Journal of Research in Environmental Health

Volume 9, Issue 1, Spring 2023

Prediction of daily $PM_{2.5}$ concentration using support vector training combination (SVM) - Adaptive and principal component analysis (PCA)

Amir Zarei
* Ph.D. Student, Faculty of Earth Sciences, Shahid Beheshti University, Tehran, Iran.
amir69zarey@yahoo.com
Sirvan Zarei
MSc, Department of Health, Safety and Environment (HSE), Workplace Health Promotion Research Center, School of Public Health and Safety, Shahid Beheshti University of Medical Sciences, Tehran, Iran.
Hossein Aghighi
PhD, Assistant Professor, Research Center of Remote Sensing and GIS, Shahid Beheshti University, Tehran, Iran.
Mohammad Hossein Vaziri
PhD, Assistant Professor, Department of Health, Safety and Environment (HSE), Workplace Health Promotion Research Center, School of Public Health and Safety, Shahid Beheshti University of Medical Sciences, Tehran, Iran.
Eghbal Mohamadi
PhD Student, Department of Watershed Management, Faculty of Natural Resources and Environment, Azad University of Tehran, Iran
Vahid kakapor
Ph.D Student, Department of GIS & RS, Faculty of Planning and Environmental Sciences, Tabriz University, Tabriz, Iran.

Received: 2022/12/06
Accepted: 2023/03/16

Document Type: Research article

Doi:10.22038/jreh.2023.65531.1516

ABSTRACT

Background and purpose: Air quality control is an inevitable issue at the forefront of national concerns. The aim of this study was to predict the daily concentration of $PM_{2.5}$.

Materials and Methods: According to the objective, the type of research can be considered practical, and the statistical population of the research includes meteorological and pollution measuring stations within the 22 districts of Tehran. However, the statistical sample (synoptic geophysical station and Tarbiat Modares measuring station) was selected using a non-random sampling method. The desired statistical year for the study included the daily data from the selected stations for one year. Eleven input variables were used, which included meteorological data from the geophysical synoptic station (maximum and minimum temperature, minimum and maximum relative humidity, rainfall, maximum wind speed, and wind direction) and pollution data of $PM_{2.5}$ concentration from the Tarbiat Modares station (daily concentrations of $PM_{2.5}$ and the previous day). The support vector machine (SVM) model was used for prediction in this step.

Results: The model was able to predict the daily concentration values of the $PM_{2.5}$ pollutant for the upcoming days with a detection coefficient $R^2 = 0.611$ and RMSE = 10.87. In the second method, the support vector machine (SVM) model was combined with principal component analysis (PCA) to reduce the number of variables and perform modeling.

Conclusion: The results of this study show that the performance of the combined model is superior to the previous model, as the coefficient of determination R^2 increased to 0.65 and the error value decreased to 10.37 RMSE (root mean square error). This hybrid model (PCA-SVM) can assist city managers and decision-makers in controlling and reducing the amount of $PM_{2.5}$ pollutants.

Keywords: Suspension of PM2.5particles, Support Vector Machine (SVM), Principal Coompnent Analysis (PCA)

Citation: Zarei A, Zarei S, Aghighi H, Vaziri MH, Mohamadi E, kakapor V. Prediction of daily $PM_{2.5}$ concentration using support vector training combination (SVM) - Adaptive and principal component analysis (PCA). *Journal of Research in Environmental Health*. 2023; 9(1):108-121.

پیش‌بینی غلظت روزانه $PM_{2.5}$ با استفاده از ترکیب آموزش بردار پشتیبان تطبیقی و آنالیز مؤلفه‌های اصلی

امیر زارعی

✱ دانشجوی تخصصی، دانشکده علوم زمین، دانشگاه شهید بهشتی، تهران، ایران.

Amir69zarey@yahoo.com

سیروان زارعی

کارشناسی ارشد، گروه سلامت، ایمنی و محیطزیست (HSE)، مرکز تحقیقات ارتقاء سلامت محیط کار، دانشکده بهداشت و ایمنی، دانشگاه علوم پزشکی شهید بهشتی، تهران، ایران.

حسین عقیقی

استادیار، مرکز مطالعات سنجش از دور و سیستم اطلاعات جغرافیایی (GIS & RS)، دانشکده علوم زمین، دانشگاه شهید بهشتی، تهران، ایران.

محمدحسین وزیری

استادیار، گروه سلامت، ایمنی و محیطزیست (HSE)، مرکز تحقیقات ارتقاء سلامت محیط کار، دانشکده بهداشت و ایمنی، دانشگاه علوم پزشکی شهید بهشتی، تهران، ایران.

اقبال محمدی

دانشجوی دکترای تخصصی، گروه مهندسی آبخیزداری، دانشکده منابع طبیعی و محیطزیست، واحد علوم و تحقیقات، دانشگاه آزاد اسلامی، تهران، ایران.

وحید کاکاپور

دانشجوی دکترای تخصصی، گروه سیستم اطلاعات جغرافیایی و سنجش از دور (GIS & RS)، دانشکده برنامه‌ریزی و علوم محیطی، دانشگاه تبریز، تبریز، ایران.

تاریخ دریافت:۱۴۰۱/۰۹/۱۵

تاریخ پذیرش:۱۴۰۱/۱۲/۲۵

نوع مقاله: پژوهشی

چکیده

زمینه و هدف: امروزه کنترل کیفیت هوا به‌صورت امری گریزناپذیر در رأس مسائل ملی مطرح شود. مطالعه حاضر با هدف پیش‌بینی مقدار غلظت روزانه $PM_{2.5}$ انجام شد.

مواد و روش‌ها: در این مطالعه کاربردی که از اول فرودین ۱۴۰۰ تا آخر فروردین ۱۴۰۱ با هدف پیش‌بینی غلظت روزانه $PM_{2.5}$ در محدود ایستگاه‌های شهر تهران انجام شد، جامعه آماری، ایستگاه‌های سنجش آلودگی و هواشناسی محدوده مناطق ۲۲گانه تهران بود و نمونه آماری (ایستگاه سینوپتیک ژئوفیزیک و ایستگاه سنجش تربیت مدرس) با توجه هدف، به روش نمونه‌گیری غیرتصادفی انتخاب شدند. ۱۱ متغیر ورودی که شامل داده‌های هواشناسی ایستگاه سینوپتیک ژئوفیزیک (دمای ماکزیمم و مینیمم، رطوبت نسبی کمینه و بیشینه، بارندگی، سرعت حداکثر باد و جهت باد) و داده‌های آلودگی غلظت ذرات معلق $PM_{2.5}$ ایستگاه تربیت مدرس (غلظت‌های روزانه $PM_{2.5}$ یک و روز قبل) بود، استفاده شد.

یافته‌ها: مدل PCA توانست مقادیر غلظت روزانه آلاینده $PM_{2.5}$ را برای روزهای آتی با ضریب تشخیص R^2=۰/۶۱۱ و RMSE=۱۰/۸۷ پیش‌بینی نماید. در روش دوم، مدل ماشین بردار پشتیبان (SVM) با آنالیز مؤلفه‌های اصلی (PCA) ترکیب گردید. شرط اساسی استفاده از مدل PCA، کافی بودن نمونه‌ها می‌باشد که این شرط با استفاده از آزمون بارتلت انجام گرفت.

نتیجه‌گیری: با این تعداد متغیر و روش SVM مدل‌سازی انجام گرفت که نتایج این عمل نشان داد عملکرد مدل ترکیبی از مدل قبلی بهتر است، به این دلیل که مقدار ضریب تعیین R^2 افزایش پیدا کرد و به مقدار ۰/۶۵ رسید و مقدار خطا نیز کاهش یافت و به مقدار RMSE=۱۰/۳۷ (جذر میانگین مربعات خطا) رسید. این مدل ترکیبی (PCA-SVM) به مدیران و تصمیم‌گیران شهری جهت کنترل و کاهش میزان آلاینده $PM_{2.5}$ کمک می‌کند.

کلیدواژه‌ها: آنالیز مؤلفه‌های اصلی (PCA)، پیش‌بینی ذرات معلق $PM_{2.5}$، ماشین بردار پشتیبان (SVM)

◀ **استناد:** زارعی الف، زارعی س، عقیقی ح، وزیری م ح، محمدی الف، کاکاپور و. پیش‌بینی غلظت روزانه $PM_{2.5}$ با استفاده از ترکیب آموزش بردار پشتیبان تطبیقی و آنالیز مؤلفه‌های اصلی. *فصلنامه پژوهش در بهداشت محیط*. بهار ۱۴۰۲؛ ۹ (۱): (۱۲۱-۱۰۸).

مقدمه

در دهه‌های اخیر، افزایش تراکم جمعیت و فعالیت‌های اقتصادی و صنعتی در کلان‌شهرها، باعث افزایش حجم ترافیک و در نتیجه، بالا رفتن سطح آلودگی هوا شده است. عمده‌ترین منبع آلوده‌کننده هوا در شهرهای بزرگ در حال توسعه مربوط به حمل‌ونقل انبوه خودروهایی است که بیش از حد استاندارد، سوخت و انرژی مصرف می‌کنند و بارِ سنگین ترافیک خیابان‌های این شهرها، بیشتر ریشه در معضلاتی مانند ضعف مدیریت ترافیک و فرهنگ ترافیکی دارد. یکی از مهم‌ترین عواملی که بر کیفیت زندگی انسان اثر می‌گذارد و اثرات نامطلوبی بر سلامت انسان می‌گذارد، آلودگی هوا است. این اثرات باعث تغییرات بیوشیمیایی و فیزیولوژیکی در بدن انسان می‌شود و در نهایت به بیماری شدید و مرگ منتج می‌گردد (۲). تعداد آلاینده‌های هوا به ۱۸۰ نوع می‌رسد، ممکن است طبیعی یا ساخته دست بشر بوده و به اشکال مختلف مانند ذرات جامد، قطرات مایع و یا گاز وجود داشته باشند (۲۵). دو گروه اصلی از این انواع آلاینده‌ها شامل: آلاینده‌های اولیه و آلاینده‌های ثانویه می‌باشند. گروه اول آن‌هایی هستند که به‌طور مستقیم از منابع آلودگی نشأت می‌گیرند مانند مونواکسید کربن، دی‌اکسید گوگرد، اکسیدهای نیتروژن، هیدروکربن‌ها و ذرات معلق (دوده، گردوغبار و مه دود). دسته دوم در اثر برهم‌کنش عوامل محیطی (نور خورشید، رطوبت و سایر آلاینده‌ها) با آلاینده‌های اولیه ایجاد شده و شامل آلاینده‌های ازن، آلدئیدها، اسیدسولفوریک و پراکسی استیل نیترات[1] (PNA) می‌باشند. آلودگی هوای شهرها شامل هر دو نوع آلاینده اولیه و ثانویه است (۱). مطالعات نشان داده‌اند آلاینده‌هایی همانند دی‌اکسید نیتروژن و ذرات معلق منجر به بیماری‌های قلبی و عروقی، تنفسی و سرطان می‌گردند (۳). مطالعات زیادی در زمینه بررسی ارتباط بین آلودگی هوا و بیماری قلبی- عروقی و تنفسی انجام گرفته (۹) که بیانگر تأثیرات منفی آلودگی هوا در سلامت افراد است. هر ۱۰ میکروگرم افزایش ذرات معلق موجب ۳-۱٪ افزایش مرگ‌ومیر خواهد شد. عملکرد این ذرات به این‌گونه است که ذراتی که در قسمت گلو و حلق گرفته می‌شوند، وارد دستگاه هاضمه شده و در مدت نسبتاً کوتاهی دفع می‌گردند، مگر آنکه وارد خون شوند. ذراتی که وارد نای می‌شوند به‌وسیله موی مانندها و مخاط از جریان هوای تنفسی جدا می‌شوند و در نهایت به دستگاه هاضمه راه می‌یابند. ذراتی که به برونش‌ها برسند، خیلی کندتر حذف می‌شوند. این امر باعث شده است در سال‌های اخیر، پژوهشگران به‌دنبال طراحی مدل‌هایی جهت پیش‌بینی آلودگی هوای شهری در مناطق مختلف برآیند تا به کمک آن بتوانند میزان خطر آلودگی هوا را پیش‌بینی کنند و نتایج این کارها رو در اختیار مدیران شهری و مردم عادی قرار داده و از خسارات بیشتر جلوگیری نمایند. آلودگی هوا، پدیده‌ای پیچیده است، زیرا عوامل و پارامترهای متفاوتی در آن تأثیرگذارند و همچنین تغییرات آن به‌صورت زمانی و مکانی است. این دلایل باعث می‌شود پیش‌بینی و مدل‌سازی آلاینده‌های هوا قابلیت اطمینان بالایی نداشته باشد و با خطا انجام گیرد (۲۶). یکی دیگر از مشکلات پیش‌بینی آلودگی هوا این است که در بحث مدل‌سازی، به تعداد پارامترهای ورودی زیادی نیاز است که همین عامل، زمان پردازش را افزایش می‌دهد؛ بنابراین، انتخاب عوامل ورودی تأثیرگذار و کافی در بحث زمان مدل‌سازی و دقت کار اهمیت بسیار زیادی دارد. یکی از روش‌های مهم و قابل‌اطمینان که برای کاهش ابعاد پارامترهای ورودی[2] در زمینه‌های مختلف مورد استفاده پژوهشگران قرار گرفته است، روش آنالیز مؤلفه‌های اصلی[3] می‌باشد. به کمک این روش می‌توان متغیرهای مستقل اولیه که همبستگی زیادی هم ممکن است بین آن‌ها وجود داشته باشد را با تعداد متغیرهای کمتری که مستقل از هم هستند، جایگزین نمود که به این متغیرهای ورودی جدید، مؤلفه‌های اصلی می‌گویند (۳۰).

پیش‌بینی غلظت روزانه آلاینده‌های هوا، اولین گام اساسی در برنامه‌ریزی کاهش اثرات آن‌ها است. برای این منظور تاکنون روش‌های زیادی برای پیش‌بینی غلظت آلاینده‌های هوا ارائه شده است که آن‌ها را می‌توان به دو دسته روش‌های قطعی و آماری تقسیم نمود. مدل‌های قطعی[4] آلودگی هوا که اساساً حالت پایه

[1] Peroxyacetyl nitrate
[2] Data reduction
[3] Principal component analysis
[4] Determinstic models

انتقال آشفتگی در اتمسفر را منعکس می‌کنند، به‌عنوان ابزاری خبره جهت مدل‌سازی آلاینده‌های گازی و ذرات به‌شمار می‌روند؛ اما نتایج آن‌ها همیشه توسط مقدار قابل‌توجهی خطا تحت تأثیر قرار می‌گیرد. همچنین روش‌های آماری علاوه بر این‌که به اطلاعات انتشار و ضرایب انتشار نیازی ندارند، از ساختاری ساده‌تر نیز نسبت به مدل‌های قطعی برخوردارند (۲۳). تابه‌حال روش‌های آماری متعددی برای پیش‌بینی غلظت آلاینده‌های هوا مورد استفاده قرار گرفته‌اند که در این راستا می‌توان به مدل‌های رگرسیون خطی و غیرخطی (۱۷، ۲۰)، شبکه عصبی مصنوعی و همچنین استفاده از دیگر روش‌های آماری مانند ماشین بردار پشتیبان (SVM)[1] که اولین بار توسط وپنیک ریاضیدان روسی ارائه شد، در پژوهش‌های مربوط به مسائل زیست‌محیطی و به‌تبع آن امر آلودگی هوا استفاده کرد که طی چند سال اخیر مورد توجه برخی از محققین قرار گرفته است (۷، ۱۸).

پیچیدگی رفتار آلودگی از سوی دیگر، سبب می‌شود روش‌های آماری معمول، قادر به مدل‌سازی این پدیده چندوجهی و غیرخطی نباشند (۴، ۱۴)، به همین دلیل دانشمندان به‌دنبال روش‌های جدیدتر و مدرن‌تر برای انجام این مدل‌سازی می‌باشند. در این میان، روش‌های مبتنی بر هوش مصنوعی (AI)[2]، به‌دلیل قابلیت بالایی که در شبیه‌سازی مسائل پیچیده و غیرخطی دارند، در بحث پیش‌بینی آلودگی هوا مورد استفاده زیادی قرار گرفته‌اند (۱۰). برای نمونه، در مطالعه فرناندو و همکاران از شبکه عصبی مصنوعی برای طراحی سیستم هشدار آلودگی هوا استفاده شد (۷). در این سیستم، غلظت آلاینده PM_{10} با استفاده داده‌های هواشناسی و مقادیر PM_{10} در روزهای گذشته، پیش‌بینی شده و نتایج حاصل با سیستم محلی موجود در منطقه مقایسه شده است که عملکرد بهتر شبکه عصبی در پیش‌بینی میزان غلظت PM_{10} را بیان می‌کند. پرزو همکاران از داده‌های سرعت و جهت باد به‌همراه داده‌های مربوط به غلظت آلاینده‌ها در روزهای قبل، برای پیش‌بینی غلظت PM_{10}، با استفاده از شبکه عصبی بهره بردند. در این تحقیق که سه روش ماندگاری[3]، رگرسیون خطی و شبکه عصبی چندلایه با یکدیگر مقایسه شدند، نتایج نشان داد شبکه عصبی با سه لایه دقتی بیشتر از دو مدل دیگر حاصل می‌کند (۲۵).

یکی از مطالعات انجام شده در این زمینه برای پیش‌بینی تأثیر ترافیک جاده‌ای بر سطح غلظت ذرات معلق (PM_{10}) در شهر لندن بود که داده‌های جمع‌آوری شده غلظت (PM_{10}) از ایستگاه‌های نظارت کیفیت هوا را تحت تأثیر متغیرهای ترافیکی و هواشناسی مورد بررسی قرار داد. در این مطالعه، از روش مدل شبکه عصبی مصنوعی برای تخمین و برآورد تأثیر ترافیک جاده‌ای استفاده شد. نتایج حاصل از این بررسی نشان داد که مدل‌های شبکه عصبی مصنوعی در پیش‌بینی سهم و مشارکت ترافیک ساعتی جاده‌ای خوب عمل می‌کنند و پیش‌بینی‌ها به‌خوبی با مشاهدات هماهنگی دارند (۳۱)؛ اما با وجود توانایی بالای شبکه‌های عصبی در مدل‌سازی مسائل غیرخطی، این روش‌ها همچنان با محدودیت‌هایی همچون ناهمگرایی به بهینه جهانی[4] مواجه‌اند (۱۵، ۲۹). این مشکل به‌دلیل استفاده از روش‌های سعی و خطا برای انتخاب لایه‌های پنهان و گره‌ها ایجاد می‌شود. همچنین، شبکه‌های عصبی قادر به تفسیر و تحلیل اطلاعات زبانی[5] نیستند (۲۱). نوری و همکاران طی تحقیقی با استفاده از مدل SVM اقدام به پیش‌بینی میزان هفتگی زباله شهر تهران نمودند (۱۹). همچنین نوری و همکاران برای پیش‌بینی ضریب انتشار طولی در رودخانه‌های طبیعی از SVM استفاده نموده و نتایج این مدل را در مقایسه با مدل‌های کلاسیک رگرسیونی بهتر گزارش کردند (۱۷). در مطالعه ی لیو و همکاران که به مقایسه SVM و شبکه عصبی تابع پایه شعاعی (RBF)[6] برای مدل‌سازی کیفی هوا در منطقه مرکز شهر هنگ کنگ پرداختند، در نهایت برتری مدل SVM نسبت به مدل RBF گزارش شد (۱۲). ساهو و همکاران برای پیش‌بینی ماکزیمم غلظت روزانه ازن تروپوسفریک در ناحیه‌ای از کشور آمریکا از مدل SVM استفاده نمودند (۲۷).

در حالت کلی نتایج به‌دست آمده از مدل SVM در زمینه پیش‌بینی آلودگی هوا امیدوار کننده بوده و روزبه‌روز بهبودهایی در این زمینه توسط محققین مختلف ارائه می‌شود که می‌تواند در کشور ایران نیز با برنامه‌ریزی صحیح از این ابزار قدرتمند،

[1] Support vector machine
[2] Artificial intelligence
[3] Persistence
[4] Global minimum
[5] Linguistic information
[6] Radial base function

راه‌کارهای مناسبی جهت مدیریت آلودگی هوا در اختیار مدیران ذی‌ربط قرار گیرد. در همین راستا در مطالعه حاضر جهت بررسی عملکرد مدل SVM در پیش‌بینی آلودگی هوای شهر تهران و با توجه به اهمیت آلاینده گازی CO در این شهر، پیش‌بینی مقدار میانگین روزانه این آلاینده با استفاده از اطلاعات هواشناسی و آلودگی هوا در ایستگاه قلهک مدنظر قرار گرفت. همچنین تأثیر عملکرد الگوریتم انتخاب پیشرو (FS)[1] در گزینش ورودی به مدل SVM از اهداف دیگر این تحقیق است.

روش‌کار

محدوده مورد مطالعه

کلان‌شهر تهران، پایتخت ۲۰۰ ساله ایران در کوهپایه‌های جنوبی رشته‌کوه البرز در حد فاصل طول ۵۱ درجه و ۵ دقیقه شرقی تا ۵۱ درجه و ۵۳ دقیقه شرقی و عرض جغرافیایی ۳۵ درجه و ۳۴ دقیقه شمالی تا ۳۵ درجه و ۵۹ دقیقه شمالی با حدود ۷۰۰ کیلومتر مربع مساحت گسترده شده است. این شهر از شمال به سلسله جبال البرز، از شرق به لواسانات، از غرب به کرج و از جنوب به ورامین محدود می‌شود. ارتفاع شهر در بلندترین نقاط شمال به ۲۰۰۰ متر و در جنوبی‌ترین نقاط به ۱۰۵۰ متر از سطح دریا می‌رسد. شهر تهران با جمعیتی قریب به ۱۲ میلیون نفر (به همراه شهرهای اقماری خود)، ۱۲٪ جمعیت کل کشور را به خود اختصاص داده است. با توجه به اینکه شهر تهران دارای موقعیت خاص جغرافیایی است (اختلاف ارتفاع زیاد در شمال و جنوب آن) و از شرایط نامناسب بافت شهری برخوردار است و وسایل نقلیه زیادی در طول شبانه‌روز در آن به فعالیت مشغولند، بادهای غربی در تمام طول سال دود کارخانجات و سایر عوامل آلاینده را به سطح شهر تهران وارد می‌سازند، در مجموع دارای شرایط نامساعد زیست‌محیطی بوده و آلودگی هوای آن در سال‌های اخیر با محتوای گازهای سمّی به‌صورت خطرناک عمل می‌نماید که تغییرات بسیار چشمگیر محیطی و اقلیمی را در آن موجب شده است (۱۳). محدوده مورد مطالعه دارای محدودیت زمانی و مکانی می‌باشد که باید به آن‌ها پرداخته شود. محدودیت مکانی به‌این‌علت که در نزدیکی تمامی ایستگاه‌های سنجش آلودگی هوا، ایستگاه سینوپتیک (هواشناسی) موجود نمی‌باشد و تعداد ایستگاه‌های هواشناسی در محدوده کلان‌شهر تهران که پارامترهای جوی را اندازه‌گیری می‌کند، در کل ۳ ایستگاه (ژئوفیزیک، مهرآباد و چیتگر) می‌باشد. محدودیت زمانی هم شامل نداشتن پایه آماری مشترک بین ایستگاه‌ها می‌باشد؛ به‌همین علت در این تحقیق تنها امکان بررسی و مقایسه بین ایستگاه ژئوفیزیک (هواشناسی) و تربیت مدرس (سنجش آلودگی) فراهم بود. پایه آماری مشترک بین دو ایستگاه از اول فروردین ۱۴۰۰ تا آخر فروردین ۱۴۰۱ انتخاب گردید. داده‌های مورد استفاده در این پژوهش شامل: حداقل و حداکثر دما، حداقل و حداکثر رطوبت نسبی، بارش، سرعت و جهت باد (حداکثر) از ایستگاه هواشناسی ژئوفیزیک و همچنین مقدار غلظت روزانه PM_{10} از ایستگاه تربیت مدرس بود.

یافته ها

داده‌های ورودی

همان‌طور که ذکر شد در محدوده شهر تهران سه ایستگاه سینوپتیک که داده‌های هواشناسی را اندازه‌گیری می‌کنند، وجود دارد. در این مطالعه ایستگاه ژئوفیزیک انتخاب شده و داده‌های جوی (دمای ماکزیمم و مینیمم، رطوبت نسبی کمینه و بیشینه، بارندگی، سرعت حداکثر باد و جهت باد) آن به‌عنوان مؤلفه‌های تأثیرگذار در مقدار غلظت آلاینده $PM_{2.5}$ مورد استفاده قرار گرفتند. همچنین تحقیقات قبلی نشان داده است که به‌جز پارامترهای جوی، مقدار غلظت روزانه $PM_{2.5}$ در ۱ و ۲ روز گذشته در مقدار روز بعد آن مؤثر است (۸). به‌همین دلیل داده‌های غلظت روزانه $PM_{2.5}$ در ۲ روز گذشته همراه با پارامترهای هواشناسی ایستگاه سینوپتیک (ژئوفیزیک)، در مجموع ۹ پارامتر به‌عنوان متغیرهای ورودی مدل انتخاب شدند.

تحلیل مؤلفه‌های اصلی (PCA)[2]

تحلیل مؤلفه اصلی و تحلیل عاملی، از روش‌های آماری چندمتغیره هستند که می‌توان از آن‌ها برای کاهش پیچیدگی تحلیل متغیرهای اولیه مسئله در مواردی که با حجم زیادی از اطلاعات روبرو هستیم و همچنین برای تفسیر بهتر اطلاعات استفاده نمود.

زمانی که تعداد ورودی‌های مدل کم باشد، مدل به‌راحتی و بدون صرف زمان زیاد اجرایی می‌شود؛ اما وقتی‌که تعداد متغیرهای

[1] Forward selection

[2] Principal component analysis

ورودی بیشتر شود، پیچیدگی مدل هم زیاد می‌شود و زمان انجام پردازش نیز به‌شدت افزایش پیدا می‌کند. با این روش، متغیرهای اولیه به مؤلفه‌های جدید و مستقل (با ضرایب همبستگی صفر برای هر دو مؤلفه) تبدیل می‌شوند و سپس از این مؤلفه‌ها به‌جای متغیرهای اولیه استفاده می‌گردد. مؤلفه‌های جدید، ترکیب خطی از متغیرهای اولیه هستند. به‌علاوه چون در تشکیل مؤلفه‌ها از تمام متغیرها استفاده می‌گردد، درنتیجه اطلاعات متغیرهای اولیه با کمترین تلفات به‌وسیله مؤلفه‌های حاصل ارائه می‌شود و باعث از دست دادن جنبه‌های اطلاعاتی داده‌های اصلی نمی‌شود.

روش کاربری ایجاد مؤلفه‌های اصلی و تعیین متغیرهای اصلی به‌صورت زیر می‌باشد:

الف- آزمون بارتلت[1] و KMO

برای تحلیل عاملی در ابتدا باید مطمئن شد که آیا تعداد نمونه برای تحلیل عاملی هر متغیر کافی است یا نه؟ برای این کار از شاخص KMO استفاده می‌شود؛ درصورتی‌که مقدار شاخص KMOبرای یک متغیر بیشتر از ۰/۶ باشد، در این صورت تعداد نمونه برای تحلیل عاملی آن متغیر مناسب است (۳۲). در گام بعدی باید از وضعیت واریانس سؤالات هر متغیر مطمئن شد؛ به‌عبارتی‌دیگر، برای تحلیل عاملی لازم است که واریانس سؤالات یک متغیر با یکدیگر برابر نباشند. برای این‌که بتوان به این نتیجه رسید که واریانس سؤالات یک متغیر با یکدیگر برابر هستند یا نه؟ از آزمون بارتلت استفاده می‌شود؛ درصورتی‌که مقادیر سطح معناداری برای متغیری در آزمون بارتلت کمتر از سطح خطای ۰/۵ باشد، در این صورت واریانس سؤالات آن متغیر با یکدیگر برابر نیستند و امکان استفاده از تحلیل عاملی اکتشافی و تأییدی وجود دارد (۳۴). به‌طور کلی از دو آزمون بارتلت و KMO برای بررسی مناسب بودن داده‌ها برای تحلیل مؤلفه اصلی استفاده می‌شود. در مطالعه حاضر با ۹ پارامتر ورودی به جهت اینکه مقدار sig کمتر از ۰/۰۵ بود، با سطح خطای ۵٪ (سطح اطمینان ۹۵٪)، فرض صفر (کافی نبودن تعداد نمونه‌ها) رد شد و در مقابل فرض یک که مبنی بر مناسب بودن تعداد نمونه‌هاست، پذیرفته شد. در مطالعه حاضر مقدار KMO برابر ۰/۷۱۲ و در محدوده قابل‌قبول بود که داده‌های جدول ۱ نیز همین را نشان می‌دهد.

جدول ۱. نتایج آزمون KMO و کرویت بارتلت

	آزمون KMO	۰/۷۱۲
آزمون بارتلت	مقدار کی دو	۳۴۹۹/۲
	درجه آزادی	۳۶
	سطح معناداری	۰/۰۰۰

ج- محاسبه ماتریس همبستگی (R) برای متغیرهای اولیه

ماتریس همبستگی که ماتریسی متقارن است، میزان تغییرات در نمونه و میزان همبستگی P متغیر را با هم نشان می‌دهد. عضوهای روی قطر اصلی این ماتریس، واریانس متغیرهای ورودی و بقیه درایه‌های این ماتریس، کوواریانس بین متغیرهای ورودی است. چون برای تشکیل این ماتریس از داده‌های استاندارد شده استفاده شده است، به‌همین دلیل این ماتریس، معادل ماتریس همبستگی بین متغیرهای ورودی است.

د- تخمین تعداد مؤلفه‌های اصلی

هرچه کمّیت عددی مقادیر ویژه بزرگ‌تر باشد، بیانگر این است که مؤلفه ایجادی از آن نیز درصد بیشتری از اطلاعات متغیرهای اولیه را دربرمی‌گیرد. اولین مؤلفه بیشترین واریانس و آخرین آن، کمترین مقدار واریانس را نشان می‌دهد. انتخاب چند مؤلفه اول که بیشترین مقدار واریانس را دارند و به‌عنوان مؤلفه‌های اصلی شناخته می‌شوند، از اساسی‌ترین اقدامات در تجزیه و تحلیل مؤلفه‌های اصلی می‌باشد. با انتخاب چند مؤلفه اصلی اول، سایر مؤلفه‌ها از محاسبات بعدی حذف می‌شوند، لذا باید دقت زیادی در انتخاب آستانه کرد. Screet plot یکی از روش‌های تشخیص آستانه حذف می‌باشد که در آن مقادیر ویژه در مقابل شماره مؤلفه‌ها رسم می‌شود. در این روش، مرز بین مؤلفه‌های اصلی و غیراصلی محلی است که نمودار میل به خطی شدن می‌نماید؛ یعنی محلی که مقادیر ویژه در مقابل تغییر شماره مؤلفه، تغییر چندانی ننماید. با توجه به اینکه مقدار حد آستانه[2] در این تحقیق ۱ در نظر گرفته شد، بنابراین مقادیر بالاتر از ۱ به‌عنوان مؤلفه‌های اصلی انتخاب می‌شوند. همان‌طور که داده‌های جدول ۳ نشان می‌دهد، تعداد مؤلفه‌های اصلی سه دسته می‌باشند.

[1] Bartlett

[2] Eigenvalue

جدول ۲. ماتریس همبستگی بین متغیرهای ورودی

Correlation PM_{i-1}	PM_{i-1}	PM_i	دمای حداقل	دمای حداکثر	کمینه رطوبت نسبی	بیشینه رطوبت نسبی	بارش	سرعت حداکثر باد	جهت باد
PM_{i-1}	۱								
PM_i	۰/۷۵۹	۱/۰۰۰							
دمای حداقل	-۰/۳۱۳	-۰/۳۴۷	۱/۰۰۰						
دمای حداکثر	-۰/۳۰۱	-۰/۳۲۷	۰/۹۷۴	۱/۰۰۰					
کمینه رطوبت نسبی	۰/۸۰۰	۰/۹۵۱	-۰/۴۶۰	-۰/۴۵۱	۱/۰۰۰				
بیشینه رطوبت نسبی	۰/۱۹۹	۰/۲۳۷	-۰/۷۵۶	-۰/۷۹۵	۰/۳۴۲	۱/۰۰۰			
بارش	-۰/۱۱۳	-۰/۰۳۳	-۰/۲۰۳	-۰/۲۴۱	-۰/۱۱۸	۰/۴۱۰	۱/۰۰۰		
سرعت حداکثر باد	-۰/۴۵۳	-۰/۴۵۷	۰/۱۱۳	۰/۱۱۹	-۰/۴۹۸	-۰/۰۴۲	۰/۰۸۶	۱/۰۰۰	
جهت باد	-۰/۷۸	-۰/۰۹۴	-۰/۲۰۱	-۰/۱۸۰	-۰/۰۰۱	۰/۱۹۳	۰/۰۸۰	۰/۳۳۴	۱/۰۰۰

جدول ۳. مؤلفه‌های ایجادی از متغیرهای اولیه (ورودی‌ها)

متغیرها	واحد	مؤلفه‌ها	مقدار هر مؤلفه از ۹	درصد اطلاعات متغیرهای اولیه	درصد تجمعی اطلاعات متغیرهای اولیه
PM_{i-1}	$\mu g/m^3$	اول	۳/۸۷۵	۴۳/۰۵۸	۴۳/۰۵۸
PM_i	$\mu g/m^3$	دوم	۲/۲۴۲	۲۴/۹۱۰	۶۷/۹۶۸
دمای حداقل	°F	سوم	۱/۰۳۹	۱۱/۵۴۵	۷۹/۵۱۳
دمای حداکثر	°F	چهارم	۰/۷۶۴	۸/۴۸۶	۸۷/۹۹۹
کمینه رطوبت نسبی	٪	پنجم	۰/۵۱۸	۵/۷۵۳	۹۳/۷۵۱
بیشینه رطوبت نسبی	٪	ششم	۰/۲۷۴	۳/۰۴۱	۹۶/۷۹۳
بارش	mm	هفتم	۰/۲۴۴	۲/۷۱۶	۹۹/۵۰۸
سرعت حداکثر باد	Km/hr	هشتم	۰/۰۲۷	۰/۳۰۳	۹۹/۸۱۱
جهت باد	-	نهم	۰/۰۱۷	۰/۱۸۹	۱۰۰/۰۰

سیستم بردار پشتیبان تطبیقی SVM-PCA

شبکه‌های عصبی مصنوعی، ابزارهایی مهم در میان مباحث هوش محاسباتی[1] به‌حساب می‌آیند. انواع مختلفی از شبکه‌های عصبی مصنوعی معرفی شده‌اند که به‌طور عمده در کاربردهایی همچون طبقه‌بندی، خوشه‌بندی، تشخیص الگو، مدل‌سازی و تقریباً توابع (رگرسیون)، کنترل، تخمین و بهینه‌سازی مورد استفاده قرار می‌گیرند. ماشین بردار پشتیبان (SVM)[2] نوع خاصی از شبکه‌های عصبی هستند که برخلاف سایر انواع شبکه عصبی مانند پرسپترون چندلایه[3] (MLP) و توابع پایه شعاعی (RBF) به‌جای کمینه کردن خطا، اقدام به کمینه کردن ریسک عملیاتی طبقه‌بندی یا مدل‌سازی می‌کند. این ابزار، بسیار قدرتمند است و در زمینه‌های مختلفی چون طبقه‌بندی، خوشه‌بندی و رگرسیون می‌تواند مورد استفاده قرار گیرد. ماشین بردار پشتیبان، یک سیستم یادگیری کارآمد برای بهینه‌سازی مفید است که از اصل استقرای کمینه‌سازی خطای ساختاری استفاده نموده و منجر به یک جواب بهینه کلی می‌گردد. این روش در مینیمم‌های محلی گیر نمی‌افتد، آموزش ساده‌ای دارد و برای داده‌های با ابعاد بالا هم معمولاً نتایج خوبی ارائه می‌دهد. ماشین

[1] Computational intelligence
[2] Support vector machines
[3] Multilayer perceptron

بردار پشتیبان، یکی از قوی‌ترین الگوریتم‌های یادگیری ماشین در حوزه‌های کاربردی متعدد و ناهمگن است که عملکرد بسیار خوبی برای گستره‌ای از برنامه‌های کاربردی مانند پردازش اطلاعات، تشخیص ارقام دست‌نویس، تشخیص چهره، مهندسی مالی، تحلیل پایگاه داده، بیوانفورماتیک و ... دارد. بیش از ده‌ها سال است که پژوهش‌های زیادی در مورد یادگیری ماشین روی ماشین‌های بردار متمرکز شده است. این مفهوم به‌دلیل پایه نظری قدرتمند خود که بر اساس نظریه یادگیری آماری است، از عملکرد عمومی بالایی برخوردار است (۵). باید گفت که ماشین‌های بردار پشتیبان، مدل‌های یادگیری تحت نظارتی هستند که با الگوریتم‌های یادگیری که داده را تحلیل و الگوها را شناسایی می‌کنند، همکاری می‌نماید و برای دسته‌بندی و تحلیل رگرسیون استفاده می‌شوند که به‌ترتیب، دسته‌بندی بردار پشتیبان [1](SVC) و رگرسیون بردار پشتیبان[2](SVR) نام دارند. با داشتن مجموعه‌ای از نمونه‌های آموزشی که مشخص شده است هرکدام به چه دسته‌ای از دو دسته موجود متعلق هستند، یک الگوریتم آموزشی SVM مدلی را می‌سازد که نمونه‌های جدید را به دسته اول یا دوم تخصیص دهد. نسخه‌ای از SVM به نام رگرسیون بردار پشتیبان، برای حل رگرسیون توسط والدیمیر واپنیک، هریس دروکر، کریستوفر برگرز، لیندا کافمن و الکساندر سمولو ارائه شد. مدل تولید شده توسط این روش، تنها وابسته به زیرمجموعه‌ای از داده‌های آموزشی است؛ زیرا تابع هزینه برای ساخت مدل، هر داده آموزشی که (با توجه به یک بازه -ε) نزدیک به مدل پیش‌بینی باشد را در نظر نمی‌گیرد (۳۳).

پیش‌بینی $PM_{2.5}$ با ترکیب SVM و PCA

الف- تعریف و تعیین متغیرهای مدل

هدف پژوهش حاضر، پیش‌بینی روزانه غلظت‌های $PM_{2.5}$ با استفاده از مدل ماشین بردار پشتیبان (SVM) و ترکیب الگوی ترکیبی PCA-SVM بود؛ که داده‌های ورودی در دو مدل متفاوت بود. از آنجایی‌که پارامترهای جوی بر میزان غلظت‌های روزانه ذرات معلق با قطر کمتر از ۲/۵ میکرون ($PM_{2.5}$) تأثیرگذار هستند، لذا باید اثر آن‌ها در مدل‌سازی مشاهده شود. پارامترهای هواشناسی در نظر گرفته شده در این پژوهش شامل: حداقل و حداکثر دما، بیشینه و کمینه رطوبت نسبی، بارش، سرعت و جهت باد بود. این پارامترها در محدوده مکانی (ایستگاه سینوپتیک) و زمانی (سال آماری تحقیق) از سازمان هواشناسی کشور دریافت شدند. غلظت‌های روزانه پس‌زمینه (روزهای قبل) $PM_{2.5}$ ایستگاه تربیت مدرس (ایستگاه سنجش آلودگی) که توسط شرکت کنترل کیفیت هوای تهران ثبت می‌شود، دریافت و در محدوده زمانی مشترک مورد استفاده قرار گرفت.

ب-توابع کرنل SVM

الگوریتم‌های SVM از مجموعه‌ای از توابع ریاضی که به‌عنوان کرنل تعریف می‌شوند، استفاده می‌کنند. وظیفه کرنل این است که داده‌ها را به‌عنوان ورودی گرفته و آن‌ها را به شکل موردنیاز تبدیل کند. الگوریتم‌های مختلف SVM، از انواع مختلف توابع کرنل استفاده می‌کنند. این توابع می‌توانند انواع متفاوتی داشته باشند. توابع کرنل برای داده‌های ترتیبی، نمودارها، متن‌ها، تصاویر و همچنین بردارها معرفی می‌شوند. پرکاربردترین نوع تابع کرنل، RBF است؛ زیرا دارای پاسخ محلی و متناهی در کل بازه محور x است (۲۸).

بحث

در این تحقیق برای توسعه مدل SVM از تابع پایه شعاعی RBF به‌دلیل عملکرد بهتر آن، برای تابع هسته استفاده شد. پس از یافتن مقادیر مناسب پارامترها، برای توسعه مدل SVM در مرحله اول از داده‌های آموزش برای انجام فرآیند آموزش استفاده شد و در گام بعدی برای درستی آزمایی مدل و برآورد دقت پیش‌بینی آن، داده‌های آزمون به‌کار گرفته شدند که نتایج آن در شکل ۲ نشان داده شده است که نمایانگر ۲ نمودار است؛ یکی نمودار مقادیر مشاهده شده (داده‌های اندازه‌گیری در روز موردنظر توسط شرکت کنترل کیفیت) و دیگری نمودار مربوط به مقادیر پیش‌بینی غلظت روزانه آلاینده $PM_{2.5}$ که با مدل SVM ساخته شد. داده‌های مورد استفاده مدل در مرحله آموزش و آزمون، ۳۸۹ مورد بود که بر اساس جدول ۴، مقادیر پیش‌بینی شده غلظت ذرات معلق توسط مدل ساخته شده با ماشین بردار پشتیبان (SVM) در مرحله آموزش، دارای ضریب همبستگی بیشتری نسبت به مرحله آزمون بود که این امر مشخص می‌کند که داده‌های پیش‌بینی شده غلظت روزانه آلاینده در مرحله

[1] Support vector clustering

[2] Support vector regression

آموزش، هماهنگی و نزدیکی بیشتر با مقادیر مشاهداتی (اندازه‌گیری) غلظت‌های روزانه دارند.

پیش‌پردازش متغیرهای ورودی به مدل ماشین بردار پشتیبان (SVM) با PCA

همان‌طور که قبلاً ذکر شد، برای بررسی امکان اجرای روش آنالیز مؤلفه‌های اصلی از آزمون بارتلت استفاده شد؛ و به‌دلیل اینکه مقدار KMO=۰/۷۱۲ در محدوده قابل قبول جهت انجام PCA قرار داشت، لذا امکان اجرای این روش در این تحقیق با پارامترهای مذکور مورد تأیید بود. برای اجرای این روش، پس از استاندارد کردن متغیرهای ورودی، ماتریس متقارن همبستگی از مرتبه ۹ (برابر با تعداد متغیرهای ورودی) تشکیل گردید که نتایج ماتریس همبستگی در جدول ۲ نیز مشخص است. پس از این مرحله، نوبت به محاسبه مقدار ویژه به ازای هر متغیر ورودی می‌رسد (λ) که این مقادیر ویژه در جدول ۳ آورده شده است. اهمیت این مقادیر ویژه به این دلیل است که با استفاده از مقدار آن برای هر متغیر می‌توان تعداد مؤلفه‌های اصلی را پیدا کرد. بر اساس داده‌های جدول ۳ ، مقدار اولین مؤلفه برابر با ۳/۸۷۵ بود که ۴۳/۰۵۸٪ از کل واریانس موجود در سری داده‌ها را توجیه می‌کرد. دومین مقدار ویژه نیز با مقدار ۲/۲۴۲، ۲۴/۹۱٪ از کل واریانس را توجیه می‌کرد. همچنین عدد مقدار ویژه برای سومین مؤلفه ۱/۰۳۹ بود که این مؤلفه نیز ۱۱/۵۴۵٪ واریانس را برقرار می‌کرد. در مجموع این سه مؤلفه حدود ۸۰٪ (۷۹/۵۱۳٪) کل پراکندگی داده‌ها را بیان می‌کردند. بنابراین تقریباً می‌توان سه مؤلفه اول را به‌عنوان مؤلفه‌های اصلی انتخاب نمود. انتخاب آستانه حذف جهت انتخاب مؤلفه‌های اصلی، یک روش تخمینی و غیردقیق می‌باشد، لذا ضروری است که آزمون‌های دیگری نیز جهت انتخاب مؤلفه‌های اصلی انجام گیرد. در واقع هدف از آزمون‌های بعدی، یافتن مبنای فیزیکی برای هر یک از مؤلفه‌های اصلی می‌باشد.

برای پیدا نمودن آستانه حذف، تعداد ۲، ۳ و ۴ مؤلفه اصلی (مؤلفه‌های داوطلب) در نظر گرفته شد و آزمون لازم برای تشخیص تعداد مؤلفه‌های مطلوب برای نگهداری به‌عمل آمد. برای این منظور، عوامل بارگذاری[1] مربوط به هر یک از مؤلفه‌ها به‌روش وریماکس دوران داده شدند و وابستگی هر یک از مؤلفه‌های دوران داده شده به $PM_{2.5}$ مورد ارزیابی قرار گرفت. در این پژوهش از دوران ۳ مؤلفه اول نتایج مناسبی به‌دست آمد که نتایج آن در جدول ۴ نشان داده شده است. با توجه به نتایج جدول مورد نظر، مقادیر دوران داده شده بارگذاری مؤلفه اول نشان داد که متغیرهای کمینه رطوبت نسبی، مقدار غلظت آلاینده ۱ و ۲ روز قبل، دارای بالاترین مقدار همبستگی و بارش و جهت باد دارای کمترین مقدار بودند. در مقابل، بردارهای دوران داده شده بارگذاری متناظر با مؤلفه دوم همبستگی زیادی با متغیرهای کمینه و بیشینه رطوبت نسبی و مقدار غلظت آلاینده ۱ و ۲ روز قبل با غلظت مقدار آلاینده در روز موردنظر داشتند و متغیرهای حداقل و حداکثر دما، دارای کمترین مقدار همبستگی بودند و در نهایت مقادیر دوران داده شده بارگذاری مؤلفه سوم نشان داد که متغیرهای جهت و سرعت باد، دارای بیشترین و متغیرهای رطوبت نسبی مینیمم و بارش دارای کمترین مقدار همبستگی با غلظت آلاینده در روز موردنظر بودند.

با توجه به جدول (جهت تعیین مؤلفه‌ها)، برای تشکیل مؤلفه اول بایستی مقادیر متغیر مینیمم رطوبت نسبی (RH_{min}) را در ضریب ۰/۹۴۵، مقادیر متغیر غلظت ۱ روز قبل (PM_i) را در ضریب ۰/۸۹۹ و به‌همین ترتیب مقادیر بقیه متغیرها را در ضرایب مربوطه ضرب کرد؛ بدین‌ترتیب مؤلفه‌هایی حاصل می‌شوند که می‌توان آن‌ها را به‌جای متغیرهای اولیه، به‌عنوان ورودی به مدل SVM استفاده کرد. طبق مطالب گفته شده، از بردارهای ویژه به‌دست آمده در جدول، ۳ مؤلفه اصلی به‌صورت روابط زیر استخراج شدند.

رابطه (۱)

$$PC_1=(PM_i \times ۰/۸۹۹)+(PM_{i-1}\times ۰/۸۷۲)-(T_{max} \times ۰/۳۰۶)-(T_{min} \times ۰/۳۳۱)+(RH_{max}\times ۰/۱۴۱)+(RH_{min} \times ۰/۹۴۵)-(WS_{max} \times ۰/۵۷)-(WD \times ۰/۰۴۸)-(RF \times ۰/۳۴۵)$$

رابطه (۲)

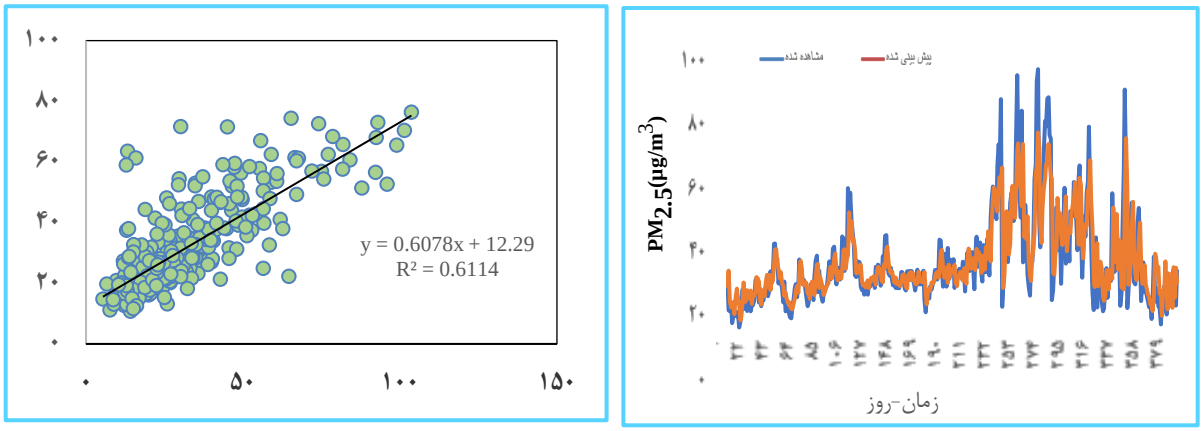
$$PC2= (PMi \times ۰/۱۵۵) + (PMi\text{-}1\times ۰/۰۹۴) - (Tmax \times ۰/۸۷۹) -(Tmin\times ۰/۸۴۹) + (RHmax\times ۰/۸۹۶) + (RHmin\times ۰/۲۲۵) -(WSmax\times ۰/۰۱۷) - (WD\times ۰/۱۶) - (RF\times ۰/۶۲۳)$$

رابطه (۳)

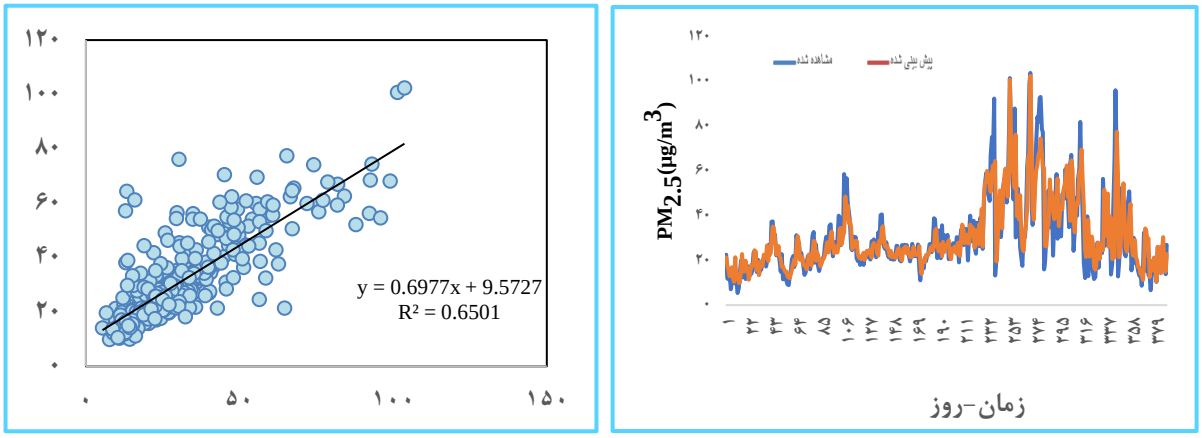
[1] Loading

$PC_3=(PMi\times-0/103)+(PM_{i-1}\times0/068)-(T_{max}\times0/159)-(T_{min}\times0/194)+(RH_{max}\times0/089)+(RH_{min}\times0/002)-(WS_{max}\times0/57)-(WD\times0/867)-(RF\times0/267)$

از میان مؤلفه‌های اصلی سه‌گانه، مؤلفه PC_1 اهمیت بیشتری دارد؛ به‌این علت که در این مؤلفه سه متغیر اول جمعاً حدود ۸۰٪ از کل واریانس متغیرها را بیان می‌کنند، اما در مؤلفه‌های دوم و سوم به‌ترتیب سه متغیر اول در حدود ۵۵٪ و ۹٪ واریانس کل را تشریح می‌نمایند؛ به همین دلیل متغیرهای اصلی از میان متغیرهای این مؤلفه انتخاب شدند.



شکل ۱. نمودار مقایسه تغییرات مقادیر غلظت مشاهده‌ای و پیش‌بینی PM2.5 توسط مدل SVM



شکل ۲. نمودار مقایسه تغییرات مقادیر غلظت مشاهده‌ای و پیش‌بینی PM2.5 توسط مدل PCA-SVM

تعیین متغیرهای اصلی

در روش مؤلفه‌های اصلی (PCA) پس از تشکیل ماتریس ضرایب مؤلفه، اقدام به تعیین پارامترهای اصلی شد؛ و با استفاده از چرخش وریماکس روی ماتریس مؤلفه، طبق جدول ۴ بردارهای ویژه به‌دست آمدند. در این روش پارامترهای اصلی، پارامترهایی هستند که حداقل یکی از ضرایب آن‌ها که برای تشکیل عامل مربوطه استفاده می‌شود، دارای مقدار بالایی باشد. تعیین مقدار این ضریب به شرایط تحقیق، وسعت منطقه مورد مطالعه و

پارامترهای ورودی بستگی دارد. هرچه منطقه مورد مطالعه به لحاظ وسعت بزرگ‌تر باشد، مقادیر کمتری از این ضرایب را در نظر می‌گیرند، ولی برای مسائلی که ساده و کوچک باشند، معمولاً مقادیری بیشتری را در نظر می‌گیرند. در این تحقیق به‌دلیل وسعت منطقه و تعداد پارامترهای ورودی محدود، این معیار ۰/۶ در نظر گرفته شد؛ به این‌معنی که متغیرهایی که مقادیر مساوی یا بالاتر از این مقدار را در جدول چرخش وریماکس دارا بودند، به‌عنوان متغیر اصلی انتخاب شدند که در این تحقیق پارامترهایی نظیر کمینه رطوبت نسبی، مقدار غلظت روزانه آلاینده $PM_{2.5}$ در ۱ و ۲ روز پیش به‌عنوان متغیر اصلی انتخاب شدند و مدل‌سازی ماشین بردار پشتیبان (SVM) با این سه متغیر ورودی که حدود ۸۰٪ واریانس کل را نیز تشریح می‌کنند، انجام گردید.

جدول ۴. پارامترهای ارزیابی به‌دست‌آمده برای داده‌ها به دو روش SVM و PCA-SVM

معیار	مرحله آموزش		مرحله آزمون	
	PCA-SVM	SVM	PCA-SVM	SVM
R	۰/۸۴۹	۰/۸۳۱	۰/۶۲۱	۰/۶۱۴
RMSE	۹/۲۵	۹/۷۴	۱۳/۸۸	۱۴/۱۳
MSE	۸۵/۶۳	۹۴/۸۹	۱۹۲/۶۶	۲۰۴/۸۳

معیارهای ارزیابی مدل‌ها

برای تعیین میزان دقت مدل‌ها از مقادیر مجذور میانگین مربع خطا[1] (RMSE) ،میانگین قدرمطلق خطا[2] (MAE) و ضریب تعیین یا امتیاز [3] (R^2) استفاده شد. هرچه مقادیر RMSE و MSE به صفر نزدیک‌تر و مقدار R^2 به ۱ نزدیک‌تر باشد، دقت مدل در شبیه‌سازی بیشتر است.

رابطه (۴): جذر میانگین مربعات

$$RMSE = \sqrt{\frac{\sum_{=1}^{n}(\widehat{Q_i} - Q_i)^2}{n}}$$

رابطه (۵): ضریب تبیین

$$Rsqrt=1-\frac{\sum_{i=1}^{n}(Q_i^o-Q_i^m)^2}{\sum_{i=1}^{n}(Q_i^o-\overline{Q^o})^2}$$

رابطه (۶): خطای میانگین مربعات

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left|(Q_i^m - Q_i^{o)2}\right|$$

که در این روابط، Qo داده‌های مشاهداتی، $\overline{Q_o}$ میانگین داده‌های مشاهداتی و Q_m داده‌های شبیه‌سازی شده می‌باشد.

بر اساس جدول ۵ و شکل‌های ۶ و ۷، ضریب رگرسیون در روش PCA-SVM بیشتر بود (در هر دو قسمت آموزش و آزمون)؛ یعنی این مدل توانست مقادیر غلظت روزانه را با دقت بیشتری پیش‌بینی نماید. می‌توان برای تمامی ایستگاه‌های منطقه تحقیق موردنظر را انجام داد و پهنه‌بندی آلودگی شهر تهران را برای کل منطقه برآورد نمود.

در تحقیق نوری و همکاران که به پیش‌بینی غلظت روزانه مونوکسید کربن در هوای شهر تهران با استفاده از روش ماشین بردار پشتیبان پرداختند، عملکرد مدل SVM از مدل‌های آماری کلاسیک مانند رگرسیون خطی چندمتغیره برتر بود و با اندکی اختلاف، قابل ارزیابی با مدل‌های هوشمند شبکه عصبی مصنوعی و سیستم استنتاج فازی تطبیقی بود (۲۲).

در مطالعه لیو و همکاران با تجزیه و تحلیل الگوی هواشناسی جهت پیش‌بینی روزانه $PM_{2.5}$ با استفاده از SVM بهینه شده، PSO-SVM عملکرد بهتری نسبت به مدل‌های Adaboost و ANN برای تحلیل الگوی هواشناسی کاربردی به پیش‌بینی درجه‌های $PM_{2.5}$ نشان داد و بالاترین دقت و کارایی را ارائه و زمان پیش‌بینی به‌طور قابل‌توجهی ۲۵٪ کاهش یافت؛ که با نتایج پژوهش حاضر هم‌راستا بود(۱۲).

ووکانتسیس و همکاران نیز در پژوهشی با مقایسه متقابل داده‌های کیفیت هوا با استفاده از تحلیل مؤلفه‌های اصلی و پیش‌بینی غلظت PM_{10} و $PM_{2.5}$ با استفاده از شبکه‌های عصبی مصنوعی گزارش دادند که ماشین بردار پشتیبان (SVM)، یک نوع جدید از ماشین یادگیری مبتنی بر تئوری یادگیری آماری، می‌تواند برای پیش‌بینی رگرسیون و سری‌های زمانی استفاده شود و با برخی از نتایج امیدوار کننده گزارش شده است که عملکرد خوبی دارد و با نتایج تحقیق حاضر همخوانی داشت(۶).

[1] Roote mean squar error
[2] Mean absolute error
[3] Coefficient of determination

لیو و همکاران نیز با بررسی امکان استفاده از SVM برای پیش‌بینی سطوح آلاینده هوا در پیشبرد سری‌های زمانی بر اساس پایگاه داده آلاینده هوای نظارت شده در منطقه مرکز شهر هنگ‌کنگ با مقایسه تجربی بین مدل SVM و شبکه تابع پایه شعاعی کلاسیک (RBF) دریافتند که SVM نسبت به شبکه RBF معمولی در پیش‌بینی پارامترهای کیفیت هوا با سری‌های زمانی مختلف و عملکرد تعمیم بهتری نسبت به مدل RBF برتری دارد که با نتایج پژوهش حاضر هم‌راستا بود(۱۲).

نتیجه‌گیری

با توجه به اینکه در حال حاضر مسئله آلودگی هوا، یکی از مشکلات زیست‌محیطی کشور محسوب می‌شود و همچنین با اثرات سوئی که آلاینده $PM_{2.5}$ بر سلامتی انسان دارد، لذا استفاده از مدل ماشین بردار پشتیبان ضروری به‌نظر می‌رسد. استفاده از تحلیل مؤلفه‌های اصلی (PCA) در کنار مدل اصلی موجب گردید میان متغیرهای ورودی در حالت اول (استفاده از SVM) با دقت بیشتر و خطای کمتری، مقدار غلظت آلاینده پیش‌بینی شود. این عملکرد بهتر به این دلیل است که روش PCA، همبستگی میان پارامترهای ورودی را حذف کرده و همچنین در عین اینکه تعداد داده‌های ورودی را کاهش می‌دهد، پراکندگی روی داده‌ها را نیز نگه می‌دارد. علاوه بر این، کاهش ابعاد متغیرهای ورودی، جهت اجرای مدل زمان محاسبات را کاهش می‌دهد. نتایج مقایسه مدل ترکیبی به‌وسیله مدل‌های ارزیابی خطا با روش بردار پشتیبان در پیش‌بینی بیانگر این است که ضریب رگرسیون بالاتر (دقت بالا) و سرعت پردازش مدل ترکیبی PCA-SVM نسبت به‌روش SVM بیشتر است. البته یکی از محدودیت‌های مدل ترکیبی PCA-SVM این است که برای تعداد نمونه‌های کم، عملکرد ضعیف‌تری را نشان می‌دهد که راهکار بعدی برای رفع مشکل، استفاده از دیگر روش‌ها جهت پیش‌بینی می‌باشد.

ملاحظات اخلاقی

نویسندگان تمام نکات اخلاقی شامل عدم سرقت ادبی، انتشار دوگانه، تحریف داده‌ها و داده‌سازی را در این مقاله رعایت کرده‌اند. همچنین هرگونه تضاد منافع حقیقی یا مادی که ممکن است بر نتایج یا تفسیر مقاله تأثیر بگذارد را رد می‌کنند.

تشکر و قدردانی

بدین‌وسیله از مسئولین محترم شرکت کنترل کیفیت هوا شهرداری تهران، به‌دلیل در اختیار داده‌های پژوهش، صمیمانه تشکر و قدردانی می‌گردد.

References

1.Adams, K.J.Exercise Physiology: ACSM Resourse Manual for Guidelines for Exercise testing and prescription.6th ed": In:Ehrman JK,editor. Philadelphia.Lippincott Williams and Wilkin.2010; 73-4.
2. Arnesano M, Revel G M, Seri, F A. Tool for the optimal sensor placement to temperature monitoring in large sports spaces.Journal of Automation in Construction.2016; 68:223–234.
3. Bono R, Raffaella D, Marco P, Valeria R, Renato R. Benzene and formaldehyde in air of two winter Olympic venues of Torino.Journal of Environment International. 2010;36(3):269-275.
4. Boznar M, Lesjak M, Mlakar, P. 1993, A Neural Network-Based Method for ShortTerm Predictions of Ambient SO2 Concentrations in Highly Polluted Industrial Areas of Complex Terrain, Atmospheric Environment, Part B. Journal of Urban Atmosphere.1993; 27(2): 221- 230.
5. Chen ST, Yu PS. Pruning of support vector networks on flood forecasting. Journal of Hydrology. 2007;347(1-2):67-78.
6. Dimitris V, Kostas K, Jaakko K, Teemu R, Ari K, Mikko K.Intercomparison of air quality data using principal component analysis, and forecasting of PM_{10} and $PM_{2.5}$ concentrations using artificial neural networks, in Thessaloniki and Helsinki. Journal of Science of The Total Environment.2011; 409(7):1266-1276.
7. Fernando H.J, Mammarella M, Grandoni, G, Fedele P, Di Marco R, Dimitrova R, Hyde P. Forecasting PM $_{10}$ in Metropolitan Areas: Efficacy of Neural Networks.Journal of Environmental Pollution. 2012; 163: 62- 67.
8. Gardner MW, Dorling SR. Neural network modelling and prediction of hourly NO_x and NO_2 concentrations in urban air in London.Journal of Atmospheric Environment. 1999;33(5):709–719.
9.Ghaemi, Z, Farnaghi M, Alimohammadi A. An Online Approach for SpatioTemporal Prediction of Air Pollution

in Tehran Using Support Vector Machine. Journal of Geospatial Information Technology.2016; 3(4): 43- 63.
10. Gorai AK, Tuluri F, Tchounwou PB. A GIS based approach for assessing the association between air pollution and asthma in New York State, USA. Int J Environ Res Public Health. 2014;11(5):4845-69.
11. Kolehmainen M, Martikainen H, Hiltunen T, Ruuskanen, J. Forecasting Air Quality Parameters Using Hybrid Neural Network Modelling, Urban Air Quality: Measurement, Modelling and Management.2000; 277-286.
12. Kumar A,Goyal P.Forecasting of Air Quality in Delhi Using Principal Component Regression Technique, Atmospheric Pollution Research.2011;2(4): 436- 444.
13. Lu WZ, Wang WJ. Potential assessment of the support vector machine method in forecasting ambient air pollutant trends. Journal of Chemosphere. 2005;59:693-701.
14. Mostafaeipour A, Zarezade M, Goudarzi H, Rezaei-Shouroki M, Qolipour M. Investigating the Factors on Using the Solar Water Heaters for Dry Arid Regions: A Case Study.Journal of Renewable and Sustainable Energy Reviews. 2017; 78: 157-166.
15. Moussiopoulos N, Sahm P, Kessler C. Numerical Simulation of Photochemical Smog Formation in Athens, Greece a Case Study.Journal of Atmospheric Environment.1995;29(24): 3619- 3632.
16. Niska H, Hiltunen T, Karppinen A, Ruuskanen J, Kolehmainen M. Evolving the Neural Network Model for Forecasting Air Pollution Time Series.Journal of Engineering Applications of Artificial Intelligence.2004; 17(2): 159 -167.
17. Noori R, Abdoli MA, Ameri- Ghasrodashti A, JaliliGhazizade M. Prediction of municipal solid waste generation with combination of support vector machine and principal component analysis: a case study of Mashhad. Journal of Environmental Progress & Sustainable Energy. 2009;28(2):249-58.
18. Noori R, Ashrafi K, Azhdarpour A. Comparison of ANN and PCA based multivariate linear regression applied to predict the daily average concentration of CO: a case study of Tehran. Journal of the Earth Space Physics. 2008;34(1):135-52.
19. Noori R, Hoshyaripour G, Ashrafi K, NadjarAraabi B. Uncertainty analysis of developed ANN and ANFIS models in prediction of carbon monoxide daily concentration. Journal of Atmosphric Environment. 2010;44(4):476,82.
20. Noori R, Karbassi A, Farokhnia A, Dehghani M. Predicting the longitudinal dispersion coefficient using support vector machine and adaptive neuro-fuzzy inference system techniques. Environmental Engineering Science. 2009;26(10):1503-10.
21. Nunnari G, Dorling S, Schlink U, Cawley G, Foxall R, Chatterton T. Modelling SO2 concentration at a point with statistical approaches, Environmental Modelling & Software. 2004;19(10):887–905.
22. Nunnari G. Simplified Fuzzy Modelling of Pollutant Time Series, Neural Network World. 2000; 10(6): 983-1000.
23. Osowski S. Forecasting of the daily meteorological pollution using wavelets and support vector machine.Journal of Engineering Applications of Artificial Intelligence. 2001;15(3):208-16.
24. Pelliccioni A, Tirabassi T. Air dispersion model and neural network: a new perspective for integrated models in the simulation of complex situations. Journal of Environmental Modelling & Software. 2006;21(4):539-46
25. Pérez P, Trier A, Reyes J. Prediction of PM $_{2.5}$ Concentrations Several Hours in Advance Using Neural Networks in Santiago, Chile.Journal of Atmospheric Environment. 2000; 34(8): 1189- 1196.
26. Qu Y, Liu Y, Nayak R, Li, M.Sustainable development of eco-industrial parks in China: effects of managers' environmental awareness on the relationships between practice and performance**.** Journal of Cleaner Production. 2015; 87:328-338
27. Sahoo M M, Patra K, Khatua K.Inference of Water Quality Index Using ANFIA and PCA.Journal of Aquatic Procedia. 2015; 4:1099 -1106.
28. Salazar-Ruiz E, Ordieres JB, Vergara EP, CapuzRizo SF. Development and comparative analysis of tropospheric ozone prediction models using linear and artificial intelligence-based models in Mexicali, Baja California (Mexico) and Calexico, California (US). Journal of Environmental Modelling & Software. 2008;23:1056- 69.
29. Schlkopf C, Smola A. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond, MA, USA, MIT Press Cambridge. 2018.
30. Singh K P, Gupta S, Rai P. Identifying Pollution Sources and Predicting Urban Air Quality Using Ensemble Learning Methods, Journal of Atmospheric Environment. 2013; 80: 426 437.
31. Statheropoulos M, Vassiliadis N, Pappa, A.Principal Component and Canonical Correlation Analysis for Examining Air Pollution and Meteorological Data.Journal of Atmospheric Environment. 1998; 32(6):1087-1095.
32. Suleiman A, Tight M R, Quinn A D. Assessment and prediction of the impact of road transport on ambient concentrations of particulate matter PM_{10}. Journal of Transportation Research Part D. 2016; 49:301-312.
33. Williamsn B, Onsman A, Brown T. Exploratory Factor Analysis: A Five-Step Guide for Novices, Australasian Journal of Paramedicine. 2010;8(3):1-13.
34. Wei L, Geng G, Fuji C, Yihui C.Meteorological pattern analysis assisted daily PM $_{2.5}$ grades prediction using SVM optimized by PSO algorithm,Journal of

Atmospheric Pollution Research, 2019;10(5): 1482-1491.
35. Wei-Zhen Lu, Wen-Jian Wang.Potential assessment of the â€œsupport vector machineâ€ method in forecasting ambient air pollutant trends.Journal of Chemosphere.2005;59(5):693-701.
36. Yu PS, Chen ST, Chang IF. Support vector regression for real-time flood stage forecasting. Journal of Hydrology. 2006;328(3-4):704-16.
37. Zhang J, Wang C, Liu L, Guo H, Liu G, Li Y, Deng S.Investigation of Carbon Dioxide Emission in China by Primary Component Analysis.Journal of Science of The Total Environment. 2014; 472:239- 247.