’s chi-square
5. Training set
6. Testing set
7. Over fitting
8. Two-step
9. Multilayer Perceptron (MLP)
10. Buffer Time

٦. منابع

- شرکت قطارهای مسافری رجا (۱۳۹٦) "پایگاه داده‌های تأخیرات قطارهای مسافری"، تهران: شرکت قطارهای مسافری رجا.

- یقینی، مسعود، خوشرفتار، محمدمهدی و سیدآبادی، سیدمسعود (۱۳۸۹) "پیش‌بینی تأخیر قطارهای مسافری با استفاده از شبکه‌های عصبی"، پژوهشنامه حمل و نقل، دوره ۷ ، شماره ۳، صفحه ۲۹۱ تا ۳۰۳.

- Van Oort, N. (2011) "Service reliability and urban public transport design service reliability", Ph.D. thesis TRAIL Research School.

- Wen, C., Li, Z., Lessan, J., Fu, L., Huang, P., & Jiang, C. (2017) "Statistical investigation on train primary delay based on real records: Evidence from Wuhan–Guangzhou HSR", International Journal of Rail Transportation, Vol. 5, No.3, pp.170-189

- Zhang, H., Li, S., & Yang, L. (2018) "Real-time optimal train regulation design for metro lines with energy-saving", Computers & Industrial Engineering, Vol. 127, pp.1282-1296.

- IBM SPSS Modeler 18.0 Algorithms Guide, (2016).

- Chen, M., Liu, X., Xia, J., Chien, S. (2004) "A Dynamic Bus-Arrival Time Prediction Model Based on APC Data", Computer-Aided Civil and Infrastructure Engineering, Vol. 19, No.5, p.p. 364–376.

- Yuan, J. (2006) "Stochastic Modelling of Train

Delays and Delay Propagation in stations", PhD dissertation, Delft University of Technology, Faculty of Civil Engineering and Geosciences, Department of Transportation and Planning.

- Meer, D.J., Goverde, R.M.P., Hansen, I.A. (2010) "prediction of Train running Times and conflicting using track occupation data", 12th WCTR-World Congress of Transportation Research, Lisbon, Portugal, July 2010.

- Hansen, Ingo A., Rob MP Goverde, and Dirk J. van der Meer. (2010) "Online train delay recognition and running time prediction", 13th International IEEE Conference on Intelligent Transportation Systems, pp. 1783-1788.

- Clue, B., Goethals, B., Tassenoy, S., & Verboven, S. (2011) "Mining train delays", International Symposium on Intelligent Data Analysis, pp. 113-124.

- Yaghini, M., Khoshraftar, M. M., & Seyedabadi, M. (2013) "Railway passenger train delay prediction via neural network model", Journal of advanced transportation, Vol. 47, No.3, pp.355-368.

- Kecman, P., & Goverde, R. M. (2015b) "Predictive modelling of running and dwell times in railway traffic", Public Transport, Vol. 7, No.3, pp.295–319.

- Lessan, J., Fu, L., & Wen, C. (2018) "A hybrid Bayesian network model for predicting delays in train operations", Computers & Industrial Engineering, Vol. 127, pp.1214-1222.

- Seo, S. (2006) "A review and comparison of methods for detecting outliers in univariate data sets" (Doctoral dissertation, University of Pittsburgh).

مسعود یقینی، درجه دکتری در رشته مهندسی حمل و نقل ریلی در سال ۱۳۸۲ از دانشگاه جیائوتونگ پکن اخذ نمود. زمینه‌های پژوهشی مورد علاقه ایشان تکنیک‌های بهینه‌سازی و داده‌کاوی در حوزه حمل و نقل ریلی است.

حامد زارعی، درجه کارشناسی در رشته مهندسی صنایع را در سال ۱۳۹٤ از دانشگاه صنعتی سجاد و درجه کارشناسی ارشد در رشته مهندسی حمل و نقل ریلی در سال ۱۳۹۷ را از دانشگاه علم و صنعت ایران اخذ نمود. زمینه‌های پژوهشی مورد علاقه ایشان مسائل داده‌کاوی و تکنیک‌های آن در حوزه حمل و نقل ریلی است.