Comparing machine learning algorithms for estimating PM_{10} particle concentration using AOD and selected meteorological parameters

ABSTRACT

Monitoring and controlling the level and sources of dust are crucial in the face of climate change and the development of suitable predictive approaches that directly impact the environment and human health. This study aims to estimate the concentration of PM10 in the city of Ahvaz using various machine learning models. Climate variables and the Aerosol Optical Depth (AOD) index, derived from the MODIS sensor at a wavelength of 476 nanometers, were used as influential variables in estimating PM_{10} concentration in three scenarios: combining AOD with PM_{10} (scenario 1), combining climate variables with PM_{10} (scenario 2), and combining climate variables and AOD with PM_{10} (scenario 3) .Using six machine learning algorithms, namely Random Forest Regression (RFR), Gradient Boosting Regression (GBR), Artificial Neural Networks (ANN), AdaBoostR with DTR, Support Vector Regression (SVR), and Decision Tree Regression (DTR), the PM_{10} concentration was estimated in different scenarios, considering accuracy and precision coefficients. The most influential variables in estimating PM_{10} concentration were determined to be sunshine hours, minimum visibility, maximum wind speed, and the AOD index. The GBR linear regression model, with R^2, MAE, RMSE, and IOA coefficients of 0.76, 0.31, 0.49 and 0.93 respectively, was found to be the most suitable model for estimating PM_{10} concentration in scenario 3. the results showed that incorporating the AOD index alongside climate variables improved the model's performance in estimating PM_{10} concentration. The proposed final model can be used for daily estimation of PM_{10} particles.

Keywords: Machine Learning Algorithms, Climatic Variables, AOD index, PM_{10}

مقایسه الگوریتم‌های یادگیری ماشین به‌منظور تخمین غلظت ذرات PM_{10} با استفاده از شاخص AOD و برخی پارامترهای هواشناسی

چکیده

نظارت و کنترل بر میزان و منابع گردوغبار تحت تأثیر تغییرات اقلیمی و توسعه رویکردهای پیش‌بینی مناسب که تأثیرات مستقیمی بر محیط‌زیست و سلامت انسان دارد بسیار حائز اهمیت هستند. این مطالعه باهدف تخمین غلظت ذرات کوچکتر از ۱۰ میکرومتر (PM_{10}) در شهر اهواز، با استفاده از مدل‌های مختلف یادگیری ماشین انجام‌شده است. از متغیرهای اقلیمی و شاخص عمق بصری (AOD) محصول باند ۴۷۶ نانومتر سنجنده مودیس به‌عنوان متغیرهای مؤثر در برآورد غلظت ذرات PM_{10} در قالب سه سناریو شامل: ترکیب شاخص AOD با PM_{10} (سناریو اول)، ترکیب متغیرهای اقلیمی با PM_{10} (سناریوی دوم) و ترکیب متغیرهای اقلیمی و شاخص AOD با PM_{10} (سناریوی سوم) استفاده گردید. با استفاده از شش الگوریتم مدل یادگیری ماشین شامل: Random Forest Regression (RFR) ، Gradient Boosting Regression (GBR)، Artificial Neural Networks (ANN)، AdaBoostR with DTR ، Support Vector Regression (SVR) و Decision Tree (DTR)، میزان غلظت ذرات (PM_{10}) در سناریوهای مختلف با در نظر گرفتن ضرایب صحت و دقت تعیین و مقایسه شدند. مهم‌ترین متغیرهای تأثیرگذار در برآورد میزان PM_{10}: ساعت آفتابی، حداقل دید افقی ، ماکزیمم سرعت باد و شاخص AOD تعیین گردید. مدل رگرسیون خطی GBR با مقادیر ضرایب R^2، MAE، RMSE و IOA به ترتیب برابر با ۰/۷۶، ۰/۳۱، ۰/۴۹ و ۰/۹۳ مناسب‌ترین مدل در تخمین غلظت ذرات (PM_{10}) بوده، که در سناریوی سوم بدست آمد. نتایج نشان داد که استفاده از ترکیب شاخص AOD در کنار متغیرهای اقلیمی منجر به بهبود عملکرد مدل در برآورد غلظت ذرات PM_{10} شده است. مدل نهائی پیشنهادی می تواند به منظور تخمین روزانه ذرات PM_{10} استفاده شود.

کلیدواژه: الگوریتم‌های یادگیری ماشین، متغیرهای اقلیمی ، عمق نوری آئروسل، ذرات معلق با قطر آئرودینامیکی کمتر از ۱۰ میکرومتر.

مقدمه

در دهه‌های اخیر، رشد جمعیت و استفاده روزافزون از انرژی، به همراه پدیده‌هایی مانند خشکسالی و انتشار ذرات و گازهای آلوده در جو، منجر به کاهش کیفیت هوا و افزایش ذرات معلق شده است (محمودی سراب،۱۳۹۷). یکی از پیامدهای بارز تغییرات آب و هوایی، وقوع گردوغبار است که به‌ویژه در مناطق خشک و بیابانی، پیامدهای اقتصادی و زیست‌محیطی جدی به دنبال دارد (Zarei et al, 2022). این مناطق تقریباً یک‌سوم مساحت زمین را پوشش می‌دهند و از جمله حساس‌ترین نواحی نسبت به فرسایش بادی به شمار می‌آیند (Reynolds et al,2007).

تحقیقات نشان داده‌اند که در آفریقا و آسیا، فرسایش بادی منجر به کاهش سطح زمین‌های زراعی و در نتیجه کاهش تولید محصولات کشاورزی شده است (Gonzalez et al., 2018). خاورمیانه، به‌ویژه ایران، یکی از تولیدکنندگان اصلی گردوغبار در جهان است و تقریباً ۲۵ درصد از گردوغبار جهانی در این منطقه تولید می‌شود (Ginoux et al, 2004). ایران دارای زمین‌های بیابانی وسیع است که حدود ۹۰۷ کیلومتر مربع از مساحت کل کشور را شامل می‌شود (Khosroshahi et al, 2009). مناطق حساس به فرسایش بادی مانند بیابان‌های لوت، کویر و تالاب‌های خشک‌شده هورالعظیم و شادگان در خوزستان، به‌دلیل تغییرات اقلیمی و کاهش آب، به‌طور قابل توجهی خشک شده و باعث افزایش فرسایش خاک و تولید گردوغبار در مناطق همجوار شده‌اند (Middleton ،Rashki et al, 2017;Alizadeh-Choobari et al, 2014,2016; Rezaei 1986; et al,2019).

فرسایش بادی و طوفان‌های گردوغبار تهدیدی جدی برای سلامت اکوسیستم‌ها و ساکنان این مناطق محسوب می‌شوند (Faghihinia and Afzali, 2013; Tahbaz, 2016; Sahebzadeh et al., 2019) و خسارت‌های جبران‌ناپذیری به زیرساخت‌ها وارد می‌کنند (Miri et al., 2009) . طوفان‌های گردوغبار یکی از دلایل اصلی انتشار ذرات معلق با قطر آئرودینامیکی کمتر از ۱۰ میکرومتر (PM_{10}) در هوای مناطق خشک، به‌ویژه

استان‌های بیابانی ایران هستند. افزایش غلظت این ذرات در هوای مناطق مختلف اثرات زیست‌محیطی نامطلوبی به همراه دارد (Modarres et al, 2018). پژوهش‌های متعدد رابطه پیچیده‌ای بین طوفان‌های شن، گردوغبار و تغییرات آب و هوایی را نشان می‌دهند که شامل عوامل مختلفی از جمله گرمایش جهانی و تخریب جنگل‌ها است (Daryanoosh et al,2018; Gautam et al,2018). شناسایی عوامل تأثیرگذار بر تغییرات در غلظت ذرات معلق می‌تواند گامی مؤثر در کاهش اثرات منفی بر محیطزیست و سلامت انسان‌ها باشد (Ashpole &,Washington, 2013).

سطح فعالیت طوفان‌های گردوغبار معمولاً بر اساس معیارهای مختلفی ارزیابی می‌شود، از جمله غلظت گردوغبار(Shao,2003)، فراوانی روزهای گردوغبار (Ekhtesasi & Sepehr, 2009) شاخص عمق نوری آئروسل‌ها (Butt et al, 2017) و شاخص طوفان گردوغبار (O'Loingsigh et al, 2014) ارزیابی شده است. یکی از داده‌های کلیدی در مدل‌های پیش‌بینی غلظت$_{2.5}$ PM و PM_{10}، شاخص(AOD(Aerosol Optical Depth است (Clarke et al, 2001; Holben et al,2001). AOD نمایانگر میزان توزیع ذرات معلق در جو از سطح زمین تا ارتفاعات بالاتر است و میزان تضعیف تابش ورودی به جو به دلیل جذب و پراکنش ذرات معلق را اندازه‌گیری می‌کند. معمولاً این شاخص با استفاده از داده‌های ماهواره‌ای محاسبه می‌شود. مقادیر AOD بین ۰/۱ تا ۱ متغیر است؛ به‌طوری که در شرایط آسمان صاف، مقدار آن ۰/۱ می‌رسد و با افزایش غلظت هواویزها، AOD به سمت ۱ نزدیک‌تر می‌شود (حسینی تابش و همکاران، ۱۴۰۰) . همچنین، با توجه به نقطه‌ای بودن اطلاعات ایستگاه‌های زمینی، استفاده از ترکیب مشاهدات ماهواره‌ای با مدل‌های زمینی می‌تواند به‌طور کارآمدی به تعیین شاخص‌های کیفیت هوا با هزینه کم کمک کند. در حوزه مطالعات هواشناسی، پیش‌بینی غلظت PM_{10} با استفاده از تصاویر ماهواره‌ای به‌طور فزاینده‌ای مورد توجه قرار گرفته است. تحقیقات انجام‌شده در نقاط مختلف جهان نشان داده‌اند که عمق نوری آئروسل‌ها (AOD) می‌تواند به‌عنوان یک نشانگر معتبر برای تخمین غلظت‌های $PM_{2.5}$و PM_{10} عمل کند. در این زمینه، همبستگی قوی بین AOD و غلظت‌های مذکور وجود دارد که ضریب تبیین آن بین۰/۴ تا ۰/۹ و RMSE کمتر از ۲۶/۲۲ و ۷۱/۷۵ میکروگرم بر متر مکعب گزارش شده است. با این حال، این همبستگی در مقیاس جهانی ضعیف است و بنابراین کاربرد آن در سطح جهانی محدود می‌باشد (حسینی تابش و همکاران، ۱۴۰۰).

استفاده از روش‌های یادگیری ماشین در پیش‌بینی غلظتPM_{10} ءعنوان یک راهکار کارآمد برای مواجهه با چالش‌هایی نظیر عدم دقت در پیش‌بینی‌های سنتی، پیچیدگی‌های متغیرهای تأثیرگذار و تغییرات ناگهانی در شرایط محیطی مطرح شده است. این روش‌ها به پژوهشگران این امکان را می‌دهند که به‌طور خودکار متغیرهای ورودی مرتبط را شناسایی کرده و یک ساختار مدل بهینه را توسعه دهند. مدل‌های یادگیری ماشین برای استخراج روابط یا الگوها بین یک مجموعه متغیر و یک عامل هدف مفید هستند و می‌توانند به شناسایی الگوهای پنهان در داده‌ها کمک کنند(Olden et al. 2008; Naghibi et al. 2016) . با کاهش خطای پیش‌بینی و افزایش دقت نتایج، این روش‌ها می‌توانند به تصمیم‌گیری‌های بهتری در زمینه مدیریت کیفیت هوا و سلامت عمومی منجر شوند. این روش‌ها به‌ویژه در مناطقی که با چالش‌های زیست‌محیطی نظیر آلودگی شدید و تغییرات اقلیمی مواجه‌اند، بسیار کارآمدند و می‌توانند به بهبود سیاست‌های مدیریتی و پیش‌بینی‌های بلندمدت کمک کنند.(Kumar et al, 2021)

Thomas Plocoste and Sylvio Laventure (2023) هدف این مطالعه پیش‌بینی غلظت‌هایPM_{10} با استفاده از شش مدل یادگیری ماشین (ML) بوده است: رگرسیون بردار پشتیبانی(SVR) ، رگرسیون همسایه‌های(kNN) k ، رگرسیون جنگل تصادفی(RFR) ، رگرسیون تقویت تدریجی(GBR) ، رگرسیون توئیدی (TR) و رگرسیون ریج بیزی (BRR) . به‌طور کلی، نتایج نشان داده که GBR بهترین عملکرد را با استفاده از سه متغیر ورودی (دما، میزان بارش روزانه و PM_{10}) ارائه داده است($r = 0.7831$ ، $R^2 = 0.6132$، $MAE = 6.8479$، $RMSE = 10.4400$)و تمامی این نتایج مربوط به ویژگی‌های غلظت‌هاPM_{10} در منطقه کارائیب بوده است. در مطالعات دیگری، کاویانی‌راد و همکاران (۲۰۲۲) تأثیر تعامل بین آلاینده‌های هوا و پارامترهای محیطی را در سه شهر ایران با استفاده از تکنیک‌های یادگیری ماشین مورد بررسی و تحلیل قرار دادند. آن‌ها گزارش کردند که باوجود وجود رابطه معنی‌دار بین آلاینده‌های هوا و عوامل اقلیمی و گیاهی، مدل تدوین‌شده دقت پایینی در پیش‌بینی داشته است. نتایج ضریب تبیین برای آلاینده‌ها به صورت زیر است: ($R^2PM2.5 = 0.36$، $R^2PM10 = 0.27$، $R^2NO2 =$

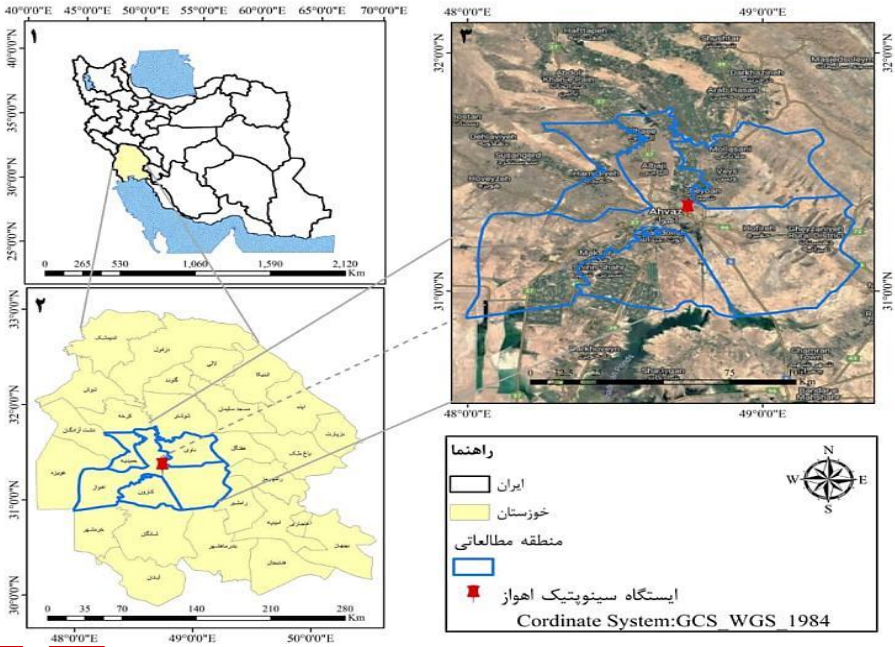
0.46، $R^2SO2 = 0.41$، $R^2O3 = 0.52$ و $R^2CO = 0.38$). حسینی تابش و همکاران (۱۴۰۰) به ارزیابی داده‌های سنجنده مودیس (MODIS) در پایش غلظت آلاینده‌های PM2.5 و PM10 با تأکید بر متغیرهای هواشناسی پرداختند. در این مطالعه، از داده‌های شش ایستگاه هواشناسی و آلودگی سنجی زمینی استفاده شد و مدل‌های خطی و غیرخطی برای برآورد غلظت هواویزها ارائه گردید. نتایج این پژوهش نشان داد که مدل رگرسیون خطی که شامل متغیرهای عمق نوری هواویزها، بارش ۲۴ ساعته، میانگین فشار بخار آب و ساعت آفتابی است، در مقیاس کل شهر تهران با ضریب تبیین ۰/۷۵ (R^2) و RMSE برابر با ۷/۳۴۷ $\mu g/m^3$ ، مناسب‌ترین مدل است . در این مدل، غلظت $PM_{2.5}$ با ساعت آفتابی رابطه معکوس و با سایر متغیرها رابطه مستقیم داشت. این محققان همچنین گزارش کردند که استفاده از متغیرهای هواشناسی و توجه به پدیده‌های جوی موجب بهبود عملکرد داده‌های سنجنده مودیس در برآورد غلظت آلاینده $PM_{2.5}$ می‌شود و مدل ارائه‌شده می‌تواند مکمل مناسبی برای ایستگاه‌های زمینی پایش آلودگی هوا باشد و نواقص آن‌ها را تا حد زیادی برطرف سازد. Bozdag et al. (2020) به بررسی غلظت ذرات PM_{10} با در نظر گرفتن مهم‌ترین پارامترهای آلودگی هوا شامل ذرات معلق(PM) ، دی‌اکسید گوگرد(SO_2) ، منواکسید کربن(CO) ، دی‌اکسید کربن (CO_2)، اوزون(O_3) ، اکسیدهای نیتروژن (NOx) و هیدروکربن‌ها (HC) پرداختند. داده‌ها از ۷ ایستگاه در استان آنکارا در ترکیه جمع‌آوری و با استفاده از الگوریتم‌های یادگیری ماشین شامل LASSO ، SVR، RF، kNN، XGBoostو ANN تحلیل شدند. در این مطالعه، غلظت ذرات PM_{10}و پارامترهای هواشناسی در سال‌های ۲۰۰۹ تا ۲۰۱۷ از ۶ ایستگاه در آنکارا به‌عنوان ورودی مورد استفاده قرار گرفت و غلظت ذرات PM_{10}در ایستگاه هفتم برای سال ۲۰۱۸ پیش‌بینی شد. نتایج نشان داد که الگوریتم ANN بهترین عملکرد را با R^2 برابر با ۰/۵۸، RMSEبرابر با ۲۰/۸ و MAE برابر با ۱۴/۴ ارائه داده است.. مسعودی و همکاران (۲۰۱۸) به ارزیابی کیفیت هوا در شهر اهواز، واقع در جنوب ایران، با استفاده از داده‌های مربوط به ذرات معلق (PM_{10}) پرداختند. در این مطالعه، اندازه‌گیری‌ها در دو مکان مختلف انجام شد و میانگین غلظت از سال ۲۰۰۹ تا ۲۰۱۰ برای هر ۲۴ ساعت به‌طور ماهانه و فصلی محاسبه گردید. نتایج نشان داد که بیشترین غلظتPM_{10} در ساعات صبح و کمترین غلظت آن در ساعات بعدازظهر مشاهده شد. همچنین، بیشترین غلظت ماهانهPM_{10} در ماه جولای و کمترین آن در ژانویه گزارش گردید، در حالی که بیشترین غلظت فصلی نیز مربوط به تابستان بود. محمودی سراب و همکاران (۱۳۹۷) رابطه بین متغیرهای آب و هوایی شامل دمای هوا (حداکثر، متوسط و حداقل)، رطوبت نسبی (حداکثر، متوسط و حداقل)، بارندگی روزانه، دید افقی، جهت و سرعت باد با داده های متغیر آلودگی هوا (PM_{10}) طی دوره آماری چهار ساله (۱۳۸۷ تا ۱۳۹۰) بررسی کردند. نتایج این مطالعه با استفاده از روابط همبستگی و برآورد PM_{10} با مدل آماری جنگل تصادفی نشان داد که متغیرهای دید افقی و دمای حداقل به ترتیب با ضریب همبستگی ۰/۳۷۶− و ۰/۳۴۹+ بیشترین همبستگی را با PM_{10} دارند. بارندگی نیز کمترین همبستگی (۰/۰۷۷−) را نشان داد. بر اساس نتایج مدل رگرسیون جنگل تصادفی، متغیر دید افقی به‌عنوان مهم‌ترین عامل تأثیرگذار در برآورد میزان PM_{10} و متغیر دمای حداقل در رتبه بعدی گزارش شد. این محققان همچنین ضریب تبیین (R^2) برابر با ۰/۴۷ در سطح معنی‌داری ۹۹ درصد به‌دست آوردند و نتیجه‌گیری کردند که می‌توان از داده‌های دید افقی و دمای حداقل برای پیش‌بینی میزان PM_{10} استفاده کرد و مدل رگرسیون جنگل تصادفی نتایج مناسبی را در این زمینه ارائه داد.

برتری این پژوهش در انتخاب بهینه‌ترین ترکیب از متغیرهای ورودی با هدف ارائه مدلی محلی با نتایج بهبودیافته نسبت به مطالعات گذشته در راستای سنجش و ارزیابی عملکرد بهترین مدل برای پیش‌بینی غلظت PM_{10} می‌باشد. هدف اصلی این مطالعه بررسی میزان غلظت ذرات معلقPM_{10} با استفاده از ترکیب PM_{10} با پارامترهای ورودی متفاوت شامل: ۱) شاخص AOD ۲) پارامترهای هواشناسی ۳) ترکیبی از شاخص AOD و پارامترهای هواشناسی در شهر اهواز بوده است. در پژوهش حاضر، نمونه‌ها به‌صورت مکرر توسط روش‌های پایه‌ای رگرسیون مانند، رگرسیون بردار پشتیبان (SVR)، رگرسیون درخت تصمیم (DTR)، رگرسیون جنگل تصادفی (RFR)، رگرسیون افزایشی گرادیان (GBR)، شبکه عصبی مصنوعی (ANN) و روش‌های یکپارچه شامل الگوریتم AdaBoost (Ada) آموزش دیده شدند (Hou et al, 2016 ., 2014;Song yand lu., 2015; Michele Fratello and Roberto Tagliaferri, 2019; Jiang et al., 2021; Xia et al, 2016; Li et al,2022). نتایج الگوریتم‌های یادگیری ماشین با استفاده از شاخص‌های آماری مقایسه شده، توانایی آن‌ها برای پایش غلظت PM_{10} ارزیابی شده است.

مواد و روشها

منطقه موردمطالعه

شهر اهواز، مرکز استان خوزستان، بزرگ‌ترین شهر جنوب غربی ایران است که در عرض جغرافیایی ۳۱ درجه و ۱۹ دقیقه شمالی و طول جغرافیایی ۴۸ درجه و ۴۰ دقیقه شرقی واقع شده و ارتفاع آن حدوداً ۲۰ متر بالاتر از میانگین سطح دریا می‌باشد (شکل ۱). میزان بارندگی سالانه اهواز حدود ۲۳۰ میلی‌متر است. این شهر دارای آب و هوای خشک و همواره یکی از گرم‌ترین شهرهای کره زمین در طول تابستان است. بر اساس مطالعاتی که توسط اداره کل زمین‌شناسی و اکتشافات معدنی انجام شده است، ۹ درصد از مساحت دشت خوزستان که معادل ۳۴۹۲۵۴ هکتار است، منشأهای تولید ریزگرد در استان خوزستان هستند. از این میان، ناحیه جنوب و جنوب شرق اهواز با بیشترین مساحت از بین کانون‌های منشأ ریزگرد در استان خوزستان شناخته شده‌اند (اژدری، ۱۳۹۶). بر اساس گزارش‌های سازمان بهداشت جهانی، میزان سالانه آلودگی در شهر اهواز ۳۷۲ میکروگرم در هر مترمکعب است. شهر اهواز با وجود کارخانه‌های بزرگ صنعتی، تأسیسات اداری و به‌خصوص خشکسالی‌های اخیر، یکی از مهم‌ترین شهرهای آلوده ایران و به عنوان آلوده‌ترین شهر دنیا، مقام اول را بین ۱۱۰۰ شهر کسب کرده است (Sadeghi & Khaksar, 2015).



شکل ۱. موقعیت جغرافیایی منطقه موردمطالعه

داده‌های مورد استفاده

در این تحقیق، از پارامترهای هواشناسی شامل حداقل دید افقی، ماکزیمم و مینیمم سرعت باد، دما (ماکزیمم، مینیمم و میانگین)، فشار میانگین، رطوبت (مینیمم و میانگین روزانه)، ساعت آفتابی، تابش ۲۴ ساعته، تبخیر، میانگین رطوبت نسبی در (ساعت ۳ گرینویچ، ساعت ۹ گرینویچ، ساعت ۱۵ گرینویچ) ماکزیمم فشار تراز ایستگاه، مینیمم فشار سطح ایستگاه، میانگین فشار بخار، عمق نوری آئروسل (AOD) و میزان غلظت ذرات کوچکتر از ۱۰ میکرومتر (PM_{10}) مربوط به آمار بلندمدت (۲۰۱۴-۲۰۰۸) ایستگاه هواشناسی مربوط به فرودگاه اهواز و اداره محیطزیست

استفاده شده است. شاخص عمق نوری آئروسل (AOD) محصول روزانه باند ۴۷۶ نانومتر سنجنده مودیس در محیط پلتفرم گوگل ارث انجین مربوط به محدوده ایستگاه سینوپتیک فرودگاه شهر اهواز با مشخصات 'MODIS/006/MCD19A2_GRANULES' ، جهت تخمین میزان ذرات کوچکتر از ۱۰ میکرومتر (PM_{10}) استفاده شده است(Cao et al., 2015; Asl et al., 2019; Bozdag et al., 2020). مجموعه داده‌های اخذ شده در این پژوهش مربوط به ۲۷۷ نمونه است که به‌صورت روزانه از سال‌های ۲۰۰۸ تا ۲۰۱۴ میلادی جمع‌آوری شده‌اند. این داده‌ها از سال ۱۹۹۴ تاکنون موجود بوده‌اند؛ اما از آنجا که داده‌های محیط‌زیست و هواشناسی در بازه سال‌های ۲۰۰۸ تا ۲۰۱۴ هم‌پوشانی زمانی داشتند، این دوره مورد بررسی قرار گرفت. روزهایی که نقص آماری داشتند در محاسبات در نظر گرفته نشده‌اند.

روش پژوهش

در این پژوهش با استفاده از زبان برنامه‌نویسی Python ، مدل‌هایی برای تخمین میزان غلظت ذرات معلق در هوا با استفاده از الگوریتم‌های یادگیری ماشین مختلف توسعه داده شد. در شکل ۲، فرایند پژوهش نشان داده شده است. برای توسعه مدل‌های مختلف، مجموعه داده‌ها به دو گروه تقسیم شدند: گروه آموزش، برای آموزش مدل‌ها و مجموعه داده‌های آزمون، جهت ارزیابی عملکرد، در نسبت ۷۰–۳۰. به دلیل متفاوت بودن محدوده‌ها و واحدهای متغیرهای اولیه، پارامترهای ورودی استاندارد شده‌اند تا یک توزیع نرمال معمولی را بازتاب دهند. در این راستا، تابع 'Standard Scaler'از کتابخانه scikit-learn استفاده شد. این تابع مدل را تغییر داده تا اطمینان حاصل شود که میانگین در مرکز صفر و انحراف معیار به یک تنظیم شده باشد. این فرایند تمام ویژگی‌ها را به یک مقیاس متناسب تبدیل می‌کند و همگرایی را در طول مرحله آموزش بهبود می‌بخشد. جهت استاندارد کردن داده‌ها، از Z-Score رابطه (۱) استفاده شد.

رابطه ۱)

$$z = (x - \mu)/\sigma$$

که در آن، :Z مقادیر استانداردشده داده‌ها، X : ماتریس داده‌های ورودی، μ : میانگین داده‌ها، σ : واریانس برای هر متغیر است (Jolliffe, 2002). پیش‌بینی عملکرد یادگیری ماشین با استفاده از شاخص‌های آماری دقت و صحت شامل ضریب همبستگی (R)، ضریب تبیین (R^2)، خطای مطلق میانگین (MAE) و ریشه میانگین مربعات خطا (RMSE) و شاخص توافق (IOA) ارزیابی شد. ضرایبR^2 و IOA به ترتیب نشان‌دهنده میزان رابطه خطی بین مقادیر مشاهده‌شده و پیش‌بینی‌شده و درجه تقارب میان این دو مجموعه داده هستند. R^2 نمایانگر نسبت واریانس مدل که به درستی توضیح داده شده است و IOA (Index of Agreement) معیاری است برای ارزیابی دقت پیش‌بینی‌ها نسبت به مقادیر واقعی. عملکرد مدل زمانی معنی‌دار تلقی می‌شود که مقادیر IOA و R^2 نزدیک به ۱ باشند؛ این به معنای آن است که پیش‌بینی‌ها به‌خوبی با مشاهدات واقعی مطابقت دارند. مقادیر نزدیک به ۰ نشان‌دهنده عملکرد ضعیف مدل و عدم تطابق با داده‌های واقعی است، بنابراین تجزیه و تحلیل دقیق این ضرایب می‌تواند به بهبود مدل‌های پیش‌بینی کمک کند. ضرایب MAE و RMSE برای اندازه‌گیری خطاهای پیش‌بینی مدل استفاده شده‌اند که در شرایط بهینه نزدیک به صفر هستند. پارامترهای ذکر شده با استفاده از فرمول‌های زیر محاسبه شدند.

(Ul-Saufie et al., 2011. Willmott et al., 2012. Fu et al., 2015)

رابطه۲)

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{n=1}^{n}(\hat{y}\,i - yi)^2}$$

رابطه۳)

$$MAE = \frac{1}{n}\sum_{n=1}^{n} |\hat{y}\,i - yi|$$

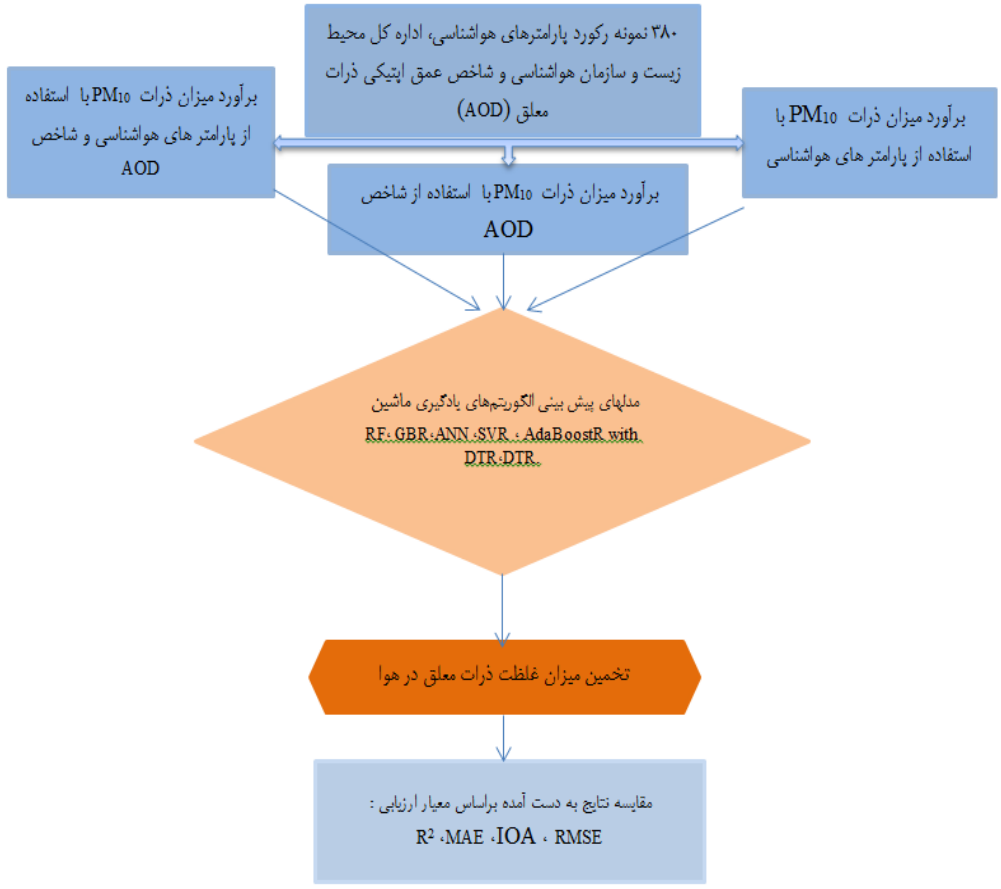
رابطه۴)

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}\,i - yi)^2}{\sum_{i=1}^{n}(yi - \bar{y})^2}$$

رابطه۵)

$$IOA = 1 - \frac{\sum_{i=1}^{n}(\hat{y}\,i - yi)2}{\sum_{i=1}^{n}(|\hat{y}\,i - \bar{y}| + |y_i - \bar{y}|)2}$$

که در آن، "y_i" به مقدار واقعی PM_{10} و "$\hat{y}_i$" مقدار PM_{10} پیش‌بینی‌شده برای مشاهده iام است که از مدل به‌دست‌آمده است، $\bar{y}$ میانگین مقدار واقعی و n تعداد نمونه‌ها است. (Ul-Saufie et al., 2011. Willmott et al., 2012. Fu et al., 2015)



شکل ۲. فلوچارت فرآیند اجرای طرح پژوهش

الگوریتم‌های یادگیری ماشین

چندین تکنیک پیش‌بینی یادگیری ماشین برای سری‌های زمانی وجود دارد. در این مطالعه، شش روش قوی استفاده شد: رگرسیون بردار پشتیبان (SVR)، رگرسیون درخت تصمیم (DTR)، رگرسیون جنگل تصادفی (RFR)، رگرسیون افزایشی گرادیان (GBR)، شبکه‌های عصبی مصنوعی(ANN) و ترکیب رگرسیون درخت تصمیم با AdaBoostRegressor تمام این روش‌ها در کتابخانه scikit-learn در زبان پایتون توسعه داده شده‌اند.

رگرسیون افزایشی گرادیان(GBR)

ماشین‌های یادگیری گرادیانی یک خانواده از تکنیک‌های قدرتمند یادگیری ماشین هستند که نتایج خوبی را در یک طیف وسیعی از برنامه‌های عملی نشان داده‌اند. روش Boosting یک روش است که چندین مدل پایه را ترکیب می‌کند تا یک خوشه ایجاد کند که عملکرد آن ممکن است به طرز قابل توجهی بهتر از هر مدل پایه دیگری باشد. به این منظور، مدل GBR مراحل را با استفاده از چندین درخت تصمیم با عملکرد ضعیف ساخته و آن‌ها را ترکیب می‌کند تا یک مدل پرکارایی به وجود آید. برای هر مرحله، GBR معادله زیر را حل می‌کند (Keprate and Ratnayake, 2017).

$$f_m(x) = f_{m-1}(x) + v_m h_m(x)$$ رابطه۶)

در اینجا $f_{m-1}(x)$ پیش‌بینی قبلی است، v_mنرخ یادگیری است که اثر هر درخت قبلی را بر درخت بعدی کاهش می‌دهد (معمولاً $0 < v_m < 1$)و h_m تابعی است که بر پایه باقیمانده‌هایی که از یک تابع خطا ساخته شده‌اند، ایجاد می‌شود.

رگرسیون بردار پشتیبان(SVR)

SVR یک الگوریتم یادگیری ماشین مبتنی بر بردار پشتیبان برای مسائل رگرسیون است. هدف آن کاهش خطا با تعیین هایپرپلین و کمینه کردن فاصله بین مقادیر پیش‌بینی شده و مشاهده شده است که با حل معادله زیر انجام می‌شود(Singh et al, 2020). پارامترهای اصلی SVR شامل C و γ هستند که با تابع هسته ارتباط دارند. ابتدا، C با کمینه کردن معادله رگرسیون زیر تعیین می‌شود.

$$\frac{1}{2}||w||^2 + C\sum_{i=1}^{N}\epsilon_i + \epsilon_i^* \rightarrow min$$ رابطه۷)

در اینجا w مجموع وزن‌هایی است که این مدل را تنظیم می‌کند، و C پارامتر ثابت برای تعریف ($C > 0$) است که تعیین کننده تعادل بین هایپرپلین و خطای تخمین است.

رگرسیون جنگل تصادفی(RFR)

الگوریتم RF که توسط بریمن توسعه یافت (Breiman et al,2001) یکی از روش‌های پرکاربرد یادگیری ماشین برای ساخت مدل‌های پیش‌بینی است. رگرسیون جنگل تصادفی شامل محاسبه میانگین پیش‌بینی‌های به‌دست‌آمده با در نظر گرفتن تمام پیش‌بینی‌های حاصل از درخت‌های تصمیم است. همچنین، طبقه‌بندی جنگل تصادفی بر اساس روش بسته‌بندی انجام می‌شود، اما تخمین نهایی با انتخاب متداول‌ترین دسته پاسخ به‌جای استفاده از همه نتایج به‌دست‌آمده انجام می‌گیرد (Ghunimat et al,2023).

درخت‌های تصمیم ایجاد شده توسط طبقه‌بندی جنگل تصادفی با استفاده از ناخالصی جینی(Gini Impurity) آموزش داده می‌شوند. این ناخالصی برای شکافتن شاخه‌ها و انتخاب گره‌هایی که عدم قطعیت در درخت‌های تصمیم را کاهش می‌دهند، مورد استفاده قرار می‌گیرد. بنابراین، بهترین تقسیم با به حداقل رساندن ناخالصی جینی هنگام تقسیم هر گره انتخاب می‌شود. ناخالصی جینی یک گره nبه عنوان فرمول زیر تعریف می‌شود.

$$I_G(n) = 1 - \sum P_i^2$$

رابطه۸)

که P_i احتمال مربوط به کلاس i در یک گره معین است. مقدار کم ناخالصی جینی به این معنی است که گره‌ها خالص هستند و احتمال اشتباه در طبقه‌بندی نمونه‌ها وجود ندارد.(Chai,2023)

رگرسیون درخت تصمیم (DTR)

درخت تصمیم (DTR) یک تکنیک یادگیری ماشین قابل تنظیم و قابل فهم برای کاربردهای طبقه‌بندی و رگرسیون است. این درخت با پیش‌بینی در برگ‌های خود، داده‌ها را به‌صورت بازگشتی بر اساس مقادیر ویژگی‌ها تقسیم می‌کند. مجموعه داده کامل در گره ریشه قرار دارد، زمانی که فرآیند شروع می‌شود. سپس بهترین ویژگی برای تقسیم داده‌ها انتخاب می‌شود، تا اطلاعات را بیشینه کند. تقسیم از طریق بازگشتی انجام می‌شود تا مجموعه‌ای از شرایط توقف برآورده شود. پیش‌بینی در هر گره برگ، کلاس بیشترین (برای طبقه‌بندی) یا میانگین مقدار هدف (برای رگرسیون) است. سادگی و قابل فهم درخت تصمیم، آن را یک انتخاب محبوب در حوزه‌های مختلف می‌کند. پیش‌بینی‌ها بر اساس مجموعه‌ای از قوانین if-else که در طول ساخت درخت تعیین شده‌اند، استوار می‌باشد (Li et al,2022; Feng et al,2022).

رگرسیون به دنبال پیش‌بینی یک مقدار پیوسته برای یک مجموعه داده با بررسی یک نمونه آموزشی از داده‌ها است (Ren et al,2016):

رابطه۹)

$$O = f(x, \theta), O \in R$$

که در آن، x مشاهده جدید، O خروجی، f(.) تابع رگرسیون و θ مجموعه پارامترهای تابع رگرسیون است. درخت تصمیم (DT) برای رگرسیون مشابه درخت‌های طبقه‌بندی است با این تفاوت که در برگ‌ها مقادیر یا مدل‌های تکه‌ای وجود دارد و نه برچسب‌های کلاس. این مقادیر می‌توانند نتیجه هر آزمایش یا نتیجه هر عملی باشند.

تکنیک یادگیری شناخته شده (AdaBoost)

تکنیک یادگیری شناخته شده به نام AdaBoost، یا افزایش تطبیقی، چندین یادگیرنده ضعیف مانند درختان تصمیم را به یک مدل پیش‌بینی قدرتمند ترکیب می‌کند. الگوریتم با افزایش برچسب وزن نقاط داده اشتباه دسته‌بندی شده در هر دور به صورت تکراری عمل می‌کند، و باعث تولید یادگیرندگان ضعیف جدیدی می‌شود که بر روی نمونه‌های قبلاً اشتباه دسته‌بندی شده تمرکز می‌کنند. سپس بر اساس دقت آن‌ها در انجام پیش‌بینی‌ها ، به هر یادگیرنده ضعیف وزنی اختصاص داده می‌شود. در گروه نهایی، پیش‌بینی‌های تمام یادگیرنده‌های ضعیف ترکیب می‌شوند و دسته‌بندهای دقیق‌تر در پیش‌بینی نهایی نقش بیشتری دارند. AdaBoost در کنترل مجموعه داده‌های پیچیده و دستیابی به دقت بالا مؤثر است، به ویژه در شرایطی کارآمد است که یادگیرنده‌های ضعیف فردی به خوبی عمل نکنند. به طور کلی، AdaBoost یک انتخاب محبوب برای وظایف طبقه‌بندی است و انعطاف‌پذیری و توانایی بهبود عملکرد یادگیرندگان ضعیف آن را به یک اضافه‌کردن ارزشمند به جعبه ابزارهای روش‌های توامان می‌کند(Feng et al ,2020;DeRousseau et al, 2019)

این روش (Oyewola et al,2021) با برازش یک رگرسور به مجموعه داده اولیه آغاز می‌شود. سپس نسخه‌های اضافی از رگرسور را به همان مجموعه داده برازش می‌کند. تنها استثنا این است که وزن‌های نمونه‌ها بر اساس خطای آخرین پیش‌بینی تغییر می‌کند. فرض کنید یک مجموعه داده $S = (x1, y1), \ldots, (xn, yn)$ که از یک سری زمانی استخراج شده است. این مجموعه داده شامل n جفت مشاهدات است و هر مشاهده وزنی به نام wi دارد. احتمال اینکه یک مشاهده در مجموعه آموزش در تکرار k گنجانده شود، برای هر مشاهده i بر اساس وزنی که به آن داده شده، تعیین می‌شود. سپس مجموع وزنی احتمالات برای محاسبه میانگین ضرر (l_k) مدلk بر روی تمام مشاهدات i استفاده می‌شود. فرمول‌های ریاضی میانگین ضرر (l_k)و احتمال p^k در روابط (۱۰) تا (۱۲) نشان داده شده است.

رابطه۱۰)

$$p_k = \frac{w_i}{\sum w_i}$$

رابطه۱۱)

$$l_k = \sum_{i=1}^{n} l_k p_k$$

رابطه۱۲)

$$w_i^{k+1} = w_i^k \beta_k (1 - l_k)$$

شبکه عصبی مصنوعی (ANN)

Artificial Neural Networks یا ANNs، یک دسته از الگوریتم های یادگیری ماشین (یا به دقت یادگیری عمیق) هستند. شبکه های عصبی معمولا در زبان برنامه‌نویسی پایتون برای پیاده سازی آن‌ها از کتابخانه Tensorflow استفاده می‌شود. نحوه عملکرد آن به این صورت است که در وسط یک ورودی و یک خروجی ANN ،لایه‌های پنهانی وجود دارد که داده‌های اطلاعات را پردازش می‌کنند و نتیجه را به لایه بعدی ارسال می‌کنند، و هر لایه به لایه بعدی، تا رسیدن به لایه نهایی که همان خروجی است ادامه پیدا می‌کند. در واقع می توان آنها را به عنوان یک لایه خوشه بندی و طبقه بندی در بالای داده هایی که ذخیره و مدیریت می شوند در نظر گرفت. شبکه‌های عصبی به گروه‌بندی داده‌های بدون برچسب بر اساس شباهت‌های میان ورودی‌های نمونه کمک می‌کنند، زمانی که یک مجموعه داده برچسب‌دار برای آموزش دارند داده‌ها را طبقه‌بندی می‌کنند. این گره‌ها معمولاً در یک سیستم کاملاً متصل از سه یا چند لایه، یعنی لایه ورودی، لایه (های پنهان) و لایه خروجی روی هم قرار می‌گیرند. این نوع معماری شامل تعداد لایه‌های مخفی، تعداد گره‌ها در هر لایه و تابع فعال‌سازی مرتبط با هر گره است. (Gardner, M. and Dorling, S,1998)

لایه‌های ورودی شبکه‌های عصبی مصنوعی (ANN) با وزن‌های مربوطه و خروجی به صورت ریاضی، به شکل زیر نمایش داده شده است:

$$\sum = (y_i \times w_i) + (y_2 \times w_2) + \cdots + (y_n \times w_n)$$

رابطه۱۳)

جایی که yiورودی‌ها و wiوزن‌های تخصیص‌یافته هستند. مقادیر وزن‌ها نقش حیاتی در خروجی ایفا می‌کنند. بردارهای سطری برای ورودی به صورت $y=[y_1,y_2,\ldots,y_n]$ و $w=[w_1,w_2,\ldots,w_n]$نمایش داده می‌شوند و ضرب نقطه‌ای آن‌ها به صورت زیر نمایش داده می‌شود :

$$y.w = (y_i \times w_i) + (y_2 \times w_2) + \cdots + (y_n \times w_n)$$

رابطه۱۴)

ضرب نهایی نقطه‌ای به صورت زیر نمایش داده می‌شود:

$$\sum = y.w$$

رابطه۱۵)

با اضافه کردن بایاس cبه مجموع ضرب نقطه‌ای، مقدار pبه دست می‌آید و معادله به صورت زیر نوشته می‌شود:

$$p = y.w + c$$

رابطه۱۶)

مقدار pمی‌تواند به تابع فعال‌سازی منتقل شود؛ در اینجا ما تابع فعال‌سازی سیگموید را پیاده‌سازی می‌کنیم:

$$\bar{x} = \sigma(p) = \frac{1}{1+e^{-p}}$$

رابطه۱۷)

که در آن σتابع فعال‌سازی و مقدار پیش‌بینی شده $\bar{x}$ارزش خروجی است.(Mustaqeem et al,2023)

نتایج و بحث

پس از جمع آوری داده ها و حذف نواقص مربوط به پارامترهای اقلیمی، شاخص (AOD) و غلظت ذرات کوچکتر از ۱۰ میکرومتر (PM_{10}) مجموع داده‌های به‌کاربرده‌شده در این تحقیق ۲۷۷ نمونه در نظر گرفته شد. نتایج آمار توصیفی پارامترهای اقلیمی ، عمق نوری آئروسل (AOD) و میزان غلظت ذرات کوچکتر از ۱۰ میکرومتر (PM_{10}) مربوط به آمار بلندمدت (۲۰۱۴–۲۰۰۸) ایستگاه هواشناسی مربوط به فرودگاه اهواز و

اداره محیطزیست در جدول (۱) نشان داده شده است. اطلاعات مربوط به غلظت آلاینده‌ها نیز از اداره محیطزیست شهرستان اهواز به‌طور روزانه و ساعتی دریافت شد. نتایج مربوط به بررسی توصیفی داده‌ها در شهرستان اهواز و برای محدوده زمانی تعریف‌شده نشان داد که حداقل، حداکثر و میانگین میزان ذرات PM_{10} به ترتیب برابر با ۳۳/۷۱، ۱۴۹۱/۶۷و ۱۸۶/۳ میکروگرم بر مترمکعب بوده است. طبق استاندارد سازمان حفاظت محیط زیست ایران، میزان غلظت حد استاندارد ذرات (PM_{10}) ۱۵۰ میکروگرم بر مترمکعب گزارش شده است. همچنین، حداقل، حداکثر و میانگین میزان شاخص (AOD) به ترتیب برابر با ۰/۰۳، ۴ و ۰/۲۹ بوده است (جدول ۱). محدوده شاخص AOD بین (۵ تا ۰/۰۵-) می‌باشد. با افزایش غلظت ذرات گردو غبار و PM شاخص AOD افزایش می‌یابد (Lanzaco et al., 2016).

به‌منظور بررسی ضرورت وجود هریک از متغیرهای مستقل اقلیمی در مدل، آزمون هم‌خطی انجام شد تا مدلی ایجاد شود که عملکرد بهتری با تعداد کمتری از متغیرها داشته باشد. هم‌خطی (Collinearity) به وضعیتی اشاره دارد که در آن یک متغیر توصیفی در رگرسیون چندگانه، رابطه خطی با یک یا چند متغیر دیگر دارد. به عبارت دیگر، می‌توان این متغیر را به‌عنوان ترکیبی خطی از سایر متغیرها در نظر گرفت. از سوی دیگر، هم‌خطی چندگانه (Multicollinearity) وضعیتی است که در آن چندین متغیر توصیفی دارای رابطه‌های خطی با یکدیگر هستند و می‌توان آنها را به‌عنوان ترکیب‌های خطی از یکدیگر بیان کرد. متغیرهای مهم با استفاده از عامل تورم واریانس (VIF) و تلرانس طبق روابط ۱۸ و ۱۹ در نرم‌افزار پایتون انتخاب شدند. بر این اساس متغیرهای با VIF > 10 و تلرانس کمتر از ۰/۱ حذف شدند (Shrestha ,2020). طبق جدول (۲)، هفت متغیر برای اجرای رویکرد مدل‌سازی انتخاب شده که شامل: حداقل دید افقی، بیشینه سرعت باد، میانگین ابرناکی روزانه، ساعت آفتابی، میزان تابش کلی ۲۴ ساعته، شاخص AOD و تبخیر می باشند.

رابطه ۱۸)
$$VIF = \frac{1}{1-r^2}$$

رابطه ۱۹)
$$\text{تلرانس} = \frac{1}{VIF}$$

که در آن r^2 نشان‌دهنده ضریب تبیین مدل رگرسیونی با متغیرهای مستقل مدل است.

جدول ۱. آماره توصیفی داده‌های پژوهش

نام متغیر	واحد	تعداد داده	حداقل آماره	حداکثر آماره	میانگین آماره	انحراف معیار	چولگی	کشیدگی
حداقل دید افقی (vvmin)	m	۲۷۷	۱۰۰	۱۰۰۰۰	۶۰۴۶/۶	۲۹۰۸/۲	-۰/۲	-۱/۰
ماکزیمم سرعت باد(ff_max)	m/s	۲۷۷	۰	۱۱	۳/۷	۱/۷	۰/۸	۲/۱
دمای ماکزیمم (tmax)	°C	۲۷۷	۱۰/۶	۴۴/۲	۲۹/۲	۸/۳	۰/۰۳	-۱/۳
دمای مینیمم (tmin)	°C	۲۷۷	-۰/۶۰	۳۷	۱۵/۴	۶/۲	-۰/۲	-۰/۹
دمای میانگین (tm)	°C	۲۷۷	۵	۳۴/۲	۲۱/۸	۷/۲	-۰/۰۴	-۱/۲
فشار میانگین سطح ایستگاه (p0m)	hPa	۲۷۷	۱۰۰۱/۹	۱۰۲۳/۸	۱۰۱۲/۳	۴/۸	-۰/۰۲	-۰/۷
میانگین ابرناکی روزانه (nm)	-	۲۷۷	۰	۵/۵	۰/۹	۱/۲	۱/۵	۱/۵
مینیمم رطوبت نسبی (umin)	٪	۲۷۷	۷	۷۰	۲۶/۵	۱۳/۳	۱/۰	۰/۷
میانگین رطوبت نسبی (um)	٪	۲۷۷	۱۴/۸	۸۵/۸	۴۶/۴	۱۶/۰	۰/۲	-۰/۵
ساعت آفتابی (sshn)	H	۲۷۷	۰	۱۱/۲	۸/۸	۱/۷	-۲/۳	۷/۶
میزان تابش کلی ۲۴ ساعته (radglo24)	$j/cm^2/day$	۲۷۷	۰	۸۳۰۳	۱۴۸۰/۱	۵۸۵/۳	۵/۴	۶۶/۸
تبخیر (evt)	mm	۲۷۷	۰/۳	۱۶/۵	۶/۰	۳/۲	۰/۷	۰/۱
میانگین رطوبت نسبی در ساعت ۳(hrel_03)	٪	۲۷۷	۲۰	۹۸	۶۴/۶۶	۱۹/۱۴	-۰/۳۰	-۰/۸۷

میانگین رطوبت نسبی در ساعت ۹(hrel_09)	٪	۲۷۷	۷	۹۳	۳۲/۹	۱۵/۶	۰/۹	۰/۵
میانگین رطوبت نسبی در ساعت ۱۵(hrel_15)	٪	۲۷۷	۸	۸۶	۳۸/۹	۱۷/۰	۰/۵	−۰/۲
ماکزیمم فشار سطح ایستگاه (p0max)	hPa	۲۷۷	۱۰۰۲/۷	۱۰۲۵/۶	۱۰۱۴/۰	۴/۸	−۰/۱	−۰/۶
مینیمم فشار سطح ایستگاه (p0min)	hPa	۲۷۷	۱۰۰۰/۶	۱۰۲۲/۵	۱۰۱۰/۹	۴/۸	۰/۰۱	−۰/۷
میانگین فشار بخار (ewsm)	hPa	۲۷۷	۸/۹	۵۵/۶	۲۹/۳	۱۲/۶	۰/۴	−۱/۲
عمق نوری آئروسل (AOD)	-	۲۷۷	۰/۰۳	۴	۰/۳	۰/۳	۸/۶	۱۰۲/۰
ذرات معلق با قطر کوچکتر از ۱۰ میکرون (PM_{10})	$\mu g/m^3$	۲۷۷	۳۳/۷	۱۴۹۱/۷	۱۸۶/۳	۱۸۵/۶	۴/۵	۲۵/۰
تعداد داده معتبر		۲۷۷						

جدول ۲. ضرائب آزمون هم‌خطی

نام متغیر	عامل تورم واریانس (VIF)	تلرانس	R^2
حداقل دید افقی (vvmin)	**۱/۷۴۸۸**	**۰/۵۷۱**	**۰/۴۲۶۲**
ماکزیمم سرعت باد(ff_max)	**۱/۶۳۱۷**	**۰/۶۱۲**	**۰/۳۸۷۲**
دمای ماکزیمم (tmax)	۱۲۷/۰۹۲۷	۰/۰۰۷۸	۰/۹۹۲۱
دمای مینیمم (tmin)	۴۶/۱۱۶۹	۰/۰۲۱۶	۰/۹۷۸۳
دمای میانگین (tm)	۲۹۴/۵۵۳۰	۰/۰۰۳۳	۰/۹۹۶۳
فشار میانگین سطح ایستگاه (p0m)	۴۴۲/۱۸۲۵	۰/۰۰۲۲	۰/۹۹۷۷
میانگین ابرناکی روزانه (nm)	**۲/۳۶۱۹**	**۰/۴۲۳۳**	**۰/۵۷۶۶**
مینیمم رطوبت نسبی (umin)	۲۲/۰۷۱۹	۰/۰۴۵۳	۰/۹۵۴۷
میانگین رطوبت نسبی (um)	۵۸/۲۰۷۹	۰/۰۱۷۱	۰/۹۸۲۸
ساعت آفتابی (sshn)	**۴/۸۰۲۶**	**۰/۲۰۸۲**	**۰/۷۹۱۸**
میزان تابش کلی ۲۴ ساعته (radglo24)	**۱/۳۲۱۹**	**۰/۷۵۶۴**	**۰/۲۴۳۵**
تبخیر (evt)	**۳/۵۷۰۱**	**۰/۲۸۰۰**	**۰/۷۱۹۹**
میانگین رطوبت نسبی در ساعت ۳(hrel_03)	۱۳/۵۹۳۵	۰/۰۷۳۵	۰/۹۲۶۴
میانگین رطوبت نسبی در ساعت ۹(hrel_09)	۱۵/۶۶۳۱	۰/۰۶۳۸	۰/۹۳۶۲
میانگین رطوبت نسبی در ساعت ۱۵(hrel_15)	۱۲/۷۵۳۹	۰/ ۰۷۸۴	۰/۹۲۱۶
ماکزیمم فشار سطح ایستگاه (p0max)	۱۴۳/۶۰۶۰	۰/۰۰۶۹	۰/۹۹۳۰
مینیمم فشار سطح ایستگاه (p0min)	۱۱۹/۴۰۵۸	۰/۰۸۳	۰/۹۹۱۶
میانگین فشار بخار (ewsm)	۴۹/۰۰۱۲	۰/۰۲۰۴	۰/۹۷۹۶
عمق نوری آئروسل (AOD)	**۲/۷۴۸۳**	**۰/۳۶۳۸**	**۰/۶۳۶۱**

برآورد میزان ذرات PM_{10} با استفاده از پارامترهای هواشناسی و شاخص AOD

پس از آزمایش‌های گسترده پارامترهای max_depth : حداکثر عمق درخت تصمیم، random_state : برای تکرارپذیری نتایج، tree number: تعداد بهینه درختان، max_features: استفاده از همه ویژگی‌ها، min_samples_split: حداقل نمونه‌ها برای انشعاب، n_iter_no_change: تعداد تکرار بدون بهبود، learning_rate: نرخ یادگیری، kernel: کرنل تابع پایه شعاعی، C: تنظیم پارامتر، epochs: تعداد دوره‌های آموزش، batch_size: اندازه دسته‌های آموزشی، Loss Function: تابع خطای میانگین مربعات، Optimization: الگوریتم بهینه‌سازی، Activation

function: تابع فعال‌سازی، برای ۶ الگوریتم مختلف یادگیری ماشین که بهترین نتایج را ارائه دادند به شرح ذیل بهینه شدند(جدول ۳). بهینه‌سازی هایپرپارامترها در این سناریو و دیگر سناریوها از روش نزول گرادیان(Gradient Descent) استفاده شده است.

در مدلDTR ، حداکثر عمق ۵ موجب آموزش سریع مدل و جلوگیری از overfitting می‌شود. همچنین، استفاده از "best" به عنوان بهترین تقویت‌کننده برای هر مرحله، امکان پیش‌بینی‌های دقیق‌تری را با تمرکز بر خطاهای پیشین فراهم می‌کند. مقدار ثابت random_state این امکان را می‌دهد که نتایج در تکرارهای مختلف قابل بررسی باشند برایRFR ، پارامترها با توجه به تعادل بین زمان پردازش و دقت انتخاب شدند. به این ترتیب، برای بهترین دقت، ۹۰ درخت در جنگل مشخص شد و عمق حداکثر تا زمانی که تمامی برگ‌ها خالص بودند، تعیین گردید. همچنین، با تعیین random_state به مقدار ۱۳، هر بار که مدل اجرا شود، نتایج مشابهی تولید خواهد شد. پارامترmax_features به تعیین حداکثر تعداد ویژگی‌هایی که در هر تقسیم از درختان تصمیم استفاده می‌شود، کمک می‌کند. مقدارmax_features=1.0 به این معنی است که در هر تقسیم، از تمامی ویژگی‌ها استفاده خواهد شد و با تنظیم min_samples_split به ۲، مدل به تعداد کافی داده برای تقسیم گره‌ها نیاز دارد و از تقسیم‌های غیرضروری جلوگیری می‌کند. برایAdaboost+DTR ، همان پارامترهای DTR به کار رفته است. در GBR نیز به این صورت است که ۹۰ درخت با نرخ یادگیری ۰/۰۵انتخاب شد تا تأثیر درخت اول در درخت تصمیم کاهش یابد. همچنین، n_iter_no_change به جلوگیری از زمان‌های طولانی آموزش و همچنین overfitting کمک می‌کند، به این معنی که اگر دقت متوقف شود، مدل به طور خودکار متوقف می‌شود. در SVR ، از هسته RBF با مقدار C حدود ۱۰ استفاده شد. نهایتا در ANN ، epochs تعداد مراحل آموزش شبکه عصبی را تعیین می‌کند و batch_size تعداد نمونه‌هایی است که برای هر بار به‌روزرسانی وزن‌ها استفاده می‌شود. برای محاسبه خطا و به‌روزرسانی وزن‌ها، از تابع خطای MSE و برای بهبود وزن‌ها از الگوریتم بهینه‌سازی Adam استفاده شد. در ادامه، نتایج پارامترهای ارزیابی الگوریتم‌های یادگیری ماشین جهت برآورد میزان ذرات PM_{10} با استفاده از پارامترهای هواشناسی و شاخص AOD، در جدول ۴ ارائه‌شده است. با توجه به این‌که در این سناریو، هدف تخمین مقدار ذرات PM_{10} با در نظر گرفتن داده‌های AOD در کنار متغیرهای هواشناسی بوده است. شاخص AOD در کنار پارامترهای اقلیمی موجب بهبود عملکرد مدل‌ها شده است. بهترین نتایج تخمین ذرات PM_{10} در الگوریتم GBR با بالاترین ضرایب دقت با مقادیر IOA=۰/۹۳ , R^2 =۰/۷۶ و کمترین ضرایب خطای تخمین به ترتیب با مقادیر RMSE= ۰/۴۹, MAE= ۰/۳۱ برای مجموعه داده های آزمون، نشان‌دهنده توافق خوب مقادیر تخمین با داده‌های واقعی و توانایی تعمیم‌پذیری قابل‌قبول آن نسبت به سایر مدل‌ها است. این نکته نشان‌دهنده اثربخشی ادغام داده‌های AOD با متغیرهای هواشناسی در بهبود قابلیت‌های پیش‌بینی مدل برای غلظت‌های PM_{10} می‌باشد. این نتایج با نتایج تحقیقات Pahlavan et al(2014) , Lee et al., (2016),Hadjimitsis et al(2004),Zieger 2011،و., You et al (2016) و حسینی تابش و همکاران (۱۴۰۰) همخوانی داشته است. Sotoudeheian and Arhami (2017) و Faraji and Nadi (2018) در تحقیقات خود، با ترکیب داده‌های AODو پارامترهای هواشناسی در شهرهای تهران و اصفهان با تعداد ایستگاه‌های بیشتر نسبت به پژوهش حاضر، نتایج مشابه با ضرایب دقت و صحت پایین تر را گزارش کردند. عواملی از قبیل نوع باند سنجنده جهت تعیین شاخص AOD ، نوع و تعداد پارامترهای هواشناسی همچنین نوع مدل استفاده شده، از عوامل مهم و موثر در اختلاف نتایج تحقیقات مشابه انجام شده بوده است. ترتیب و اهمیت نسبی عوامل تأثیرگذار پارامترهای اقلیمی و شاخص AOD در تخمین میزان ذرات PM_{10} در این مدل(GBR) در شکل ۳ نشان داده‌شده است. اهمیت نسبی، نشان‌دهنده مشارکت نسبی یک ویژگی درنتیجه پیش‌بینی‌شده بر اساس اختلاف تنوع تقسیم هر برگ (ویژگی داده) در فرایند رشد هر گره از درخت است. نتایج تعامل داده‌های حاضر و متغیرهای مستقل در این مدل نشان داد که پارامترهای ساعت آفتابی(حدود ۰/۵۵)، حداقل دید افقی(حدود ۰/۲۵)، ماکزیمم سرعت باد(حدود ۰/۱۵) و شاخصAOD با اهمیت نسبی حدود ۰/۰۸ بالاترین تأثیر بر میزان پیش‌بینی ذرات (PM10) را داشته‌اند، و به عنوان مهمترین متغیرهای موثر تعریف شده‌اند. همچنین، میزان میانگین ابرناکی روزانه با کمترین اهمیت نسبی کمتر از ۰/۰۲ نیز کمترین تأثیر را بر عملکرد مدل داشته و درصد کمی از تغییرات را توجیه می‌کند. در تحقیق (Afzali et al, 2014)، ارتباط بین PM_{10} و پارامترهای هواشناسی ازجمله شدت تابش، رطوبت نسبی و جهت باد نشان داده شده است که تقریبا مشابه با نتایج حاضر است. Akbari

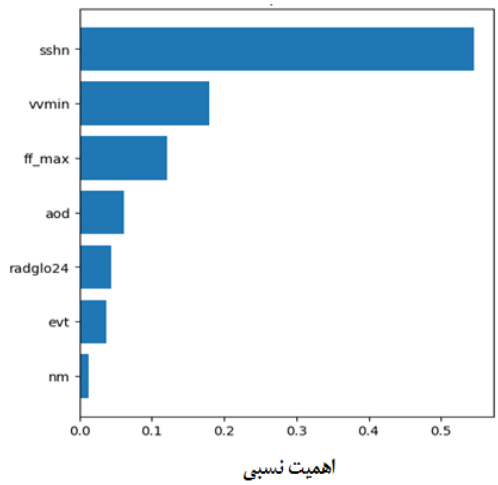
و همکاران (۲۰۱۶) نیز ثابت کردند که رابطه معنی‌داری بین پارامترهای اقلیمی و آلودگی هوا وجود دارد . همچنین در تحقیقات (Cao et al., 2015; Asl et al., 2019) نشان دادند که، شاخص AOD حاصل از تصاویر ماهواره‌ای پتانسیل لازم برای تعیین توزیع مکانی $PM_{2.5}$ و PM_{10} را دارد .

جدول ۳. هایپرپارامترهای یادگیری ماشین برای مدل‌ها

الگوریتم	پارامتر۱	پارامتر۲	پارامتر۳	پارامتر۴	پارامتر۵
DTR	max_depth=5	best	random_state=0	-	-
RFR	max_depth=5	tree number=90	random_state=13	max_features=1.0	min_samples_split=2
Adaboost+DTR	max_depth=5	best	random_state=13	-	-
GBR	max_depth=6	tree number=90	random_state=20	n_iter_no_change=1000	learning_rate=0.05
SVR	'kernel='rbf	C=10	-	-	-
ANN	epochs=10	batch_size=15	Loss Function= MSE	Optimization= Adam	Activation function= Relu

جدول ۴. پارامترهای ارزیابی الگوریتم‌های مورد استفاده در تخمین مقدار PM_{10} ($\mu g/m^3$)

الگوریتم	مجموعه داده	MAE	RMSE	IOA	R^2
DTR	آموزش	۰/۳۱	۰/۳۱	۰/۹۷	۰/۹۱
	آزمون	۰/۳۶	۰/۵۹	۰/۸۹	۰/۶۵
RFR	آموزش	۰/۱۳	۰/۲۵	۰/۹۸	۰/۹۴
	آزمون	۰/۳۷	۰/۵۶	۰/۸۷	۰/۶۸
Adaboost+DTR	آموزش	۰/۱۵	۰/۱۹	۰/۹۹	۰/۹۶
	آزمون	۰/۳۲	۰/۵۳	۰/۹۱	۰/۷۱
GBR	**آموزش**	**۰/۰۷**	**۰/۱۴**	**۱**	**۰/۹۸**
	آزمون	**۰/۳۱**	**۰/۴۹**	**۰/۹۳**	**۰/۷۶**
SVR	آموزش	۰/۱۲	۰/۲۳	۰/۹۹	۰/۹۵
	آزمون	۰/۳۹	۰/۸۰	۰/۶۷	۰/۳۵
ANN	آموزش	۰/۲۸	۰/۵۰	۰/۹۲	۰/۷۵
	آزمون	۰/۳۶	۰/۶	۰/۸۵	۰/۶۴



شکل ۳. اهمیت نسبی پارامترهای هواشناسی و شاخص AOD در الگوریتم GBR جهت برآورد میزان PM_{10}($\mu g/m^3$)

برآورد میزان ذرات PM_{10} با استفاده از پارامترهای هواشناسی

هایپرپارامترهای الگوریتم های یادگیری ماشین در این سناریو، که بهترین نتایج را نشان داده‌اند در جدول ۵ به شرح زیر ارائه شده است. در مدلDTR ، حداکثر عمق ۴ موجب آموزش سریع مدل و جلوگیری از overfitting می‌شود. مقدار ثابت random_state این امکان را می‌دهد که نتایج در تکرارهای مختلف قابل بررسی باشند. برایRFR ، پارامترها با توجه به تعادل بین زمان پردازش و دقت انتخاب شدند. به این ترتیب، برای بهترین دقت، ۹۰ درخت در جنگل مشخص شد و عمق حداکثر تا زمانی که تمامی برگ‌ها خالص بودند، تعیین گردید. همچنین، با تعیین random_state به مقدار ۵، هر بار که مدل اجرا شود، نتایج مشابهی تولید خواهد شد. با تنظیم min_samples_split به ۳، مدل به تعداد کافی داده برای تقسیم گره‌ها نیاز دارد و از تقسیم‌های غیرضروری جلوگیری می‌کند وmin_samples_leaf شامل حداقل تعداد نمونه‌ها در درخت تصمیم است. max_samples با مقدار ۱/۰نشان می‌دهد که تمامی داده‌های آموزشی برای ساخت هر درخت استفاده می‌شود. برای Adaboost+DTR، همان پارامترهای DTR به کار رفته است و همچنین، استفاده از "best" به عنوان بهترین تقویت‌کننده برای هر مرحله، امکان پیش‌بینی‌های دقیق‌تری را با تمرکز بر خطاهای پیشین فراهم می‌کند. در GBR نیز به این صورت که ۹۰ درخت با نرخ یادگیری ۰/۱انتخاب شد تا تأثیر درخت اول در درخت تصمیم کاهش یابد. همچنین، n_iter_no_change به جلوگیری از زمان‌های طولانی آموزش و همچنین overfitting کمک می‌کند. در SVR ، از هسته RBF با مقدار C حدود ۱۰ استفاده شد. در نهایت در الگوریتم ANN ، epochs تعداد مراحل آموزش شبکه عصبی را تعیین می‌کند و batch_size تعداد نمونه‌هایی است که برای هر بار به‌روزرسانی وزن‌ها استفاده می‌شود. برای محاسبه خطا و به‌روزرسانی وزن‌ها، از تابع خطای MSE و برای بهبود وزن‌ها از الگوریتم بهینه‌سازی Adam استفاده شده است. با توجه به نتایج پارامترهای ارزیابی الگوریتم‌های یادگیری ماشین جهت برآورد میزان ذرات PM_{10} با استفاده از پارامترهای هواشناسی، الگوریتم AdaBoostRegressor+DTR با بالاترین مقادیر ضرایب دقت به ترتیب ۰/۸۸= IOA-۰/۶۲=R^2 و کمترین مقادیر ضرایب خطا به ترتیب۰/۶۱= RMSE، ۰/۳۷=MAE نسبت به سایر مدل‌ها، نتایج بهتری در تخمین ذرات PM_{10} داشته است. عملکرد خوب الگوریتم AdaBoostRegressor+DTR می‌تواند به واسطه توانایی مؤثر آن در شناسایی الگوهای محلی در داده‌ها باشد. با توجه به نتایج ضرایب صحت و دقت، الگوریتم SVR با داشتن کمترین مقادیر ضرایب دقت R^2 -IOA به ترتیب ۰/۶۳-۰/۳۲و بیشترین مقادیر ضرایب خطا به ترتیب۰/۸۲= RMSE -۰/۴۳ =MAE نسبت به سایر الگوریتم‌ها، پایین‌ترین عملکرد پیش‌بینی مقادیر PM_{10} را داشته است (جدول ۶). این نتایج با مطالعات انجام‌شده توسط Bozdağ et al, 2020 در آنکارا (ترکیه)، Suleiman et al,2019 در لندن و Thomas Plocoste and Svivio Laventure;2023 که در تحقیقات خود از آمار داده و ایستگاههای بیشتری استفاده نموده اند، هماهنگ و سازگار است. همچنین محمودی سراب (۱۳۹۷) در تحقیقی مشابه، با ترکیب داده های هواشناسی و آلودگی هوا (PM10)، نتایج مشابهی را با ضرایب دقت و صحت پایین‌تری نسبت به نتایج این پژوهش ثبت و گزارش کرده است.

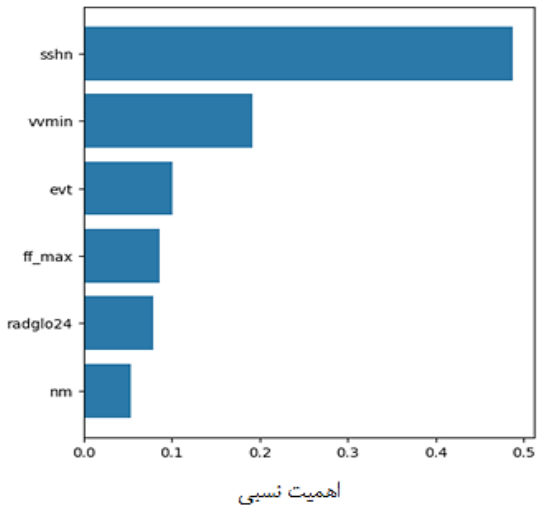
ترتیب و اهمیت نسبی عوامل تأثیرگذار (پارامترهای اقلیمی) در تخمین میزان ذراتPM_{10} در مدل AdaBoostRegressor+DTR در شکل ۴ نشان داده شده است. پارامتر ساعت آفتابی با اهمیت حدود ۰/۵، به‌طور قابل توجهی مهم‌تر از سایر متغیرها بوده و مهم‌ترین ویژگی است که بیشترین تأثیر را در پیش‌بینی‌های مدل داشته است. حداقل دید افقی با اهمیت حدود ۰/۲، دومین متغیر مهم و تأثیرگذار می باشد. میزان تبخیر با اهمیت حدود ۰/۱ و ماکزیمم سرعت باد با اهمیت کمتر از ۰/۱ به ترتیب در رتبه های سوم و چهارم قرار گرفتند. همچنین، میزان میانگین ابرناکی روزانه با کمترین اهمیت (حدود ۰/۰۵) کمترین تأثیر را بر عملکرد مدل داشته و درصد کمی از تغییرات را توجیه می‌کند.

جدول ۵. هایپرپارامترهای یادگیری ماشین برای مدل‌ها

الگوریتم	پارامتر۱	پارامتر۲	پارامتر۳	پارامتر۴	پارامتر۵	پارامتر۶
DTR	max_depth=4	random_state=0	-	-	-	-
RFR	max_depth=5	tree number=90	random_state=5	min_samples_leaf=4	max_samples=1.0	min_samples_split=3
Adaboost+DTR	max_depth=5	best	random_state=0	-	-	-
GBR	max_depth=5	tree number=90	random_state=80	n_iter_no_change=100	learning_rate=0.1	min_samples_split=2
SVR	'kernel='rbf	C=10	-	-	-	-
ANN	epochs=10	batch_size=15	Loss Function= MSE	Optimization= Adam	Activation function= Relu	-

جدول ۶. پارامترهای ارزیابی الگوریتم‌های مورد استفاده در تخمین مقدار PM_{10} ($\mu g/m^3$)

الگوریتم	مجموعه داده	MAE	RMSE	IOA	R^2
DTR	آموزش	۰/۲۳	۰/۳۱	۰/۹۷	۰/۹۰
	آزمون	۰/۳۶	۰/۶۳	۰/۸۷	۰/۶۰
RFR	آموزش	۰/۲۹	۰/۵۶	۰/۸۸	۰/۶۹
	آزمون	۰/۴۳	۰/۷۴	۰/۷۵	۰/۴۵
Adaboost+DTR	**آموزش**	**۰/۱۵**	**۰/۲۰**	**۰/۹۹**	**۰/۹۶**
	آزمون	**۰/۳۷**	**۰/۶۱**	**۰/۸۸**	**۰/۶۲**
GBR	آموزش	۰/۱۱	۰/۲۳	۰/۹۹	۰/۹۵
	آزمون	۰/۴۷	۰/۸۰	۰/۷۵	۰/۳۵
SVR	آموزش	۰/۲۹	۰/۳۹	۰/۹۵	۰/۸۵
	آزمون	۰/۴۳	۰/۸۲	۰/۶۳	۰/۳۲
ANN	آموزش	۰/۳۱	۰/۵۳	۰/۹۱	۰/۷۲
	آزمون	۰/۳۹	۰/۶۷	۰/۷۸	۰/۵۴



شکل ۴. اهمیت نسبی متغیرهای اقلیمی در مدل AdaBoostRegressor+DTR جهت برآورد میزان PM_{10} ($\mu g/m^3$)

برآورد میزان ذرات PM_{10} با استفاده از شاخص AOD

هایپرپارامترهای الگوریتم های یادگیری ماشین در این سناریو، که بهترین نتایج را نشان داده‌اند در جدول ۷ به شرح زیر ارائه شده است.

در الگوریتمDTR ، حداکثر عمق ۳ موجب آموزش سریع مدل و جلوگیری از overfitting می‌شود. همچنین، استفاده از "best" به عنوان بهترین تقویت‌کننده برای هر مرحله، امکان پیش‌بینی‌های دقیق‌تری را با تمرکز بر خطاهای پیشین فراهم می‌کند. مقدار ثابت random_state این امکان را می‌دهد که نتایج در تکرارهای مختلف قابل بررسی باشند. برای الگوریتمRFR ، پارامترها با توجه به تعادل بین زمان پردازش و دقت انتخاب شدند. به این ترتیب، برای بهترین دقت، ۲۰ درخت در جنگل مشخص شد و عمق حداکثر تا زمانی که تمامی برگ‌ها خالص بودند، تعیین گردید. همچنین، با تعیین random_state به مقدار ۱۳، هر بار که مدل اجرا شود، نتایج مشابهی تولید خواهد شد. پارامترmax_features به تعیین حداکثر تعداد ویژگی‌هایی که در هر تقسیم از درختان تصمیم استفاده می‌شود، کمک می‌کند. مقدارmax_features=1.0 به این معنی است که در هر تقسیم، از تمامی ویژگی‌ها استفاده خواهد شد. با تنظیم min_samples_split برابر با ۳، مدل به تعداد کافی داده برای تقسیم گره‌ها نیاز دارد و از تقسیم‌های غیرضروری جلوگیری می‌کند وmin_samples_leaf شامل حداقل تعداد نمونه‌ها در درخت تصمیم بوده که برابر با ۲ تعیین گردید. برایAdaboost+DTR ، مشابه پارامترهای DTR به کار رفته است. در GBR نیز به این صورت که ۸۰ درخت با نرخ یادگیری ۰/۰۵انتخاب شد تا تأثیر درخت اول در درخت تصمیم کاهش یابد. همچنین n_iter_no_change با مقادیر برابر با ۱۰۰، به جلوگیری از زمان‌های طولانی آموزش و همچنین overfitting کمک می‌کند. در SVR ، از هسته RBF با مقدار C حدود ۵ استفاده شد. در نهایت در الگوریتم ANN، epochs تعداد مراحل آموزش شبکه عصبی را تعیین می‌کند و batch_size تعداد نمونه‌هایی است که برای هر بار به‌روزرسانی وزن‌ها استفاده می‌شود. برای محاسبه خطا و به‌روزرسانی وزن‌ها، از تابع خطای MSE و برای بهبود وزن‌ها از الگوریتم بهینه‌سازی Adam استفاده می‌شود. نتایج پارامترهای ارزیابی(جدول ۸) الگوریتم‌های یادگیری ماشین نشان داد که، الگوریتم DTR در مقایسه با دیگر الگوریتم‌ها نتایج مطلوب‌تری جهت برآورد میزان ذرات PM_{10} با استفاده از شاخص AOD داشته است. الگوریتم DTR با داشتن بالاترین مقادیر ضرایب دقت به ترتیب R^2 = ۰/۵۹– IOA= ۰/۸۲ و کمترین مقادیر ضرایب خطا به ترتیب MAE=۰/۴– RMSE=۰/۶۴ نسبت به سایر مدل‌ها، نتایج بهتری در تخمین ذرات PM_{10}داشته است. همچنین با توجه به نتایج ضرایب صحت و دقت، الگوریتم SVR با داشتن کمترین مقادیر ضرایب دقت R^2 -IOA به ترتیب ۰/۴۵–۰/۱و بیشترین مقادیر ضرایب خطا MAE -RMSE به ترتیب۰/۹۵–۰/۴۲، پایین‌ترین عملکرد پیش‌بینی مقادیرPM_{10} را در میان سایر الگوریتم‌ها داشته است (جدول ۸). نتایج به‌دست‌آمده در این بخش از پژوهش با یافته‌های ثبت‌شده در مطالعات پهلوان و همکاران (۱۳۹۳) همخوانی و مشابهت دارد.

جدول ۷. هایپرپارامترهای یادگیری ماشین برای مدل‌ها

الگوریتم	پارامتر۱	پارامتر۲	پارامتر۳	پارامتر۴	پارامتر۵	پارامتر۶
DTR	max_depth=3	best	random_state=80	-	-	-
RFR	max_depth=5	tree number=20	random_state=13	min_samples_leaf=2	max_features=1.0	min_samples_split=3
Adaboost+DTR	max_depth=2	best	random_state=80	-	-	-
GBR	max_depth=2	tree number=80	random_state=5	n_iter_no_change=100	learning_rate=0.05	-
SVR	'kernel='rbf	C=5	-	-	-	-
ANN	epochs=20	batch_size=10	Loss Function= MSE	Optimization= Adam	Activation function= Relu	

جدول ۸. پارامترهای ارزیابی الگوریتم‌های مورد استفاده در تخمین مقدار PM_{10} $(\mu g/m^3)$

الگوریتم	مجموعه داده	MAE	RMSE	IOA	R^2

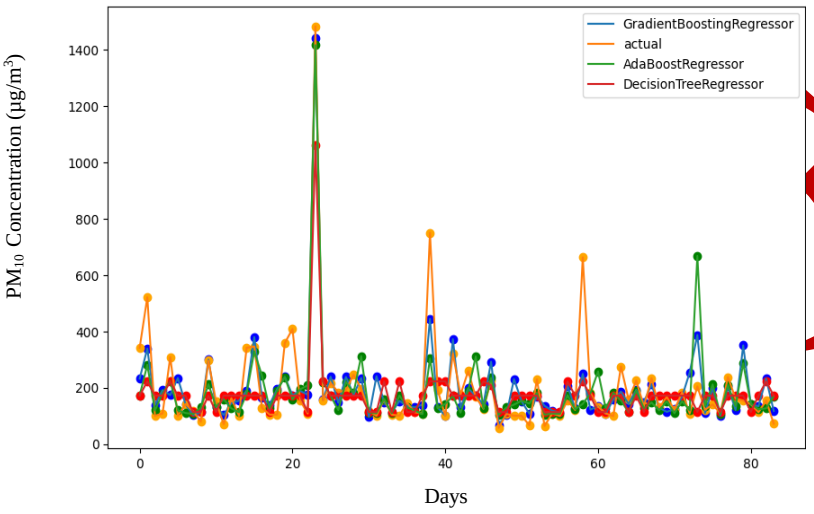
DTR	**آموزش**	**۰/۳۷**	**۰/۶۳**	**۰/۸۶**	**۰/۶۱**
	آزمون	**۰/۴**	**۰/۶۴**	**۰/۸۲**	**۰/۵۹**
RFR	آموزش	۰/۳۵	۰/۶۱	۰/۸۷	۰/۶۳
	آزمون	۰/۴۳	۰/۷۱	۰/۸	۰/۴۹
Adaboost+DTR	آموزش	۰/۵۳	۰/۷۵	۰/۷۶	۰/۴۵
	آزمون	۰/۵۹	۰/۸۳	۰/۷۱	۰/۳
GBR	آموزش	۰/۳۶	۰/۶۷	۰/۸	۰/۵۵
	آزمون	۰/۴۴	۰/۸۴	۰/۵۳	۰/۲۸
SVR	آموزش	۰/۲۸	۰/۶	۰/۸۶	۰/۶۴
	آزمون	۰/۴۲	۰/۹۵	۰/۴۵	۰/۱
ANN	آموزش	۰/۴۱	۰/۷۲	۰/۷۶	۰/۴۸
	آزمون	۰/۴۴	۰/۷۷	۰/۶۸	۰/۴

مقایسه نتایج تخمین ذرات PM_{10} با استفاده از الگوریتم‌های یادگیری ماشین در سه رویکرد (سناریو)

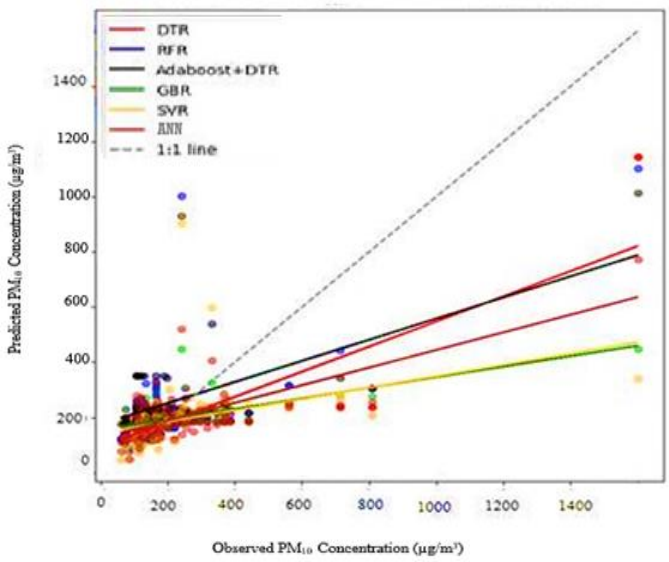
در این مطالعه، نقش متغیرهای هواشناسی و ارتباط بین رخدادهای جوی و اثرات آن‌ها بر آلودگی هوا به‌روشنی تبیین شده است. با توجه به اینکه مدل‌های برآورد غلظت آلودگی هوا با استفاده از تصاویر ماهواره‌ای در مقیاس منطقه‌ای معتبر هستند، در این پژوهش یک مدل منطقه‌ای برای شهر اهواز ارائه شد. همچنین، نقش متغیرهای هواشناسی در این مدل‌ها تشریح و ارتباط بین این متغیرها و غلظت آلودگی مورد بررسی قرار گرفته شده است. بر اساس سه سناریو: ۱) ترکیب شاخص AOD باPM_{10} (سناریو اول)، ۲) ترکیب متغیرهای اقلیمی باPM_{10} (سناریو دوم)، و ۳) ترکیب متغیرهای اقلیمی و شاخص AOD باPM_{10} (سناریو سوم)، عملکرد بهترین الگوریتم‌های سه سناریو در پیش‌بینی غلظتPM_{10} با مقادیر مشاهده شده مورد ارزیابی قرار گرفته شد (شکل ۵). نتایج به‌طور واضح نشان‌دهنده برتری الگوریتم GBR در سناریوی سوم می باشد. این الگوریتم به‌دلیل توانایی خود در مدیریت داده‌های غیرخطی و پیچیده، به‌ویژه در مواجهه با الگوهای تغییرات زمانی و مکانی غلظت آلودگی، بسیار کارآمد است. این یافته‌ها با مطالعات قبلی که به کارایی مدل‌های رگرسیونی پیشرفته اشاره دارند، همخوانی دارد (Friedman, 2001).

مدل GBR با استفاده از ویژگی‌های ورودی شامل شاخص AOD و پارامترهای هواشناسی ، عملکرد قابل‌توجهی در پیش‌بینی غلظت ذراتPM_{10} ارائه داد. شاخص AOD به‌عنوان نماینده‌ای از ذرات معلق در جو عمل می‌کند و اطلاعات مهمی درباره وضعیت آلودگی را در اختیار مدل قرار می‌دهد. در کنار آن، پارامترهای هواشناسی مانند تابش آفتاب، دید افقی، و سرعت باد نیز تأثیر زیادی بر دینامیک آلودگی دارند. نتایج این تحقیق تأکید می‌کند که ترکیب این دو دسته از داده‌ها، منجر به پیش‌بینی‌های دقیق‌تری می‌شود. از سوی دیگر در رویکرد برآورد میزان PM_{10} صرفا با استفاده از شاخص AOD ، تفاوت بیشتری بین مقادیر پیش‌بینی و مشاهده شده ذرات PM_{10} در مقایسه با دو رویکرد قبل وجود دارد. به‌عبارت‌دیگر مدل‌هایی که فقط از شاخص AOD استفاده می‌کنند، دقت کمتری در پیش‌بینی میزانPM_{10} داشته، نمی توانند مانند سایر مدلها مقادیر غلظت ذرات PM_{10} را به خصوص در زمانهای گردو غبار پیش بینی و تخمین بزنند. بیشترین و کمترین اریب پیش بینی غلظت ذرات PM_{10} در الگوریتم ها، به ترتیب در سناریوهای یک و سه مشاهده شد(شکل ۶). عواملی از جمله عدم وجود اطلاعات کافی در مورد ذرات معلق در هوا و همچنین اندازه پیکسل تصاویر ماهواره‌ای مودیس که یک کیلومتر بوده و در مقایسه با ابعاد ایستگاه‌های آلودگی‌سنجی زمینی مساحت بزرگتری دارند، همین اختلاف مساحت می‌تواند باعث ایجاد عدم قطعیت در نتایج مدل‌هایی شوند که صرفاً از داده‌های تصاویر ماهواره استفاده می‌کنند. نتایج مطالعه Chelani (۲۰۱۹) نشان داد که پارامترهای هواشناسی قابلیت بهبود عملکرد رگرسیون چندگانه در تخمین غلظت ذرات آلاینده‌های هوا را دارند؛ بنابراین، استفاده از این پارامترها در مدل‌های پیش‌بینی توصیه می‌شود. همچنین طبق مطالعات (Paciorek and Liu, 2009) ، با توجه به همبستگی آماری بین شاخص AOD و غلظت ذرات PM می‌توان اذعان کرد که علاوه بر پارامترهای هواشناسی شاخص

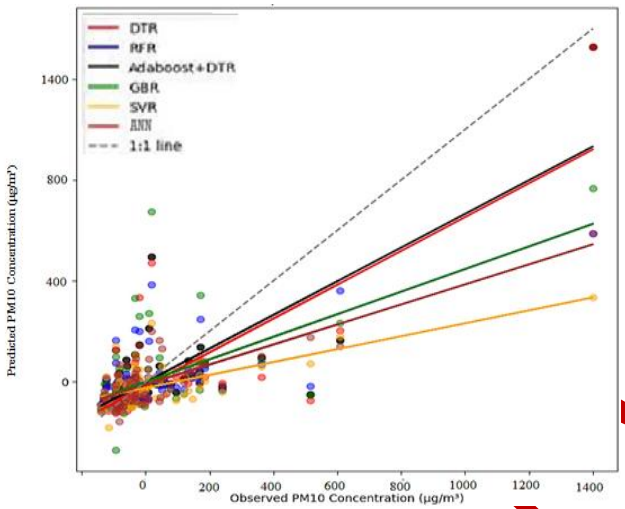
AODتأثیر قابل‌توجهی در پیش‌بینی آلاینده‌های هوا را دارد. با توجه به نتایج این پژوهش، تأثیر قابل‌توجه متغیرهای هواشناسی و شاخص AOD در پایش آلودگی هوا نشان داده شده است. به‌طور کلی با اضافه کردن متغیرهای هواشناسی علاوه بر شاخص AOD ، نتایج بهبود چشمگیری پیدا کرده‌اند. دسترسی آسان به محصولات رایگان شاخصAOD و همچنین پوشش مکانی و زمانی مناسب آن، امکان پایش گسترده (و نه به‌صورت نقطه‌ای) آلودگی هوا را فراهم می‌کند. این روش می‌تواند کاستی‌های اندازه‌گیری‌های زمینی را جبران کرده و مکمل آن باشد تا نتایج بهتری حاصل شود.



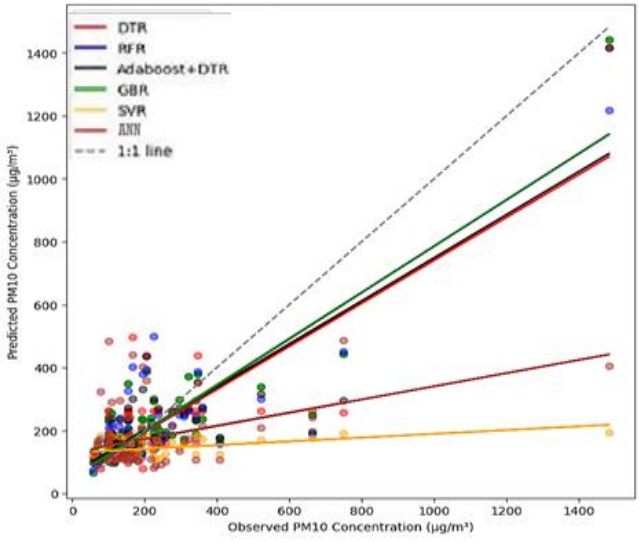
شکل ۵. مقایسهٔ مقادیر مشاهده و پیش‌بینی شده PM_{10} بهترین الگوریتم‌ها برای سه سناریو برای نمونه های آزمون



(a)



(b)



(c)

شکل ۶. مقایسهٔ مقادیر مشاهده و پیش‌بینیPM_{10} و الگوریتم‌ها ی مختلف برای نمونه های آزمون : (a)سناریو ۱ ، (b)سناریو ۲ و (c) سناریو ۳

نتیجه‌گیری

نتایج این پژوهش در پایش غلظت آلایندهPM_{10} در شهر اهواز با توجه به داده‌های رایگان و در دسترس شاخصAOD ، و همچنین اهمیت پایش این آلاینده و معضل آلودگی هوا، گامی مؤثر برای بهره‌برداری از شاخص AOD در کنار داده‌های زمینی است.این اقدام درنهایت منجر به پایش بسیار دقیق‌تر و به‌صرفه‌تر آلودگی هوا می‌شود. ازبین پارامترهای اقلیمی استفاده شده در این پژوهش ساعت آفتابی ، حداقل دید افقی و ماکزیمم سرعت باد به‌عنوان مهمترین متغیرهای هواشناسی تأثیرگذار شناسایی شده، که در کنار شاخصAOD می‌توانند عملکرد چشمگیری در مدلسازی غلظت آلاینده PM_{10} داشته باشند. تابش خورشید به‌طور مستقیم بر واکنش‌های فتوشیمیایی و تشکیل ذرات ثانویه تأثیر می‌گذارد و افزایش دما ناشی از تابش خورشید، تغییراتی در الگوی جریان همرفتی و پراکنش عمودی آلاینده‌ها ایجاد می‌کند. همچنین، حداقل دید افقی رابطه معکوس با غلظتPM_{10} داشته؛ به‌طوری که در سرعت‌های پایین باد، تجمع آلاینده‌ها مشاهده شده و با افزایش سرعت باد، احتمال تولید گرد و غبار افزایش می‌یابد. به‌علاوه، شاخص AOD همبستگی قوی با غلظتPM_{10} دارد که نشان‌دهنده میزان جذب و پراکنش نور توسط آئروسل‌ها است(Tuna Tuygun et al, 2021).مقایسه پیش بینی مقادیرPM10 بهترین الگوریتم در سه سناریو شامل: ترکیب شاخص AOD با PM_{10} (سناریو اول)، ترکیب متغیرهای اقلیمی با PM_{10} (سناریوی دوم) و ترکیب متغیرهای اقلیمی و شاخص AOD با PM_{10} (سناریوی سوم) با مقادیر مشاهده شده ذرات PM10 نشان داد که، در سناریوی سوم با استفاده از الگوریتم GBR، بهترین عملکرد با بالاترین ضرایب دقت و کمترین ضرایب خطا (IOA=0.93, R^2 = 0.76 ،MAE = 0.31 و RMSE = 0.49) به دست آمد. در حالی که در سناریوی اول با استفاده تنها از شاخص AOD ، پایین‌ترین ضرایب دقت و صحت در پیش بینی مقادیر PM10 در الگوریتم DTR (IOA=0.82 , R^2 =0.59 ،MAE = 0.40 و RMSE = 0.64) حاصل شد. بیشترین و کمترین اریب پیش بینی غلظت ذرات PM_{10} در الگوریتم ها نسبت به مقدار مشاهده شده، به ترتیب در سناریوهای یک و سه مشاهده شد. با توجه به نتایج این تحقیق، به خوبی می توان با استفاده از الگوریتمهای یادگیری ماشین و داده های پارامترهای اقلیمی و شاخص AOD ، میزان غلظت آلودگی هوا (PM10) را براورد نمود. در مطالعات آینده، پیشنهاد می‌شود با استفاده از مدل‌های هیبریدی در زمینه یادگیری ماشین برای پیش‌بینی‌های دقیق‌تر استفاده گردد. این اقدامات می‌تواند به ایجاد یک سیستم پایش منطقه‌ای کارآمد منجر شود.

"هیچ‌گونه تعارض منافع بین نویسندگان وجود ندارد"

منابع

اژدری، علی؛ حیدریان، پیمان؛ جودکی، محمد.؛ درویشی خاتونی، جواد و شهبازی، رضا. (۱۳۹۶). شناسایی منشاهای داخلی توفان‌های گردوغبار با استفاده از سنجش‌ازدور، GIS و زمین‌شناسی) مطالعه موردی: استان خوزستان، فصلنامه علمی- پژوهشی علوم زمین. سازمان زمین‌شناسی و اکتشافات معدنی کشور، ۲۷ (۱۰۵): ۴۶-۳۳.

اکبری، الهه؛ فاخری، معصومه؛ پورغلامحسن، عفت و اکبری، زهرا. (۱۳۹۴). پهنه بندی ماهانهٔ میزان آلودگی هوا و بررسی نحوهٔ ارتباط آن با عوامل اقلیمی (مطالعهٔ موردی: شهر مشهد). نشریه محیط زیست طبیعی، ۶۸(۴): ۵۳۳-۵۴۷.

پهلوان، احمد؛ اسماعیلی، علی و پهلوان، راضیه. (۱۳۹۳). برآورد غلظت آالاینده‌های PM_{10} و $PM_{2.5}$در کلان شهر تهران با استفاده از داده‌های سنجنده مودیس ماهواره‌های آکوا و ترا. مجله علمی و ترویجی نیوار، ۳۸(۸۵): ۶۸-۵۸.

حسینی تابش، صبا؛ آقاشریعتمداری، زهرا و حجابی، سمیه.(۱۴۰۰). ارزیابی داده‌های سنجنده مودیس (MODIS) در پایش غلظت آلاینده‌های$PM_{2.5}$ و PM10 با تأکید بر متغیرهای هواشناسی. تحقیقات آب‌وخاک ایران؛ ۵۲ (۱۲)، ۲۹۶۷-۲۹۸۳.

ستوده یان، سعید و ارحامی، محمد. (۱۳۹۶). بهره‌گیری از مدل اثرات اختلاط خطی جهت پیش بینی غلظت ذرات معلق در سطح زمین: مطالعه موردی در تهران .سلامت و محیط زیست، ۱۰ (۲) :۲۲۴-۲۱۳.

صادقی،حسین و خاکسار آستانه، سمانه. (۱۳۹۳). پیش‌بینی کوتاه‌مدت آلودگی ذرات معلق شهر اهواز با کمک شبکه‌های عصبی. پژوهش های محیط زیست, ۵(۹): ۱۷۷-۱۸۶.

محمودی سراب، علی؛ معیری، سجاد؛ معیری، محمدهادی؛ شتایی جویباری، شعبان و راشکی، علیرضا. (۱۳۹۷). برآورد میزان آلودگی هوا با استفاده از داده‌های آب و هوایی (مطالعه موردی: شهرستان اهواز). محیطزیست طبیعی. منابع طبیعی ایران، ۷۱(۳): ۳۸۵ تا ۳۹.

REFERENCES

Afzali, A., Rashid, M., Sabariah, B., Ramli, M., (2014). PM10 Pollution: Its Prediction and Meteorological Influence in PasirGudang, Johor.8th International Symposium of the Digital Earth (ISDE8). IOP Conf. Series: Earth and Environmental Science 18: 012100. doi:10.1088/1755-1315/18/1/012100.

Akbari, A., Fakheri, M., Poorgholamhossin, A., Akbari, Z. (2016). Monthly Zoning of the Air Pollution and Surveying its Relationship with Climatic Factors (Case Study: Mashhad City). Journal of Natural Environment. 68(4): 533- 547. (In Persian)

Alimahmoodi Sarab, S., Shataee Jouybari, S., & Rashki, A. (2018). The Estimate of Dust Concentration Using of Weather Variable (A Case study: Ahvaz City). Journal of Natural Environment, 71(3), 385-397. (In Persian).

Alizadeh-Choobari O, Ghafarian P, Owlad E. (2016).Temporal variations in the frequency and concentration of dust events over Iran based on surface observations, https://doi.org/10.1002/joc.4479.

Alizadeh-Choobari, O., Zawar-Reza, P., Sturman, A. (2014). The “wind of 120days” and dust storm activity over the Sistan Basin. Atmospheric Research, 143: 328–341. https://doi.org/10.1016/j.atmosres.2014.02.001

Ashpole I, Washington R . (2013). A new high-resolution central and western Saharan summertime dust source map from automated satellite dust plume tracking. J Geophys Res 118:6981–6995. doi:10.1002/jgrd.50554.

Asl, S. Z., Farid, A., & Choi, Y. S. (2019). Assessment of CALIOP and MODIS aerosol products over Iran to explore air quality. Theoretical and Applied Climatology, 137(1-2), 117-131.

Azhdari, A., Heidarian, P., Jodaki, M., Darvishi Khatooni, J., & Shahbazi, R. (2017). Identifying interior sources of dust storms using remote sensing, GIS and geology (case study: Khuzestan province). Scientific Quarterly Journal of Geosciences, 27(105), 33-4. (In Persian).

Bozdağ A, Dokuz Y, Begüm Gökçek Ö .(2020). Spatial prediction of PM_{10} concentration using machine learning algorithms in Ankara, Turkey, Environmental PollutionVolume 263, Part A, August 2020, 114635.

Breiman, L.(2001). Random forests. Mach. Learn, 45, 5–32.

Butt M.J, Ebraheem Assiri M, Md. Arfan A. (2017). Assessment of AOD variability over Saudi Arabia using MODIS Deep Blue products, Environmental PollutionVolume 231, Part 1, December 2017, Pages 143-153.

Cao, H., Amiraslani, F., Liu, J., & Zhou, N. (2015). Identification of dust storm source areas in West Asia using multiple environmental datasets. Science of the Total Environment, 502, 224-235.

Chai, C. P. (2023). “Comparison of text preprocessing methods”, Natural Language Engineering, 29 (3), 509-553.

Chelani, A. B. (2019). Estimating PM2. 5concentrations from satellite derived aerosol optical depth and meteorological variables using a combination model. Atmospheric Pollution Research, 10(3), 847-857.

Clarke, A. D., Collins, W. G., Rasch, P. J., Kapustin, V. N., Moore, K., Howell, S. and Fuelberg, H. E. (2001). Dust and pollution transport on global scales: Aerosol measurements and model predictions. Journal of Geophysical Research: Atmospheres, 106(D23), 32555-32569.

Daryanoosh, M., Goudarzi, G., Rashidi, R., Keishams, F., Hopke, P.K., Mohammadi, M.J. (2018). Risk of morbidity attributed to ambient PM10 in the western cities of Iran. Toxin Reviews 37, 313-318.

DeRousseau, M.A.Laftchiev, E.Kasprzyk, J.R.Rajagopalan, B.Srubar, W.V., (2019). A comparison of machine learning methods for predicting the compressive strength of field-placed concrete, Constr. Build. Mater. vol. 228, 116661.

Ekhtesasi M. R , Sepehr A.(2009). Investigation of wind erosion process for estimation, prevention, and control of DSS in Yazd–Ardakan plain, Environ Monit Assess Environmental researches. 5(9): 177- 186. DOI 10.1007/s10661-008-0628-4.

Faghihinia and Afzali .(2013). Effects of wind erosion on soil organic carbon dynamics and other soil properties: Dejgah catchment, Farashband County, Shiraz Province, Iran, September 2013,African Journal of Agricultural Research 8(34):4452-4459

Faraji, M. and Nadi, S. (2018). Assessment of aerosol optical depth of MODIS sensor data by using PM2.5meteorological data in urban area. In proceeding of 3th spatial data of technology of engineering. Khaje Nasir Toosi University of technology, Tehran. (In Persian)

Feng, D.C. et al. (2020),, Machine learning-based compressive strength prediction for concrete: an adaptive boosting

approach, Constr. Build. Mater. vol. 230.
Feng, J. Zhang, H. Gao, K. Liao, Y. Gao, W. Wu, G. (2022). Efficient creep prediction of recycled aggregate concrete via machine learning algorithms, Constr. Build.Mater. vol. 360 129497.
Friedman. J H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. Institute of Mathematical Statistics Vol. 29, No. 5 (Oct., 2001), pp. 1189-1232 (44 pages).
Fu, M.; Wang, W.; Le, Z.; Khorram, M.S.(2015). Prediction of particular matter concentrations by developed feed-forward neural network with rolling mechanism and gray model. Neural Comput. Appl, 26, 1789–17.
Gardner, M. and Dorling, S.. (1998),\Arti_cial neural networks (the multilayer perceptron) { a review of applications in the atmospheric sciences," At-mospheric Environment.
Gautam, S., Patra, A.K., Sahu, S.P., Hitch, M. (2018). Particulate matter pollution in opencast coal mining areas: a threat to human health and environment. International Journalof Mining, Reclamation and Environment 32, 75-92.
Ginoux P, Prospero J.M. , Torres O, Chin M,(2004). Long-term simulation of global dust distribution with the GOCART model: correlation with North Atlantic Oscillation, Environmental Modelling & SoftwareVolume 19, Issue 2, February, Pages 113-128.
Gonzalez ,P., Wang ,F., Notaro, M., Vimont ,D. J. and Williams ,J.W.(2018). Disproportionate magnitude of climate change in United States national parks. , Published by IOP Publishing Ltd. Published 24 September 2018.
Ghunimat, D., Alzoubi, A. E., Alzboon, A., & Hanandeh, S. (2023). “Prediction of concrete compressive strength with GGBFS and fly ash using multilayer perceptron algorithm, random forest regression and k-nearest neighbor regression”, Asian Journal of Civil Engineering, 24 (1), 169-177.
Hadjimitsis, D. G., Clayton, C. R. I. and Hope, V. S. (2004). An assessment of the effectiveness ofatmospheric correction algorithms through theremote sensing of some reservoirs. InternationalJournal of Remote Sensing, 25(18), 3651-3674.
Holben, B. N., Tanre, D., Smirnov, A., Eck, T. F., Slutsker, I., Abuhassan, N., and Kaufman, Y. J. (2001). An emerging ground-based aerosol climatology: Aerosol optical depth from AERONET. Journal of Geophysical Research: Atmospheres, 106(D11), 12067-12097.
Hoseini Tabesh, S., Aghashariatmadari, Z., & Hejabi, S. (2022). Assessment of MODIS Data in Monitoring the Concentrations of PM2.5 and PM10 Pollutants with Emphasis on Meteorological Variables. *Iranian Journal of Soil and Water Research*, *52*(12), 2967-2983. doi: 10.22059/ijswr.2022.330907.66908(In Persian).
Hou,W.; Li, Z.; Zhang, Y.; Xu, H.; Zhang, Y.; Li, K.; Li, D.;Wei, P.; Ma, Y.(2014).Using support vector regression to predict PM10 and PM2.5. IOP Conf. Ser. Earth Environ. Sci. 17, 012268.
Jiang, T., Chen, B., Nie, Z., Ren, Z., Xu, B., & Tang, S. (2021). Estimation of hourly full-coverage PM2.5 682 concentrations at 1-km resolution in China using a two-stage random forest model. 683 Atmospheric Research, 248. https://doi.org/10.1016/j.atmosres.2020.105146
Jolliffe I. (2002).Principal component analysis: Wiley Online Library.
Kaviani Rad.A, Redmond R. Shamshiri , Naghipour.A, Odeen Razmi .S , Shariati.M, Golkar.Sh and Siva K. Balasundram, (2022), Machine Learning for Determining Interactions between Air Pollutants and Environmental Parameters in Three Cities of Iran, sustainability, https://doi.org/10.3390/su14138027.
Keprate, A.; Ratnayake, R.C. (2017). Using gradient boosting regressor to predict stress intensity factor of a crack propagating in small bore piping. In Proceedings of the 2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), Singapore, 10–13 December 2017; IEEE: Piscataway, NJ, USA,; pp. 1331–1336.
Kumar.d, singh.m, kushwaha.m, makarana.v and yadav. M.r. (2021). Integrated use of organic and inorganic nutrient sources influences the nutrient content, uptake and nutrient use efficiencies of fodder oats (Avena sativa). ICAR-National Dairy Research Institute, Karnal, Haryana 132 , February 2021; Revised accepted: November 2021.
Khosroshahi, m, kashaki,m, ensafi moghadam,t.(2009). Determination of climatological deserts in Iran, International Journal of Phytoremediation Vol, 16(1): 96-11.
Lanzaco, B. L., Olcese, L. E., Palancar, G. G. and Toselli. B. M. (2016). Aerosol and Air Quality Research. 16 1509-1522.
Lee, H. J., Chatfield, R. B. and Strawa, A. W. (2016). Enhancing the applicability of satellite remote sensing for PM2. 5 estimation using MODIS deep blue AOD and land use regression in California, United States. Environmental Science & Technology, 50(12), 6546-6555.
Li, H. Lin, J. Lei, X. Wei, T. (2022). Compressive strength prediction of basalt fiber reinforced concrete via random forest

algorithm, Mater. Today Commun. vol. 30, 103117, https://doi.org/10.1016/J.MTCOMM.2021.103117.
Masoudi M, Asadifard E. and Rastegar M.(2018). Status of PM10 as an air pollutant and its prediction using meteorological parameters in Ahvaz, Iran, Environmental Resources Research Vol. 6, No. 2.
Middleton, N. J. (1986a). Dust storms in the Middle East. – Journal of Arid Environments 10 (2): 83–96.
Fratello, M, Tagliaferri, R.(2019). Decision Trees and Random Forests. Encyclopedia of Bioinformatics and Computational Biology (1) 2019: 374-383
Miri, A., Hasan, A., Ekhtesasi, M. R., Panjehkeh, N., and Ghanbari, A. (2009). Environmental and socio-economic impacts of dust storms in Sistan Region, Iran. The International Journal of Environmental Studies 66, 343–355.
Mustaqeem,M., Siddiqui,T., Ahmad Khan,.N., Kumar,.D.2023. In-Depth Analysis of Various Artificial Intelligence Techniques in Software Engineering: An Experimental Study. Journal of Information Technology Management, 2023, Vol. 15, Issue 3, pp. 162-181
Modarres, R., Sadeghi, S. (2018). Spatial and temporal trends of dust storms across desert areas of Iran. Natural Hazards 90, 101-114.
Naghibi, S.A., Pourghasemi, H.R., Dixon, B. (2016). GIS-based groundwater potential mapping using boosted regression tree, classification and regression tree, and randomforestmachine learning models in Iran. Environ. Monit. Assess. 188, 44. https://doi.org/10.1007/s10661-015-5049-6.
O'LoingsighT, McTainsh G.H., TewsE. K , StrongC.L , Leys J.F., ShinkfieldP. Tapper, N.J. (2014).The Dust Storm Index (DSI): A method for monitoring broadscale wind erosion using meteorological records, Aeolian Research Volume 12, March 2014, Pages 29-40.
Olden, J.D., Lawler, J.J., Poff, N.L. (2008). Machine learning without tears: a primer for ecologists.Q. Rev. Biol. 83 (2), 171–193.
Oyewola, D.O.; Dada, E.G.; Misra, S.; Damaševi˘cius, R. Predicting COVID-19 Cases in South Korea with All K-Edited Nearest Neighbors Noise Filter and Machine Learning Techniques. Information 2021, 12, 528.
Paciorek, C. J. and Liu, Y. (2009). Limitations of remotely sensed aerosol as a spatial proxy for fine particulate matter. Environmental health perspectives, 117(6), 904-909.
Pahlavan, A. Pahlavan, R. and Esmaeli, A. (2014). Estimating PM10 and PM2.5 in Tehran mega city using MODIS data of Terra and Aqua satellites. In: Proceedings of the first International Congress on Application of advanced models of spatial analysis (remote sensing and GIS) in land management, 24-25 Oct. Azad University, Yazd, Iran, pp.125138.(In Persian)
Plocoste T and Laventure S. (2023). Forecasting PM10 Concentrations in the Caribbean Area Using Machine Learning Models, Atmosphere 2023, 14(1), 134; https://doi.org/10.3390/atmos14010134.
Rashki, A., Arjmand, M, Kaskaoutis, D. G. (2017). Assessment of dust activity and dust-plume pathways over Jazmurian Basin, southeast Iran. Aeolian Research Journal, 24: 145–160.
Reynolds , James f. , stafford smith, d. Mark, .Lambin, eric f, . Turner, b. L. ii, mortimore, michael , Batterbury , simon p. J.,Downing, thomas e. , dowlatabadi, hadi , . Fernández, robertoj, and walkr, brian . (2007). Global Desertification: Building a Science for Dryland Development, SCIENCE, 11 May 2007,Vol 316, Issue 5826,pp. 847-851DOI: 10.1126/science.1131634.
Rezaei M ، J.P.M M. Riksen , Sirjani E , Sameni A, Geissen V.(2019). Wind erosion as a driver for transport of light density microplastics, Science of The Total EnvironmentVolume 669, 15 June 2019, Pages 273-281.
Sadeghi, H., khaksar, S.(2015). Neural Network Model for Short Term Prediction of PM10 Pollution in Ahvaz City. (In Persian).
Sahebzadeh B, Shabani-Goraji K, Shoaei Z, Afshari M. (2019). Statistical study of eolian sediment risk in human ecosystems on the health of respiratory system and the eyes of inhabitants of Sistan, Iran, Arabian Journal of Geosciences.
Shao, Y., and others (2003). Northeast Asian dust storms: Real-time numerical prediction and validation. Journal of Geophysical Research: Atmospheres, vol. 108, No. D22
Shrestha .N. 2020. Detecting Multicollinearity in Regression Analysis. American Journal of Applied Mathematics and Statistics, 8, 39-42.
Singh, A.; Kotiyal, V.; Sharma, S.; Nagar, J.; Lee, C.C. (2020). A machine learning approach to predict the average localization error with applications to wireless sensor networks. IEEE Access, 8, 208253–208263.

SONG Yand LU Y. (2015). Decision tree methods: applications for classification and prediction, Shanghai Arch Psychiatry. 2015 Apr 25; 27(2): 130–135.
Sotoudeheian, S., and Arhami, M. (2017). Using linear mixed effect model to estimate ground-level PM2.5: case study for Tehran. Iranian Journal of Health and Environment. 10 (2), 213-224. (In Persian)
Suleiman A., Tight M.R., Quinn A.D. (2019). Applying machine learning methods in managing urban concentrations of traffic-related particulate matter (PM_{10} and $PM_{2.5}$), Atmospheric Pollution ResearchVolume 10, Issue 1, January 2019, Pages 134-144.
Tahbaz Geophys Res M. J. (2016).summertime dust source map from automated satellite dust plume tracking. Environmental Challenges in Today's Iran, Iranian Studies, November 2016.
Tuna Tuygun,G., Gündoğdu,S., Elbir1,T,.(2021). Estimation of ground-level particulate matter concentrations based on synergistic use of MODIS, MERRA-2 and AERONET AODs over a coastal site in the Eastern Mediterranean, Version of Record: https://www.sciencedirect.com/science/article/pii/S1352231021003848
Ul-Saufie, A.Z.; Yahya, A.S.; Ramli, N.A.; Hamid, H.A. (2011). Comparison between multiple linear regression and feed forward back propagation neural network models for predicting PM10 concentration level based on gaseous and meteorological parameters.Int. J. Appl. Sci. Technol., 1, 42–49.
Willmott, C.J.; Robeson, S.M.; Matsuura, K. (2012).A refined index of model performance. Int. J. Climatol., 32, 2088–2094.
Xia, L.; Bai, R. (2008). Freight Vehicle Travel Time Prediction Using Gradient Boosting Regression Tree. In Proceedings of the IEEE International Conference on Machine Learning & Applications, Anaheim, CA, USA, 18–20 December 2016.
You, W., Zang, Z., Zhang, L. et al. (2016). Estimating national-scale ground-level PM25 concentration in China using geographically weighted regression based on MODIS and MISR AOD. Environ Sci Pollut Res, 23(9), 8327–8338.
Ye Ren, Le Zhang, and Ponnuthurai N Suganthan (2016) Ensemble classification and regression-recent developments, applications and future directions, IEEE Computational Intelligence Magazine, 11(1), pp. 41–53.
Zarei, T., Abdolzadeh, M., Yaghoubi, M. (2022). Comparing the impact of climate on dustaccumulation and power generation of PV modules: A comprehensive review.Energy Sustain. Dev. 66, 238–270.
Zieger, P., Weingartner, E., Henzing, J., Moerman, M., de Leeuw, G., Mikkilä, J., Ehn, M., Petäjä, T., Clémer, K., van Roozendael, M., Yilmaz, S., Frieß, U., Irie, H., Wagner, T., Shaiganfar, R., Beirle, S., Apituley, A., Wilson, K. and Baltensperger, U. (2011). Comparison of ambient aerosol extinction coefficients obtained from in- situ, MAX-DOAS and LIDAR measurements at Cabauw, Atmos. Chem. Phys., 11, 2603–2624.

Comparing machine learning algorithms for estimating PM_{10} particle concentration using AOD and meteorological parameters

EXTENDED ABSTRACT

Introduction

Monitoring and controlling the levels and sources of dust, influenced by climate change, and developing appropriate predictive approaches that have direct impacts on the environment and human health are of great importance. Dust storms are one of the main reasons for the dispersion of airborne particles with an aerodynamic diameter of less than 10 micrometers (PM_{10}) in the air of dry regions worldwide, including the dry and desert regions of Iran. These storms occur more severely and with higher concentrations than in the past, leading to adverse environmental effects. One of the negative consequences of increased particulate matter concentration is the health risks posed to residents in these areas. Among these, the southern and southeastern regions of Ahvaz are recognized for having the largest area of origin centers of dust storms in Khuzestan Province. This study was conducted with the aim of estimating the PM_{10} particle concentrations in the city of Ahvaz using various machine learning models.

Methodology

In this research, climate variables and the Aerosol Optical Depth (AOD) index, derived from the MODIS sensor at a wavelength of 476 nanometers, were used as influential variables in estimating PM10 concentration in three scenarios: combining AOD with PM10 (scenario 1), combining climate variables with PM10 (scenario 2), and combining climate variables and AOD with PM10 (scenario 3) .Using six machine learning algorithms, namely Random Forest (RF), Gradient Boosting Regression (GBR), Artificial Neural Networks (ANN), AdaBoostR with DTR, Support Vector Regression (SVR), and Decision Tree Regression (DTR), the PM10 concentration was estimated in different scenarios, considering accuracy and precision coefficients.

Results and Discussion

The most influential variables in estimating PM10 concentration were determined to be sunshine hours, minimum visibility, maximum wind speed, and the AOD index. The results indicated that the method of combining the input variables of the AOD index and meteorological parameters using the GBR algorithm showed the best performance with the highest accuracy and precision coefficients, including MAE = 0.31, RMSE = 0.49, IOA=0.93, R^2 = 0.76 . The approach using only the AOD index showed the worst performance in estimating PM_{10} levels, with accuracy and precision coefficients, including MAE = 0.40, RMSE = 0.64, IOA=0.82, R^2 = 0.59. The approach using only meteorological parameters showed intermediate performance in estimating PM_{10} levels, with accuracy and precision coefficients, including MAE = 0.37, RMSE = 0.61, IOA=0.88, R^2 = 0.62. The pixel size of the satellite image used in the MODIS sensor is one kilometer, which in comparison to the dimensions of ground-based PM10 pollution monitoring stations, has a larger area. This difference in area creates uncertainty in the results of models that solely rely on satellite image data.The use of influential meteorological variables alongside the AOD index for modeling PM_{10} pollutant concentrations has shown remarkable performance. In the field of machine learning studies, utilizing multiple models for making predictions is crucial. By providing more accurate predictions in complex processes, focusing on improving these models for optimal management can ultimately contribute to advancing more sustainable and cost-effective operations. The proposed final model can be used for daily estimation of PM10 particles.

Keywords: Machine Learning Algorithms, Climatic Variables, AOD index, PM_{10}

Author Contributions

Conceptualization, F.Kh. and E.Kh.; methodology, F.Kh. and E.Kh.; software, E.Kh.; validation, E.Kh.; formal analysis, F.Kh. and E.Kh.; investigation, F.Kh. and E.Kh.; resources F.Kh. and E.Kh.; data curation, F.Kh. and E.Kh.; writing—original draft preparation, F.Kh. and E.Kh.; writing—review and editing, M.R.A and S.H; visualization, M.R.A., S.H. and E.Kh.; supervision, M.R.A., S.H. and E.Kh.; project administration, M.R.A., S.H. and E.Kh.; funding acquisition, M.R.A.
All authors have read and agreed to the published version of the manuscript.

Data Availability Statement
Data is available on reasonable request from the authors.

Acknowledgements
The authors would like to thank Agricultural Sciences and Natural Resources University of Khuzestan, for providing all the needed facilities.

Ethical considerations
The authors avoided data fabrication, falsification, plagiarism, and misconduct.