

مقایسه کاربرد شبکه عصبی مصنوعی، درخت تصمیم، رگرسیون مؤلفه‌های اصلی و

رگرسیون خطی چندگانه در مدل‌سازی شاخص کیفیت هوای شهری

علیرضا احسان زاده^{۱*}، فرهاد نژاد کورکی^۲، علی طالبی^۳

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 پذیرش مقاله: ۱۳۹۵/۰۸/۸

تاریخ وصول مقاله: ۱۳۹۴/۰۹/۱۷

چکیده

شاخص کیفیت هوا ابزاری کلیدی برای آگاهی از کیفیت هوا، نحوه اثر آلودگی هوا بر سلامت و روش‌های محافظتی در برابر آلودگی هواست. هدف اصلی این تحقیق مدل‌سازی و برآورد شاخص کیفیت هوا از طریق شبکه 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 محاسبه شاخص کیفیت هوا و از داده‌های هواشناسی و آلودگی هوای ثبت‌شده در ایستگاه تجریش و قلپک شهر تهران در دوره زمانی ۱۳۸۵ تا ۱۳۹۰ به منظور توسعه مدل‌ها و به منظور ارزیابی عملکرد مدل‌های برآوردگر از شاخص‌های آماری خطا، همبستگی و صحت استفاده شد. نتایج تحقیق نشان داد که مدل شبکه عصبی در هر دو ایستگاه نسبت به سایر مدل‌ها از عملکرد بهتری برخوردار است، به نحوی که در ایستگاه قلپک $RMSE=0/006$ ، $MAE=0/004$ ، $IA=0/99$ و در ایستگاه تجریش $RMSE=0/004$ ، $MAE=0/002$ ، $IA=1$ بود. مدل درخت تصمیم بعد از مدل شبکه عصبی عملکرد مطلوبی از خود نشان داد و مدل رگرسیون خطی چندگانه بعد از مدل شبکه 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، شبکه عصبی مصنوعی، درخت تصمیم، رگرسیون مؤلفه‌های اصلی، مدل‌سازی.

۱. سرآغاز

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^۱ (US-EPA) توسعه داد. این شاخص غلظت آلاینده‌های اصلی هوا نظیر کربن مونوکسید (CO)، دی‌اکسید گوگرد (SO₂)، ذرات معلق کوچک‌تر از ده

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 در حال توسعه 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

روش انتخاب متغیرهای تصادفی به منظور پیش‌بینی کیفیت هوای شهری (Russo et al., 2013)، پیش‌بینی غلظت روزانه مونوکسید کربن با استفاده از روش رگرسیون خطی چندگانه براساس تحلیل مؤلفه‌های اصلی، همچنین استفاده از شبکه عصبی (نوری و همکاران، ۱۳۸۷) و پیش‌بینی و مدل‌سازی غلظت آلاینده مونوکسید کربن با تلفیق شبکه 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 زمینه 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، پیش 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 به موقع 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).

میکرون (PM₁₀)، ازن (O₃) و دی‌اکسید نیتروژن (NO₂) را به شاخص استاندارد آلودگی هوا تبدیل می‌کند. در سال ۱۹۹۹ شاخص PSI را US-EPA کامل‌تر کرد و شاخصی به‌نام شاخص کیفیت هوا^۱ (AQI) جایگزین آن شد (Sowlat et al., 2011).

امروزه، پیش‌بینی و برآورد مشخصه‌های کیفیت هوا در نواحی شهری به دلیل 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 به دنبال 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. دسته اول مدل‌های پیش‌ساخته، دسته دوم مدل‌های آماری و دسته سوم مدل‌های هوش محاسباتی است (Zhang et al., 2012). مدل‌های هوش محاسباتی شامل شبکه‌های عصبی مصنوعی^۲ (ANN) و درختان تصمیم^۳ (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 برده شده می‌توان به موارد زیر اشاره کرد: استفاده از روش تحلیل مؤلفه‌های اصلی^۴ و استفاده از شبکه عصبی مصنوعی در پیش‌بینی شاخص روزانه کیفیت هوا (Kumar and Goyal, 2013)، استفاده از سیستم استنتاج فازی^۵ و مدل اتورگرسیون^۶ به منظور ارزیابی و پیش‌بینی کیفیت هوا (Carbajal-Hernández et al., 2012)، کاربرد رگرسیون خطی چندگانه^۷ مبتنی بر تحلیل مؤلفه‌های اصلی برای پیش‌بینی کوتاه‌مدت شاخص کیفیت هوا (Kumar and Goyal, 2011)، کاربرد الگوریتم یادگیری جمعی^۸ برای شناسایی منابع آلاینده هوا و مقدار شاخص کیفیت هوا (Singh et al., 2013)، کاربرد شبکه عصبی مصنوعی و

ازن، دی اکسید گوگرد، مونوکسید نیتروژن، هیدروکربن های بدون متان (NMHC) و متان (CH₄). همچنین، شامل برخی پارامترهای هواشناسی است، نظیر سرعت باد (WS)، جهت باد (WD)، دمای هوا (T)، فشار (P) و رطوبت هوا (H). داده های موجود مربوط به دوره زمانی ۱۳۸۵ تا ۱۳۹۰ است.

۲.۲. شاخص کیفیت هوا

مدیریت پایش و نظارت بر کیفیت هوا در شهرهای بزرگ داده های خام اندازه گیری شده با دستگاه های سنجش غلظت آلاینده های هوا را به AQI تبدیل می کند. اطلاعات مورد نیاز مربوط به سطوح سلامتی انسان و آثار بهداشتی آلاینده های هوا را شاخص AQI در اختیار مردم قرار می دهد. AQI برای پنج آلاینده اصلی هوا - یعنی ذرات معلق، دی اکسید نیتروژن، ازن سطح زمین، مونوکسید کربن و دی اکسید گوگرد - محاسبه می شود (Zhang et al., 2012). به منظور محاسبه این شاخص از معادله (۱) استفاده می شود که غلظت آلاینده ها و نقاط شکست در محاسبه طبقه کیفیت هوا مطابق با دستورالعمل مرکز سلامت و محیط کار ایران است. US-EPA به منظور درک راحت تر مقدار شاخص کیفیت هوا و سطوح بهداشتی مختلف با آن، همچنین دستورالعمل های کنترلی مربوط با مقادیر مختلف AQI را به شش دسته طبقه بندی می کند و هر دسته را به سطوح مختلف سلامت انسان مربوط می سازد (جدول ۱).

$$Ip = \frac{I_{Hi} - I_{Lo}}{BP_{Hi} - BP_{Lo}} (C_p - BP_{Lo}) + I_{Lo} \quad (1)$$

I_p = شاخص کیفیت هوا برای آلاینده P ، C_p = غلظت اندازه گیری شده آلاینده P ، BP_{Hi} = نقطه شکستی و بزرگ تر یا مساوی C_p ، BP_{Lo} = نقطه شکست کوچک تر یا مساوی C_p ، I_{Hi} = مقدار AQI منطبق با BP_{Hi} ، I_{Lo} = مقدار AQI منطبق با BP_{Lo} .

لذا، هدف اصلی این تحقیق مقایسه کارایی استفاده از روش شبکه عصبی مصنوعی، درخت تصمیم، رگرسیون خطی چندگانه و رگرسیون مؤلفه های اصلی در مدل سازی و برآورد شاخص کیفیت هوای شهری است.

۲. مواد و روش ها

در این تحقیق، از داده های ساعتی غلظت آلاینده های هوا و پارامترهای هواشناسی مربوط به ایستگاه های تجریش و قلهک شهر تهران در برآورد و مدل سازی شاخص AQI استفاده شد. هدف نخست، استفاده از دستورالعمل US-EPA و مرکز سلامت و محیط کار ایران در محاسبه شاخص کیفیت هوا براساس غلظت های ساعتی مربوط به تک تک آلاینده ها است. در مرحله بعد، با استفاده از سری های زمانی مربوط به داده های هواشناسی، آلودگی هوا و میزان AQI محاسبه شد. به منظور ایجاد و توسعه مدل های برآوردگر و شبیه ساز کیفیت هوا با استفاده از روش های درخت تصمیم، رگرسیون مؤلفه های اصلی، رگرسیون خطی چندگانه و شبکه عصبی مصنوعی از نرم افزار MATLAB استفاده شده است. در اولین مرحله، غلظت تک تک آلاینده ها ورودی الگوریتم محاسبه AQI بود. خروجی این الگوریتم که شاخص کیفیت هوای مربوط به هر آلاینده و شاخص کلی کیفیت هواست، همراه با داده های هواشناسی برای توسعه مدل ها استفاده شد. هدف نهایی، شبیه سازی و برآورد شاخص کیفیت هوا در ایستگاه های مورد مطالعه در شهر تهران است. در آخر مقایسه ای بین مدل های به کار رفته در تحقیق صورت گرفته است و مدل دارای نتایج بهتر به منظور برآورد و شبیه سازی معرفی می شود.

۱.۲. داده های تحقیق

داده های مورد استفاده در تحقیق شامل غلظت های ساعتی مربوط به آلاینده های هواست، شامل مونوکسید کربن، ذرات معلق کوچک تر از ۱۰ میکرون، اکسیدهای نیتروژن،

جدول ۱. نقاط شکست در محاسبه AQI

نقاط شکست							AQI I _{Lo} -I _{Hi}
BP _{Lo} -BP _{Hi}							
O ₃ ⁽¹⁾ (ppm) ۸ ساعته	O ₃ (ppm) ۱ ساعته	PM _{2.5} (µg/m ³) ۲۴ ساعته	PM ₁₀ (µg/m ³) ۲۴ ساعته	CO (ppm) ۸ ساعته	SO ₂ (ppm) ۲۴ ساعته	NO ₂ (ppm) ۱ ساعته	
۰-۰/۰۵۹	-	۰/۰-۱۵/۴	۰-۵۴	۰/۰-۴/۴	۰/۰-۰/۰۳۴	۰-۰/۰۵۳	۰-۵۰
۰/۰۶۰-۰/۰۷۵	-	۱۵/۵-۳۵	۵۵-۱۵۴	۴/۵-۹/۴	۰/۰۳۵-۰/۱۴۴	۰/۰۵۴-۰/۱	۵۱-۱۰۰
۰/۰۷۶-۰/۰۹۵	۰/۱۲۵-۰/۱۶۴	۳۵/۱-۶۵/۴	۱۵۵-۲۵۴	۹/۵-۱۲/۴	۰/۱۴۵-۰/۲۲۴	۰/۱۰۱-۰/۲۶۰	۱۰۱-۱۵۰
۰/۰۹۶-۰/۱۱۵	۰/۱۶۵-۰/۲۰۴	۶۵/۵-۱۵۰/۴	۲۵۵-۳۵۴	۱۲/۵-۱۵/۴	۰/۲۲۵-۰/۳۰۴	۰/۳۶۱-۰/۶۴۰	۱۵۱-۲۰۰
۰/۱۱۶-۰/۳۷۴	۰/۲۰۵-۰/۴۰۴	۱۵۰/۵-۲۵۰/۴	۳۵۵-۴۲۴	۱۵/۵-۳۰/۴	۰/۳۰۵-۰/۶۰۴	۰/۶۵-۱/۲۴	۲۰۱-۳۰۰
(۲)	۰/۴۰۵-۰/۵۰۴	۲۵۰/۵-۳۵۰/۴	۴۲۵-۵۰۴	۳۰/۵-۴۰/۴	۰/۶۰۵-۰/۸۰۴	۱/۲۵-۱/۶۴	۳۰۱-۴۰۰
	۰/۵۰۵-۰/۶۰۴	۳۵۰/۵-۵۰۰/۴	۵۰۵-۶۰۴	۴۰/۵-۵۰/۴	۰/۸۰۵-۱/۰۰۴	۱/۲۶-۲/۰۴	۴۰۱-۵۰۰

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 AQI 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. شبکه عصبی مصنوعی

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

$$y_i = \sum_{j=1}^n w_{i,j} x_{i,j} + \beta_i \quad (2)$$

که در آن، $x_{i,j}$ سیگنال ورودی از زمین نورون (در لایه ورودی)، $w_{i,j}$ وزن اتصال نورون j به نورون i (در لایه 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، به طور پی‌درپی تغییر می‌کند تا خطا به کمترین مقدار خود برسد. پس از محاسبه 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، شبکه عصبی پرسپترون چندلایه به خوبی 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 به صورت پیوسته و بدون هر گونه بازخوردی به سمت خروجی انتقال می‌یابد (Zhang et al. 2012). فرایند آموزش در شبکه‌های عصبی در واقع به معنای روزآمد کردن اتصالات بین نورون‌هاست. تاکنون الگوریتم‌های گوناگونی برای آموزش شبکه عصبی عرضه شده است که معروف‌ترین آن‌ها الگوریتم و قاعده انتشار به عقب^۱ است (Moustris et al., 2012). برای اصلاح وزن‌ها در شبکه و برای به حداقل رساندن میانگین مربعات خطا که از معادله (۳) به دست می‌آید، از راه الگوریتم انتشار به عقب استفاده می‌شود (Zhang et al., 2012).

الف) محاسبه فاکتور KMO^{۱۱}

از آنجا که روش PCA مستلزم وجود و قبول فرضیه‌هایی درباره جامعه مورد مطالعه نیست، از روش‌های ناپارامتری است. لازم است امکان استفاده از این روش و نتایج به دست آمده از آن به وسیله عامل KMO مشخص شود. مقدار KMO بین صفر تا یک تغییر می‌کند. این عامل با استفاده از ضرایب همبستگی ساده و جزئی طبق معادله (۵) محاسبه می‌شود. در معادله (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 کوچک بودن ضرایب همبستگی جزئی و بیانگر دقت محاسبات مربوط، با استفاده از روش PCA است (Abdul-Wahab et al., 2005).

$$KMO = \frac{\sum_{i=1}^p \sum_{j=1}^p r_{ij}^2}{\sum_{i=1}^p \sum_{j=1}^p r_{ij}^2 + \sum_{i=1}^p \sum_{j=1}^p a_{ij}^2} \quad i \neq j \quad (5)$$

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

ب) استاندارد کردن متغیرهای ورودی

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

$$Z = \frac{X - \mu}{\sigma} \quad (6)$$

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

ج) محاسبه 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

$$MSE = \frac{1}{N} \sum_{n=1}^N (t_i - td_i)^2 \quad (3)$$

N تعداد نورون‌های لایه خروجی و متناسب با تعداد مشاهدات تابع هدف، t_i مقدار مشاهده شده برای آامین رکورد، و td_i مقدار خروجی شبکه برای آامین رکورد است (Yadav and Naresh, 2010). به منظور بررسی دقت شبکه آموزش یافته لازم است آزمون شبکه صورت گیرد. این کار با دادن زوج داده‌های مجموعه آزمون به شبکه و محاسبه میزان خطای شبکه صورت می‌گیرد. به طور کلی، ویژگی‌های شبکه عصبی مصنوعی، ساختار شبکه و روش آموزش شبکه، با نوع تابع محرک نورون‌ها مشخص می‌شود (Zhang et al., 2012). در این تحقیق شبکه عصبی پرسپترون سه لایه با الگوریتم‌های آموزش لوبزبرگ مارکوارت (Lm)، پس انتشار ارتجاعی (Rb) و شیب توأم مقیاس شده (Scg) با تابع محرک تانژانت هایپربولیک در لایه پنهان و تابع محرک خطی در لایه 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 به جای 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, \dots, X_n برای ایجاد n مؤلفه مستقل PC_1, PC_2, \dots, PC_n برقرار می‌شود. در این روش اطلاعات متغیرهای اصلی با کمترین تلفات در مؤلفه‌ها محفوظ می‌ماند. هر مؤلفه اصلی با دنباله زیر مشخص می‌شود.

$$PC_i = a_{i1}X_1 + a_{i2}X_2 + \dots + a_{in}X_n \quad (4)$$

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

ورودی و بقیه درایه‌های این ماتریس، کوواریانس بین متغیرهای ورودی است. چون برای تشکیل این ماتریس از داده‌های استاندارد شده استفاده شده است، به همین دلیل این ماتریس، معادل ماتریس همبستگی بین متغیرهای ورودی است.

(د) محاسبه مقادیر ویژه (λ) و بردارهای ویژه مربوط از ماتریس همبستگی

بدین منظور معادله (۷) حل می‌شود.

$$|R - \lambda I_p| = 0 \quad (7)$$

I_p ماتریس واحد با بعد $p \times p$ است. بنابراین، می‌توان p مقدار ویژه مرتب شده $p\lambda \geq \lambda_2 \geq \dots \geq \lambda_p$ را به دست آورد، به طوری 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 بیان شده 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 مؤلفه 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 به وجود 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 معادله (۸) نشان داد.

$$Y = X\beta + e \quad (8)$$

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

$$\beta = (X'X)^{-1}(X'Y) \quad (9)$$

که در آن X' ترانزاده ماتریس X است. برای محاسبه 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 گرد کردن 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 نشانه 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 قابل درک برای انسان، از روابط موجود در مجموعه داده‌ای است و برای وظایف دسته‌بندی و پیش‌بینی به کار 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 $(X, Y) = (X_1, X_2, X_3, \dots, X_k, Y)$ داده می‌شود. با استفاده از متغیرهای X_1, X_2, \dots, X_k سعی در درک، دسته‌بندی یا عام‌سازی متغیر وابسته Y داریم. انواع صفات در درخت

این شاخص به صورت ساعتی در محیط نرم افزار 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. تقسیم بندی دادهها به صورت تصادفی و بعد از حذف دادههای گمشده و پرت انجام شد.

۸.۲. ارزیابی اعتبار مدلها

برای ارزیابی عملکرد مدلها و مقایسه نتایج به دست آمده 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 (معادله های ۱۲-۱۸).

(۱۲)

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

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

$$FB = 1/N \sum_{i=1}^N P_i - O_i / (P_i + O_i) / 2 \quad (13)$$

$$RMSE = (1/N \sum_{i=1}^N (P_i - O_i)^2)^{0.5} \quad (14)$$

$$MAE = 1/N \sum_{i=1}^N |P_i - O_i| \quad (15)$$

$$MSE = 1/N \sum_{i=1}^N (P_i - O_i)^2 \quad (16)$$

(۱۷)

$$R = \frac{\sum_{i=1}^N (P_i - \bar{O})^2 - \sum_{i=1}^N (P_i - O_i)^2}{\sum_{i=1}^N (P_i - \bar{O})^2}$$

$$R^2 = (R)^2 \quad (18)$$

N تعداد کل دادهها، P_i مقادیر پیش بینی شده، O_i مقادیر مشاهده یا محاسبه شده، و \bar{O} میانگین مقادیر مشاهده یا محاسبه است.

تصمیم به دو نوع صفات دسته ای و صفات حقیقی است. صفات دسته ای، صفاتی است که دو یا چند مقدار گسسته می پذیرد (یا صفات سمبلیک)، در حالی که صفات حقیقی مقادیر خود را از مجموعه اعداد حقیقی می گیرد (Breiman et al., 1993). درخت $CART^{13}$ نامی است که به هر دو روال بالا اطلاق می شود. نام CART سرنام کلمات درختان رگرسیون و دسته بندی است. در این تحقیق از الگوریتم درخت تصمیم CART برای مدل سازی و برآورد شاخص کیفیت هوا استفاده شده است. ساختار این درخت بر سه اصل استوار است:

۱. مجموعه ای از سؤالها به شکل $x \leq d?$ که در آن x متغیر مستقل و d مقدار ثابت و جواب هر سؤال بله/خیر است.

۲. بهترین معیار شاخه زدن در انتخاب بهترین متغیر مستقل برای ایجاد شاخه جدید. در این تحقیق از روش انحراف حداقل مربعات^{۱۴} برای ایجاد درخت رگرسیونی استفاده شد (معادله ۱۰).

$$SS(t) = \sum_{i=1}^{Ntt} (y_i(t) - \check{y}(t))^2 \quad (10)$$

Ntt تعداد رکوردها (دادهها) در گره برگ t، $Y_i(t)$ مقدار خروجی (متغیر هدف در گره برگ)، و $\check{y}(t)$ میانگین مقادیر متغیر هدف برای همه گرههاست.

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

$$Q(s,t) = SS(t) - SS(tR) - SS(tL) \quad (11)$$

که در آن $SS(tR)$ و $SS(tL)$ به ترتیب میزان $SS(t)$ در شاخه 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 به عنوان ورودی مدلها در مراحل آموزش و آزمون استفاده شد. همچنین، خروجی تمامی مدلها مقدار شاخص کیفیت هوا در نظر گرفته شد.

۳. نتایج و یافته‌ها

شهری در دو ایستگاه پایش کیفیت هوای قلهک و تجریش ارائه شده است.

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

جدول ۲. مشخصات آماری هر یک از پارامترهای ایستگاه قلهک

نام پارامتر	WS	WD	Temp	Press	HUM	THC	CH ₄	NO	O ₃	CO	SO ₂	NO ₂	PM ₁₀	واحد
حداقل	۰/۰۰۳	۰/۰	-۷/۶۲۹	۴۰۳/۹	۰/۳۷	۰/۰۰۱	۰/۰۰۱	۰/۰۰۱	۰/۰۰۱	۰/۰۱۸	۰/۰۰	۰/۰۰۱	۰/۰۰	ug/m3
حداکثر	۴/۵	۳۲۵/۷	۳۹/۵۱	۹۵۷/۶	۹۵/۳۴	۱۷/۳۳	۴/۷۹۰	۰/۳۸۶	۰/۰۵۰	۹/۰۹۲	۰/۰۸	۰/۴۲۸	۳۱۶/۸	
میانگین	۰/۹۰	۱۱۶/۳	۱۸/۲۸	۸۷۳/۹	۴۴/۳۳	۳/۵۳	۲/۱۱۲	۰/۰۸۸	۰/۰۱۵	۳/۵۵	۰/۰۳۸	۰/۱۲۰	۸۹/۶۹	
چولگی	۰/۹۷	۰/۰۶	-۰/۲۴	-۲/۲۴	۰/۳۲	۰/۹۷	۰/۰۴۴	۱/۰۲	۱/۲۱	۰/۵۶۲	۰/۰۵	۱/۲۴	۱/۳۹	
کشیدگی	۱/۱۳	-۱/۳۱	-۰/۷۸	۲۱/۷	-۰/۶۳	۲/۹	-۰/۱۹	۰/۷۹	۰/۸۸	۰/۲۰	۰/۲۳	۱/۶۵	۲/۰۵	

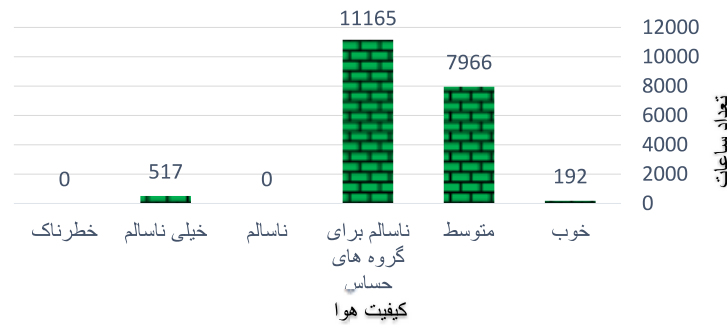
جدول ۳. مشخصات آماری هر یک از پارامترهای ایستگاه تجریش

نام پارامتر	WS	WD	Temp	Press	HUM	NMHC	THC	NO _x	CO	SO ₂	NO ₂	PM ₁₀	واحد
حداقل	۰/۰۰۱	۰/۰	-۹/۹۷	۴۱۲/۳	۲/۷۶	۰/۰۱۵	۰/۰۱۷	۰/۰۰۱	۰/۰۰۱	۰/۰۰۱	۰/۰	۰/۰	ug/m3
حداکثر	۱۷/۰۹	۳۵۵	۶۱/۲۱	۸۴۷/۸	۹۹/۸۵	۸/۹۲	۱۴/۹۴	۰/۲۹	۱۳/۴۴	۰/۰۵۴	۰/۲۴	۵۱۵	
میانگین	۱/۰۵	۲۱۳/۶۵	۱۹/۲	۸۳۳	۳۴/۷	۲/۰۱	۵/۴	۰/۰۶	۴/۰۱	۰/۰۱۷	۰/۰۴	۱۱۷/۵	
چولگی	۲/۶۵	-۰/۲۲	۰/۳۴	-۱۶/۷	۱/۲۶	۰/۹۷	۰/۲۳	۱/۱۷	۰/۸۵	۰/۶۴	۱/۳۶	۱/۴	
کشیدگی	۳۹/۳	-۰/۱۱	۰/۶۶	۲۸۵/۵۸	۰/۹۶	۱/۶۵	۰/۰۴	۱/۴۴	۰/۳	۰/۷۶	۳/۳۶	۴	

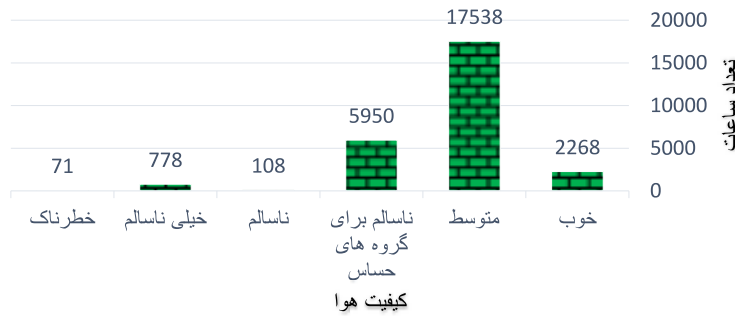
۱.۳. برآورد شاخص کیفیت هوا

اصلی آن حجم ترافیک بیشتر خودروها، جمعیت بیشتر و احتراق سوخت در موتور خودروها در محدوده این ایستگاه است. بیشترین طبقه کیفیت هوا در ایستگاه تجریش مربوط به طبقه متوسط با تعداد ساعات ۱۷۵۳۸ و آلاینده مسئول ایجاد کیفیت نامطلوب در اکثر اوقات ذرات معلق کوچک‌تر از ۱۰ میکرون است.

نتایج تحلیل حاصل از محاسبه شاخص کیفیت هوا در ایستگاه قلهک در شکل ۱ و ایستگاه تجریش در شکل ۲ نشان داده شده است. بیشترین طبقه کیفیت هوا در ایستگاه قلهک مربوط به طبقه ناسالم برای گروه‌های حساس جامعه با تعداد ساعات ۱۱۱۶۵ و آلاینده مسئول ایجاد کیفیت نامطلوب در اکثر اوقات دی‌اکسید نیتروژن است. عامل



شکل ۱. مقدار شاخص و طبقه کیفیت هوای ایستگاه قلهدک



شکل ۲. مقدار شاخص و طبقه کیفیت هوای ایستگاه تجریش

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 انجام شده روی درخت به مقدار $0/03$ در تعیین بهترین ساختار درخت مشخص شد. نتایج به دست آمده حاکی از این است که الگوریتم CART در شبیه سازی شاخص کیفیت هوا عملکرد بسیار بالایی دارد، به طوری که همبستگی بین مقادیر شبیه سازی شده و مشاهده شده بسیار نزدیک به عدد ۱ است. مقادیر خطای درخت در سه مرحله آزمون و آموزش بسیار ناچیز و مشابه است. مهم ترین پارامترهای تأثیرگذار در برآورد AQI در ایستگاه قلهدک PM_{10} با $18/5$ درصد تأثیر و NO_2 با $32/2$ درصد تأثیر است. مهم ترین عوامل در برآورد AQI در ایستگاه تجریش ذرات معلق کوچک تر از 10 میکرون با 47 درصد، دی اکسید نیتروژن با $2/1$ و مونوکسید کربن با $3/1$ درصد تأثیر است.

۲.۳. مدل سازی شاخص کیفیت هوا

نتایج مدل سازی نشان داد که کارایی روش های به کاررفته در تحقیق در برآورد شاخص کیفیت هوای شهری عملکرد متفاوتی دارد. تعداد نمونه های ایستگاه قلهدک در مرحله آموزش مدل ها 15872 و در مرحله آزمون 3968 و شامل سیزده پارامتر ورودی و یک پارامتر خروجی است. تعداد نمونه های ایستگاه تجریش در مرحله آموزش 21370 و در مرحله آزمون 5343 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 مدل سازی و برآورد شاخص کیفیت هوای ایستگاه قلهدک و تجریش شهر تهران ارائه شده است. در جدول ۴ مقادیر شاخص های آماری حاصل از سه مرحله آموزش، آزمون

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
۱	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
۲	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 ایده آل هر شاخص را نشان می دهد

۴.۳. مدل سازی با شبکه عصبی مصنوعی

تعداد سیزده پارامتر ورودی در مدل شبکه عصبی در ایستگاه قلهک ۵۴۰۰۰ و در ایستگاه تجریش با توجه به تعداد دوازده پارامتر ورودی در مدل شبکه عصبی ۵۰۰۰۰ دور در نظر گرفته شد. در جدول ۵ مقادیر شاخص های آماری حاصل از دو مرحله آموزش و آزمون شبکه عصبی با یک لایه پنهان و با الگوریتم های آموزش مختلف و در بهترین حالت ارائه شده است.

پس از ایجاد مدل شبکه عصبی مصنوعی با داده های آموزش برای اعتبارسنجی مدل از داده های آزمون استفاده شد. همچنین، تعداد نورون لایه پنهان از طریق رابطه $(2n+1)$ (n: تعداد ورودی) و بهترین تعداد نورون از طریق کمترین خطا و بیشترین همبستگی بین متغیر هدف و مقادیر برآورد شده تعیین شد. تعداد دور بذای آموزش شبکه ۲۰۰۰ در نظر گرفته شد و در مجموع با توجه به

جدول ۵. نتایج حاصل از مراحل آموزش و آزمون شبکه عصبی در بهترین حالت

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

با روش سعی و خطا مشخص شد که شبکه عصبی مصنوعی پرسپترون با یک لایه پنهان و الگوریتم آموزش لونیگ مارکوآرت با بیست نورون در لایه پنهان در ایستگاه قلهک و ۲۵ نورون در ایستگاه تجریش بهترین عملکرد را برای برآورد و مدل سازی شاخص کیفیت هوا به دست می دهد. همچنین، یک شبکه عصبی پرسپترون با ۲۵ نورون در لایه پنهان در ایستگاه قلهک و ۲۴ نورون در ایستگاه تجریش و الگوریتم شیب توأم مقیاس شده نسبت به شبکه عصبی پرسپترون با الگوریتم آموزش پس انتشار ارتجاعی عملکرد بهتری داشت.

۳/۱۳۴ است که ۲۴/۱۰۹ درصد از کل واریانس موجود در سری داده ها را توجیه می کند. با توجه به جدول ۸، مقادیر ویژه چهار مؤلفه بعدی در ایستگاه قلهک نیز بزرگ تر از ۱ و به صورت جمعی با مؤلفه اول دارای واریانس جمعی ۷۲/۱۳۶ درصد کل داده هاست. همچنین، مقدار اولین مؤلفه برابر ۳/۴۵۹ است که ۲۸/۸۲۷ درصد از کل واریانس موجود در سری داده ها را توجیه می کند. مقادیر ویژه سه مؤلفه بعدی در ایستگاه تجریش نیز بزرگ تر از ۱ و به صورت جمعی با مؤلفه اول دارای واریانس جمعی ۶۵/۷۱۴ درصد کل داده هاست. بدین ترتیب، مؤلفه اول تا پنجم در ایستگاه قلهک و مؤلفه اول تا چهارم در ایستگاه تجریش مقادیر ویژه بزرگ تر از ۱ دارد. ضرایب بردارهای ویژه ایستگاه قلهک و ایستگاه تجریش در جدول ۹ آمده است. در جدول ۹ مقادیر بردارهای ویژه و ایجادکننده ضرایب هر یک از پارامترهای تشکیل دهنده هر مؤلفه بیانگر آن است که در مؤلفه اول ضرایب مربوط به متغیرهای CO، SO₂ و NO₂ مقادیر بیشتر و تأثیر بسزایی در تشکیل این مؤلفه دارد و با قلم پرننگ مشخص شده است. در مؤلفه دوم متغیرهای رطوبت و دمای هوا عوامل اصلی ایجاد این مؤلفه است. در مؤلفه سوم مقدار PM₁₀، SO₂ و فشار هوا عوامل اصلی ایجاد این مؤلفه است. همچنین، در تشکیل مؤلفه چهارم متغیرهای جهت باد، سرعت باد و O₃ و در مؤلفه پنجم متان و کل هیدروکربن ها عوامل اصلی تشکیل این مؤلفه است. در ایستگاه تجریش نیز مشاهده می شود که در مؤلفه اول ضرایب مربوط به متغیرهای PM₁₀، NO₂، SO₂، CO و NOx مقادیر بیشتری و تأثیر بسزایی در تشکیل این مؤلفه دارد و با قلم پرننگ مشخص شده است. در مؤلفه دوم متغیرهای کل هیدروکربن ها و هیدروکربن های بدون متان عوامل اصلی ایجاد این مؤلفه است. در مؤلفه سوم مقدار رطوبت، دمای هوا و سرعت باد عوامل اصلی ایجاد این مؤلفه است. همچنین، در تشکیل مؤلفه چهارم متغیرهای جهت باد و فشار هوا عوامل اصلی تشکیل این مؤلفه است.

۵.۳. مدل سازی با رگرسیون مؤلفه های اصلی و رگرسیون چندگانه

۱.۵.۳. اجرای روش PCA در پیش پردازش داده های ورودی

بررسی اولیه نشان داد که بین متغیرهای ورودی مورد استفاده در ایستگاه قلهک و تجریش همبستگی معناداری وجود دارد. برای از بین بردن این مشکل، از روش تحلیل مؤلفه های اصلی استفاده شد. برای بررسی امکان اجرای تحلیل مؤلفه های اصلی از آزمون KMO استفاده شد. مقدار $KMO = 0/581$ در ایستگاه قلهک و $KMO = 0/606$ در ایستگاه تجریش امکان اجرای روش PCA را تأیید کرد. برای اجرای این روش، پس از استاندارد کردن متغیرهای ورودی ماتریس همبستگی تشکیل شد (جدول ۶ و ۷). با حل دستگاه معادله (۷)، سیزده مقدار ویژه و به ازای هر مقدار ویژه سیزده بردار ویژه برای ایستگاه قلهک و دوازده مقدار ویژه و به ازای هر مقدار ویژه دوازده بردار ویژه برای ایستگاه تجریش حاصل شد که با استفاده از آن ها، مؤلفه های اصلی از متغیرهای اولیه به دست آمد. مقدار عددی هر مؤلفه با تقسیم مقادیر ویژه به دست آمده بر تعداد متغیرهای مورد استفاده به دست 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

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

												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	-۰/۰۱	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	
												WS	۱	-۰/۰۶	-۰/۱۵	-۰/۰۴	-۰/۰۱	-۰/۰۵	-۰/۱۹	-۰/۱۵	-۰/۱۶	-۰/۱	-۰/۲۴	-۰/۳۱	WD
												۱	-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
درصد	واریانس	تجمعی	درصد	واریانس	تجمعی	
۳/۱۳۴	۲۴/۱۰۹	۲۴/۱۰۹	۳/۴۵۹	۲۸/۸۲۷	۲۸/۸۲۷	۱
۲/۰۵۶	۱۵/۸۱۴	۳۹/۹۲۳	۱/۷۲۵	۱۴/۳۷۶	۴۳/۲۰۳	۲
۱/۶۶۷	۱۲/۸۲۱	۵۲/۷۴۴	۱/۵۴۳	۱۲/۸۵۵	۵۶/۰۵۸	۳
۱/۵۰۱	۱۱/۵۴۳	۶۴/۲۸۸	۱/۱۵۹	۹/۶۵۶	۶۵/۷۱۴	۴
۱/۰۲	۷/۸۴۸	۷۲/۱۳۶	۰/۹۸۶	۸/۲۱۴	۷۳/۹۲۸	۵
۰/۹۱۹	۷/۰۷۳	۷۹/۲۰۸	۰/۸۷۹	۷/۳۲۶	۸۱/۲۵۴	۶
۰/۶۵۵	۵/۰۴۱	۸۴/۲۵	۰/۷۳۲	۶/۰۹۷	۸۷/۳۵۲	۷
۰/۵۶۲	۴/۳۲۲	۸۸/۵۷۲	۰/۶۱۳	۵/۱۱	۹۲/۴۶۱	۸
۰/۵۱۶	۳/۹۶۹	۹۲/۵۴۱	۰/۴۰۲	۳/۳۴۹	۹۵/۸۱	۹
۰/۴۵۵	۳/۴۹۶	۹۶/۰۳۸	۰/۳۲۸	۲/۷۳۱	۹۸/۵۴۱	۱۰
۰/۲۴۷	۱/۹۰۱	۹۷/۹۳۹	۰/۱۰۴	۰/۸۶۴	۹۹/۴۰۵	۱۱
۰/۲۰۷	۱/۵۹۶	۹۹/۵۳۵	۰/۰۷۱	۰/۵۹۵	۱۰۰	۱۲
۰/۰۶	۰/۴۶۵	۱۰۰				۱۳

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	NO _x	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	-۰/۰۸۶	-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 رگرسيون مبتنى بر مؤلفه‌هاى اصلى و

رگرسيون چندگانه

به منظور توسعه مدل رگرسيون مؤلفه‌هاى اصلى، پنج مؤلفه ايجادشده ايستگاه قلحك و چهار مؤلفه ايجادشده ايستگاه تجریش ورودى مدل رگرسيون استفاده شد. از آزمون دوربين-واتسون به منظور بررسى پيش فرض نرمال بودن توزيع خطاى باقىمانده‌هاى مدل رگرسيون استفاده شد. در صورتى كه مقدار آين آماره بين عدد ۱/۵ تا ۲/۵ باشد، نشان دهنده نرمال بودن توزيع خطاست. در مدل رگرسيون ايستگاه قلحك مقدار آين آماره ۲/۰۰۷۹ و در مدل ايستگاه تجریش ۱/۹۹ به دست آمد كه نشان از نرمال بودن توزيع خطا داشت. در بررسى پيش فرض عدم هم خطى (استقلال خطاها) از آماره تورم واريانس استفاده شد كه مقدار ایده آل آن ۱ و مقادير بيشتر از ۱۰ نشان دهنده مشكل ساز بودن هم خطى بين متغيرهاى مستقل در مدل رگرسيون است. با استفاده از روش PCA مشكل همبستگى بين متغيرهاى مستقل نيز رفع شد. پس از تأييد نرمال بودن توزيع مقادير خطا و رفع مشكل همبستگى در متغيرهاى مستقل، مدلى مناسب با استفاده از روش رگرسيون خطى چندگانه براى

برآورد مقدار شاخص كيفيت هوا بسط يافت. نتايج ورود مؤلفه‌هاى اصلى به مدل رگرسيون ايستگاه قلحك در جدول ۱۰ و ايستگاه تجریش در جدول ۱۱ آمده است. پنج مؤلفه ايجادشده در ايستگاه قلحك و چهار مؤلفه ايجادشده تجریش ورودى مدل رگرسيون استفاده شد. نتايج ارزيابى مدل رگرسيون مؤلفه‌هاى اصلى در جدول ۱۲ آمده است. در جدول ۱۴ نتايج ورود تمامى متغيرهاى هواشناسى و آلودگى هوا به مدل رگرسيون ارائه شده است. در جدول ۱۳ نتايج ارزيابى مدل رگرسيون چندگانه آمده است. نتايج ارزيابى مدل‌ها نشان مى دهد كه مدل رگرسيونى ايجادشده توسط تمامى پارامترهاى هواشناسى و آلودگى هوا نسبت به مدل ايجادشده توسط مؤلفه‌هاى اصلى در برآورد مقدار شاخص كيفيت هوا عملکرد بهتري دارد. در معادله (۱۹) مدل رگرسيون مؤلفه‌هاى اصلى برآوردگر AQI و در معادله (۲۰) مدل رگرسيون چندگانه ايستگاه قلحك ارائه شده است. همچنين، در معادله (۲۱) مدل رگرسيون مؤلفه‌هاى اصلى برآوردگر AQI و در معادله (۲۲) مدل رگرسيون چندگانه ايستگاه تجریش ارائه شده است.

جدول ۱۰. نتايج ورود مؤلفه‌هاى اصلى به مدل رگرسيون چندگانه ايستگاه قلحك

مؤلفه‌ها	ضرايب	فاكتور تورم واريانس	سطح معنادارى
(Constant)	-۶۳/۷۴	—	•
PC1	۹/۸۹	۱/۱۰	•
PC2	۰/۲۰	۱/۰۴	•
PC3	۰/۱۹	۱/۱۵	•
PC4	-۰/۰۹۴	۱/۱۲	•
PC5	-۱/۰۹	۱/۰۹	•

جدول ۱۱. نتايج ورود مؤلفه‌هاى اصلى به مدل رگرسيون چندگانه ايستگاه تجریش

مؤلفه‌ها	ضرايب	فاكتور تورم واريانس	سطح معنادارى
(Constant)	۲۸/۲۳	—	•
PC1	۰/۹۳۳	۱/۰۴	•
PC2	۰/۲۴۱۵	۱/۰۵	•
PC3	۰/۰۳۳۶	۱/۰۲	•
PC4	-۰/۰۰۸۸	۱/۰۱	•

جدول ۱۲. نتایج ارزیابی مدل رگرسیون مؤلفه‌های اصلی

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

جدول ۱۳. نتایج ارزیابی مدل رگرسیون چندگانه

ایستگاه	IA (1)	FB (0)	R ² (1)	R (1)	MAE (0)	RMSE (0)	MSE (0)	MLR مدل
قلهک	۱	۰	۰/۸۱۱۲	۰/۹۰۰۷	۸/۷۴	۱۰/۸۶	۱۱۷/۹۷	مرحله آموزش
	۱	-/۰۰۱	۰/۸۱۹۰	۰/۹۰۵۰	۸/۶۹	۱۰/۷۷	۱۱۶/۱۷	مرحله آزمون
تجریش	۱	۰	۰/۹۰۲۱	۰/۹۴۹۸	۵/۷	۱۰/۵	۱۱۱/۲	مرحله آموزش
	۱	۰/۰۰۲	۰/۸۹۵۰	۰/۹۴۶۰	۵/۷	۱۰/۴	۱۰۹/۱	مرحله 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	NO _x	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				

$$AQI = -63/74 + 9/89 \times PC_1 + 1/2 \times PC_2 + 1/19 \times PC_3 - 1/094 \times PC_4 - 1/09 \times PC_5 \quad (19)$$

$$AQI = 23/47 + 224 \times PM_{10} + 257/3 \times NO_2 + 43/29 \times SO_2 + 178 \times CO + 5/34 \times O_3 + \quad (20)$$

$$28/99 \times NO - 368 \times CH_4 - 625 \times THC - 0/529 \times HUM + 0/344 \times Press - 0/87 \times Temp - 0/33 \times WD - 0/21 \times WS$$

$$AQI = 28/23 + 9/33 \times PC_1 + 2/415 \times PC_2 + 0/336 \times PC_3 - 0/088 \times PC_4 \quad (21)$$

$$AQI = 26/65 + 48 \times PM_{10} + 111/92 \times NO_2 - 160/5 \times SO_2 + 1/66 \times CO - 38/30 \times NO_x + \quad (22)$$

$$0/21 \times THC - 2 \times NMHC + 0/47 \times HUM - 0/008 \times Press - 0/23 \times Temp - 0/02 \times WD + 0/9 \times WS$$

۴. بحث

Carbajal et al. (2011) و Sing et al. (2013) مطابقت

دارد و نتایج بهتری نسبت به مدل‌های رگرسیون خطی حاصل شده است. مقدار آماره‌های خطا نظیر RMSE و MAE شبکه عصبی مصنوعی پرسپترون در بهینه‌ترین حالت در مرحله آموزش به ترتیب در ایستگاه تجریش برابر ۰/۰۰۴۴ و ۰/۰۰۲۶ و در مرحله آزمون به ترتیب ۰/۰۰۴۸ و ۰/۰۰۲۸ و مقدار آماره‌های خطا نظیر RMSE و MAE در مرحله آموزش به ترتیب در ایستگاه قلهک برابر ۰/۰۰۶۸ و ۰/۰۰۴۳ و در مرحله آزمون به ترتیب ۰/۰۰۸۲ و ۰/۰۰۴۵ است. مقایسه این آماره‌ها در دو ایستگاه نشان می‌دهد که مدل شبکه عصبی مصنوعی در ایستگاه تجریش نسبت به مدل شبکه عصبی مصنوعی در ایستگاه قلهک عملکرد بهتری دارد.

آماره‌های ضریب همبستگی R و ضریب تبیین R² در هر دو ایستگاه در مرحله آموزش برای ایستگاه تجریش برابر ۰/۹۹۹۶ و ۰/۹۹۹۳ و در مرحله آزمون ۰/۹۹۹۵ و ۰/۹۹۹۱ و در ایستگاه قلهک به ترتیب در مرحله آموزش برابر ۰/۹۹۹۷ و ۰/۹۹۹۴ و در مرحله آزمون ۰/۹۹۹۲ است. معیارهای همبستگی نشان از این دارد که بین مقادیر برآورد شده مدل و محاسبه شده کیفیت هوا در مدل شبکه عصبی در ایستگاه قلهک نسبت به مدل شبکه عصبی در ایستگاه تجریش همبستگی بیشتری است و این اختلاف بسیار ناچیز و ۰/۰۰۰۱ است.

آماره‌های ضریب همبستگی R و ضریب تبیین R² در هر دو مدل نزدیک به عدد ۱ است که نشان از توانایی

مقدار آماره‌های خطا نظیر RMSE و MAE مدل رگرسیون درخت تصمیم در بهینه‌ترین حالت در مرحله آموزش به ترتیب در ایستگاه تجریش برابر ۰/۴۴۲۸ و ۰/۱۱۵۷ و در مرحله آزمون به ترتیب ۰/۶۸۸۸ و ۰/۱۹۵۱ و مقدار آماره‌های خطا نظیر RMSE و MAE در مرحله آموزش به ترتیب در ایستگاه قلهک برابر ۰/۲۵۹۳ و ۰/۰۴۰۸ و در مرحله آزمون به ترتیب ۰/۷۵ و ۰/۱۰۱۲ است. مقایسه 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 R² در هر دو ایستگاه در مرحله آموزش برای ایستگاه تجریش برابر ۰/۹۹۹۹ و ۰/۹۹۹۸ و در مرحله آزمون ۰/۹۹۹۸ و ۰/۹۹۹۶ و در ایستگاه قلهک به ترتیب در مرحله آموزش برابر ۰/۹۹۹۹ و ۰/۹۹۹۹ و در مرحله آزمون ۰/۹۹۹۶ و ۰/۹۹۹۱ است. معیارهای همبستگی نشان از این دارد که بین مقادیر برآورد شده مدل و محاسبه شده کیفیت هوا در مدل درخت تصمیم در هر دو ایستگاه اختلاف بسیار ناچیز است. آماره‌های ضریب همبستگی R و ضریب تبیین R² 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)،

آماره‌های ضریب همبستگی R هر دو ایستگاه در مرحله آموزش در ایستگاه تجریش برابر ۰/۹۴۹۸ و در مرحله آزمون ۰/۹۴۶۰ و در ایستگاه قلهک به ترتیب در مرحله آموزش برابر ۰/۹۰۰۷ و در مرحله آزمون ۰/۹۰۵۰ است. معیار ضریب همبستگی نشان از این دارد که بین مقادیر برآوردشده مدل و محاسبه شده 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 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; Kumar & Goyal (2013, 2011), Voukantsis et al. (2011), نوری و همکاران (2012)، 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. نتیجه گیری

نتایج ارزیابی تمامی مدل‌های مورد استفاده در مدل‌سازی و برآورد شاخص کیفیت هوای ایستگاه تجریش و قلهک نشان داد که مدل شبکه عصبی مصنوعی با الگوریتم آموزش لوبنبرگ 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. به دلیل اهمیت بسیار زیاد موضوع کیفیت هوا در شهرهای آلوده نظیر تهران پیشنهاد

بالای مدل شبکه عصبی مصنوعی در برآورد شاخص کیفیت هوای شهری در دو ایستگاه مورد مطالعه دارد. نتایج حاصل از مدل شبکه عصبی مصنوعی با تحقیقات قبلی نظیر Sing et al. (2013; 2012), Kurt et al. (2008), Russo et al. (2013), Kumar and Goyal (2013, 2011) و Carbajal et al. (2011) مطابقت دارد و مدل شبکه عصبی مصنوعی عملکرد بهتری نسبت به مدل‌های رگرسیون خطی داراست. مقدار آماره‌های خطا نظیر RMSE و MAE رگرسیون مؤلفه‌های اصلی در مرحله آموزش به ترتیب در ایستگاه تجریش برابر ۱۱/۱۲ و ۵/۷ و در مرحله آزمون به ترتیب ۱۰/۷ و ۵/۵ و مقدار آماره‌های خطا نظیر RMSE و MAE در مرحله آموزش به ترتیب در ایستگاه قلهک برابر ۱۹/۳۸ و ۱۵/۳۸ و در مرحله آزمون به ترتیب ۱۹/۵۵ و ۱۵/۵۵ است. مقایسه 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 مرحله آموزش در ایستگاه تجریش برابر ۰/۹۴۲۵ و در مرحله آزمون ۰/۹۴۴۷ است. در ایستگاه قلهک به ترتیب در مرحله آموزش برابر ۰/۶۳۳۲ و در مرحله آزمون ۰/۶۲۹۸ است. معیار ضریب همبستگی نشان از این دارد که بین مقادیر برآوردشده مدل و محاسبه شده 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 مرحله آموزش به ترتیب در ایستگاه تجریش برابر ۱۰/۵ و ۵/۷ و در مرحله آزمون به ترتیب ۱۰/۴ و ۵/۷ و مقدار آماره‌های خطا نظیر RMSE و MAE در مرحله آموزش به ترتیب در ایستگاه قلهک برابر ۱۰/۸۶ و ۸/۷۴ و در مرحله آزمون به ترتیب ۱۰/۷۷ و ۸/۶۹ است. مقایسه این آماره‌ها در دو ایستگاه نشان می‌دهد که مدل MLR در ایستگاه تجریش نسبت به مدل MLR در ایستگاه قلهک عملکرد بهتری دارد.

تشکر و قدردانی

بدین وسیله از سازمان حفاظت محیط‌زیست به دلیل در اختیار گذاشتن داده‌های تحقیق تشکر و قدردانی می‌نمایم.

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

می‌شود در تحقیقات مشابه آتی مطالعه در سطح بسیار گسترده‌تر انجام شود و برای درک بهتر موضوع کیفیت هوا در نقاط بسیار آلوده سطح شهر نقشه‌های پهنه‌بندی کیفیت هوا از طریق روش‌های زمین‌آمار و مدل‌های برآوردگر کیفیت هوا نظیر شبکه عصبی مصنوعی و درخت تصمیم در نرم‌افزار سامانه 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 درجه سلامت هوای تنفسی به عموم مردم در مناطق آلوده شهری محسوب می‌شود.

منابع

- امیدوار، ک. شفیعی، ش. تقی‌زاده، ز. ۱۳۹۳. «ارزیابی کارایی مدل درخت تصمیم در پیش‌بینی بارش ایستگاه سینوپتیک کرمانشاه». نشریه تحقیقات کاربردی علوم جغرافیایی. شماره ۳۴، ۸۹-۱۱۰.
- خزاعی، ا. آل‌شایخ، ع. کریمی، م. وحیدنیا، م. ۱۳۹۱. «پیش‌بینی و مدل‌سازی غلظت آلاینده مونوکسیدکربن با تلفیق شبکه عصبی-فازی تطبیقی و سیستم اطلاعات جغرافیایی». مجله سنجش از دور و سامانه اطلاعات جغرافیایی در منابع طبیعی (کاربرد سنجش از دور و GIS در علوم منابع طبیعی). شماره ۳، ۲۱-۳۳.
- صدرموسوی، م. رحیمی، ا. ۱۳۸۹. «مقایسه نتایج شبکه‌های عصبی پرسپترون چندلایه با رگرسیون چندگانه در پیش‌بینی غلظت ازن در شهر تبریز». مجله پژوهش‌های جغرافیای طبیعی، شماره ۷۱، ۶۵-۷۲.
- نوری، ر. اشرفی، خ. اژدرپور، ا. ۱۳۸۷. «مقایسه کاربرد روش‌های شبکه عصبی مصنوعی و رگرسیون خطی چندمتغیره بر اساس تحلیل مؤلفه‌های اصلی برای پیش‌بینی غلظت میانگین روزانه کربن مونوکسید: بررسی موردی شهر تهران». مجله فیزیک زمین و فضا، شماره 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.
- Čamdevýren, 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, 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.