ارایه یک روش ترکیبی تحلیل پوششی داده‌ها[1] و درخت تصمیم جهت سنجش کارایی واحدها (مطالعه موردی: ناوگان اتوبوسرانی شهر تهران)

مهسا صفی*، دانش آموخته کارشناسی‌ارشد، دانشکدهٔ مهندسی صنایع، دانشگاه خواجه نصیرالدین طوسی، تهران، ایران
سمیه علیزاده، استادیار، دانشکدهٔ مهندسی صنایع، دانشگاه خواجه نصیر الدین طوسی، تهران، ایران
عماد روغنیان، استادیار، دانشکدهٔ مهندسی صنایع، دانشگاه خواجه نصیر الدین طوسی، تهران، ایران
پست الکترونیکی نویسنده مسئول: m.safi@sina.kntu.ac.ir
دریافت: ۱۳۹۳/۰۶/۲۵- پذیرش: ۱۳۹۳/۱۰/۱۵

چکیده

سنجش عملکرد و کارایی ناوگان اتوبوسرانی از جمله مسائلی است که به جهت نیاز به ارتقا هر چه بیشتر سطح خدمت رسانی به شهروندان همواره مورد توجه سیاست‌گذاران و متولیان مدیریت شهری بوده است. از طرفی دستیابی به این دانش که چه شرایطی در افزایش کارایی خطوط اتوبوسرانی مؤثرند، از اهمیت بالایی برخوردار می‌باشد. داده‌های تاریخی مربوط به عملکرد خطوط اتوبوسرانی همه روزه توسط شبکه ارتباطی بیسیم در سطح شهر جمع آوری می‌شوند. در این پژوهش داده‌های مذکور به کمک یک روش ترکیبی جدید مورد بررسی و پردازش قرار گرفته اند. کارایی خطوط توسط روش DEA سنجیده شده و در ادامه بر اساس کارایی یا عدم کارایی خطوط، به وسیله روش درخت تصمیم قواعد مستتر در داده‌ها که نشان دهنده‌ی وضعیت کارایی خطوط می‌باشند، استخراج شده است. در نهایت توانایی پیش‌بینی کارایی یا ناکارایی خطوط توسط الگوریتم‌های کارت[2]، C5.0 و مدل گروهی[3] مورد ارزیابی قرار گرفته است. این روش ترکیبی تاکنون در حوزه حمل و نقل استفاده نشده است و در عین حال در این روش از قابلیت الگوریتم‌های گروهی جهت افزایش صحت[4] مدل در پیش بینی کارایی خطوط استفاده شده است. به این ترتیب به وسیله‌ی ترکیب روش DEA و درخت تصمیم به عنوان دو روش ناپارامتریک، نقشه ی راهی جهت ارتقا کارایی ناوگان حمل و نقل در این حوزه، در اختیار مسئولان امر قرار می‌گیرد.

واژه‌های کلیدی: حمل و نقل درون شهری، کارایی[5]، تحلیل پوششی داده‌ها، داده کاوی[6]، درخت تصمیم[7]

۱- مقدمه

گسترش و بهبود حمل و نقل عمومی تأثیر بسزایی در توسعه پایدار کشورها دارد به طوری که امروزه هزینه و انرژی زیادی صرف تحقیق در این زمینه می‌شود تا روز به روز راهکارهای مناسب‌تری در این حوزه ارایه شوند. در این راستا سنجش عملکرد و اندازه گیری کارایی در این حوزه جهت ایجاد بهبود و توسعه سامانه‌های حمل و نقل، به صورت گسترده‌ای مورد استفاده قرار می‌گیرد. در این میان سامانه اتوبوسرانی کلان شهرها نقش بسزایی را در جابه‌جایی مسافر دارا می‌باشند. اتوبوسرانی شهرها جهت خدمت‌رسانی به مناطق مختلف شهری، با اختصاص خطوط اتوبوسرانی که هریک منابع خاص خود را دارا

می‌باشند، اقدام به پوشش بخش عظیمی از حمل و نقل درون شهری می‌نماید. با توجه به تأثیر گسترده ناوگان حمل و نقل اتوبوسرانی در جابه‌جایی مسافر، سنجش عملکرد و کارایی این بخش در بهبود عملکرد آن نقش بسزایی ایفا می‌کند. در این خصوص روش‌های مختلفی جهت سنجش کارایی واحدهای تصمیم گیرنده[8] تحت بررسی، وجود دارد. روش ارایه شده در این مقاله روشی ترکیبی از داده‌کاوی و تحلیل پوششی داده است.

روش تحلیل پوششی داده‌ها به عنوان یک ابزار مدیریتی قوی و روشی غیر پارامتریک در سال ۱۹۷۸ توسط چارنز[9]، کوپر[10] و رودز[11]، معرفی شد. به وسیله این روش کارایی یک مجموعه واحد تصمیم‌گیرنده، اندازه‌گیری می‌شود, (Cooper and Rodes,1978)

داده‌کاوی به علت در دسترس بودن مقادیر عظیم داده و نیاز حتمی به استفاده از این داده‌ها جهت تبدیل به دانش، توجهات زیادی را به خود جلب نموده است. به بیان ساده داده‌کاوی اشاره به استخراج یا کاوش دانش و اطلاعات از داده‌هایی با اندازه‌های بزرگ دارد (Han and Kamber, 2006).

در بخش بعد مروری بر مطالعات انجام گرفته در زمینه مدل‌های ترکیبی داده‌کاوی و تحلیل پوششی داده‌ها خواهیم داشت. پس از آن هر یک از روش‌های درخت تصمیم و تحلیل پوششی داده‌ها توضیح داده خواهند شد. در ادامه روش کار مورد نظر ارایه می‌شود. در بخش پنجم مدل ارایه شده با استفاده از داده‌های واقعی در حوزه حمل و نقل اجرا می‌شود. در پایان نتیجه‌گیری از دستاوردهای مقاله ارایه خواهد شد.

۲- پیشینه تحقیق

کارایی در واقع بیانگر این مفهوم است که یک مجموعه به چه میزان از منابع خود در راستای تولید، نسبت به بهترین عملکرد در مقطعی از زمان استفاده کرده است (مهرگان،۱۳۸۳). در طول ۲۰ سال گذشته تحقیقات بسیاری در خصوص ارزیابی عملکرد حمل و نقل عمومی به‌وسیله DEA، صورت گرفته است. به طور نمونه چو و همکاران، یک شاخص منفرد برای اندازه گیری کارایی و اثربخشی خدمات در شرکت‌های ارایه دهنده خدمات با توجه به پژوهش‌های پیشین (جدول۱) که از روش‌های ترکیبی استفاده شده بود، هیچگاه از مدل‌های ترکیبی درخت تصمیم که دو یا چند الگوریتم درخت تصمیم را ترکیب کند، استفاده نشده است. از طرفی روش‌های ترکیبی داده‌کاوی و DEA تاکنون در حوزه حمل و نقل، اجرا نشده‌اند. در این پژوهش، دو نکته‌ی مذکور لحاظ شده است و از این نظر نسبت به مطالعات پیشین تفاوت دارد. در یک نمونه، یک شاخص منفرد برای اندازه‌گیری کارایی و اثربخشی خدمات در شرکت‌های ارایه‌دهنده خدمات حمل و نقل عمومی ارایه شد ,Chu). (Fielding and Lamar,1992. در موردی دیگر، با استفاده از داده‌های گروهی از شرکت‌های حمل و نقل اتوبوسرانی در فاصله ۱۹۸۹ تا ۱۹۹۳ توسط روش DEA کارایی فنی شرکت‌ها سنجیده شده است. در این مقاله با استفاده از رگرسیون، ارتباط میان امتیاز کارایی و ویژگی‌های شرکت‌ها ارزیابی شده است (Nolan,1996). همین طور در سال۲۰۰۱ ناکارایی مربوط به عوامل خارجی و متغیرهای ورودی و خروجی در مدل‌های DEA مشخص شده و در واقع این نظر وجود داشته که حل کامل‌تر مسئله DEA علاوه بر ورودی‌ها و خروجی‌های مسئله، نیازمند در نظر گرفتن عوامل بیرونی یا جانبی مؤثر نیز می‌باشد (Boile',2001).

کارلافتیس در پژوهش خود نشان داد که مقیاس بهینه عملیاتی در سامانه‌های حمل و نقل به صورت مشخصی متغیر است و این موضوع ارتباط مستقیمی با نحوه مشخص کردن خروجی‌های مدل DEA دارد (Karlaftis,2004) همچنین، با ترکیب DEA و داده‌های استخراجی از سامانه‌های اطلاعات جغرافیایی[12]، ارزیابی خطوط اتوبوسرانی با در نظر گرفتن اطلاعات دموگرافیک مناطق و مدنظر قرار دادن عوامل جانبی محیط عملیاتی در ارزیابی کارایی واحدها، انجام گرفته است Lao (and Liu,2009).

در عین حال در خصوص صنعت حمل و نقل ایران نیز مطالعاتی با استفاده از تکنیک‌های داده‌کاوی انجام گرفته است (یقینی و همکاران، ۱۳۸۹) علاوه بر موارد ذکر شده، رویکردی که اخیراً در برخی کارهای پژوهشی به صورت موردی دیده می‌شود، استفاده از ترکیب تکنیک‌های داده‌کاوی به همراه DEA است.

جدول ۱. کارهای پژوهشی با استفاده از مدل ترکیبی DEA و داده کاوی

موضوع	ویژگی	مدل DEA	مدل داده کاوی		
			C5.0	کارت	مدل گروهی
تجاری‌سازی فن‌آوری (Sohn and Moon, 2004)	نخستین کار پژوهشی با استفاده از روش ترکیبی	BCC (مدل ۲)	✓		
ارزیابی فرآیندها و روش‌ها (Seol et al., 2007)	استفاده از نتایج مدل جهت بهینه کاوی[۱۳]	CCR (مدل ۱)		✓	
انتخاب تأمین‌کنندگان (Wu, 2009)	مقایسه هر دو مدل DEA و در نظر گرفتن واحدهای کارا جهت استخراج قواعد	BCC-CCR (مدل ۱و۲)		✓	
ارایه پیشنهادات نظارتی در کسب و کار (Lee, 2010)	در نظر گرفتن واحدهای کارا جهت استخراج قواعد	CCR (مدل ۱)		✓	
بانکداری (Emrouznejad and Anouze,2010)	لحاظ‌کردن داده‌های محیطی علاوه بر ورودی‌ها وخروجی‌ها	BCC (مدل ۲)		✓	
مراکز درمانی (Nicola,Gitto and Mancuso,2012)	استفاده از روش بوت استرپ[۱۴] در DEA	BCC (مدل ۲)		✓	
حمل و نقل درون شهری (پژوهش حاضر)	استفاده از روش گروهی جهت افزایش صحت	BCC (مدل ۲)	✓	✓	✓

جدول ۲. نمونه داده‌های ساختاری

تاریخ	شماره اتوبوس	کد راننده	نام خط	سریال کارت	نوع کارت	مبلغ باقی مانده	اعتبار باقی مانده	نرخ کرایه	تعداد ایستگاه	تعداد کیوسک بلیت فروشی	میانگین فاصله ایستگاه‌ها
18-Jun-11	۱۱۲۱۵۶۲	۵۰۸۰۱	شهرک رسالت–م. محمدیه	۳۵۲۸۰۰۳۰۷٤	اعتباری	۱٦۰۰۰		۱۲٦۰	۲۱	۱	٦۱۵
1-Jun-11	۳٤۲۱۸۸۵	۵۹۰۷۷	مترو شهرری– پارکینگ شهرری	۲۲۹۳۵٤٦٤۵۹	اعتباری	۳٦٤۵۰		٤۲۰	٤	۱	۳۸۱
15-Jun-11	۳۲۰۰۷۲۷	۳۸۱۳۱	پ.معین–پ.تجریش	۲۵۷۹٤۸۱۲	٦ماهه	۰	20-Jun-11	۲۰۰۰	۳۵	۰	۵۰۲
21-Jun-11	۱۰۷۵	۱٦۵۱۷	چهارراه تهرانپارس– پ.آزادی	۷٦۸۷۵۳٤۲۰	اعتباری	۱۰٤٤۰		۱٦۸۰	۳۱	۰	٦۱۳
2-Jun-11	۱۳۲۱۵٤۲	۵۵۵۳٦	م. کلانتری– م. بهارستان	۲۲۹۲٤۲۳٦۲۷	نقدی	–	–	٤۲۰	۱۰		٤۹۵

۳- تشریح روش‌های تحلیل پوششی داده‌ها و درخت تصمیم

۳-۱-تحلیل پوششی داده‌ها

تحلیل پوششی داده‌ها روشی ناپارامتریک است که در ابتدا یک مرز کارا را با استفاده از مجموعه‌ای از واحدهای تصمیم گیرنده تشکیل داده سپس سطوح کارایی را نسبت به سایر واحدهایی که روی مرز کارا قرار نگرفته‌اند، بر طبق فاصله‌ی آن‌ها نسبت به مرزکارا تخصیص می‌دهد (Liu et al.,2013).

یکی از مدل‌های کلاسیک DEA، مدل CCR می‌باشد که نام آن از ابتدای اسامی چانز، کوپر و رودز گرفته شده است. فرض کنید m ورودی و s خروجی داریم. v_i وزن ورودی iام و u_r وزن خروجی rام است x_{ij} مقدار ورودی iام DMU jام و y_{rj} خروجی rام DMU_j ام است. مدل CCR ورودی محور به شرح زیر می باشد (مهرگان،۱۳۸۳):

$$\text{Max} Z_0 = \sum_{r=1}^{s} u_r y_{rj0}$$

Subject to

$$\sum_{i=1}^{m} v_i x_{ij0} = 1 \qquad (۱)$$

$$\sum_{r=1}^{s} u_r y_{rj} - \sum_{i=1}^{m} v_i x_{ij} \le 0$$

$u_r, v_i \ge \varepsilon$ i=1,….,m , j = 1,……,n , r = 1,……,s

این مدل تنها زمانی مناسب است که همه‌ی واحدها در مقیاس بهینه فعالیت کنند. اما گاهی اوقات این امکان با توجه به ورودی‌ها و خروجی‌های مسئله فراهم نمی‌باشد. در این حالت مدل دیگری با نام BCC که از حرف ابتدای اسامی بنکر، چارنز و کوپر گرفته شده است، مطرح شد. تفاوت آن نسبت به مدل CCR یا (۱) بازده به مقیاس متغیر[۱۵] بودن آن می‌باشد. در حالت بازده به مقیاس ثابت[۱۶] هر مضربی از ورودی‌ها همان مضرب از خروجی‌ها را تولید می‌کنند. اما در حالت بازده به مقیاس متغیر این حالت وجود ندارد. و تغییرات خروجی‌ها می‌توانند به همان نسبت تغییرات ورودی‌ها نباشند. در مدلBCC یا (۲) علامت w_0 وضعیت بازده به مقیاس بودن مدل را مشخص می‌کند. اگر $w_0 > 0$ باشد بازده به مقیاس مثبت است، اگر $w_0 = 0$ معادل بازده به مقیاس ثابت است و اگر $w_0 < 0$ باشد نشان دهنده بازده به مقیاس منفی است. مدل خروجی محور BCC نیز به شکل زیر است (مهرگان،۱۳۸۳):

$$\text{Min} Z_0 = \sum_{i=1}^{m} v_i x_{ij0} + w_0$$

Subject to

$$\sum_{r=1}^{s} u_r y_{rj0} = 1 \qquad (۲)$$

$$\sum_{r=1}^{s} u_r y_{rj} - \sum_{i=1}^{m} v_i x_{ij} + w_0 \ge 0$$

$u_r, v_i \ge \varepsilon$ i=1,….,m , j = 1,……,n r = 1,……,s,

W متغیر آزاد در علامت؛

مدل DEA مورد استفاده در این مقاله یک مدل خروجی محور BCC (مدل۲) است. هدف کلی یک خط اتوبوسرانی، جابه‌جایی هر چه بیشتر مسافران (خروجی) است. دلیل استفاده از مدل BCC این است که در این حالت فرض بازده به مقیاس متغیر بودن در نظر گرفته می‌شود. این موضوع بیانگر این است که کارایی ممکن است با تغییر در اندازه خروجی یا ورودی، افزایش یا کاهش یابد. حالت بازده به مقیاس متغیر تناسب بیشتری با وضعیت خطوط اتوبوسرانی دارد که علت آن تغییرپذیری بالای کارایی در این حوزه است (Lao and Lio, 2009).

۳-۱-۱- ورودی‌ها و خروجی‌های مدل

پس از انتخاب نوع مدل DEA، گام بعدی انتخاب ورودی‌ها و خروجی‌هاست. رایج‌ترین عواملی که بعنوان ورودی و خروجی جهت سنجش شرکت‌های حمل و نقل در نظر گرفته می‌شوند، عبارتند از: نیروی کار، سرمایه و انرژی به عنوان ورودی و مسافران یا تعداد کیلومتر- صندلی استفاده شده توسط مسافر، به عنوان خروجی. در این مقاله تعداد اتوبوس‌های فعال و طول مسیری که یک خط طی می‌کنند، به عنوان ورودی در نظر گرفته

شده‌اند. تعداد اتوبوس فعال در واقع بیانگر امکانات سخت‌افزاری آن خط و طول مسیرش بیانگر قابلیت بالقوه‌ی آن خط در جذب هر چه بیشترخروجی (مسافر) می‌باشد. تعداد مسافران جابه‌جا شده در یک خط می‌توانند ملموس‌ترین خروجی ایجاد شده توسط یک خط اتوبوسرانی باشد؛ اما این نکته نیز باید در نظر گرفته شود که در معیار تعداد مسافر جابه‌جا شده مدت زمان خدمت‌رسانی آن خط نیز باید مدنظر قرار گیرد. چرا که ممکن است دو خط یک تعداد مسافر جابه‌جا کرده باشند، اما مدت زمان خدمت‌رسانی آن‌ها متفاوت بوده باشد. این حالت وقتی پیش می‌آید که اتوبوس‌های یک خط بنا به دلایلی نظیر خرابی ناوگان، مدت زمانی خارج از خدمت قرار گرفته باشد.

۳-۲- درخت تصمیم

درخت تصمیم یک روش یادگیری سمبلیک است و اطلاعاتی که از یک مجموعه داده آموزشی استخراج شده را در یک ساختار سلسله‌مراتبی که از یکسری گره و شاخه تشکیل شده، سازماندهی می‌کند.در همه الگوریتم‌های درخت تصمیم دو متغیر قابل تعریف هستند: متغیرهای مستقل و متغیر وابسته. متغیر وابسته یا متغیر هدف[17]در یک درخت دسته‌بندی مقادیرش را از مقادیرگسسته و در یک درخت رگرسیون از مقادیر پیوسته، انتخاب می‌کند (Nie et al. ,2011). برای ایجاد یک درخت تصمیم، مجموعه داده‌ها حداقل به دو دسته تقسیم می‌شوند: داده‌های آموزشی و داده‌های آزمون. درخت تصمیم یک مدل را بر اساس داده‌های آموزشی می‌سازد و در ادامه جهت جلوگیری از بیش برازش مدل، هرس درختان انجام می‌گیرد (Emrouznejad and Anouze,2010). دو نمونه از رایج‌ترین الگوریتم‌های درخت تصمیم، کارت و $C_{5.0}$ می‌باشند.

۳-۲-۱- درخت C5.0

روش دسته‌بندی $C_{5.0}$ اخیراً در سال ۲۰۰۷ توسط کوئینلن توسعه یافته است(Quinlan, 2007). این الگوریتم قادر به ارتقاء[18] مدل جهت افزایش صحت در نمونه داده‌های تحت بررسی می‌باشد. الگوریتم $C_{5.0}$ امکان ایجاد درخت چندتایی را دارد. مجموعه آموزشی داده‌ها به صورت بازگشتی به زیر مجموعه‌های کوچک‌تر تقسیم می‌شوند و به این ترتیب درخت توسعه می‌یابد. راهکار ارایه شده توسط الگوریتم جهت محدود کردن رشد درخت، هرس درخت تصمیم می‌باشد (Chou, 2012).

۳-۲-۲- مدل گروهی

یک مدل گروهی به عنوان مجموعه‌ای از دسته‌بندهای[19] آموزش دیده تعریف می‌شود که هنگام مواجه با داده‌ی جدید جهت دسته‌بندی، با هم ترکیب می‌شوند. مدل گروهی خروجی چندین دسته‌بند را به صورت یک دسته‌بند مرکب ترکیب می‌کند. هدف از ترکیب همه‌ی دسته‌بندها با هم، ایجاد یک مدل گروهی است که صحت دسته‌بندی را در مقایسه با تک‌تک دسته بندها ارتقا دهد (Pujari and Gupta,2012).

٤- روش پیشنهادی

شِمای کلی چارچوب ارایه شده در این پژوهش در شکل ۱ نمایش داده شده است. در این طرح شماتیک ارتباط روش‌های DEA و درخت تصمیم و داده‌های مورد استفاده نشان داده شده است. با توجه به ورودی‌ها و خروجی‌های خطوطِ تحت بررسی که اطلاعات آماری مربوط به آن‌ها در جدول ٤ نمایش داده شده است، مدل (۲) اجرا می‌شود تا نتایج آن در مرحله بعد مورد استفاده قرار گیرد.

٤-۱- جمع‌آوری و شناخت داده‌ها

داده‌های این مقاله مربوط به سامانه بلیت الکترونیک ناوگان اتوبوسرانی شهر تهران می‌باشد. هر مسافر در هر بار استفاده از ناوگان اتوبوسرانی به‌وسیله‌ی کارت بلیت الکترونیک، هزینه کرایه را از طریق دستگاه‌های کارتخوان پرداخت می‌کند. در پروژه بلیت الکترونیک داده‌های مربوط به تراکنش‌های حاصل از استفاده‌ی مسافران از ناوگان اتوبوسرانی شهر تهران به صورت روزانه ثبت

و ذخیره می‌شود. داده‌های مربوط به هر تراکنش، از طریق شبکه ارتباطی بیسیم موجود در سطح شهر جمع‌آوری شده و در مرکز داده‌ی شهرداری تهران جمع‌آوری می‌شوند. جهت جمع‌آوری داده‌های مورد نظر این پژوهش، از پایگاه داده‌ی پروژه‌ی بلیت الکترونیک، مجموعه داده‌ی مورد نیاز استخراج شده است. این داده‌ها جهت انجام مطالعه حاضر ادغام شده و به‌صورت یکپارچه درآمده‌اند. حجم داده‌های تحت بررسی در حدود ۱/۵ میلیون رکورد می‌باشد و از نظر زمانی مربوط به خرداد ماه سال ۱۳۹۰ می‌باشد. داده‌ها به دو دسته کلی تقسیم می‌شوند.

داده‌های ساختاری که مربوط به مشخصات خطوط می‌باشند ورودی‌ها و خروجی‌های مدل DEA را به ما می‌دهند و دیگری داده‌های مربوط به تراکنش‌های ثبت شده است. داده‌های ساختاری از طریق اتوبوسرانی تهران در قالب مشخصات خطوط اتوبوسرانی دریافت شده است. این داده‌ها شامل فیلدهایی نظیر کد راننده، شماره اتوبوس، زمان خدمت‌رسانی یا حمل مسافر توسط اتوبوس، نام خط و تعداد مسافر می‌باشد. داده‌های مذکور به عنوان ورودی و خروجی‌های DEA استفاده خواهند شد. داده‌های تراکنشی نیز از پایگاه داده‌ی شهرداری تهران استخراج شده‌اند. این داده‌ها در استخراج قواعد بوسیله درخت تصمیم مورد استفاده قرار خواهند گرفت و شامل فیلدهایی نظیر تاریخ وقوع تراکنش، شماره اتوبوس، کدراننده، نام خط، شماره سریال کارت، نوع کارت مورد استفاده، ساعت وقوع تراکنش، مبلغ باقی مانده کارت، اعتبار زمانی کارت، نرخ کرایه، تعداد ایستگاه‌ها، تعداد کیوسک بلیت فروشی و میانگین فاصله میان ایستگاه‌های یک خط می‌باشد.

خطوط تحت بررسی در این تحقیق ۱۱۷ مورد می‌باشند. نمونه داده‌های ساختاری مربوط به خطوط اتوبوسرانی تهران مطابق جدول ۲ می‌باشد. این جدول نمایشگر داده‌های مربوط به مدت زمان خدمت‌رسانی هر اتوبوس در ناوگان حمل و نقل اتوبوسرانی می‌باشد. به این صورت که زمان کارکرد هر اتوبوس با توجه به شماره پلاک آن، در طول زمان تحت بررسی که خرداد ۱۳۹۰ می‌باشد؛ محاسبه شده و جمع‌آوری می‌شود تا از مجموع فیلدهای زمان خدمت‌رسانی و تعداد مسافر به عنوان خروجی‌های DEA استفاده شود. نمونه داده‌های تراکنشی نیز در جدول ۳ آمده است.

٤-۲- اجرای روش DEA

در گام بعدی باید با استفاده از ورودی‌ها و خروجی‌های خطوط اتوبوسرانی، با استفاده از مدل (۲)، کارایی خطوط محاسبه شود.

٤-۳- اجرای روش درخت تصمیم

در ادامه جهت اجرای درخت تصمیم از نتیجه‌ی اجرای روشDEA، که عنوان کارا یا ناکارا برای هر خط اتوبوس می‌باشد، برچسب دسته های مربوط به هر خط به آن تخصیص می‌یابد. پس از آن به کمک مجموعه داده‌های تراکنشی که فیلدهای آن‌ها در جدول ۳ نمایش داده شده‌اند؛ قواعد مربوط به عملکرد خطوط استخراج می‌شوند.

۵- اجرای روش پیشنهادی و تجزیه و تحلیل نتایج

توصیف آماری ارایه شده در جدول ٤ با استفاده از داده‌های ساختاری و تراکنشی ارایه شده است. به این صورت که میانگین طول مسیر از داده‌های ساختاری و بقیه موارد از داده‌های تراکنشی استخراج شده است.

۵-۱- نتایج

در این مقاله سه مدل کارت، $C_{5.0}$ و مدل گروهی این دو درخت بر روی مجموعه داده‌ها، اجرا شده‌اند. قواعد حاصل از اجرای درخت تصمیم $C_{5.0}$ مطابق جدول ۵ می‌باشد.

قاعده اول بیانگر این موضوع است که در مسیرهایی که از نظر مسافت متوسط می‌باشند، اگر باجه بلیت فروشی وجود نداشته باشد، آن خط ناکارا خواهد بود. قانون دوم نیز به نوعی تأیید کننده قانون اول می باشد. قانون سوم نیز مزیت استفاده از بلیت الکترونیک را نشان می‌دهد. در همین شرایط اگر مسافران از

کارت بلیت استفاده کنند، خط کارا خواهد بود. قواعد استخراجی حاصل از به کارگیری الگوریتم کارت نیز در جدول ٦ نشان داده است.

قواعد اول و دوم بیان کننده این نکته هستند که در مسیرهای کوتاه و متوسط، در صورتی که تعداد باجه‌های بلیت فروشی نسبت به ایستگاه‌ها بیشتر از نصف باشند، خط کارا خواهد بود. به عبارت دیگر در مسیرهای نه چندان طولانی نیاز به باجه‌های بلیت فروشی بیشتر است. قاعده سوم نشان‌دهنده‌ی این نکته است که در مسیرهای طولانی، بهتر است تعداد باجه‌های بلیت فروشی نسبت به ایستگاه‌ها، بسیار کم باشند.

۵-۲- اعتبارسنجی

جهت ارزیابی عملکرد الگوریتم مورد استفاده، از شاخص صحت استفاده می‌شود. این شاخص نشان دهنده‌ی میزان پیش بینی صحیح الگوریتم دسته‌بندی می‌باشد. جهت محاسبه‌ی این شاخص ذکر برخی تعاریف ضروری است. "صحیح مثبت" داده‌های مثبتی هستند که بدرستی توسط الگوریتم پیش‌بینی شده‌اند درحالی که "صحیح منفی" داده‌های منفی هستند که بدرستی پیش‌بینی شده‌اند. "غلط مثبت" داده‌های منفی هستند که به اشتباه پیش‌بینی شده‌اند. "غلط‌های منفی" داده‌های مثبتی هستند که به اشتباه پیش‌بینی شده‌اند. به این ترتیب میزان صحت یک الگوریتم دسته‌بندی به صورت زیر محاسبه می‌شود: (Han and Kamber,2006)

$$\text{صحت} = \frac{\text{صحیح مثبت + صحیح منفی}}{\text{داده منفی + داده مثبت}} \quad (۳)$$

با تقسیم مجموعه داده به دو بخش داده آموزشی (۷۰ درصد) و داده آزمون (۳۰ درصد)، صحت مدل‌های درخت تصمیم، مطابق جدول ۹ به دست می‌آید.

جهت اعتبارسنجی مدل بجز شاخص ریاضی"صحت"، نتایج حاصله به خبرگان کسب و کار حاضر نیز عرضه شده است. با ارایه نتایج و قواعد حاصل به مدیر پروژه بلیت الکترونیک شهر تهران و کارشناسان امر به عنوان افراد خبره‌ی حوزه حمل و نقل درون شهری اعتبار نتایج، قواعد و ملاحظات ارایه شده در خصوص قواعد از نظر کاربردی بودن و تطابق با واقعیت ازجانب ایشان مورد تأیید قرار گرفت. البته مجموعه داده‌های مورد استفاده در این پروژه پیش از کاربرد به تأیید مدیر پروژه رسیده است. از طرفی استاد راهنمای این پروژه نیزصحت نتایج را معتبر دانسته است. مشخصات افراد خبره مورد مشورت مطابق جدول ۷ می‌باشد.

٦- مقایسه نتایج پیش‌بینی روش‌های پیشین با روش پیشنهادی

نتایج برای مدل پیشنهادی و همچنین سایر روش‌های پیش‌بینی با استفاده از یک دستگاه رایانه با مشخصات CPU 2GHz 32Bits با٤ گیگا بایت حافظه‌رم و همچنین به کمک نرم‌افزار SPSS Clementine12.0 به دست آمده است.

بالاترین صحت در داده‌های آزمون مربوط به مدل گروهی شامل ترکیب درخت تصمیم کارت و$C_{5.0}$ می باشد. این مدل با ترکیب دو الگوریتم کارت و $C_{5.0}$ صحت بالاتری جهت پیش‌بینی کارایی یا ناکارایی خطوط اتوبوسرانی نسبت به هر یک از درخت‌های تصمیم نشان می‌دهد. پایین‌ترین صحت داده‌های آزمون مربوط به درخت $C_{5.0}$ می باشد.

۷- بحث

در حوزه‌ی حمل و نقل درون شهری، علاوه بر کارا بودن خطوط، لزوم خدمت‌رسانی به همه شهروندان در تمامی نقاط شهر اهمیت پیدا می‌کند. ممکن است برخی خطوط ناکارا بوده و صرفه اقتصادی نداشته باشند. اما مدیریت شهری خدمت‌رسانی به کلیه شهروندان اعم از ساکنین مرکز و حومه را جزء اهداف خود می‌داند. قاعده ٦ درخت $C_{5.0}$ و قواعد ٤ و ٥ کارت مصادیق چنین مناطقی می‌باشند.

از قواعد تولید شده توسط درخت‌های تصمیم کارت و $C_{5.0}$ نکات مهمی می‌توان استخراج نمود. در قواعد حاصل، کرایه هر

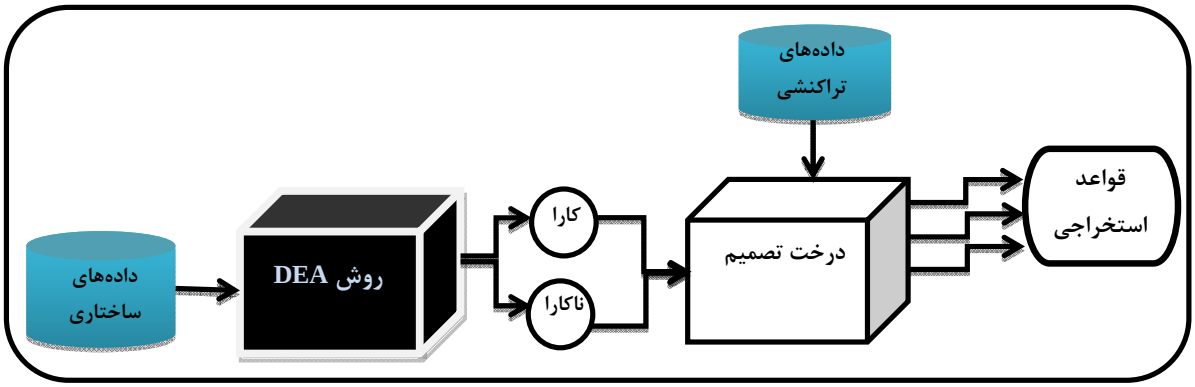
خط اتوبوس، به نوعی متناسب با طول مسیر است و تعداد باجه بلیت فروشی به ایستگاه، نشان‌دهنده میزان دسترسی شهروندان به بلیت الکترونیک است. قواعد ٤ و ٥ درخت $C_{5.0}$ نشانگر نقش استفاده از بلیت الکترونیک در ناوگان اتوبوسرانی می‌باشد. چرا که معمولاً تعداد پرداخت‌های نقدی مطابق با واقعیت نبوده و همیشه بخشی از آن ثبت نمی‌شود. این شرایط یعنی عدم وجود شفافیت در ثبت تعداد مسافران که در حالت پرداخت نقدی رخ می‌دهد، منجر به ثبت کمتر از واقعیت تعداد مسافران شده و امکان برآورد نیازهای یک خط از نظر تعداد اتوبوس، باجه بلیت فروشی و سایر امکانات را مقدور نمی‌سازد.

در طرف مقابل پرداخت الکترونیک تعداد مسافران را مشخص نموده و امکان برنامه‌ریزی بر اساس نیاز واقعی را به واحد سیاست‌گذار ارایه می‌دهد. از سویی دیگر استفاده از بلیت الکترونیک زمان توقف اتوبوس در ایستگاه‌ها را به جهت عدم نیاز به پرداخت مابقی کرایه مسافران کاهش می‌دهد که این نکات در افزایش کارایی خطوط به جهت کاهش زمان خدمت‌رسانی و افزایش تعداد مسافر ثبت شده، موثر می‌باشد.

در برخی نواحی و شهرک‌ها که در حومه شهر تهران می‌باشند، اغلب بافت جمعیتی دانشجویی و کارمندی دارند، بنابراین, در اوایل صبح و پایان وقت اداری تراکم مسافر بالاتر از سایر زمان‌ها می‌باشد. اما بعلت استقرار خط در حومه‌ی شهر تعداد باجه بلیت فروشی نسبت به ایستگاه‌ها بسیار کم می‌باشد. این خطوط می‌توانند مصادیق قاعده ٤ درخت $C_{5.0}$ باشد. برای یک چنین خطوطی استقرار باجه‌هایی که تنها در ساعات ابتدایی و انتهایی وقت اداری فعال باشند، می‌تواند موجب ارتقا عملکرد خطوط شود.

۸- نتیجه‌گیری

در این پژوهش به پیش بینی عملکرد خطوط با استفاده از الگوریتم‌های درخت تصمیم و مدل گروهی آن‌ها پرداخته شد. روال کار به این شکل است که روش‌های تحلیل پوششی داده‌ها و درخت تصمیم جهت سنجش کارایی خطوط اتوبوسرانی و استخراج قواعد از داده‌های موجود، با یکدیگر ترکیب شدند و کارایی هر یک از خطوط با استفاده از مدل (۲) محاسبه شد. نتایج مرحله فوق به عنوان یک ویژگی، به مجموعه داده تحت بررسی اضافه می‌شوند. در مرحله بعد، با استفاده از مدل‌های درخت تصمیم کارت و $C_{5.0}$، قواعد مستتر در مجموعه داده‌ها استخراج می‌شوند. جهت تشخیص مدل مناسب‌تر، شاخص صحت برای مقایسه سه مدل درخت تصمیم استفاده می‌شود. بر این اساس مدل گروهی عملکرد بهتری نسبت به دو مدل دیگر نشان می‌دهد. در واقع با استفاده از مدل گروهی، صحت پیش بینی کارایی یا ناکارایی یک خط، بالاتر از دو مدل کارت و $C_{5.0}$ خواهد بود. کاربرد قواعد استخراج شده می‌تواند در تغییر وضعیت فعلی خطوط اتوبوسرانی جهت ارتقا عملکرد آن‌ها به مسئولین امر یاری رساند. شاید بدیهی‌ترین راه حل مورد استفاده مدیران جهت بهبود عملکرد خطوط، افزایش تعداد اتوبوس‌های خط باشد اما قواعد استخراجی و تأیید خبرگان این حوزه بر قواعد مذکور، نشان می‌دهد که چه بسا الزاماً در همه مواقع نیاز به افزایش ناوگان حمل و نقل جهت افزایش کارایی نباشد. بلکه استفاده از فن‌آوری‌های نوین نظیر سامانه بلیت الکترونیک جهت جمع‌آوری خودکار کرایه و یا رعایت نسبت تعداد باجه‌های بلیت فروشی به ایستگاه‌ها با توجه به نوع مسیر، می‌توانند مؤثر واقع شوند. البته لزوم خدمت‌رسانی به دورترین مناطق شهر علی رغم نبود صرفه اقتصادی نیز باید مد نظر قرار گیرد.



شکل ۱. طرح چارچوب پیشنهادی

جدول ۳. نمونه داده‌های تراکنشی

کد راننده	شماره اتوبوس	زمان خدمت رسانی	تعداد مسافر	نام خط
۵۰۵۱۳	۳٤۲۱۸۵۲	۳۰۲٤۰	۷۰	عباس آباد- مترو شهرری
۵۵۰۱٦	۱٦۲۱۵۵۲	۲۲٦٤٦	٦۷	میدان رسالت- میدان بهارستان
۵۰۷۳۸	۳۳۲۱۸۵۸	۱٤۲۲٤	۷٦	م. راه آهن- م. توحید
۱۲۱۹۸۰	۳۹۲۱٦٤۸	۵۹۲۲٦	۲۳۵	م. راه آهن-تجریش

جدول ٤. توصیف آماری ورودی‌ها و خروجی‌های DEA برای هر خط

متغیرها	میانگین	انحراف معیار	میانه
میانگین طول مسیر	۱۰۲۱۹/۰۷	٤٤۲۹/٦	۹۱۰۰
تعداد اتوبوس	۳٦/۰٤	٤۲/۸۲	۲۷
تعداد مسافر	۹۳/۱۵	۹۲/۸	۷۷
معکوس زمان خدمت رسانی	۰/۰۰۰۰٤۰	۰/۰۰۰۰٤۳	۰/۰۰۰۰۳۱

جدول ۵. قواعد استخراجی از الگوریتم C5.0

ردیف	مقدم	تالی
۱	اگر کرایه کمتر از ۱۲٦۰ریال باشد و به ازای یک باجه حداقل ۱٦ایستگاه، وجود داشته باشد.	خط ناکاراست.
۲	اگر کرایه کمتر از ۸٤۰ ریال باشد و به ازای هر باجه حداکثر ۳ ایستگاه وجود داشته باشد.	خط ناکاراست.
۳	اگر کرایه بین ۸٤۰ و۱۲٦۰ریال باشد و به ازای هر باجه حداکثر ۳ ایستگاه وجود داشته باشد.	خط کاراست.
٤	اگر کرایه بیشتر از ۱۲٦۰ریال باشد، مسافران بیشتر از پرداخت نقدی به جای کارت استفاده کنند و حداکثر یک باجه در طول مسیر وجود داشته باشد.	خط ناکاراست.
۵	اگر کرایه بیشتر از ۱۲٦۰ریال باشد، مسافران بیشتر از کارت بلیت به جای پرداخت نقدی استفاده کنند و حداکثر یک باجه در طول مسیر وجود داشته باشد.	خط ناکاراست.
٦	اگر کرایه بیشتر از ۱٦۸۰ریال باشد و حداکثر یک باجه در طول مسیر وجود داشته باشد.	خط ناکاراست.

جدول شماره ٦. قواعد استخراجی از کارت

ردیف	مقدم	تالی
١	اگر کرایه کمتر از ١٤٧٠ریال باشد و به ازای هر باجه حداقل ٢ ایستگاه وجود داشته باشد.	خط ناکاراست.
٢	اگر کرایه کمتر از ١٤٧٠ریال باشد و به ازای هر باجه حداکثر ٢ ایستگاه وجود داشته باشد.	خط کاراست.
٣	اگر کرایه بین١٤٧٠و ١٨٤٠ریال باشد و در طول مسیر حداکثر یک باجه باشد.	خط کاراست.
٤	اگر کرایه بیشتر از ١٨٤٠ریال باشد و در طول مسیر حداکثر یک باجه باشد.	خط ناکاراست.
٥	اگر کرایه بیشتر از ١٤٧٠ریال باشد و در طول مسیر حداقل یک باجه باشد.	خط ناکاراست

جدول ٧. مشخصات افراد خبره

ویژگی	کارشناسی	کارشناسی ارشد	دکتری
تحصیلات	٪٣٣	٪٣٣	٪٣٣
سن	کمتر از ٤٠	٤٠ و بیشتر	
	٪ ٦٦	٪ ٣٣	
سابقه حضور در حوزه حمل و نقل	١٠ و کمتر	بین ١٠ تا ٢٠	٢٠ و بالاتر
	٪٦٦	٪٣٣	٠

جدول شماره ٨. مقایسه مدل‌های اجرا شده

مدل	زمان اجرا (ثانیه)	زمان CPU (ثانیه)
کارت	٩٢٧.٦	٩٢٢.٧٩
$C_{5.0}$	١٨٩٩.٣٨	١٨٨١.٢٥
مدل گروهی	٢١٠٨.٣٧	١٠٣١.٢١

جدول شماره ٩. صحت مدل‌های درخت تصمیم

مدل	آموزش	آزمون
کارت	٪٩٧.٧٦	٪٩٧.٧٣
$C_{5.0}$	٪٨٨.٤١	٪٨٨.٣٤
مدل گروهی کارت و $C_{5.0}$	٪٩٨	٪٩٧.٩٦

٩- سپاسگزاری

لازم است در این بخش از کمیته هماهنگی نظارت بلیت الکترونیکی شهر تهران به جهت در اختیار نهادن داده‌ها، در راستای انجام تحقیقات قدردانی به عمل آید. در ضمن شرکت فن‌آسا و مدیریت آن نیز به عنوان مجری پروژه بلیت الکترونیک تهران، کمال همکاری را به جهت ارایه راهنمایی‌های فنی در این خصوص ارایه نمودند.

١٠- پی‌نوشت‌ها

1-Data Envelopment Analysis (DEA)
2-Classification and Regression Tree (CART)
3- Ensemble
4- Accuracy
5-Efficiency
6-Data mining
7- Decision Tree
8- Decision Making Unit (DMU)
9-Charnes
10-Cooper
11-Rohdes
12-Geographical Information Systems (GIS)
13-Benchmarking
14-Bootstrap
15-Variable Return to Scale
16- Constant Return to Scale
17- Target Variable
18-Boosting
19-Classifier

۱۱- مراجع

- مهرگـــان، ن. (۱۳۸۳) "مدل‌های کمی برای ارزیابی عملکرد سازمان‌هاDEA"، تهران، دانشگاه تهران دانشکده **مدیریت**.

- یقینـــی،م.، خوشـــرفتار، م.م. و ســـیدآبـــادی ، س.م. (۱۳۸۹) "پیش‌بینی تأخیر قطارهای مسافری با استفاده از شبکه‌های عصبی" پژوهشنامه حمل و نقل، سال هفتم، شماره سوم.

- Boilé, M. P. (2001) “Estimating technical and scale inefficiencies of public transit systems”, Journal of Transportation Engineering, 127(3), 187–194.

- Charnes, A., Cooper, W.W., Rhodes, E., 1978. “Measuring the efficiency of decision making units”, European Journal of Operational Research 2, 429–444.

-Chou, J. (2012) “Comparison of multi label classification models to forecast project dispute resolutions”, Expert Systems with Applications 39, 10202–10211.

-Chu, X., Fielding, G., & Lamar, B. W. (1992) “Measuring transit performances using data envelopment analysis”, Transportation Research A, 3, 223–230.

-Han, Jiawei and Kamber, Micheline (2006)”Data Mining: Concepts and Techniques”, 2nd Edition, USA: Elsevier Inc.

-Karlaftis, M. G (2004) “A DEA approach for evaluating the efficiency and effectiveness of urban transit systems”, European Journal of Operations Research, 152, 354–364.

-Lao.Y , Liu. L (2009) “Performance evaluation of bus lines with data envelopment analysis and geographic information systems”, Computers, Environment and Urban Systems 33, 24.

-Lee.S (2010) “Using data envelopment analysis and decision trees for efficiency analysis and recommendation of B2C controls”, Decision Support Systems 49, 486–497.

-Liu J, Lu L, Lu W, Lin B (2013) “A survey of DEA applications”, Omega, 41: 893–902.

-Nicola. A, Gitto.S, Mancuso.P (2012)” Uncover the predictive structure of healthcare efficiency applying a bootstrapped data envelopment analysis”, Expert Systems with Applications 39, 10495–10499.

-Nie.G, Rowe.W, Zhang.L,Tian.Y, Shi.Y (2011) “Credit card churn forecasting by logistic regression and decision tree”, Expert Systems with Applications, Volume 38, Issue 12, Pages 15273-15285.

-Nolan, JF. (1996) “Determinants of Productive Efficiency in Urban Transit”, Logistics and Transportation Review 32(3): 319-342.

-Pujari .P, Gupta. J (2012) “Improving classification accuracy by using feature selection and ensemble model”, International Journal of Soft Computing and Engineering (IJSCE), ISSN: 2231-2307, Volume-2, Issue-2.

-Seol.H, Choi. J, Park. G, Park. Y (2007) “A framework for benchmarking service process using data envelopment analysis and decision tree”, Expert Systems with Applications 32: 432–440.

-Sohn S.Y, Moon.T (2004) “Decision Tree based on data envelopment analysis for effective technology commercialization”, Expert Systems with Applications 26, 279–284.

-Quinlan, J. R. (2007). Data Mining Tools See 5 and C5.0.