محیط‌شناسی، دورهٔ ۴۲، شمارهٔ ۳، پاییز ۱۳۹۵، صفحهٔ ۴۷۳-۴۵۵

مقایسهٔ کاربرد شبکهٔ عصبی مصنوعی، درخت تصمیم، رگرسیون مؤلفه‌های اصلی و رگرسیون خطی چندگانه در مدل‌سازی شاخص کیفیت هوای شهری

علیرضا احسان زاده[1]*، فرهاد نژادکورکی[2]، علی طالبی[3]

۱. کارشناس ارشد مهندسی محیط زیست، گروه محیط زیست، دانشگاه یزد
۲. دانشیار گروه مهندسی محیط زیست، گروه محیط زیست، دانشگاه یزد
۳. . دانشیار گروه مهندسی آبخیزداری، گروه آبخیزداری، دانشگاه یزد

تاریخ وصول مقاله: ۱۳۹٤/۰۹/۱۷ تاریخ پذیرش مقاله: ۱۳۹۵/۰۸/۸

چکیده

شاخص کیفیت هوا ابزاری کلیدی برای آگاهی از کیفیت هوا، نحوهٔ اثر آلودگی هوا بر سلامت و روش‌های محافظتی در برابر آلودگی هواست. هدف اصلی این تحقیق مدل‌سازی و برآورد شاخص کیفیت هوا از طریق شبکهٔ عصبی مصنوعی، درخت تصمیم، رگرسیون خطی چندگانه و رگرسیون مؤلفه‌های اصلی است. از دو روش سازمان حفاظت محیط زیست آمریکا و مرکز سلامت و محیط کار ایران برای محاسبهٔ شاخص کیفیت هوا و از داده‌های هواشناسی و آلودگی هوای ثبت‌شده در ایستگاه تجریش و قلهک شهر تهران در دورهٔ زمانی ۱۳۸۵ تا ۱۳۹۰ به منظور توسعهٔ مدل‌ها و به‌منظور ارزیابی عملکرد مدل‌های برآوردگر از شاخص‌های آماری خطا، همبستگی و صحت استفاده شد. نتایج تحقیق نشان داد که مدل شبکهٔ عصبی در هر دو ایستگاه نسبت به سایر مدل‌ها از عملکرد بهتری برخوردار است، به نحوی که در ایستگاه قلهک RMSE=۰/۰۰۶، MAE=۰/۰۰۴، IA= ۰/۹۹و در ایستگاه تجریش RMSE=۰/۰۰۴، MAE=۰/۰۰۲، IA= ۱ بود. مدل درخت تصمیم بعد از مدل شبکهٔ عصبی عملکرد مطلوبی از خود نشان داد و مدل رگرسیون خطی چندگانه بعد از مدل شبکهٔ عصبی و درخت تصمیم عملکرد بهتری نسبت به مدل رگرسیون مبتنی بر تحلیل مؤلفه‌های اصلی ارائه کرد. روش تحلیل مؤلفه‌های اصلی علی‌رغم آنکه توانست همبستگی بین داده‌های ورودی و تعداد پارامترهای ورودی به مدل را کاهش دهد، باعث بهبود عملکرد مدل رگرسیون نشد.

کلیدواژه

شاخص کیفیت هوا، شبکهٔ عصبی مصنوعی، درخت تصمیم، رگرسیون مؤلفه‌های اصلی، مدل‌سازی.

۱. سرآغاز

آلودگی هوای شهری یکی از نگرانی‌های روزافزون جهانی است، زیرا تأثیرهای بسزایی بر محیط‌زیست، آب‌وهوا و سلامت عمومی دارد (Li et al., 2014). افزایش نرخ شهرنشینی و صنعتی‌شدن در شهرهای کشورهای توسعه‌یافته و درحال‌توسعه مثل تهران، منجر به افزایش سطح آلودگی هوا شده است. همچنین، نگرانی جهانی در مورد آثار آلودگی هوا بر سلامتی انسان افزایش یافته است. راهبردهای مختلفی برای کنترل و مدیریت آلودگی هوا وجود دارد. شاخص‌های سنجش کیفیت هوا به‌طور گسترده در طرح‌های کنترل کیفیت هوا استفاده می‌شود. این شاخص‌ها کیفیت هوا را برحسب میزان آلودگی و آلاینده‌های مختلف طبقه‌بندی می‌کند.

اولین شاخص سنجش کیفیت هوا، شاخص آلودگی هوا (PSI) نام دارد، که سازمان حفاظت محیط زیست آمریکا[1] (US-EPA) توسعه داد. این شاخص غلظت آلاینده‌های اصلی هوا نظیر کربن مونوکسید (CO)، دی‌اکسید گوگرد (SO_2)، ذرات معلق کوچک‌تر از ده

* نویسنده مسئول: Email: ehsanzadehalireza@gmail.com

میکرون (PM_{10})، ازن (O_3) و دی‌اکسید نیتروژن (NO_2) را به شاخص استاندارد آلودگی هوا تبدیل می‌کند. در سال ۱۹۹۹ شاخص PSI را US-EPA کامل‌تر کرد و شاخصی به‌نام شاخص کیفیت هوا[۲] (AQI) جایگزین آن شد (Sowlat et al., 2011).

امروزه، پیش‌بینی و برآورد مشخصه‌های کیفیت هوا در نواحی شهری به‌دلیل تأثیر آن بر سلامتی انسان، یکی از موضوع‌های مهم در تحقیقات محیط‌زیستی است. غلظت‌های زیاد آلاینده‌ها، تأثیرهای سوء و مرگ زودرس گروه‌های حساس و آسیب‌پذیر جامعه را به‌دنبال دارد، از جمله افراد مسن و کسانی که به تنگی نفس دچارند (صدرموسوی و همکاران، ۱۳۸۹). برای مدل‌سازی و برآورد مشخصه‌های کیفیت هوا، به‌طور کلی، سه دسته مدل وجود دارد. دستهٔ اول مدل‌های پیش‌ساخته، دستهٔ دوم مدل‌های آماری و دستهٔ سوم مدل‌های هوش محاسباتی است (Zhang et al., 2012). مدل‌های هوش محاسباتی شامل شبکه‌های عصبی مصنوعی[۳] (ANN) و درختان تصمیم[۴] (DT) است. مدل‌های این دسته در مدل‌سازی سیستم‌های غیرخطی از قابلیت بالایی دارد (Zhang et al., 2012).

از نمونه مطالعاتی که از روش‌های هوش مصنوعی و آماری در پیش‌بینی و مدل‌سازی پارامترهای کیفیت هوا به‌کار برده‌شده می‌توان به موارد زیر اشاره کرد: استفاده از روش تحلیل مؤلفه‌های اصلی[۵] و استفاده از شبکهٔ عصبی مصنوعی در پیش‌بینی شاخص روزانهٔ کیفیت هوا (Kumar and Goyal, 2013)، استفاده از سیستم استنتاج فازی[۶] و مدل اتورگرسیو[۷] به‌منظور ارزیابی و پیش‌بینی کیفیت هوا (Carbajal-Hernández et al., 2012)، کاربرد رگرسیون خطی چندگانه[۸] مبتنی بر تحلیل مؤلفه‌های اصلی برای پیش‌بینی کوتاه‌مدت شاخص کیفیت هوا (Kumar and Goyal 2011)، کاربرد الگوریتم یادگیری جمعی[۹] برای شناسایی منابع آلایندهٔ هوا و مقدار شاخص کیفیت هوا (Singh et al., 2013)، کاربرد شبکهٔ عصبی مصنوعی و روش انتخاب متغیرهای تصادفی به‌منظور پیش‌بینی کیفیت هوای شهری (Russo et al., 2013)، پیش‌بینی غلظت روزانهٔ مونوکسید کربن با استفاده از روش رگرسیون خطی چندگانه براساس تحلیل مؤلفه‌های اصلی، همچنین استفاده از شبکهٔ عصبی (نوری و همکاران، ۱۳۸۷) و پیش‌بینی و مدل‌سازی غلظت آلایندهٔ مونوکسید کربن با تلفیق شبکهٔ عصبی- فازی تطبیقی و سیستم اطلاعات جغرافیایی (خزاعی و همکاران، ۱۳۹۱). نتایج اکثر این تحقیقات حاکی از برتری مدل‌های هوش محاسباتی نسبت به مدل‌های آماری است.

یکی از بهترین روش‌ها که در زمینهٔ تعیین کیفیت هوا امروزه در سراسر دنیا استفاده می‌شود، تبدیل غلظت آلاینده‌ها به شاخص کیفیت هواست. در واقع، AQI شاخص مفیدی برای آگاهی از کیفیت هوا، میزان اثر آلاینده‌ها بر سلامت و روش‌های مختلف کنترلی در برابر آلودگی هوا را مشخص می‌کند (Lee et al., 2012). یکی از اقدام‌های مؤثر در کنترل، پایش و وضع اقدام‌های احتیاطی و پیشگیرانه در مواردی که کیفیت هوا نامطلوب و آلودگی از حد استاندارد فراتر می‌رود، تعیین میزان واقعی غلظت آلاینده‌ها و توصیف وضعیت کیفی هوا در مقایسه با شرایط استاندارد و اطلاع‌رسانی به‌موقع به مردم است (Zhang et al., 2012). یکی از ابزارهای مناسب در انجام این راهبرد استفاده از AQI است که تلاش می‌کند تا اطلاعات مربوط به کیفیت هوا را در سطوح بهداشتی خوب، متوسط، غیربهداشتی برای افراد حساس، غیربهداشتی، ناسالم و خطرناک به مردم آگاهی دهد (Cheng et al., 2007). کیفیت نامطلوب هوا ناشی از وجود غلظت‌های زیاد آلاینده‌ها در کلان‌شهر تهران، موجب ایجاد بیماری‌های مختلف و مشکلات فراوان برای سلامتی و رفاه عمومی این کلان‌شهر شده است. از این‌رو، برآورد و مدل‌سازی کیفیت هوای شهری و دارای ماهیت غیرخطی، همچنین تعیین عناصر مؤثر بر آن از ضروری‌ترین برنامه‌های محیط‌زیستی در کلان‌شهرهاست (Kumar and Goyal,

2013). لذا، هدف اصلی این تحقیق مقایسهٔ کارایی استفاده از روش شبکهٔ عصبی مصنوعی، درخت تصمیم، رگرسیون خطی چندگانه و رگرسیون مؤلفه‌های اصلی در مدل‌سازی و برآورد شاخص کیفیت هوای شهری است.

۲. مواد و روش‌ها

در این تحقیق، از داده‌های ساعتی غلظت آلاینده‌های هوا و پارامترهای هواشناسی مربوط به ایستگاه‌های تجریش و قلهک شهر تهران در برآورد و مدل‌سازی شاخص AQI استفاده شد. هدف نخست، استفاده از دستورالعمل US-EPA و مرکز سلامت و محیط کار ایران در محاسبهٔ شاخص کیفیت هوا براساس غلظت‌های ساعتی مربوط به تک‌تک آلاینده‌هاست. در مرحلهٔ بعد، با استفاده از سری‌های زمانی مربوط به داده‌های هواشناسی، آلودگی هوا و میزان AQI محاسبه شد. به‌منظور ایجاد و توسعهٔ مدل‌های برآوردگر و شبیه‌ساز کیفیت هوا با استفاده از روش‌های درخت تصمیم، رگرسیون مؤلفه‌های اصلی، رگرسیون خطی چندگانه و شبکهٔ عصبی مصنوعی از نرم‌افزار MATLAB استفاده شده است. در اولین مرحله، غلظت تک‌تک آلاینده‌ها ورودی الگوریتم محاسبهٔ AQI بود. خروجی این الگوریتم که شاخص کیفیت هوای مربوط به هر آلاینده و شاخص کلی کیفیت هواست، همراه با داده‌های هواشناسی برای توسعهٔ مدل‌ها استفاده شد. هدف نهایی، شبیه‌سازی و برآورد شاخص کیفیت هوا در ایستگاه‌های مورد مطالعه در شهر تهران است. در آخر مقایسه‌ای بین مدل‌های به‌کار رفته در تحقیق صورت گرفته است و مدل دارای نتایج بهتر به‌منظور برآورد و شبیه‌سازی معرفی می‌شود.

۱.۲. داده‌های تحقیق

داده‌های مورد استفاده در تحقیق شامل غلظت‌های ساعتی مربوط به آلاینده‌های هواست، شامل مونوکسید کربن، ذرات معلق کوچک‌تر از ۱۰ میکرون، اکسیدهای نیتروژن، ازن، دی‌اکسید گوگرد، مونوکسید نیتروژن، هیدروکربن‌های بدون متان (NMHC) و متان (CH_4). همچنین، شامل برخی پارامترهای هواشناسی است، نظیر سرعت باد (WS)، جهت باد (WD)، دمای هوا (T)، فشار (P) و رطوبت هوا (H). داده‌های موجود مربوط به دورهٔ زمانی ۱۳۸۵ تا ۱۳۹۰ است.

۲.۲. شاخص کیفیت هوا

مدیریت پایش و نظارت بر کیفیت هوا در شهرهای بزرگ داده‌های خام اندازه‌گیری‌شده با دستگاه‌های سنجش غلظت آلاینده‌های هوا را به AQI تبدیل می‌کند. اطلاعات مورد نیاز مربوط به سطوح سلامتی انسان و آثار بهداشتی آلاینده‌های هوا را شاخص AQI در اختیار مردم قرارمی‌دهد. AQI برای پنج آلایندهٔ اصلی هوا- یعنی ذرات معلق، دی‌اکسید نیتروژن، ازن سطح زمین، منوکسید کربن و دی‌اکسید گوگرد- محاسبه می‌شود (Zhang et al., 2012). به‌منظور محاسبهٔ این شاخص از معادلهٔ (۱) استفاده می‌شود که غلظت آلاینده‌ها و نقاط شکست در محاسبهٔ طبقهٔ کیفیت هوا مطابق با دستورالعمل مرکز سلامت و محیط کار ایران است. US-EPA به‌منظور درک راحت‌تر مقدار شاخص کیفیت هوا و سطوح بهداشتی مختلف با آن، همچنین دستورالعمل‌های کنترلی مربوط با مقادیر مختلف AQI را به شش دسته طبقه‌بندی می‌کند و هر دسته را به سطوح مختلف سلامت انسان مربوط می‌سازد (جدول ۱).

$$Ip = \frac{I_{Hi} - I_{Lo}}{BP_{Hi} - BP_{Lo}}(C_p - BP_{Lo}) + I_{Lo} \qquad (۱)$$

Ip= شاخص کیفیت هوا برای آلایندهٔ P، Cp= غلظت اندازه‌گیری‌شدهٔ آلایندهٔ P، BP_{Hi}= نقطهٔ شکستی و بزرگ‌تر یا مساوی Cp، BP_{Lo}= نقطهٔ شکست کوچک‌تر یا مساوی Cp، I_{Hi}= مقدار AQI منطبق با BP_{Hi}، I_{Lo}= مقدار AQI منطبق با BP_{Lo}.

جدول ۱. نقاط شکست در محاسبهٔ AQI

AQI I_{Lo}–I_{Hi}	نقاط شکست BP_{Lo}–BP_{Hi}						
	NO_2 (ppm) ۱ ساعته	SO_2 (ppm) ۲۴ ساعته	CO (ppm) ۸ ساعته	PM_{10} (µg/m^3) ۲۴ ساعته	$PM_{2.5}$ (µg/m^3) ۲۴ ساعته	O_3 (ppm) ۱ ساعته	O_3[1] (ppm) ۸ ساعته
۰–۵۰	۰–۰/۰۵۳	۰/۰–۰/۰۳۴	۰/۰–۴/۴	۰–۵۴	۰/۰–۱۵/۴	–	۰–۰/۰۵۹
۵۱–۱۰۰	۰/۰۵۴–۰/۱	۰/۰۳۵–۰/۱۴۴	۴/۵–۹/۴	۵۵–۱۵۴	۱۵/۵–۳۵	–	۰/۰۶۰–۰/۰۷۵
۱۰۱–۱۵۰	۰/۱۰۱–۰/۲۶۰	۰/۱۴۵–۰/۲۲۴	۹/۵–۱۲/۴	۱۵۵–۲۵۴	۳۵/۱–۶۵/۴	۰/۱۲۵–۰/۱۶۴	۰/۰۷۶–۰/۰۹۵
۱۵۱–۲۰۰	۰/۳۶۱–۰/۶۴۰	۰/۲۲۵–۰/۳۰۴	۱۲/۵–۱۵/۴	۲۵۵–۳۵۴	۶۵/۵–۱۵۰/۴	۰/۱۶۵–۰/۲۰۴	۰/۰۹۶–۰/۱۱۵
۲۰۱–۳۰۰	۰/۶۵–۱/۲۴	۰/۳۰۵–۰/۶۰۴	۱۵/۵–۳۰/۴	۳۵۵–۴۲۴	۱۵۰/۵–۲۵۰/۴	۰/۲۰۵–۰/۴۰۴	۰/۱۱۶–۰/۳۷۴
۳۰۱–۴۰۰	۱/۲۵–۱/۶۴	۰/۶۰۵–۰/۸۰۴	۳۰/۵–۴۰/۴	۴۲۵–۵۰۴	۲۵۰/۵–۳۵۰/۴	۰/۴۰۵–۰/۵۰۴	(۲)
۴۰۱–۵۰۰	۱/۲۶–۲/۰۴	۰/۸۰۵–۱/۰۰۴	۴۰/۵–۵۰/۴	۵۰۵–۶۰۴	۳۵۰/۵–۵۰۰/۴	۰/۵۰۵–۰/۶۰۴	

۱. در بیشتر مناطق AQI بر اساس مقادیر ازن ۸ ساعته گزارش می‌شود اما در برخی مناطق AQI بر اساس مقادیر ازن ۱ ساعته به احتیاط نزدیک‌تر است. در این شرایط AOI هم برای مقادیر ازن ۸ ساعته و هم برای ازن ۱ ساعته محاسبه و هر کدام بیشتر بود گزارش می‌شود.

۲. وقتی غلظت ازن ۸ ساعته از ۰/۳۷۴ ppm فراتر رود، مقادیر AQI بیش از ۳۰۰ باید با استفاده از غلظت ازن ۱ ساعته محاسبه شود.

۳.۲. شبکهٔ عصبی مصنوعی

شبکه‌های عصبی مصنوعی از چندین لایه تشکیل می‌شود. لایه‌های ابتدایی و انتهایی به ترتیب لایهٔ ورودی و لایهٔ خروجی نام دارد. همچنین، بین این دو لایه ممکن است یک یا چند لایهٔ مخفی وجود داشته باشد. خروجی هر نورون به‌صورت معادلهٔ (۲) به‌دست می‌آید.

$$y_i = \sum_{j=1}^{n} w_{i,j} x_{i,j} + \beta_i \qquad (۲)$$

که در آن، $x_{i,j}$ سیگنال ورودی از j امین نورون (در لایهٔ ورودی)، $w_{i,j}$ وزن اتصال نورون j به نورون i (در لایهٔ مخفی)، β_i اریبی نورون i، و y_i خروجی نورون است. طی فرایند آموزش، این وزن‌ها و مقادیر ثابتی که با آن‌ها جمع می‌شود و در اصطلاح اریبی نامیده می‌شود، به‌طور پی‌درپی تغیر می‌کند تا خطا به کمترین مقدار خود برسد. پس از محاسبهٔ y_i، شبکه تحت یک تابع محرک شروع به اصلاح آن می‌کند. تابع محرک نوعاً تابعی خطی یا غیرخطی است (Fausett, 1994). طی سال‌های اخیر، شبکهٔ عصبی پرسپترون چندلایه به‌خوبی قابلیتش را در مدل‌سازی و پیش‌بینی پارامترهای متفاوت اتمسفر از جمله آلاینده‌ها نشان داده است (Kumar and Goyal, 2013; Singh et al., 2012; Zhang et al., 2012). در این نوع شبکه، داده‌ها به‌صورت پیوسته و بدون هر گونه بازخوردی به سمت خروجی انتقال می‌یابد (Zhang et al. 2012). فرایند آموزش در شبکه‌های عصبی در واقع به‌معنای روزآمدکردن اتصالات بین نورون‌هاست. تاکنون الگوریتم‌های گوناگونی برای آموزش شبکهٔ عصبی عرضه شده است که معروف‌ترین آن‌ها الگوریتم و قاعدهٔ انتشار به عقب[10] است (Moustris et al., 2012). برای اصلاح وزن‌ها در شبکه و برای به‌حداقل رساندن میانگین مربعات خطا که از معادلهٔ (۳) به‌دست می‌آید، از راه الگوریتم انتشار به‌عقب استفاده می‌شود (Zhang et al., 2012).

$$MSE = \frac{1}{N}\sum_{n=1}^{N}(t_i - td_i)^2 \qquad (۳)$$

N تعداد نورون‌های لایهٔ خروجی و متناسب با تعداد مشاهدات تابع هدف، t_i مقدار مشاهده‌شده برای iامین رکورد، و td_i مقدار خروجی شبکه برای iامین رکورد است (Yadav and Naresh, 2010). به‌منظور بررسی دقت شبکهٔ آموزش‌یافته لازم است آزمون شبکه صورت گیرد. این کار با دادن زوج داده‌های مجموعه آزمون به شبکه و محاسبهٔ میزان خطای شبکه صورت می‌گیرد. به‌طور کلی، ویژگی‌های شبکهٔ عصبی مصنوعی، ساختار شبکه و روش آموزش شبکه، با نوع تابع محرک نورون‌ها مشخص می‌شود (Zhang et al., 2012). در این تحقیق شبکهٔ عصبی پرسپترون سه‌لایه با الگوریتم‌های آموزش لونبرگ مارکوارت (Lm)، پس‌انتشار ارتجاعی (Rb) و شیب توأم مقیاس‌شده (Scg) با تابع محرک تانژانت هایپربولیک در لایهٔ پنهان و تابع محرک خطی در لایهٔ خروجی استفاده شده است.

۴.۲. تحلیل مؤلفه‌های اصلی

روش تحلیل مؤلفه‌های اصلی متغیرهای مستقل اولیه را به مؤلفه‌های جدید و مستقل (بدون همبستگی) تبدیل می‌کند. سپس، از این مؤلفه‌ها به‌جای متغیرهای اولیه استفاده می‌شود. مؤلفه‌های جدید ترکیبی خطی از متغیرهای اولیه است (Çamdevýren et al., 2005). با استفاده از این روش، ترکیباتی از n متغیر $X_1, X_2, \ldots, X_n$ برای ایجاد n مؤلفهٔ مستقل $PC_1, PC_2, \ldots, PC_n$ برقرار می‌شود. در این روش اطلاعات متغیرهای اصلی با کمترین تلفات در مؤلفه‌ها محفوظ می‌ماند. هر مؤلفهٔ اصلی با دنبالهٔ زیر مشخص می‌شود.

$$PCi = a_{i1}X_1 + a_{i2}X_2 + \ldots + a_{in}X_n \qquad (۴)$$

که در آن PCi معرف مؤلفهٔ مورد نظر، a_{ij} بردار ویژهٔ مربوط و X_i نیز متغیرهای مستقل اولیه است. روش کار برای ایجاد مؤلفه‌های اصلی به صورت زیر است.

الف) محاسبه فاکتور KMO[11]

از آنجا که روش PCA مستلزم وجود و قبول فرضیه‌هایی دربارهٔ جامعهٔ مورد مطالعه نیست، از روش‌های ناپارامتری است. لازم است امکان استفاده از این روش و نتایج به‌دست آمده از آن به‌وسیلهٔ عامل KMO مشخص شود. مقدار KMO بین صفر تا یک تغییر می‌کند. این عامل با استفاده از ضرایب همبستگی ساده و جزئی طبق معادلهٔ (۵) محاسبه می‌شود. در معادلهٔ (۵) r_{ij} و a_{ij} ضرایب همبستگی ساده و جزئی بین متغیرهای i و j است. با توجه به این معادله مقادیر بزرگ‌تر KMO مستلزم کوچک‌بودن ضرایب همبستگی جزئی و بیانگر دقت محاسبات مربوط، با استفاده از روش PCA است (Abdul-Wahab et al., 2005).

$$KMO = \frac{\sum_{i=1}^{p}\sum_{i=1}^{p} r_{ij}^2}{\sum_{i=1}^{p}\sum_{i=1}^{p} r_{ij}^2 + \sum_{i=1}^{p}\sum_{i=1}^{p} a_{ij}^2} \qquad i \neq j \qquad (۵)$$

در صورتی که فاکتور KMO بزرگ‌تر از ۰/۵ به‌دست آید، نشان‌دهندهٔ امکان اجرای روش PCA روی متغیرهای مستقل است.

ب) استانداردکردن متغیرهای ورودی

در این مرحله داده‌های ورودی براساس معادلهٔ (۶) به‌نحوی استاندارد می‌شود که دارای میانگین صفر و انحراف معیار یک باشد.

$$Z = \frac{X - \mu}{\sigma} \qquad (۶)$$

در این معادله، Z معادل مقادیر استانداردشدهٔ داده‌ها، X ماتریس داده‌های ورودی، μ میانگین هر متغیر و σ نیز مقادیر انحراف معیار برای هر متغیر است.

ج) محاسبهٔ ماتریس همبستگی (R) برای متغیرهای اولیه

این ماتریس که ماتریسی متقارن است، میزان تغییرات در نمونه و میزان همبستگی N متغیر را با هم نشان می‌دهد. عضوهای روی قطر اصلی این ماتریس، واریانس متغیرهای

ورودی و بقیهٔ درایه‌های این ماتریس، کوواریانس بین متغیرهای ورودی است. چون برای تشکیل این ماتریس از داده‌های استانداردشده استفاده شده است، به‌همین دلیل این ماتریس، معادل ماتریس همبستگی بین متغیرهای ورودی است.

د) محاسبهٔ مقادیر ویژهٔ (λ) و بردارهای ویژهٔ مربوط از ماتریس همبستگی

بدین منظور معادلهٔ (۷) حل می‌شود.

$$\left|R - \lambda I_p\right| = 0 \qquad (۷)$$

I_p ماتریس واحد با بعد p×p است. بنابراین، می‌توان p مقدار ویژهٔ مرتب‌شدهٔ $\lambda_p \geq \ldots \geq \lambda_2 \geq \lambda_1$ را به‌دست آورد، به‌طوری که مجموع مقادیر ویژه برابر p باشد. هر مقدار ویژه با اطلاعات مربوط به آن (بردارهای ویژه) ویژگی‌های یک مؤلفه را ارائه می‌دهد. هر مؤلفه نیز درصدی از اطلاعات بیان‌شدهٔ متغیرهای اولیه را دربرمی‌گیرد و معادل با بخشی از اطلاعات مسئله است (Jolliffe, 2002). اولین مؤلفه بیشترین واریانس و آخرین آن کمترین مقدار واریانس را نشان می‌دهد. انتخاب چند مؤلفهٔ اول بیشترین مقدار واریانس و مؤلفه‌های اصلی شناخته می‌شود (Çamdevýren et al., 2005).

ه) اجرای چرخش وریماکس[۱۲] روی ماتریس ضرایب مؤلفه‌ها

چون در تشکیل هر مؤلفه از تمام متغیرهای اولیه استفاده می‌شود، تفسیر مؤلفه‌ها مشکل خواهد بود. به این دلیل روش‌هایی برای تفسیر ساده‌تر مؤلفه‌ها به‌وجود آمده است (Jolliffe, 2002 ؛ نوری و همکاران ۱۳۸۷).

۵،۲. رگرسیون خطی چندگانه

رگرسیون چندگانه روشی است که برای ارتباط خطی بین یک متغیر وابسته و یک یا چند متغیر مستقل استفاده می‌شود (Rawlings et al., 1998). مدل ماتریسی رگرسیون چندگانه را می‌توان به صورت معادلهٔ (۸) نشان داد.

$$Y = X\beta + e \qquad (۸)$$

که در آن β ماتریس ضرایب رگرسیون، e ماتریس خطای برازش و Y نیز ماتریس پاسخ است. با حل معادلهٔ (۸) بر حسب β خواهیم داشت:

$$\beta = (X'X)^{-1}(X'Y) \qquad (۹)$$

که در آن 'X ترانهاده ماتریس X است. برای محاسبهٔ معکوس (X'X) لازم است متغیرهای مستقل همبستگی زیادی نداشته باشند، زیرا در این صورت ماتریس (X'X) را نمی‌توان معکوس کرد و باعث افزایش خطا در اثر گردکردن داده‌ها و محاسبات می‌شود. برای رفع این مشکل باید قبل از ساخت مدل رگرسیونی، همبستگی بین متغیرهای مستقل را از بین برد. در این خصوص، روش مناسب استفاده از تحلیل مؤلفه‌های اصلی روی متغیرهای مستقل ورودی به مدل است. معیار قضاوت برای رفع این مشکل با اجرای تحلیل مؤلفه‌های اصلی روی متغیرهای ورودی، فاکتور تورم واریانس است. عدد ایده‌آل برای فاکتور تورم واریانس ۱ و مقادیر بزرگ‌تر از ۱۰ برای تورم واریانس نشانهٔ ناپایداری مدل رگرسیونی است (Chatterjee and Hadi, 2015; نوری و همکاران، ۱۳۸۷). در این تحقیق پس از رفع مشکل هم‌خطی بین متغیرهای مستقل به روش تحلیل مؤلفه‌های اصلی، مدل رگرسیون چندگانه برای مدل‌سازی و برآورد شاخص کیفیت هوا توسعه یافت.

۶،۲. درخت تصمیم

درختان تصمیم قادر به تولید توصیفات قابل‌درک برای انسان، از روابط موجود در مجموعهٔ داده‌ای است و برای وظایف دسته‌بندی و پیش‌بینی به‌کار می‌رود. این ساختار تصمیم‌گیری به شکل تکنیک‌های ریاضی و محاسباتی نیز معرفی می‌شود که به توصیف، دسته‌بندی و عام‌سازی مجموعه‌ای از داده‌ها کمک می‌کند. داده‌ها در رکوردهایی به شکل $(x, y) = (x_1, x_2, x_3\ldots, x_k, y)$ داده می‌شود. با استفاده از متغیرهای $x_1,x_2,..,x_k$ سعی در درک، دسته‌بندی یا عام‌سازی متغیر وابستهٔ Y داریم. انواع صفات در درخت

تصمیم به دو نوع صفات دسته‌ای و صفات حقیقی است. صفات دسته‌ای، صفاتی است که دو یا چند مقدار گسسته می‌پذیرد (یا صفات سمبلیک)، درحالی که صفات حقیقی مقادیر خود را از مجموعهٔ اعداد حقیقی می‌گیرد (Breiman et al., 1993). درخت CART[13] نامی است که به هر دو روال بالا اطلاق می‌شود. نام CART سرنام کلمات درختان رگرسیون و دسته‌بندی است. در این تحقیق از الگوریتم درخت تصمیم CART برای مدل‌سازی و برآورد شاخص کیفیت هوا استفاده شده است. ساختار این درخت بر سه اصل استوار است:

۱. مجموعه‌ای از سؤال‌ها به شکل x≤d? که در آن x متغیر مستقل و d مقدار ثابت و جواب هر سؤال بله/خیر است.

۲. بهترین معیار شاخه‌زدن در انتخاب بهترین متغیر مستقل برای ایجاد شاخهٔ جدید. در این تحقیق از روش انحراف حداقل مربعات[14] برای ایجاد درخت رگرسیونی استفاده شد (معادلهٔ ۱۰).

$$SS(t) = \sum_{i=1}^{Ntt}(yi(t) - \check{y}(t))^2 \qquad (۱۰)$$

Ntt تعداد رکوردها (داده‌ها) در گره برگ t، Yi(t) مقدار خروجی (متغیر هدف در گره برگ)، و ẏ(t) میانگین مقادیر متغیر هدف برای همهٔ گره‌هاست.

حال متغیر ورودی S زمانی بهترین متغیر برای ایجاد شاخه در گره t است که مقدار Q(s,t) را بیشینه کند.

$$Q(s,t)=ss(t)-ss(t\ R)-SS(t\ L) \qquad (۱۱)$$

که در آن ss(t R) و SS(t L) به‌ترتیب میزان ss(t) در شاخهٔ سمت راست و سمت چپ گره t است.

۳. ایجاد آمار خلاصه برای گره انتهایی (Breiman et al., 1993; امیدوار و همکاران.، ۱۳۹۳).

۷،۲. ورودی و خروجی مدل‌ها

در این تحقیق از داده‌های ساعتی آلاینده‌های هوا و پارامترهای هواشناسی ساعتی به‌عنوان ورودی مدل‌ها در مراحل آموزش و آزمون استفاده شد. همچنین، خروجی تمامی مدل‌ها مقدار شاخص کیفت هوا در نظر گرفته شد. این شاخص به‌صورت ساعتی در محیط نرم‌افزار MATLAB محاسبه و پارامتر هدف در مدل‌ها استفاده شد. در این پژوهش ۸۰ درصد کل داده‌ها برای آموزش و ۲۰ درصد داده‌ها برای آزمون مدل‌ها استفاده شده است. تقسیم‌بندی داده‌ها به‌صورت تصادفی و بعد از حذف داده‌های گمشده و پرت انجام شد.

۸،۲. ارزیابی اعتبار مدل‌ها

برای ارزیابی عملکرد مدل‌ها و مقایسهٔ نتایج به‌دست‌آمده در مراحل آموزش و آزمون از شاخص‌های آماری نظیر شاخص صحت (IA)، بایاس (FB)، جذر میانگین مربعات خطا (RMSE)، میانگین خطای مطلق (MAE)، میانگین مربعات خطا (MSE)، ضریب همبستگی (R) و ضریب تبیین (R^2) استفاده شده است (معادله‌های ۱۲– ۱۸).

(۱۲)

$$IA = 1 - \sum_{i=1}^{N}(P_i - O_i)^2 \Big/ \sum_{i=1}^{N}\sum_{i=1}^{N}(|P_i - \bar{O}| + |O_i - \bar{O}|)^2$$

$$FB = 1/N \sum_{i=1}^{N} P_i - O_i /(P_i + O_i)/2 \qquad (۱۳)$$

$$RMSE = (1/N \sum_{i=1}^{N}(P_i - O_i)^2)^{0.5} \qquad (۱۴)$$

$$MAE = 1/N \sum_{i=1}^{N}|P_i - O_i| \qquad (۱۵)$$

$$MSE = 1/N \sum_{i=1}^{N}(P_i - O_i)^2 \qquad (۱۶)$$

(۱۷)

$$R = \sum_{i=1}^{N}\left((P_i - \bar{O})^2 - \sum_{i=1}^{N}(P_i - O_i)^2\right) \Big/ \sum_{i=1}^{N}(P_i - \bar{O})^2$$

$$R^2 = (R)^2 \qquad (۱۸)$$

N تعداد کل داده‌ها، P_i مقادیر پیش‌بینی‌شده، O_i مقادیر مشاهده یا محاسبه‌شده، و $\bar{O}$ میانگین مقادیر مشاهده یا محاسبه است.

۳. نتایج و یافته‌ها

نتایج حاصل از بررسی آماری داده‌های ایستگاه قلهک در جدول ۲ و ایستگاه تجریش در جدول ۳ و در ادامه نتایج حاصل از مدل‌سازی و برآورد شاخص کیفیت هوای شهری در دو ایستگاه پایش کیفیت هوای قلهک و تجریش ارائه شده است.

جدول ۲. مشخصات آماری هر یک از پارامترهای ایستگاه قلهک

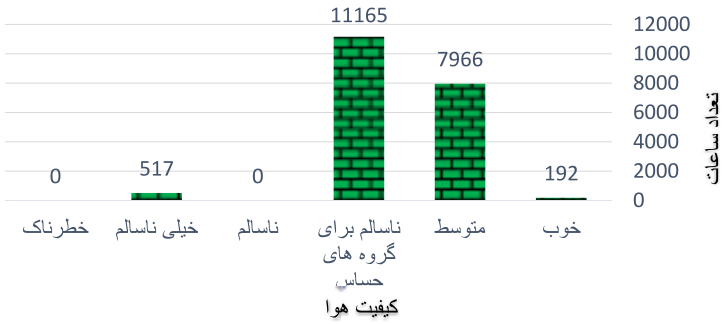
نام پارامتر	PM_{10}	NO_2	SO_2	CO	O_3	NO	CH_4	THC	HUM	Press	Temp	WD	WS
واحد	ug/m3	ppm	ppm	ppm	ppm	ppm	ppm	ppm	%RH	mBar	DegC	Deg	m/s
حداقل	۰/۰۰	۰/۰۰۱	۰/۰۰	۰/۰۱۸	۰/۰۰۱	۰/۰۰۱	۰/۰۰۱	۰/۰۰۱	۰/۳۷	۴۰۳/۹	−۷/۶۲۹	۰/۰	۰/۰۰۳
حداکثر	۳۱۶/۸	۰/۴۲۸	۰/۰۸	۹/۰۹۲	۰/۰۵۰	۰/۳۸۶	۴/۷۹۰	۱۷/۳۳	۹۵/۳۴	۹۵۷/۶	۳۹/۵۱	۳۲۵/۷	۴/۵
میانگین	۸۹/۶۹	۰/۱۲۰	۰/۰۳۸	۳/۵۵	۰/۰۱۵	۰/۰۸۸	۲/۱۱۲	۳/۵۳	۴۴/۳۳	۸۷۳/۹	۱۸/۲۸	۱۱۶/۳	۰/۹۰
چولگی	۱/۳۹	۱/۲۴	۰/۰۵	۰/۵۶۲	۱/۲۱	۱/۰۲	۰/۰۴۴	۰/۹۷	۰/۳۲	−۲/۲۴	−۰/۲۴	۰/۰۶	۰/۹۷
کشیدگی	۲/۰۵	۱/۶۵	۰/۲۳	۰/۲۰	۰/۸۸	۰/۷۹	−۰/۱۹	۲/۹	−۰/۶۳	۲۱/۷	−۰/۷۸	−۱/۳۱	۱/۱۳

جدول ۳. مشخصات آماری هر یک از پارامترهای ایستگاه تجریش

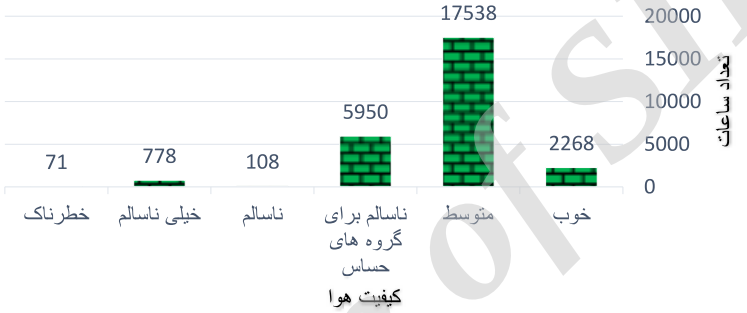
نام پارامتر	PM_{10}	NO_2	SO_2	CO	NOx	THC	NMHC	HUM	Press	Temp	WD	WS
واحد	ug/m3	ppm	ppm	ppm	ppm	ppm	ppm	%RH	mBar	DegC	Deg	m/s
حداقل	۰/۰	۰/۰	۰/۰۰۱	۰/۰۰۱	۰/۰۰۱	۰/۰۱۷	۰/۰۱۵	۲/۷۶	۴۱۲/۳	−۹/۹۷	۰/۰	۰/۰۰۱
حداکثر	۵۱۵	۰/۲۴	۰/۰۵۴	۱۳/۴۴	۰/۲۹	۱۴/۹۴	۸/۹۲	۹۹/۸۵	۸۴۷/۸	۶۱/۲۱	۳۵۵	۱۷/۰۹
میانگین	۱۱۷/۵	۰/۰۴	۰/۰۱۷	۴/۰۱	۰/۰۶	۵/۴	۲/۰۱	۳۴/۷	۸۳۳	۱۹/۲	۲۱۳/۶۵	۱/۰۵
چولگی	۱/۴	۱/۳۶	۰/۶۴	۰/۸۵	۱/۱۷	۰/۲۳	۰/۹۷	۱/۲۶	−۱۶/۷	۰/۳۴	−۰/۲۲	۲/۶۵
کشیدگی	۴	۳/۳۶	۰/۷۶	۰/۳	۱/۴۴	۰/۰۴	۱/۶۵	۰/۹۶	۲۸۵/۵۸	۰/۶۶	−۰/۱۱	۳۹/۳

۱.۳. برآورد شاخص کیفیت هوا

نتایج تحلیل حاصل از محاسبهٔ شاخص کیفیت هوا در ایستگاه قلهک در شکل ۱ و ایستگاه تجریش در شکل ۲ نشان داده شده است. بیشترین طبقهٔ کیفیت هوا در ایستگاه قلهک مربوط به طبقهٔ ناسالم برای گروه‌های حساس جامعه با تعداد ساعات ۱۱۱۶۵ و آلایندهٔ مسئول ایجاد کیفیت نامطلوب در اکثر اوقات دی‌اکسید نیتروژن است. عامل اصلی آن حجم ترافیک بیشتر خودروها، جمعیت بیشتر و احتراق سوخت در موتور خودروها در محدودهٔ این ایستگاه است. بیشترین طبقهٔ کیفیت هوا در ایستگاه تجریش مربوط به طبقهٔ متوسط با تعداد ساعات ۱۷۵۳۸ و آلایندهٔ مسئول ایجاد کیفیت نامطلوب در اکثر اوقات ذرات معلق کوچک‌تر از ۱۰ میکرون است.



شکل ۱. مقدار شاخص و طبقه کیفیت هوای ایستگاه قلهک



شکل ۲. مقدار شاخص و طبقه کیفیت هوای ایستگاه تجریش

۲.۳. مدل‌سازی شاخص کیفیت هوا

نتایج مدل‌سازی نشان داد که کارایی روش‌های به‌کاررفته در تحقیق در برآورد شاخص کیفیت هوای شهری عملکرد متفاوتی دارد. تعداد نمونه‌های ایستگاه قلهک در مرحلهٔ آموزش مدل‌ها ۱۵۸۷۲ و در مرحلهٔ آزمون ۳۹۶۸ و شامل سیزده پارامتر ورودی و یک پارامتر خروجی است. تعداد نمونه‌های ایستگاه تجریش در مرحلهٔ آموزش ۲۱۳۷۰ و در مرحلهٔ آزمون ۵۳۴۳ و شامل دوازده پارامتر ورودی و یک پارامتر خروجی است.

۳.۳. مدل درخت تصمیم

در این قسمت نتایج حاصل از کاربرد الگوریتم CART در مدل‌سازی و برآورد شاخص کیفیت هوای ایستگاه قلهک و تجریش شهر تهران ارائه شده است. در جدول ۴ مقادیر شاخص‌های آماری حاصل از سه مرحلهٔ آموزش، آزمون درخت تصمیم قبل از هرس و آزمون بعد از هرس درخت تصمیم ارائه شده است. برای هرس درخت از معیار پیچیدگی استفاده شد و مقدار آن پس از آزمون و خطای انجام‌شده روی درخت به مقدار ۰/۰۳ در تعیین بهترین ساختار درخت مشخص شد. نتایج به‌دست آمده حاکی از این است که الگوریتم CART در شبیه‌سازی شاخص کیفیت هوا عملکرد بسیار بالایی دارد، به‌طوری که همبستگی بین مقادیر شبیه‌سازی‌شده و مشاهده‌شده بسیار نزدیک به عدد ۱ است. مقادیر خطای درخت در سه مرحله آزمون و آموزش بسیار ناچیز و مشابه است. مهم‌ترین پارامترهای تأثیرگذار در برآورد AQI در ایستگاه قلهک PM_{10} با ۱۸/۵ درصد تأثیر و NO_2 با ۳۲/۲ درصد تأثیر است. مهم‌ترین عوامل در برآورد AQI در ایستگاه تجریش ذرات معلق کوچک‌تر از ۱۰ میکرون با ۴۷ درصد، دی‌اکسید نیتروژن با ۲/۱ و مونوکسید کربن با ۳/۱ درصد تأثیر است.

جدول ۴. نتایج حاصل از ارزیابی مدل درخت تصمیم

ایستگاه	IA (1)	FB (0)	R^2 (1)	R (1)	MAE (0)	RMSE (0)	MSE (0)	مدل درخت تصمیم
قلهک	۱	۰	۰/۹۹۹۹	۰/۹۹۹۹	۰/۰۴۰۸	۰/۲۵۹۳	۰/۰۶۷	**مرحله آموزش**
	۱	۰	۰/۹۹۹۱	۰/۹۹۹۶	۰/۱۰۱۲	۰/۷۵	۰/۵۶۳	**مرحله آزمون قبل از هرس**
	۱	۰	۰/۹۹۷۲	۰/۹۹۸۶	۰/۸۸۴	۱/۳۱	۱/۷۳	**مرحله آزمون بعد از هرس**
تجریش	۱	۰	۰/۹۹۹۸	۰/۹۹۹۹	۰/۱۱۵۷	۰/۴۴۲۸	۰/۱۹۶۱	**مرحله آموزش**
	۱	۰	۰/۹۹۹۶	۰/۹۹۹۸	۰/۱۹۵۱	۰/۶۸۸۸	۰/۴۷۴۴	**مرحله آزمون قبل از هرس**
	۱	۰	۰/۹۹۶۸	۰/۹۹۸۴	۱/۱۸۴	۱/۸۵	۳/۴۲	**مرحله آزمون بعد از هرس**

اعداد داخل پرانتز مقادیر ایده‌آل هر شاخص را نشان می‌دهد

۴.۳. مدل‌سازی با شبکهٔ عصبی مصنوعی

پس از ایجاد مدل شبکهٔ عصبی مصنوعی با داده‌های آموزش برای اعتبارسنجی مدل از داده‌های آزمون استفاده شد. همچنین، تعداد نورون لایهٔ پنهان از طریق رابطهٔ (2n+1) (n: تعداد ورودی) و بهترین تعداد نورون از طریق کمترین خطا و بیشترین همبستگی بین متغیر هدف و مقادیر برآوردشده تعیین شد. تعداد دور بذای آموزش شبکهٔ ۲۰۰۰ در نظر گرفته شد و در مجموع با توجه به تعداد سیزده پارامتر ورودی در مدل شبکهٔ عصبی در ایستگاه قلهک ۵۴۰۰۰ و در ایستگاه تجریش با توجه به تعداد دوازده پارامتر ورودی در مدل شبکهٔ عصبی ۵۰۰۰۰ دور در نظر گرفته شد. در جدول ۵ مقادیر شاخص‌های آماری حاصل از دو مرحلهٔ آموزش و آزمون شبکهٔ عصبی با یک لایهٔ پنهان و با الگوریتم‌های آموزش مختلف و در بهینه‌ترین حالت ارائه شده است.

جدول ۵. نتایج حاصل از مراحل آموزش و آزمون شبکهٔ عصبی در بهینه‌ترین حالت

مشخصات		ایستگاه قلهک			ایستگاه تجریش		
	الگوریتم آموزش	Lm	Scg	Rb	Lm	Scg	Rb
	تعداد لایهٔ مخفی	۱	۱	۱	۱	۱	۱
	تعداد نورون	۲۰	۲۵	۱۲	۲۵	۲۴	۱۳
مرحلهٔ آموزش	MSE	۰/۰۰۰۰۵	۰/۰۰۰۲۸	۰/۰۰۰۷۴	۰/۰۰۰۰۲۰	۰/۰۰۰۱۶	۰/۰۰۰۲۵
	RMSE	۰/۰۰۶۸	۰/۰۱۶	۰/۰۲۷۲	۰/۰۰۴۴	۰/۰۱۲۶	۰/۰۱۶۰
	R	۰/۹۹۹۷۴	۰/۹۹۸۴	۰/۹۹۵۹	۰/۹۹۹۶	۰/۹۹۷۲	۰/۹۹۵۶
	R^2	۰/۹۹۹۴۹	۰/۹۹۶۸۶	۰/۹۹۱۸۹	۰/۹۹۹۳	۰/۹۹۴۵	۰/۹۹۱۳
	MAE	۰/۰۰۴۳	۰/۰۱۱۱	۰/۰۱۹۵	۰/۰۰۲۶	۰/۰۰۸۶	۰/۰۱۰۷
	FB	۰/۰۰۰۱۳	۰/۰۰۱۸	۰/۰۰۴۱	۰	۰	۰
	IA	۰/۹۹۹۹	۰/۹۹۹۹	۰/۹۹۹۹	۱	۰/۹۹۹	۰/۹۹۹۹
مرحلهٔ آزمون	MSE	۰/۰۰۰۰۷	۰/۰۰۰۲۹	۰/۰۰۰۷۳	۰/۰۰۰۰۲۴	۰/۰۰۰۱۶	۰/۰۰۰۲۹
	RMSE	۰/۰۰۸۲۵	۰/۰۱۷۱	۰/۰۲۷۰	۰/۰۰۴۸	۰/۰۱۲۷	۰/۰۱۷۲
	R	۰/۹۹۹۶۳	۰/۹۹۸۳	۰/۹۹۵۸	۰/۹۹۹۵	۰/۹۹۷۲	۰/۹۹۴۷
	R^2	۰/۹۹۹۲۵	۰/۹۹۶۷۸	۰/۹۹۱۶	۰/۹۹۹۱	۰/۹۹۴۴	۰/۹۸۹۵
	MAE	۰/۰۰۴۵۳	۰/۰۱۱۰	۰/۰۱۹۴	۰/۰۰۲۸	۰/۰۰۸۶	۰/۰۱۱۱
	FB	۰/۴۷۴	۰/۰۲۵۴	۰/۰۰۶۳	۰	۰/۰۰۰۱	۰
	IA	۰/۹۹۹۹	۰/۹۹۹۹	۰/۹۹۹۹	۱	۰/۹۹۹	۰/۹۹۹۹

با روش سعی و خطا مشخص شد که شبکهٔ عصبی مصنوعی پرسپترون با یک لایهٔ پنهان و الگوریتم آموزش لونبرگ مارکوآرت با بیست نورون در لایهٔ پنهان در ایستگاه قلهک و ۲۵ نورون در ایستگاه تجریش بهترین عملکرد را برای برآورد و مدل‌سازی شاخص کیفیت هوا به‌دست می‌دهد. همچنین، یک شبکهٔ عصبی پرسپترون با ۲۵ نورون در لایهٔ پنهان در ایستگاه قلهک و ۲۴ نورون در ایستگاه تجریش و الگوریتم شیب توأم مقیاس‌شده نسبت به شبکهٔ عصبی پرسپترون با الگوریتم آموزش پس‌انتشار ارتجاعی عملکرد بهتری داشت.

۵.۳. مدل‌سازی با رگرسیون مؤلفه‌های اصلی و رگرسیون چندگانه

۱.۵.۳. اجرای روش PCA در پیش‌پردازش داده‌های ورودی

بررسی اولیه نشان داد که بین متغیرهای ورودی مورد استفاده در ایستگاه قلهک و تجریش همبستگی معناداری وجود دارد. برای ازبین‌بردن این مشکل، از روش تحلیل مؤلفه‌های اصلی استفاده شد. برای بررسی امکان اجرای تحلیل مؤلفه‌های اصلی از آزمون KMO استفاده شد. مقدار KMO= ۰/۵۸۱ در ایستگاه قلهک و KMO=۰/۶۰۶ در ایستگاه تجریش امکان اجرای روش PCA را تأیید کرد. برای اجرای این روش، پس از استانداردکردن متغیرهای ورودی ماتریس همبستگی تشکیل شد (جدول ۶ و ۷). با حل دستگاه معادلهٔ (۷)، سیزده مقدار ویژه و به‌ازای هر مقدار ویژه سیزده بردار ویژه برای ایستگاه قلهک و دوازده مقدار ویژه و به‌ازای هر مقدار ویژه دوازده بردار ویژه برای ایستگاه تجریش حاصل شد که با استفاده از آن‌ها، مؤلفه‌های اصلی از متغیرهای اولیه به‌دست آمد. مقدار عددی هر مؤلفه با تقسیم مقادیر ویژهٔ به‌دست آمده بر تعداد متغیرهای مورد استفاده به‌دست آمد. درصد پراکندگی نیز از تقسیم مقدار عددی هر مؤلفه بر تعداد متغیرهای مورد استفاده محاسبه شد. مشخصات هر مؤلفه در جدول ۸ آمده است. مقدار اولین مؤلفه در ایستگاه قلهک برابر ۳/۱۳۴ است که ۲۴/۱۰۹ درصد از کل واریانس موجود در سری داده‌ها را توجیه می‌کند. با توجه به جدول ۸، مقادیر ویژهٔ چهار مؤلفهٔ بعدی در ایستگاه قلهک نیز بزرگ‌تر از ۱ و به صورت جمعی با مؤلفهٔ اول دارای واریانس تجمعی ۷۲/۱۳۶ درصد کل داده‌هاست. همچنین، مقدار اولین مؤلفه برابر ۳/۴۵۹ است که ۲۸/۸۲۷ درصد از کل واریانس موجود در سری داده‌ها را توجیه می‌کند. مقادیر ویژهٔ سه مؤلفهٔ بعدی در ایستگاه تجریش نیز بزرگ‌تر از ۱ و به صورت جمعی با مؤلفهٔ اول دارای واریانس تجمعی ۶۵/۷۱۴ درصد کل داده‌هاست. بدین ترتیب، مؤلفهٔ اول تا پنجم در ایستگاه قلهک و مؤلفهٔ اول تا چهارم در ایستگاه تجریش مقادیر ویژهٔ بزرگ‌تر از ۱ دارد. ضرایب بردارهای ویژهٔ ایستگاه قلهک و ایستگاه تجریش در جدول ۹ آمده است. در جدول ۹ مقادیر بردارهای ویژه و ایجادکنندهٔ ضرایب هر یک از پارامترهای تشکیل‌دهندهٔ هر مؤلفه بیانگر آن است که در مؤلفهٔ اول ضرایب مربوط به متغیرهای CO، SO_2 و NO_2 مقادیر بیشتر و تأثیر بسزایی در تشکیل این مؤلفه دارد و با قلم پررنگ مشخص شده است. در مؤلفهٔ دوم متغیرهای رطوبت و دمای هوا عوامل اصلی ایجاد این مؤلفه است. در مؤلفهٔ سوم مقدار PM_{10}، SO_2 و فشار هوا عوامل اصلی ایجاد این مؤلفه است. همچنین، در تشکیل مؤلفهٔ چهارم متغیرهای جهت باد، سرعت باد و O_3 و در مؤلفهٔ پنجم متان و کل هیدروکربن‌ها عوامل اصلی تشکیل این مؤلفه است. در ایستگاه تجریش نیز مشاهده می‌شود که در مؤلفهٔ اول ضرایب مربوط به متغیرهای PM_{10}، NO_2، SO_2، CO و NOx مقادیر بیشتری و تأثیر بسزایی در تشکیل این مؤلفه دارد و با قلم پررنگ مشخص شده است. در مؤلفهٔ دوم متغیرهای کل هیدروکربن‌ها و هیدروکربن‌های بدون متان عوامل اصلی ایجاد این مؤلفه است. در مؤلفهٔ سوم مقدار رطوبت، دمای هوا و سرعت باد عوامل اصلی ایجاد این مؤلفه است. همچنین، در تشکیل مؤلفهٔ چهارم متغیرهای جهت باد و فشار هوا عوامل اصلی تشکیل این مؤلفه است.

جدول ۶. ماتریس همبستگی بین متغیرهای ورودی روش PCA داده‌های ایستگاه قلهک

											PM_{10}	**
										NO_2	۱	PM_{10}
									SO_2	۱	-۰/۱۸	NO_2
								CO	۱	۰/۲۱	۰/۲۷	SO_2
							O_3	۱	۰/۱۵	۰/۵۲	۰/۰۰۱	CO
						NO	۱	-۰/۲۷	-۰/۰۰۸	-۰/۲۳	۰/۰۹	O_3
					CH_4	۱	-۰/۲۸	۰/۵۳	۰/۱۳	۰/۹۱	-۰/۱۲	NO
				THC	۱	۰/۱۸	-۰/۱۸	۰/۱۱	۰/۳	۰/۲۵	۰/۰۵	CH_4
			HUM	۱	۰/۵۵	۰/۰۶	-۰/۱۷	۰/۲۸	-۰/۰۹	۰/۰۵	۰/۱۲	THC
		Press	۱	۰/۰۱	۰/۰۹	۰/۱۱	-۰/۲۸	۰/۰۲	-۰/۰۳	۰/۱۹	-۰/۱۷	HUM
	Temp	۱	-۰/۱۸	۰/۰۱	۰/۲۸	-۰/۰۲	۰/۰۵	-۰/۰۴	۰/۴۱	-۰/۰۱	۰/۳۰	Press
WD	۱	۰/۰۲	-۰/۶۳	۰/۱۳	-۰/۲۳	-۰/۲۸	۰/۲۲	۰/۰۱	-۰/۲	-۰/۴۰	۰/۱۲	Temp
۱	-۰/۰۶	۰/۱۵	-۰/۰۴	-۰/۰۱	۰/۰۵	-۰/۱۹	۰/۱۵	-۰/۱۶	۰/۱	-۰/۲۴	۰/۳۱	WD
۰/۱۶	-۰/۰۳	-۰/۰۴	-۰/۰۹	-۰/۲۶	-۰/۲۹	-۰/۱۴	۰/۴۵	-۰/۱۸	-۰/۰۷	-۰/۰۷	۰/۰۲	WS

جدول ۷. ماتریس همبستگی بین متغیرهای ورودی روش PCA داده‌های ایستگاه تجریش

											PM_{10}	**
										NO_2	۱	PM_{10}
									SO_2	۱	۰/۲۴	NO_2
								CO	۱	۰/۲۹	۰/۲۹	SO_2
							NOx	۱	۰/۴۴	۰/۵۰	۰/۳۰	CO
						THC	۱	۰/۶۲	۰/۳۸	۰/۸۷	۰/۲۶	NOx
					NMHC	۱	۰/۱۲	۰/۲۴	۰/۰۵-	۰/۱۲	۰/۱۳	THC
				HUM	۱	۰/۸۵	۰/۲۵	۰/۵۲	۰/۰۰	۰/۲۳	۰/۲۲	NMHC
			Press	۱	۰/۰۲	۰/۱۲	۰/۲۸	۰/۲۲	۰/۲۵	۰/۱۸	۰/۰۲-	HUM
		Temp	۱	۰/۱۰	۰/۰۷	۰/۱۰	۰/۰۸	۰/۰۵	۰/۰۳	۰/۰۸	۰/۰۵	Press
	WD	۱	۰/۰۲-	۰/۶۵-	۰/۰۸-	۰/۱۹-	۰/۲۷-	۰/۲۲-	۰/۲۱-	۰/۱۹-	۰/۰۱-	Temp
WS	۱	۰/۱۴	۰/۱۴	۰/۰۹-	۰/۱۱-	۰/۰۰	۰/۱۲-	۰/۲۰-	۰/۱۳-	۰/۱۰-	۰/۰۸-	WD
۱	۰/۰۵-	۰/۱۹	۰/۰۱	۰/۲۷-	۰/۰۹-	۰/۲۰-	۰/۱۸-	۰/۱۴-	۰/۱۱-	۰/۱۰-	۰/۰۶	WS

جدول ۸. مشخصات مؤلفه‌های ایجادشده با روش PCA

	ایستگاه قلهک			ایستگاه تجریش		
مؤلفه‌ها	مقادیر ویژه	درصد واریانس	واریانس تجمعی	مقادیر ویژه	درصد واریانس	واریانس تجمعی
۱	۳/۱۳۴	۲۴/۱۰۹	۲۴/۱۰۹	۳/۴۵۹	۲۸/۸۲۷	۲۸/۸۲۷
۲	۲/۰۵۶	۱۵/۸۱۴	۳۹/۹۲۳	۱/۷۲۵	۱۴/۳۷۶	۴۳/۲۰۳
۳	۱/۶۶۷	۱۲/۸۲۱	۵۲/۷۴۴	۱/۵۴۳	۱۲/۸۵۵	۵۶/۰۵۸
۴	۱/۵۰۱	۱۱/۵۴۳	۶۴/۲۸۸	۱/۱۵۹	۹/۶۵۶	۶۵/۷۱۴
۵	۱/۰۲	۷/۸۴۸	۷۲/۱۳۶	۰/۹۸۶	۸/۲۱۴	۷۳/۹۲۸
۶	۰/۹۱۹	۷/۰۷۳	۷۹/۲۰۸	۰/۸۷۹	۷/۳۲۶	۸۱/۲۵۴
۷	۰/۶۵۵	۵/۰۴۱	۸۴/۲۵	۰/۷۳۲	۶/۰۹۷	۸۷/۳۵۲
۸	۰/۵۶۲	۴/۳۲۲	۸۸/۵۷۲	۰/۶۱۳	۵/۱۱	۹۲/۴۶۱
۹	۰/۵۱۶	۳/۹۶۹	۹۲/۵۴۱	۰/۴۰۲	۳/۳۴۹	۹۵/۸۱
۱۰	۰/۴۵۵	۳/۴۹۶	۹۶/۰۳۸	۰/۳۲۸	۲/۷۳۱	۹۸/۵۴۱
۱۱	۰/۲۴۷	۱/۹۰۱	۹۷/۹۳۹	۰/۱۰۴	۰/۸۶۴	۹۹/۴۰۵
۱۲	۰/۲۰۷	۱/۵۹۶	۹۹/۵۳۵	۰/۰۷۱	۰/۵۹۵	۱۰۰
۱۳	۰/۰۶	۰/۴۶۵	۱۰۰			

جدول ۹. ضرایب بردارهای ویژه برای ایجاد مؤلفه اصلی

	مؤلفه‌های اصلی									
	ایستگاه قلهک					ایستگاه تجریش				
پارامترها	۱	۲	۳	۴	۵	پارامترها	۱	۲	۳	۴
PM_{10}	−۰/۱۴۲	−۰/۱۹۹	**۰/۵۵۱**	۰/۲۲۳	۰/۲۵۳	PM_{10}	**۰/۵۳۶**	۰/۱۷۲	−۰/۲۵۶	−۰/۰۶۷
NO_2	**۰/۹۲۱**	۰/۲۶۹	۰/۰۴۴	−۰/۰۵۳	−۰/۰۰۳	NO_2	**۰/۸۲۴**	−۰/۰۷۱	۰/۱۱۴	۰/۱۱۷
SO_2	۰/۲۰۳	۰/۱۲۷	**۰/۸۰۶**	−۰/۰۵۴	−۰/۰۸۶	SO_2	**۰/۶۲۳**	−۰/۱۹	−۰/۲۰۶	−۰/۱۲۸
CO	**۰/۷۳۱**	−۰/۱۶۲	−۰/۰۱۸	−۰/۱۵۸	۰/۲۲۷	CO	**۰/۷۳۴**	۰/۳۴۴	۰/۱۴۱	−۰/۱۱۸
O_3	−۰/۱۶۷	−۰/۲۶۵	۰/۰۳۵	**۰/۷۰۶**	−۰/۱۳۴	NOx	**۰/۸۶۵**	−۰/۰۷۳	−۰/۲۲۶	۰/۰۸۴
NO	**۰/۹۱۲**	۰/۱۵۱	۰/۰۱۹	−۰/۰۹۰	۰/۰۰۸	THC	۰/۰۱۲	**۰/۹۳۴**	۰/۱۷۷	−۰/۰۶۷
CH_4	۰/۱۴۷	۰/۲۴۱	۰/۳۳۶	−۰/۱۸۴	**۰/۶۹۵**	NMHC	۰/۲۲۵	**۰/۹۴۵**	۰/۰۰۳	−۰/۰۴۳
THC	۰/۰۹۴	−۰/۱۲۸	−۰/۱۰۴	−۰/۱۳۶	**۰/۹۱۱**	HUM	−۰/۱۵۸	−۰/۰۳۵	−۰/۸۶	−۰/۰۲۳
HUM	۰/۰۰۰	**۰/۸۳۴**	−۰/۲۰۰	−۰/۱۴۸	۰/۰۵۸	Press	−۰/۱۳۶	−۰/۰۶۸	−۰/۰۶۴	**۰/۶۹۱**
Press	−۰/۰۵۰	−۰/۰۹۷	**۰/۷۸۸**	−۰/۰۳۸	۰/۰۴۱	Temp	−۰/۱۵	−۰/۰۵۴	**−۰/۸۱۴**	−۰/۱۴۷
Temp	−۰/۱۸۷	**−۰/۹۰۸**	−۰/۱۰۰	−۰/۰۲۸	۰/۰۴۳	WD	−۰/۱۶۵	−۰/۰۶۵	−۰/۰۶	**۰/۷۶۷**
WD	−۰/۲۹۵	۰/۱۵۸	۰/۳۰۵	**۰/۵۰۲**	۰/۲۷۲	WS	−۰/۰۱۶	−۰/۱۶۵	**−۰/۵۳۲**	−۰/۱۶۱
WS	۰/۰۰۰	۰/۰۲۸	−۰/۱۰۶	**۰/۸۲۲**	−۰/۲۳۸					

۲.۵.۳. مدل رگرسیون مبتنی بر مؤلفه‌های اصلی و رگرسیون چندگانه

به‌منظور توسعهٔ مدل رگرسیون مؤلفه‌های اصلی، پنج مؤلفهٔ ایجادشدهٔ ایستگاه قلهک و چهار مؤلفهٔ ایجادشدهٔ ایستگاه تجریش ورودی مدل رگرسیون استفاده شد. از آزمون دوربین– واتسون به‌منظور بررسی پیش‌فرض نرمال‌بودن توزیع خطای باقیمانده‌های مدل رگرسیون استفاده شد. در صورتی که مقدار این آماره بین عدد ۱/۵ تا ۲/۵ باشد، نشان‌دهندهٔ نرمال بودن توزیع خطاست. در مدل رگرسیون ایستگاه قلهک مقدار این آماره ۲/۰۰۷۹ و در مدل ایستگاه تجریش ۱/۹۹ به‌دست آمد که نشان از نرمال‌بودن توزیع خطا داشت. در بررسی پیش‌فرض عدم‌هم‌خطی (استقلال خطاها) از آمارهٔ تورم واریانس استفاده شد که مقدار ایده‌آل آن ۱ و مقادیر بیشتر از ۱۰ نشان‌دهندهٔ مشکل‌سازبودن هم‌خطی بین متغیرهای مستقل در مدل رگرسیون است. با استفاده از روش PCA مشکل همبستگی بین متغیرهای مستقل نیز رفع شد. پس از تأیید نرمال‌بودن توزیع مقادیر خطا و رفع مشکل همبستگی در متغیرهای مستقل، مدلی مناسب با استفاده از روش رگرسیون خطی چندگانه برای برآورد مقدار شاخص کیفیت هوا بسط‌یافت. نتایج ورود مؤلفه‌های اصلی به مدل رگرسیون ایستگاه قلهک در جدول ۱۰ و ایستگاه تجریش در جدول ۱۱ آمده است. پنج مؤلفهٔ ایجادشده در ایستگاه قلهک و چهار مؤلفهٔ ایستگاه تجریش ورودی مدل رگرسیون استفاده شد. نتایج ارزیابی مدل رگرسیون مؤلفه‌های اصلی در جدول ۱۲ آمده است. در جدول ۱۴ نتایج ورود تمامی متغیرهای هواشناسی و آلودگی هوا به مدل رگرسیون ارائه شده است. در جدول ۱۳ نتایج ارزیابی مدل رگرسیون چندگانه آمده است. نتایج ارزیابی مدل‌ها نشان می‌دهد که مدل رگرسیونی ایجادشده توسط تمامی پارامترهای هواشناسی و آلودگی هوا نسبت به مدل ایجادشده توسط مؤلفه‌های اصلی در برآورد مقدار شاخص کیفیت هوا عملکرد بهتری دارد. در معادلهٔ (۱۹) مدل رگرسیون مؤلفه‌های اصلی برآوردگر AQI و در معادلهٔ (۲۰) مدل رگرسیون چندگانهٔ ایستگاه قلهک ارائه شده است. همچنین، در معادلهٔ (۲۱) مدل رگرسیون مؤلفه‌های اصلی برآوردگر AQI و در معادلهٔ (۲۲) مدل رگرسیون چندگانهٔ ایستگاه تجریش ارائه شده است.

جدول ۱۰. نتایج ورود مؤلفه‌های اصلی به مدل رگرسیون چندگانه ایستگاه قلهک

مؤلفه‌ها	ضرایب	فاکتور تورم واریانس	سطح معناداری
(Constant)	۶۳/۷۴-	———	۰
PC1	۹/۸۹	۱/۱۰	۰
PC2	۰/۲۰	۱/۰۴	۰
PC3	۰/۱۹	۱/۱۵	۰
PC4	۰/۰۹۴-	۱/۱۲	۰
PC5	۱/۰۹-	۱/۰۹	۰

جدول ۱۱. نتایج ورود مؤلفه‌های اصلی به مدل رگرسیون چندگانهٔ ایستگاه تجریش

مؤلفه‌ها	ضرایب	فاکتور تورم واریانس	سطح معناداری
(Constant)	۲۸/۲۳	———	۰
PC1	۰/۹۳۳	۱/۰۴	۰
PC2	۰/۲۴۱۵	۱/۰۵	۰
PC3	۰/۰۳۳۶	۱/۰۲	۰
PC4	۰/۰۰۸۸-	۱/۰۱	۰

جدول ۱۲. نتایج ارزیابی مدل رگرسیون مؤلفه‌های اصلی

ایستگاه	IA (1)	FB (0)	R^2 (1)	R (1)	MAE (0)	RMSE (0)	MSE (0)	مدل PCR
قلهک	۱	۰	۰/۴۰۰۹	۰/۶۳۳۲	۱۵/۳۸	۱۹/۳۸	۳۷۵/۷۲	مرحلهٔ آموزش
	۰/۹	۰/۰۰۳-	۰/۳۹۶۶	۰/۶۲۹۸	۱۵/۵۵	۱۹/۵۵	۳۸۲/۲۸	مرحلهٔ آزمون
تجریش	۱	۰	۰/۸۸۸۴	۰/۹۴۲۵	۵/۷	۱۱/۱۲	۱۲۵/۷	مرحلهٔ آموزش
	۱	۰/۰۰۳-	۰/۸۹۲۴	۰/۹۴۴۷	۵/۵	۱۰/۷	۱۱۵/۳	مرحلهٔ آزمون

جدول ۱۳. نتایج ارزیابی مدل رگرسیون چندگانه

IA (1)	FB (0)	R^2 (1)	R (1)	MAE (0)	RMSE (0)	MSE (0)	مدل MLR
۱	۰	۰/۸۱۱۲	۰/۹۰۰۷	۸/۷۴	۱۰/۸۶	۱۱۷/۹۷	مرحلهٔ آموزش
۱	۰/۰۰۱-	۰/۸۱۹۰	۰/۹۰۵۰	۸/۶۹	۱۰/۷۷	۱۱۶/۱۷	مرحلهٔ آزمون
۱	۰	۰/۹۰۲۱	۰/۹۴۹۸	۵/۷	۱۰/۵	۱۱۱/۲	مرحلهٔ آموزش
۱	۰/۰۰۲	۰/۸۹۵۰	۰/۹۴۶۰	۵/۷	۱۰/۴	۱۰۹/۱	مرحلهٔ آزمون

جدول ۱۴. نتایج ورود تمامی متغیرها به مدل رگرسیون چندگانه

ایستگاه قلهک			ایستگاه تجریش		
متغیرها	ضرایب	فاکتور تورم واریانس	سطح معناداری	متغیرها	ضرایب	فاکتور تورم واریانس	سطح معناداری
(Constant)	۲۳/۴۷	------	۰	(Constant)	۲۶/۶۵	------	۰
PM_{10}	۰/۲۲۴	۱/۴۰	۰	PM_{10}	۰/۴۸	۱/۲۱	۰
NO_2	۲۵۷/۳۱	۹/۴۰	۰	NO_2	۱۱۱/۹۲	۴/۵۲	۰
SO_2	۴۳/۲۹	۱/۶۹	۰	SO_2	۱۶۰/۵-	۱/۴۹	۰
CO	۰/۱۷۸	۱/۸۴	۰	CO	۱/۶۶	۳/۲۲	۰
O_3	۵/۳۴	۱/۵۱	۰/۰۱۴	NOx	۳۸/۳۰-	۵/۸۰	۰
NO	۲۸/۹۹	۷/۶۷	۰/۵۹	THC	۰/۰۲۱	۵/۷۹	۰/۷۴۲
CH_4	۰/۳۶۸-	۲/۴۲	۰	NMHC	۰/۲-	۷/۴۳	۰/۲۴
THC	۰/۶۲۵-	۲/۱۷	۰	HUM	۰/۰۴۷	۱/۹۵	۰
HUM	۰/۰۵۲۹-	۱/۹۶	۰	Press	۰/۰۰۸-	۱/۰۶	۰
Press	۰/۰۳۴۴	۱/۴۰	۰	Temp	۰/۰۲۳	۱/۸۷	۰
Temp	۰/۰۸۷-	۲/۶۸	۰	WD	۰/۰۰۲-	۱/۱۲	۰/۰۱۸
WD	۰/۰۳۳-	۱/۲۹	۰	WS	۰/۹۰	۱/۲۰	۰
WS	۰/۰۲۱-	۱/۲۱	۰				

$$AQI= -63/74 + 9/89\times PC_1 + /2\times PC2 + /19\times PC_3 - /094\times PC_4 – 1/09\times PC_5 \quad (۱۹)$$

$$AQI= 23/47+/224\times PM_{10}+257/3\times NO_2+43/29\times SO_2+/178\times CO+5/34\times O_3+ \quad (۲۰)$$
$$28/99\times NO-/368\times CH_4-/625\times THC-/0529\times HUM+/0344\times Press-/087\times Temp-/033\times WD-/021\times WS$$

$$AQI= 28/23 + /933\times PC_1 + /2415\times PC_2 + /0336\times PC_3 - /0088\times PC_4 \quad (۲۱)$$

$$AQI=26/65+/48\times PM_{10}+111/92\times NO_2-160/5\times SO_2+1/66\times CO -38/30\times NOx + \quad (۲۲)$$
$$/021\times THC-/2\times NMHC+/047\times HUM- /008\times Press-/023\times Temp-/002\times WD+0/9\times WS$$

۴. بحث

مقدار آماره‌های خطا نظیر RMSE و MAE مدل رگرسیون درخت تصمیم در بهینه‌ترین حالت در مرحلهٔ آموزش به‌ترتیب در ایستگاه تجریش برابر ۰/۴۴۲۸ و ۰/۱۱۵۷ و در مرحلهٔ آزمون به‌ترتیب ۰/۶۸۸۸ و ۰/۱۹۵۱ و مقدار آماره‌های خطا نظیر RMSE و MAE در مرحلهٔ آموزش به‌ترتیب در ایستگاه قلهک برابر ۰/۲۵۹۳ و ۰/۰۴۰۸ و در مرحلهٔ آزمون به‌ترتیب ۰/۷۵ و ۰/۱۰۱۲ است. مقایسهٔ این آماره‌ها در دو ایستگاه نشان می‌دهد که مدل درخت تصمیم در ایستگاه قلهک نسبت به مدل درخت تصمیم در ایستگاه تجریش عملکرد بهتری دارد.

آماره‌های ضریب همبستگی R و ضریب تبیین R^2 در هر دو ایستگاه در مرحلهٔ آموزش برای ایستگاه تجریش برابر ۰/۹۹۹۹ و ۰/۹۹۹۸ و در مرحلهٔ آزمون ۰/۹۹۹۸ و ۰/۹۹۹۶ و در ایستگاه قلهک به‌ترتیب در مرحلهٔ آموزش برابر ۰/۹۹۹۹ و ۰/۹۹۹۹ و در مرحلهٔ آزمون ۰/۹۹۹۶ و ۰/۹۹۹۱ است. معیارهای همبستگی نشان از این دارد که بین مقادیر برآوردشدهٔ مدل و محاسبه‌شدهٔ کیفیت هوا در مدل درخت تصمیم در هر دو ایستگاه اختلاف بسیار ناچیز است. آماره‌های ضریب همبستگی R و ضریب تبیین R^2 در هر دو مدل نزدیک به عدد ۱ است که نشان از توانایی بالای مدل درخت تصمیم رگرسیون در برآورد شاخص کیفیت هوای شهری دارد.

نتایج حاصل از مدل درخت تصمیم رگرسیون با تحقیقات قبلی نظیر Kumar & Goyal (2013, 2011)، Sing et al. (2013) و Carbajal et al. (2011) مطابقت دارد و نتایج بهتری نسبت به مدل‌های رگرسیون خطی حاصل شده است. مقدار آماره‌های خطا نظیر RMSE و MAE شبکهٔ عصبی مصنوعی پرسپترون در بهینه‌ترین حالت در مرحلهٔ آموزش به‌ترتیب در ایستگاه تجریش برابر ۰/۰۰۴۴ و ۰/۰۰۲۶ و در مرحلهٔ آزمون به‌ترتیب ۰/۰۰۴۸ و ۰/۰۰۲۸ و مقدار آماره‌های خطا نظیر RMSE و MAE در مرحلهٔ آموزش به‌ترتیب در ایستگاه قلهک برابر ۰/۰۰۶۸ و ۰/۰۰۴۳ و در مرحلهٔ آزمون به‌ترتیب ۰/۰۰۸۲ و ۰/۰۰۴۵ است. مقایسهٔ این آماره‌ها در دو ایستگاه نشان می‌دهد که مدل شبکهٔ عصبی مصنوعی در ایستگاه تجریش نسبت به مدل شبکهٔ عصبی مصنوعی در ایستگاه قلهک عملکرد بهتری دارد.

آماره‌های ضریب همبستگی R و ضریب تبیین R^2 در هر دو ایستگاه در مرحلهٔ آموزش برای ایستگاه تجریش برابر ۰/۹۹۹۶ و ۰/۹۹۹۳ و در مرحلهٔ آزمون ۰/۹۹۹۵ و ۰/۹۹۹۱ و در ایستگاه قلهک به‌ترتیب در مرحلهٔ آموزش برابر ۰/۹۹۹۷ و ۰/۹۹۹۴ و در مرحلهٔ آزمون ۰/۹۹۹۶ و ۰/۹۹۹۲ است. معیارهای همبستگی نشان از این دارد که بین مقادیر برآوردشدهٔ مدل و محاسبه‌شدهٔ کیفیت هوا در مدل شبکهٔ عصبی در ایستگاه قلهک نسبت به مدل شبکهٔ عصبی در ایستگاه تجریش همبستگی بیشتری است و این اختلاف بسیار ناچیز و ۰/۰۰۰۱ است.

آماره‌های ضریب همبستگی R و ضریب تبیین R^2 در هر دو مدل نزدیک به عدد ۱ است که نشان از توانایی

بالای مدل شبکهٔ عصبی مصنوعی در برآورد شاخص کیفیت هوای شهری در دو ایستگاه مورد مطالعه دارد. نتایج حاصل از مدل شبکهٔ عصبی مصنوعی با تحقیقات قبلی نظیر Sing et al. (2013; 2012), Kurt et al. (2008), Kumar and Goyal (2013, 2011) Russo et al. (2013)، و Carbajal et al. (2011) مطابقت دارد و مدل شبکهٔ عصبی مصنوعی عملکرد بهتری نسبت به مدل‌های رگرسیون خطی داراست. مقدار آماره‌های خطا نظیر RMSE و MAE رگرسیون مؤلفه‌های اصلی در مرحلهٔ آموزش به‌ترتیب در ایستگاه تجریش برابر ۱۱/۱۲ و ۵/۷ و در مرحلهٔ آزمون به‌ترتیب ۱۰/۷ و ۵/۵ و مقدار آماره‌های خطا نظیر RMSE و MAE در مرحلهٔ آموزش به‌ترتیب در ایستگاه قلهک برابر ۱۹/۳۸ و ۱۵/۳۸و در مرحلهٔ آزمون به‌ترتیب ۱۹/۵۵ و ۱۵/۵۵ است. مقایسهٔ این آماره‌ها در دو ایستگاه نشان می‌دهد که مدل PCR در ایستگاه تجریش نسبت به مدل PCR در ایستگاه قلهک عملکرد بهتری دارد.

آماره‌های ضریب همبستگی R هر دو ایستگاه در مرحلهٔ آموزش در ایستگاه تجریش برابر ۰/۹۴۲۵ و در مرحلهٔ آزمون ۰/۹۴۴۷ است. در ایستگاه قلهک به‌ترتیب در مرحلهٔ آموزش برابر ۰/۶۳۳۲ و در مرحلهٔ آزمون ۰/۶۲۹۸ است. معیار ضریب همبستگی نشان از این دارد که بین مقادیر برآوردشدهٔ مدل و محاسبه‌شدهٔ کیفیت هوا در مدل ایستگاه تجریش نسبت به مدل ایستگاه قلهک همبستگی بیشتری وجود دارد. در نتیجه، مدل PCR ایستگاه تجریش نسبت به مدل PCR ایستگاه قلهک عملکرد بهتری دارد.

مقدار آماره‌های خطا نظیر RMSE و MAE رگرسیون خطی چندگانه در مرحلهٔ آموزش به‌ترتیب در ایستگاه تجریش برابر ۱۰/۵ و ۵/۷ و در مرحلهٔ آزمون به‌ترتیب ۱۰/۴ و ۵/۷ و مقدار آماره‌های خطا نظیر RMSE و MAE در مرحلهٔ آموزش به‌ترتیب در ایستگاه قلهک برابر ۱۰/۸۶ و ۸/۷۴ در مرحلهٔ آزمون به‌ترتیب ۱۰/۷۷ و ۸/۶۹ است. مقایسهٔ این آماره‌ها در دو ایستگاه نشان می‌دهد که مدل MLR در ایستگاه تجریش نسبت به مدل MLR در ایستگاه قلهک عملکرد بهتری دارد.

آماره‌های ضریب همبستگی R هر دو ایستگاه در مرحلهٔ آموزش در ایستگاه تجریش برابر ۰/۹۴۹۸ و در مرحلهٔ آزمون ۰/۹۴۶۰ و در ایستگاه قلهک به‌ترتیب در مرحلهٔ آموزش برابر ۰/۹۰۰۷ و در مرحلهٔ آزمون ۰/۹۰۵۰ است. معیار ضریب همبستگی نشان از این دارد که بین مقادیر برآوردشدهٔ مدل و محاسبه‌شدهٔ کیفیت هوا در مدل ایستگاه تجریش نسبت به مدل ایستگاه قلهک همبستگی بیشتری وجود دارد. در نتیجه، مدل MLR ایستگاه تجریش نسبت به مدل PCR ایستگاه قلهک عملکرد بهتری دارد. در نهایت، با توجه به مطالب فوق مشخص است که مدل MLR ایستگاه تجریش با دوازده ورودی نسبت به مدل PCR ایستگاه تجریش و مدل PCR و MLR ایستگاه قلهک عملکرد بهتری دارد. همچنین، مشخص است که مدل PCR ایستگاه تجریش نسبت به مدل PCR ایستگاه قلهک در برآورد شاخص کیفیت هوا عملکرد بهتری دارد. نتایج حاصل از مدل رگرسیون خطی با تحقیقات قبلی نظیر Sing et al. (2013; 2012), Kumar & Goyal (2013, 2011)، Voukantsis et al. (2011)، نوری و همکاران (۱۳۸۷) و صدرموسوی و رحیمی (۱۳۸۸) مطابقت داشت و استفاده از تحلیل مؤلفه‌های اصلی باعث کاهش تعداد متغیرهای ورودی به مدل رگرسیون شد. همچنین، باعث حذف همبستگی بین متغیرهای ورودی به مدل و تفسیر آسان‌تر مدل رگرسیون خطی شد.

۵. نتیجه‌گیری

نتایج ارزیابی تمامی مدل‌های مورد استفاده در مدل‌سازی و برآورد شاخص کیفیت هوای ایستگاه تجریش و قلهک نشان داد که مدل شبکهٔ عصبی مصنوعی با الگوریتم آموزش لونبرگ مارکوآرت بهترین عملکرد را نسبت به دیگر مدل‌ها در هر دو ایستگاه از خود نشان داد. ضعیف‌ترین عملکرد مربوط به مدل رگرسیون مؤلفه‌های اصلی است. در این تحقیق، مطالعه در سطح دو ایستگاه پایش کیفیت هوا انجام شد. به‌دلیل اهمیت بسیار زیاد موضوع کیفیت هوا در شهرهای آلوده نظیر تهران پیشنهاد

می‌شود در تحقیقات مشابه آتی مطالعه در سطح بسیار گسترده‌تر انجام شود و برای درک بهتر موضوع کیفیت هوا در نقاط بسیار آلودهٔ سطح شهر نقشه‌های پهنه‌بندی کیفیت هوا از طریق روش‌های زمین‌آمار و مدل‌های برآوردگر کیفیت هوا نظیر شبکهٔ عصبی مصنوعی و درخت تصمیم در نرم‌افزار سامانهٔ اطلاعات جغرافیایی (GIS) تولید شود. دستاوردهای این تحقیق حاکی از آن است که مدل‌های مورد استفاده روش‌هایی مناسب برای ارزیابی کیفیت هوای ایستگاه‌های مورد مطالعه است و ابزاری برای پژوهشگران و برنامه‌ریزان شهری جهت آگاهی از کیفیت هوا و اتخاذ تدابیر کنترلی در کاهش و پیشگیری از آلودگی هوای شهری و اطلاع‌رسانی کیفیت و درجهٔ سلامت هوای تنفسی به عموم مردم در مناطق آلودهٔ شهری محسوب می‌شود.

تشکر و قدردانی

بدین‌وسیله از سازمان حفاظت محیط‌زیست به‌دلیل در اختیار گذاشتن داده‌های تحقیق تشکر و قدردانی می‌نماییم.

یادداشت‌ها

1. United States Environmental Protection Agency
2. Air Quality Index
3. Artificial Neural Network
4. Decision Trees
5. Principal Component Analysis
6. Fuzzy Inference System
7. Autoregressive
8. Multiple Liner Regression
9. Ensemble Learning
10. Back propagation
11. Kaiser-Meyer-Olkin
12. Varimax

منابع

امیدوار، ک. شفیعی، ش. تقی‌زاده، ز. ۱۳۹۳. «ارزیابی کارایی مدل درخت تصمیم در پیش‌بینی بارش ایستگاه سینوپتیک کرمانشاه». نشریهٔ تحقیقات کاربردی علوم جغرافیایی.شمارهٔ ۳۴، ۸۹-۱۱۰.

خزاعی، ا. آل‌شیخ، ع. کریمی، م. وحیدنیا، م. ۱۳۹۱. «پیش‌بینی و مدلسازی غلظت آلایندهٔ مونوکسیدکربن با تلفیق شبکهٔ عصبی-فازی تطبیقی و سیستم اطلاعات جغرافیایی». مجلهٔ سنجش از دور و سامانهٔ اطلاعات جغرافیایی در منابع طبیعی (کاربرد سنجش از دور و GIS در علوم منابع طبیعی). شمارهٔ ۳، ۲۱-۳۳.

صدرموسوی، م. رحیمی. ا. ۱۳۸۹. «مقایسهٔ نتایج شبکه‌های عصبی پرسپترون چندلایه با رگرسیون چندگانه در پیش‌بینی غلظت ازن در شهر تبریز». مجلهٔ پژوهش‌های جغرافیای طبیعی، شمارهٔ ۷۱، ۶۵-۷۲.

نوری، ر. اشرفی، خ. اژدرپور، ا. ۱۳۸۷. «مقایسهٔ کاربرد روش‌های شبکهٔ عصبی مصنوعی و رگرسیون خطی چندمتغیره بر اساس تحلیل مؤلفه‌های اصلی برای پیش‌بینی غلظت میانگین روزانه کربن مونوکسید: بررسی موردی شهر تهران». مجلهٔ فیزیک زمین و فضا، شمارهٔ ۳۴، ۱۳۵-۱۵۲.

Abdul-Wahab, S.A. Bakheit, C.S., Al-Alawi, S. M. 2005. "Principal component and multiple regression analysis in modelling of ground-level ozone and factors affecting its concentrations". Environmental Modelling & Software, 20(10):1263-1271.

Breiman, L. Friedman, J.H. Olshen, R. Stone, C. 1993. *Classification and regression trees*. New York: Chapman & Hall.

Çamdevýren, H. Demýr, N. Kanik, A. Keskýn, S. 2005. "Use of principal component scores in multiple linear regression models for prediction of Chlorophyll-a in reservoirs". Ecological Modelling. 181(4): 581-589.

Carbajal-Hernández, J.J. Sánchez-Fernández, L.P. Carrasco-Ochoa, J.A. Martínez-Trinidad, J.F. 2012. "Assessment and prediction of air quality using fuzzy logic and autoregressive models". Atmospheric Environment. 60:37-50.

Chatterjee, S. Hadi, A.S. 2015. *Regression analysis by example*: John Wiley & Sons.

Cheng, W.L., Chen, Y.S. Zhang, J. Lyons, T. Pai, J.L. Chang, S.H. 2007. "Comparison of the revised air quality index with the PSI and AQI indices". Science of the total environment. 382 (2):191-198.

Fauset,t L. 1994. *Fundamentals of neural networks: architectures, algorithms, and applications*: Prentice-Hall, Inc.

Jolliffe, I. 2002. *Principal component analysis*: Wiley Online Library.

Kumar, A. Goyal. P. 2013. "Forecasting of air quality index in Delhi using neural network based on principal component analysis". Pure and Applied Geophysics. 170 (4), 711-711.

Kumar, A. Goyal. P. 2011. "Forecasting of air quality index in Delhi using principal component regression technique". Atmospheric Pollution Research, 2:436-444.

Lee, C.C. Ballinger, T.J. Domino, N.A. 2012. "Utilizing map pattern classification and surface weather typing to relate climate to the Air Quality Index in Cleveland, Ohio". Atmospheric Environment. 63:50-59.

Li, L. Qian, J. Ou, C.Q. Zhou, Y.X. Guo, C. Guo, Y. 2014. "Spatial and temporal analysis of Air Pollution Index and its timescale-dependent relationship with meteorological factors in Guangzhou, China, 2001–2011". Environmental Pollution. 190:75-81.

Moustris, K. Nastos, P. Larissi, I. Paliatsos, A. 2012. "Application of multiple linear regression models and artificial neural networks on the surface ozone forecast in the greater Athens area, Greece". Advances in Meteorology.

Rawlings, J.O. Pantula, S.G. Dickey, D.A. 1998. *Applied regression analysis: a research tool*: Springer Science & Business Media.

Russo, A. Raischel, F. Lind, P.G. 2013. "Air quality prediction using optimal neural networks with stochastic variables". Atmospheric Environment .79:822-830.

Singh, K.P. Gupta, S. Rai, P. 2013. "Identifying pollution sources and predicting urban air quality using ensemble learning methods". Atmospheric Environment. 80:426-437.

Singh, K.P. Gupta, S. Kumar, A. Shukla, S.P. 2012. "Linear and nonlinear modeling approaches for urban air quality prediction". Science of the Total Environment ,426:244-255.

Sowlat, M.H. Gharibi, H. Yunesian, M. Tayefeh Mahmoudi, M. Lotfi, S. 2011. "A novel, fuzzy-based air quality index (FAQI) for air quality assessment". Atmospheric Environment. 45 (12):2050-2059.

Yadav, D. Naresh, V.S. 2010. "Artificial Neural Network based Hydro Electric Generation Modelling". International Journal of Applied Engineering Research. 1(3):343.

Zhang, Y. Bocquet, M. Mallet, V. Seigneur, C. Baklanov, A. 2012. "Real-time air quality forecasting, part I: History, techniques, and current status". Atmospheric Environment, 60:632-655.