Vol.12, No.2, Summer 2023, pp.165-186 ISSN (Print):2322-1682 ISSN (Online):2383-3076

Journal of Geography and Environmental Hazards

RESEARCH ARTICLE

DOI:10.22067/geoeh.2022.76121.1212 Open access

Development of Machine Learning Algorithms to Predict Urban Air Quality Index (Study Area: Tehran City)

Peyman Karami[a], Seyed Ahmad Eslaminezhad[b], Mobin Eftekhari[c], Faraz boroumand[d], Mohammad Akbari[e*]

[a] *MA., Faculty of Surveying and Geomatics Engineering, University of Tehran, Tehran, Iran*
[b] *MA, Faculty of Surveying and Geomatics Engineering, University of Tehran, Tehran, Iran*
[c] *Researcher, Young Researchers and Elite Club, Mashhad Branch of Islamic Azad University, Mashhad, Iran*
[d] *MA, Faculty of Surveying Engineering, K. N. Toosi University of Technology,Tehran, Iran*
[e] *Research Group of Drought and Climate Change ،University of Birjand*

Received: 8 April** 2022* ***Revised: 7 May 2022 ***Accepted: 29 May 2022***

Abstract

Considering the harms of air pollution on human health and the environment, it seems necessary to reduce and solve this problem based on accurate knowledge of pollutants and criteria affecting it and identifying polluted areas. Therefore, using mathematical models in the form of machine learning is an optimal and cost-efficient approach to air pollution modeling. This research is applied in terms of purpose and its method is descriptive-analytical. The novelty of this research is presenting a new combination approach to determine the effective criteria for predicting the amount of air pollution. Therefore, the purpose of this study was to evaluate and compare the capabilities of two machine learning models, namely Support Vector Machine (SVM) and Random Forest (RF) in combination with Genetic Algorithm (GA) to predict air pollution in Tehran. The data used in this research include particulate matter and gaseous pollutants in Tehran in 2020, which was obtained from Tehran Traffic Control Company. MATLAB and ArcMap software were used to analyze the data. The value of coefficient of determination (R^2) obtained from the combined RF-GA method was 0.997, which indicates the high compatibility of this model with the data of this study. Moreover, the Root Mean Square Error (RMSE) value from the combined RF-GA method was 0.153, which indicates high accuracy of this model. Based on the data obtained from Tehran Traffic Control Company, the results of the RF method indicate the appropriateness of selecting the model to estimate the amount of air pollution in Tehran.

Keywords: Air Pollution, Machine Learning, Random Forest, Support Vector Machine, Genetic Algorithm

* Corresponding author: Mohammad Akbari E-mail: Moakbari@birjand.ac.ir Tel: + 98 9153616696

How to cite this Article: Karami, P., Eslaminezhad, S. A., Eftekhari, M., Boroumand, F., & Akbari, M. (2023). Development of machine learning algorithms to predict urban air quality index (Study area: Tehran city). *Journal of Geography and Environmental Hazards*, *12*(2), 165-186.
DOI:10.22067/geoeh.2022.76121.1212

Geography and Environmental Hazards
Volume 12, Issue 2 - Number 46, Summer 2023
https://geoeh.um.ac.ir
https://doi.org/10.22067/geoeh.2022.76121.1212

جغرافیا و مخاطرات محیطی، سال دوازدهم، شمارهٔ چهل و ششم، تابستان ۱۴۰۲، صص ۱۸۶-۱۶۵

مقاله پژوهشی

توسعه الگوریتم‌های یادگیری ماشین جهت پیش‌بینی شاخص کیفیت هوای شهری (منطقه مطالعاتی: شهر تهران)

پیمان کرمی- کارشناسی ارشد، دانشکده مهندسی نقشه‌برداری و اطلاعات مکانی، دانشگاه تهران، تهران، ایران
سید احمد اسلامی نژاد- کارشناسی ارشد، دانشکده مهندسی نقشه‌برداری و اطلاعات مکانی، دانشگاه تهران، تهران، ایران
مبین افتخاری- پژوهشگر، باشگاه پژوهشگران جوان و نخبگان، دانشگاه آزاد اسلامی واحد مشهد، مشهد، ایران
فراز برومند- کارشناسی ارشد، دانشکده مهندسی نقشه‌برداری، دانشگاه صنعتی خواجه‌نصیرالدین طوسی، تهران، ایران
محمد اکبری[1]- گروه پژوهشی خشکسالی و تغییر اقلیم، دانشگاه بیرجند، بیرجند، ایران

تاریخ دریافت: ۱۴۰۱/۱/۱۹ تاریخ بازنگری: ۱۴۰۱/۲/۱۷ تاریخ تصویب: ۱۴۰۱/۳/۸

چکیده

با توجه به مضرات آلودگی هوا بر سلامت انسان‌ها و محیط، کاهش و حل این معضل براساس شناخت دقیق آلاینده‌ها و عوامل تأثیرگذار بر آن و مشخص نمودن پهنه‌های آلوده ضروری به نظر می‌رسد؛ بنابراین استفاده از مدل‌های ریاضی در قالب یادگیری ماشینی رویکردی بهینه و مقرون به صرفه برای مدل‌سازی آلودگی هواست. این تحقیق به لحاظ هدف کاربردی بوده و روش بررسی آن توصیفی-تحلیلی است. نوآوری تحقیق حاضر ارائه‌یک رویکرد ترکیبی جدید جهت تعیین معیارهای مؤثر در پیش‌بینی میزان آلودگی هوا می‌باشد. لذا هدف از تحقیق حاضر ارزیابی و مقایسه قابلیت دو مدل یادگیری ماشین، یعنی ماشین بردار پشتیبان (SVM) و جنگل تصادفی (RF) در ترکیب با الگوریتم ژنتیک (GA) جهت پیش‌بینی میزان آلودگی هوا در شهرستان تهران است. داده‌های مورداستفاده در این تحقیق شامل ذرات معلق و آلاینده‌های گازی شهر تهران مرتبط با سال ۱۳۹۹ می‌باشد که از شرکت کنترل ترافیک شهر تهران اخذ گردیده است. به منظور تجزیه‌وتحلیل داده‌ها از نرم‌افزارهای Matlab و ArcMap استفاده شد. مقدار

۱ نویسنده مسئول: ۰۹۱۵۳۶۱۶۶۹۶
Email: Moakbari@birjand.ac.ir

نحوه ارجاع به این مقاله:

کرمی، پیمان؛ اسلامی نژاد، سید احمد؛ افتخاری، مبین؛ برومند، فراز؛ اکبری، محمد؛ ۱۴۰۲. توسعه الگوریتم‌های یادگیری ماشین جهت پیش‌بینی شاخص کیفیت هوای شهری (منطقه مطالعاتی: شهر تهران). *جغرافیا و مخاطرات محیطی*. ۱۲(۲)صص ۱۸۶-۱۶۵
https://doi.org/10.22067/geoeh.2022.76121.1212

ضریب تشخیص (R^2) حاصل از روش ترکیبی RF-GA برابر ۰/۹۹۷ به دست آمد که نشان‌دهنده سازگاری بالای این مدل با داده‌های این تحقیق است. همچنین مقدار ریشه میانگین خطای مربعات (RMSE) برابر ۰/۱۵۳ به دست آمد که نشان‌دهنده دقت بالای این مدل می‌باشد. بر اساس اطلاعات گرفته شده از شرکت کنترل ترافیک شهر تهران، نتایج حاصل از روش RF بیانگر مناسب بودن انتخاب مدل مذکور جهت برآورد میزان آلودگی هوای شهر تهران بوده است.

کلیدواژه‌ها: آلودگی هوا، یادگیری ماشین، جنگل تصادفی، ماشین بردار پشتیبان، الگوریتم ژنتیک.

۱- مقدمه

آلودگی هوا زمانی اتفاق می‌افتد که حجم زیادی از ذرات یا مواد مضر از قبیل گازها، ذرات و مولکول‌های بیولوژیکی وارد اتمسفر کره زمین شود. آلودگی هوا مخلوطی از ذرات معلق و گازهایی است که غلظت آن به محدوده مضر برای انسان رسیده است که می‌تواند هم در داخل ساختمان و هم در خارج ساختمان باشد (اکبری و همکاران[1]، ۲۰۲۱). شاخص کیفیت هوا[2] (AQI) یا شاخص آلودگی هوا، شاخصی عددی است که توسط سازمان‌های دولتی برای سنجش آلودگی هوای یک منطقه و پیش‌بینی آینده آن به کار می‌رود (ژو و همکاران[3]، ۲۰۱۹). با افزایش عدد این شاخص، احتمال در خطر قرار گرفتن سلامت عمومی بالاتر می‌رود. با افزایش عدی شاخص کیفیت هوا، درصد بیشتری از مردم احتمالاً دچار پیامدهای بهداشتی نابجای شدید ناشی از آلودگی هوا می‌شوند (اکبری و همکاران[4]، ۲۰۲۱). کشورهای متفاوت بر حسب استانداردهای ملی متفاوتی که دارند شاخص‌های کیفیت هوای خاص خودشان را دارند. مقادیر شاخص کیفیت هوا معمولاً به صورت طیف‌هایی گروه‌بندی می‌شوند. هر طیف با یک نام توصیف‌کننده، یک کدرنگی و توصیه‌های استانداردشده بهداشت عمومی مشخص می‌شود. محاسبه AQI نیاز به اندازه‌گیری غلظت یک ماده آلاینده در طول یک دوره میانگین مشخص دارد که به وسیله ایستگاه‌های پایش هوا یا مدل‌ها به دست می‌آید (کومار[5]، ۲۰۱۸). غلظت یک ماده آلاینده و مدت زمان حضور آن بیانگر دوز یک آلاینده هوا است. شاخص کیفیت هوا بر اساس ذرات معلق (PM2.5، PM10)، اوزون (O3)، دی اکسید نیتروژن (NO2)، دی اکسید گوگرد (SO2) و انتشار کربن مونوکسید (CO) محاسبه می‌شود (لیو و همکاران[6]، ۲۰۱۹). در حال حاضر یکی از مشکلات عظیم کلان‌شهرهایی مثل تهران، وجود حجم زیادی از آلاینده‌های مختلف می‌باشد که مهمترین این آلاینده‌ها، ذرات معلق (PM2.5، PM10) و آلاینده‌های گازی اوزون (O3)، دی اکسید نیتروژن (NO2)، دی اکسید گوگرد (SO2) و انتشار کربن مونوکسید (CO) است. عدم استفاده از سیستم‌های تحلیل مکانی و توصیفی آلاینده‌های

1 Akbari et al.
2 Air quality index
3 Xue et al.
4 Akbari et al.
5 Kumar
6 Liu et al.

فوق بصورت توأم و همچنین عدم وجود یک سیستم منسجم در ذخیره سازی، بازیابی، بـه‌هنگام‌سازی، مدیریت، پردازش، نمایش، کاربرد و تبادل داده‌های زیست‌محیطی، لزوم استفاده از آخرین علوم و فناوری‌ها را در این زمینه مشخص می‌سازد. لذا هدف این تحقیق برآورد مکانی میزان آلودگی هوا شـــهر تهران بر پایه ترکیب الگوریتم‌های یادگیری ماشـین و الگوریتم ژنتیک جهت شـناسـایی آلاینده‌های گازی مؤثر می‌باشـد. با توجه به تحقیقات پیش بین، برآورد میزان تراکم آلودگی هوا (پهنه‌بندی آلودگی هوا) توسـط روش‌های ادغام در دو دسـته کلی رویکرد دانش‌محور و رویکرد داده‌محور قابل دسـته‌بندی می‌باشـند (وانگ و لیو[1]، ۲۰۱۹؛ گواوارا و همکاران[2]، ۲۰۱۹). رویکرد داده‌محور در مناطق شناخته شده یا مناطقی که از لحاظ آماری تعداد شواهد شناخته شده کافی می‌باشند، کارآیی بالایی دارد. در این روش‌ها هدف مشـخص کردن مکان‌های جدید برای کارهای تفصـیلی‌تر اسـت. از جمله این روش‌ها می‌توان به روش‌های یادگیری ماشین شـامل جنگل تصـادفی (RF)[3]، ماشین بردار پشـتیبان (SVM)[4] و ... اشاره کرد. در حالی‌که رویکرد دانش‌محور در محیط‌های که کمتر شناخته شده‌اند و یا تعداد کمی از اهداف موردنظر در محدوده وجود دارند کارآمد هستند. تخمین وزن‌ها برای نقشه‌های شاهد و تخمین کلاس‌ها در هر نقشه شاهد بر اساس قضاوت کارشناس و با توجه به ویژگی‌های نشانه‌ها است؛ بنابراین در روش‌های دانش‌محور پارامترهای تابع برای ترکیب داده‌ها بر اساس دانش تجربی تخمین زده می‌شـود (گواوارا و همکاران، ۲۰۱۹). مطالعات متعددی در خصـوص برآورد میزان تراکم آلودگی هوا (پهنه‌بندی آلودگی هوا) توسـط دو رویکرد دانش‌محور و داده‌محور انجام شـده اسـت که می‌توان به موارد زیر اشاره نمود:

خزایی و همکاران (۱۳۹۱) غلظت آلاینده مونواکسیدکربن را با تلفیق روش شبکه عصبی- فازی با GIS مدلسازی کردند؛ به عبارت دیگر در این مقاله با به‌کارگیری شـبکه عصـبی- فازی و GIS، دانش حاکم بر محیط در قالب قوانین فازی، از داده‌ها اسـتخراج شـده و با اسـتفاده از این قوانین غلظت آلاینده مونواکسیدکربن مدلسـازی شـده اسـت. منطقه مورد مطالعه در این کار تحقیقی شـهر تهران در نظر گرفته شـد. جهت پیاده‌سـازی، داده‌های هواشـناسـی شـش ایسـتگاه موجود در سـطح شـهر تهران در فصـل تابسـتان برای چهار سـال متوالی به طور جداگانه بررسـی شـده و به منظور ورود به فرآیند آموزش شـبکه عصـبی مورد اسـتفاده قرار گرفت. برای هر ایسـتگاه قوانین فازی آن اسـتخراج شـده و غلظت آلاینده تخمین زده شـد. به علت اینکه در این پژوهش پیش‌بینی در ایسـتگاه‌ها انجام می‌گیرد، برای مدلسـازی مکانی غلظت در محدوده مورد مطالعه از روش کریجینگ اسـتفاده شـده و میزان خطای مربوطه نیز محاسـبه شـد. رحیمی و همکاران (۱۳۹۲) در مقاله خود تداوم روزهای همراه با آلاینده مونواکسیدکربن را در هوای شهر تهران ارزیابی نمودند که این کار با اسـتفاده از مدل ریاضـی زنجیره مارکف انجام شـده اسـت. برای این کار اطلاعات ۵ سـاله پنج ایسـتگاه

1 Wang and Liu
2 Guevara et al.
3 Random Forest
4 Support vector machine

سنجش آلودگی شرکت کنترل کیفیت هوای تهران گردآوری و با استفاده از زنجیره مارکف مدلسازی گردید. نتایج این پژوهش نشان داد که بیشترین احتمال وقوع تداوم آلاینده CO به ترتیب در ایستگاه‌های فاطمی، بازار و اقدسیه وجود دارد و در اکثر ماه‌های سال ایستگاه فاطمی بالاترین احتمال وقوع تداوم دو روزه CO را دارد. میری و همکاران (۱۳۹٤) جهت بررسی مکانی آلودگی هوای کلان شهر مشهد از سه مدل درونیابی کریجینگ معمولی، کریجینگ عمومی و معکوس فاصله وزنی (IDW)[1] استفاده کردند. آن‌ها جهت مقایسه مدل‌ها و انتخاب بهترین مدل از ریشه میانگین خطای مربعات (RMSE)[2] و ضریب تعیین (R^2) استفاده کردند. نتایج نشان داد که مدل کریجینگ معمولی دارای کمترین مقدار RMSE و بیشترین مقدار R^2 نسبت به سایر مدل‌های استفاده شده می‌باشد. حق بیان و تشیع (۱۳۹۹) از مدل رگرسیون کاربری اراضی بهبودیافته جهت مدل‌سازی آلاینده‌های هوا به منظور مدیریت مواجهه با استفاده از داده‌های حاصل از حسگرهای همراه استفاده کردند. به منظور بهبود دقت مدل‌سازی روش موردنظر برای تخمین غلظت PM2.5 از هفت ایستگاه ثابت شهر اصفهان و چهارده حسگر همراه استفاده گردید. نتایج نشان داد که حتی با افزودن یک حسگر همراه به ایستگاه‌های ثابت میزان RMSE به مقدار ۰/۱۱۳ میکروگرم بر متر مکعب کاهش می‌یابد و با افزودن چهارده حسگر همراه به هفت ایستگاه ثابت میزان RMSE حدود سه برابر کاهش می‌یابد. نجات‌کورکی و باروتیان[3] (۲۰۱۲) در تحقیقی بر روی پیش‌بینی حداکثر غلظت PM10 در طی ۲٤ ساعت آتی در شهر تهران پرداختند. از این رو از داده‌های هواشناسی و غلظت آلاینده‌ها به عنوان پارامترهای ورودی شبکه پس‌انتشار خطا استفاده شد. نتایج پیش‌بینی شده با شاخص دقت بالای ۰/۸۳، مطلوب نشان داده شد. از طرف دیگر شبکه با عملکرد مطلوب به خوبی می‌تواند نسبت به سایت‌های سنجش انسانی در شبکه پایش کیفیت هوا برتری داشته باشد. ویمن و همکاران[4] (۲۰۱۲) در تحقیقی به مدلسازی آلودگی هوا در ایالت Saxony آلمان پرداختند. لذا از مدل IDW برای تخمین غلظت آلاینده‌های PM10 و O3 در منطقه موردمطالعه استفاده می‌کنند. مک‌کندی[5] (۲۰۱۵) از مدل‌های شبکه عصبی مصنوعی (ANN)[6] و رگرسیون خطی ساده (MLR)[7] جهت پیش‌بینی حداکثر و متوسط روزانه مقدار O3 و ذرات معلق (PM10 و PM2.5) استفاده کرد. در واقع یکی از محدودیت‌های رگرسیون خطی ساده، خطی بودن این مدل می‌باشد؛ اما ممکن است بین خروجی‌ها و ورودی‌ها رابطه غیرخطی برقرار باشد که می‌توان از شبکه عصبی مصنوعی (ANN) به این منظور استفاده کرد. نتایج نشان داد که مدل ANN توانایی بالاتری در پیش‌بینی حداکثر و متوسط روزانه O3 و ذرات معلق (PM10 و PM2.5) دارد. آدامز و کلناروگلو[8] (۲۰۱٦)، در مطالعه‌ای از

1 Inverse Distance Weighted
2 Root Mean Square Error
3 Nejadkoorki and Baroutian
4 Wiemann et al.
5 McKendry
6 Artificial neural network
7 Multiple linear regression
8 Adams and Kanaroglou

مدل شبکه عصبی برای پایش آلودگی هوای ایستگاه‌ها و برآورد مقدار AQI حاصل از دو آلاینده PM2.5 و NO2 در شهر همیلتون کانادا استفاده کردند. نتایج نشان داد که ضریب همبستگی برای آلاینده PM2.5 و NO2 به ترتیب ۰/۷۸ و ۰/۳٤ محاسبه گردید. مسعودی و گرامی[1] (۲۰۱۷) کیفیت هوای شهر اصفهان را بر اساس میزان مونوکسید کربن (CO) مورد تجزیه‌وتحلیل قرار دادند. نتایج نشان داد که بیشترین میزان غلظت CO در صبح و ابتدای شب اتفاق می‌افتد. در واقع هدف اصلی تحقیقشان استفاده از مدلی بود که بتوانند ارتباط بین غلظت آلاینده‌ها و پارامترهای هواشناسی را بررسی نمایند. از این رو از شبکه عصبی مصنوعی پرسپترون سه لایه و رگرسیون خطی برای پیش‌بینی غلظت آلاینده‌های CO و PM10 استفاده کردند و مشخص گردید که شبکه عصبی مصنوعی پرسپترون سه لایه با توجه به در نظر گرفتن روابط غیرخطی بین آلاینده‌ها، دقت بالاتری در پیش‌بینی غلظت آلاینده‌های CO و PM10 دارد. پارک و همکاران[2] (۲۰۱۸) جهت پیش‌بینی غلظت PM10 شهر سئول کشور کره جنوبی از شبکه عصبی مصنوعی (ANN) استفاده کردند. نتیجه نشان داد مدل ANN ضریب همبستگی بالایی را بین مقادیر اندازه‌گیری شده و مقادیر واقعی غلظت PM10 نشان می‌دهد. فرهادی و همکاران[3] (۲۰۲۰) از شبکه عصبی مصنوعی پرسپترون سه لایه جهت پیش‌بینی غلظت آلاینده‌های PM10 و CO هوای شهر تهران استفاده کردند. نتایج نشان داد که بیشترین مقدار R^2 برای آلاینده PM10 با مقدار ۰/۸۳ برای فصول گرم بود و هم‌چنین بیشترین مقدار R^2 برای آلاینده CO با مقدار ۰/۷٦ برای فصول سرد است. سونگ و همکاران[4] (۲۰۲۱) از مدل‌های RF و رگرسیون کاربری اراضی برای برآورد تغییرات مکانی–زمانی آلاینده‌های PM2.5 و NO2 در شهر شانگهای چین استفاده کردند. جهت پیاده‌سازی مدل‌های موردنظر، از ۸۰ متغیر پیش‌بینی‌کننده مختلف مرتبط با شرایط جوی و جغرافیایی، حمل‌ونقل، تراکم جمعیت، کاربری زمین و نقاط مورد علاقه استفاده شد. نتایج نشان داد که مدل RF دقت بالاتری در برآورد تغییرات مکانی–زمانی آلاینده‌های PM2.5 و NO2 نسبت به مدل رگرسیون کاربری اراضی دارد.

بررسی پیشینه تحقیقات نشان داد که با توجه به اهمیت موضوع آلودگی هوا، مطالعات زیادی در این زمینه انجام گرفته است که هریک تلاش نموده راه‌حل‌ها و راهکارهای پیشنهادی را ارائه نمایند. از آن‌جایی که آلودگی هوا یک مساله پیچیده و چندوجهی می‌باشد و سازوکار مدلسازی آن خود به تنهایی مسأله بسیار پیچیده‌ای می‌باشد، این تحقیق درنظر دارد بدون درگیر شدن با مفاهیم پیچیده آلودگی هوا و معادلات شیمیایی شکل‌دهنده، به پیش‌بینی آلودگی هوا از نقطه نظر مکانی به مسأله پرداخته و روابط آلودگی هوا را تنها با تکیه بر معادلات مکانی مدلسازی نماید. برآورد میزان تراکم آلودگی هوا (پهنه‌بندی آلودگی هوا) موضوعی است که تاکنون زیاد بدان پرداخته شده است؛ اما در میان مطالعات صورت پذیرفته، نکاتی وجود دارد که کم‌تر بدان توجه شده است؛ اول این‌که در هیچ یک

1 Masoudi and Gerami
2 Park et al.
3 Farhadi et al.
4 Song et al.

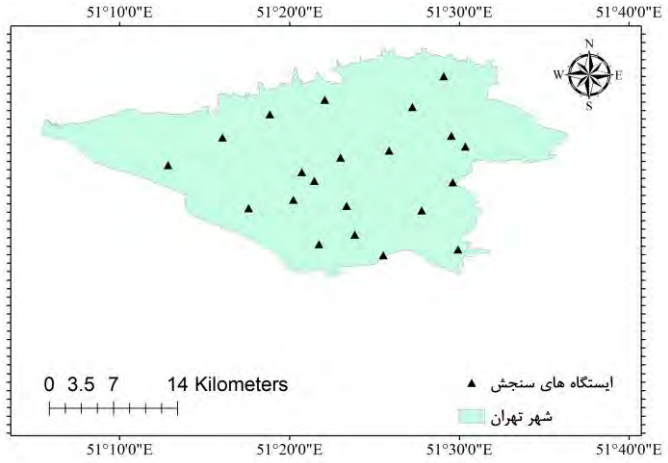
از مطالعات صورت گرفته، ترکیب مناسب و کافی از آلاینده‌های گازی برای برآورد میزان تراکم آلودگی هوا در نظر گرفته نشده است. دوم این‌که تحلیل مناسبی برای تعیین ترکیب بهینه معیارهای مؤثر و تهیه نقشه برآورد میزان تراکم آلودگی هوا بر اساس تأثیرات معیارهای مؤثر به کار برده نشده است. هدف از این پژوهش تهیه نقشه پهنه‌بندی آلودگی هوا با استفاده از روش‌های نوین یادگیری ماشین مبتنی بر الگوریتم فراابتکاری ژنتیک (GA)[1] است. بنابراین، این مطالعه از مطالعات قبلی متمایز است، زیرا در این مطالعه از مدل‌های ترکیبی یادگیری ماشین توسعه‌یافته یعنی RF-GA و SVM-GA جهت برآورد میزان تراکم آلودگی هوای شهر تهران بر مبنای تعیین ترکیب بهینه آلاینده‌های گازی استفاده شده است که نوآوری تحقیق حاضر نیز می‌باشد. در نهایت، معیار سطح زیر منحنی (AUC)[2] و معیارهای آماری شامل ضریب تشخیص (R^2) و ریشه میانگین خطای مربعات (RMSE) برای اعتبارسنجی مدل‌های پیش‌بینی آلودگی هوا در منطقه موردمطالعه مورد استفاده قرار گرفتند.

۲- مواد و روش‌ها

۲-۱- معرفی منطقه موردمطالعه

شهر تهران به عنوان پایتخت کشور، مهم‌ترین کلان‌شهر و مرکز سیاسی و تجاری کشور ایران محسوب می‌شود که بالغ بر ۲۰ درصد جمعیت کشور در آن ساکن هستند. تهران توسط رشته‌کوه‌های البرز از سمت شمال و دشت کویر از سمت جنوب احاطه شده است. آب‌وهوای شهر تهران تأثیر گرفته از موقعیت جغرافیایی آن است. جز مناطق شمالی تهران که تحت تأثیر کوهستان، تا اندازه‌ای معتدل و مرطوب هستند، آب‌وهوای دیگر مناطق شهر تقریباً گرم و خشک و در زمستان اندکی سرد است. رشته‌کوه البرز همچون سدی از نفوذ بسیاری از توده‌های هوا جلوگیری می‌کند، از همین روی سبب گردیده که تهران از آب‌وهوایی نسبتاً خشک برخوردار باشد. محصور بودن در بین کوه‌ها از سه طرف، که مانع خروج آلودگی‌ها از شهر می‌شود، از یک سو و افزایش بی‌رویه استفاده از وسایل نقلیه و گسترش صنایع از عوامل اصلی آلودگی هوا در شهر تهران می‌باشند. آلودگی هوا در شهر تهران عمدتاً مصنوعی و ناشی از فعالیت وسایل نقلیه است که سهم بالایی در آلودگی هوای شهر دارند؛ بنابراین پیش‌بینی و مدل‌سازی آلودگی هوا برای شهر تهران امری ضروری بوده تا اقدامات لازم جهت کنترل آلودگی انجام شده و مکان‌هایی که از نظر آلودگی در وضعیت خطرناکی قرار دارند، شناسایی گردند. برای این منظور جهت پیش بینی آلودگی هوا، کل شهر تهران به عنوان منطقه مطالعاتی انتخاب شده است. در شکل ۱ نقشه مناطق شهر تهران به همراه ایستگاه‌های سنجش آلودگی هوا اخذ شده از شرکت کنترل کیفیت شهرداری تهران نشان داده شده است.

1 Genetic algorithm
2 Area under the curve

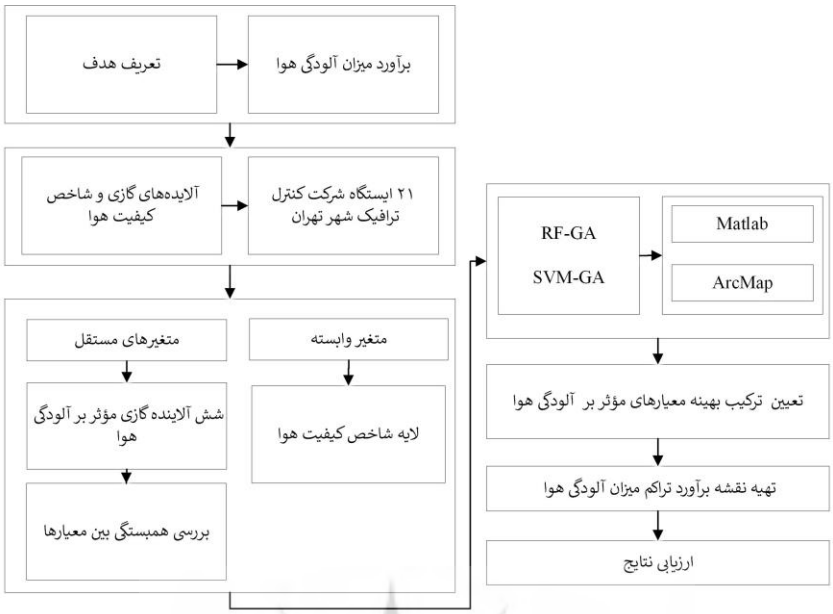


شکل ۱- محدوده منطقه مطالعاتی و ایستگاه‌های سنجش آلودگی هوای شهر تهران (مأخذ: نویسندگان)

داده‌های مورداستفاده در این تحقیق شامل ذرات معلق (PM2.5، PM10)، آلاینده‌های گازی اوزون (O3)، دی اکسید نیتروژن (NO2)، دی اکسید گوگرد (SO2) و کربن مونوکسید (CO) است. این داده‌ها مربوط به سال ۱۳۹۹ شهر تهران می‌باشد که از شرکت کنترل ترافیک شهر تهران اخذ شده است.

۲-۲- روش انجام پژوهش

روش انجام پژوهش توصیفی-تحلیلی بوده و نوع آن بر اساس هدف کاربردی است. مبانی تئوریک تحقیق بر اساس مطالعات اسنادی و کتابخانه‌ای انجام گرفته است. در راستای رسیدن به هدف این تحقیق که تعیین آلاینده‌های گازی مؤثر در برآورد میزان آلودگی هوای شهر تهران می‌باشد، از داده‌های شرکت کنترل ترافیک شهر تهران در سال ۱۳۹۹ استفاده شده است. تمامی پردازش‌های مربوط به داده‌ها در محیط نرم‌افزاری ArcMap و محاسبات کمی آن در محیط نرم‌افزاری Matlab انجام گرفته است. در این تحقیق از الگوریتم‌های یادگیری ماشین ترکیبی شامل RF-GA و SVM-GA جهت برآورد میزان آلودگی هوا استفاده شده است. نهایتاً پس از تعیین معیارهای مؤثر در برآورد میزان آلودگی هوا و تهیه نقشه معیارهای مختلف، میزان آلودگی هوای شهر تهران پیش‌بینی شده است. در شکل ۲ روند اجرایی تحقیق نشان داده شده است.



شکل ۲- روند اجرایی تحقیق (مأخذ: نویسندگان)

۲-۲-۱- الگوریتم جنگل تصادفی (RF)

الگوریتم جنگل تصادفی یکی از رایج‌ترین الگوریتم‌های به کار رفته برای بررسی مشکلات طبقه‌بندی و پیش‌بینی چندگانه است که حساسیت کمی به چندخطی بودن دارد و نتایج آن از نظر داده‌های از دست رفته و نامتعادل نسبتاً پایدار است (کویروز و همکاران[1]، ۲۰۱۸؛ دی سانتانا و همکاران[2]، ۲۰۱۸، اسلامی‌نژاد و همکاران، ۱۴۰۰ب). مدل پیش‌بینی‌کننده RF بر اساس میانگین‌گیری از نتایج حاصل از تمامی درخت‌های تصمیم مربوطه استوار است و برای بسیاری از مجموعه داده‌ها، طبقه‌بندی را با صحت بالایی انجام می‌دهد (دی سانتانا و همکاران، ۲۰۱۸). چهار مرحله زیر فرآیند الگوریتم RF را بیان می‌کند: ۱) تعریف و بازنمونه‌گیری داده‌های آموزشی؛ ۲) انتخاب مجموعه ویژگی‌های تصادفی مربوط به هر نمونه مجدد؛ ۳) اختصاص یک درخت تصمیم برای هر کدام از آن‌ها به مجموعه ویژگی‌های تصادفی و گسترده؛ ٤) ایجاد یک درخت تصمیم واحد از طریق تجمیع درخت تصمیم اختصاص‌داده‌شده به هر مثال.

1 Quiroz et al.
2 De Santana et al.

۲-۲-۲- الگوریتم ماشین بردار پشتیبان (SVM)

ماشین بردار پشتیبان (SVM)، یک روش یادگیری ماشین نسبتاً جدید و یک الگوریتم یادگیری ماشین نظارت شده است (عرب‌گل و همکاران[1]، ۲۰۱۶). الگوریتم SVM یکی از متقاعدکننده‌ترین روش‌های پیش‌بینی است که براساس روش حداقل‌سازی ریسک ساختاری می‌باشد. در مقابل، بیشتر مدل‌های هوش مصنوعی مانند شبکه‌های عصبی مصنوعی، از تکنیک‌های به حداقل رساندن ریسک تجربی استفاده می‌کنند؛ بنابراین، روش SVM می‌تواند خطای تجربی را کاهش دهد، پیچیدگی را مدل کند و احتمال را بیش از حد تنظیم کند (قربان‌زاده و همکاران[2]، ۲۰۱۹). هدف SVM پیدا کردن ابر صفحه جداساز بهینه است که بتواند حاشیه را بین کلاس‌های مختلف مشخص کرده و فاصله یک کلاس را به حداقل برساند. در بیشتر شرایط، ابر صفحه توسط یک سطح غیر خطی تعریف خواهد شد. در این مورد، عبارت ریاضیاتی زیر برای طبقه‌بندی مجموعه داده‌ها به کار گرفته خواهد شد (عرب‌گل و همکاران، ۲۰۱۶):

$$f(x) = \sum_{i=1}^{n} (a_i - a_i^*) K(x_i, x) + b \qquad (۱)$$

که در آن α_i و α_i^* ضرایب لاگرانژ، K تابع کرنل و b انحراف ابرصفحه از مبدأ است.

۲-۲-۳- الگوریتم ژنتیک (GA)

الگوریتم ژنتیک را می‌توان یک روش جستجوی کلی نامید که از قوانین تکامل بیولوژیک طبیعی تقلید می‌کند (میرجلیلی[3]، ۲۰۱۹). به منظور حل هر مسئله با استفاده از الگوریتم‌های ژنتیکی، ابتدا باید یک تابع هدف برای آن مسئله ابداع شود. برای هر کروموزوم، این تابع عددی غیر منفی را برمی‌گرداند که نشان‌دهنده شایستگی یا توانایی فردی آن کروموزوم است. در الگوریتم‌های ژنتیکی، در طی مرحله تولیدمثل از عملگرهای ژنتیکی استفاده می‌شود. با تأثیر این عملگرها بر روی یک جمعیت، نسل بعدی آن جمعیت تولید می‌شود. عملگرهای انتخاب[4]، ترکیب[5] و جهش[6] معمولاً بیشترین کاربرد را در الگوریتم‌های ژنتیکی دارند (میرجلیلی، ۲۰۱۹). به طور کلی مراحل اجرای یک مدل بهینه‌سازی توسط الگوریتم ژنتیک به شکل زیر می‌باشد (سان و همکاران[7]، ۲۰۲۰):

- ایجاد جمعیت تصادفی و ارزیابی آن‌ها
- انتخاب والدین و ترکیب آن‌ها برای ایجاد جمعیت اولیه فرزندان
- انتخاب اعضای جمعیت برای اعمال جهش و ایجاد جمعیت جهش‌یافتگان

1 Arabgol et al.
2 Ghorbanzadeh et al.
3 Mirjalili
4 Selection
5 Crossover
6 Mutation
7 Sun et al.

- ترکیب یا ادغام جمعیت اصلی، فرزندان و جهش‌یافتگان و ایجاد جمعیت اصلی جدید
- اگر شرایط خاتمه محقق نشده باشند، از مرحله ۲ تکرار می‌شود
- پایان

۲-۳- ارزیابی عملکرد و دقت مدل‌ها

۲-۳-۱- شاخص‌های آماری

خروجی الگوریتم‌های یادگیری ماشین شامل پارامترهای متعددی است که از آن میان معمولاً پارامتر R^2 برای سنجش مناسبت برازش مدل و پارامتر RMSE جهت سنجش توزیع باقیمانده‌های مدل به کار می‌روند که به ترتیب، طبق روابط (۲) و (۳) محاسبه می‌شوند (اوشان و همکاران[1]، ۲۰۱۹؛ ویلر[2]، ۲۰۱٤):

(۲)
$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

(۳)
$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$$

که n تعداد مشاهدات، y_i مشاهده i ام، $\hat{y}_i$ مقدار پیش‌بینی مشاهده i ام و $\bar{y}$ میانگین مشاهدات است.

۲-۳-۲- منحنی مشخصه عملکرد سامانه (ROC)

در این مطالعه منحنی مشخصه عملکرد سامانه[3] (ROC) برای ارزیابی عملکرد مدل‌ها به ترتیب با نرخ‌های مثبت واقعی و نرخ مثبت کاذب بر روی محور Y و محور X استفاده شده است (گورسوسکی و همکاران[4]، ۲۰۰٦). منطقه تحت منحنی ROC یعنی (AUC)[5] عملکرد مدل را به صورت کمی تعیین می‌کند (تین بوی و همکاران[6]، ۲۰۱۹). مقادیر بالاتر AUC (نزدیک به ۱) نشان‌دهنده عملکرد بهتر مدل‌ها می‌باشد (تین بوی و همکاران، ۲۰۱۹). خوبی تناسب یا قابلیت یادگیری مدل با استفاده از منحنی ROC توسط مجموعه داده‌های آموزشی مشخص می‌شود؛ در حالی که مجموعه داده‌های آزمایشی، مهارت پیش‌بینی مدل را نشان می‌دهند (فاوست[7]، ۲۰۰٦).

1 Oshan et al.
2 Wheeler
3 receiver operating characteristics curve
4 Gorsevski et al.
5 Area under the curve
6 Tien Bui et al.
7 Fawcett

۳- تجزیه‌وتحلیل داده‌ها

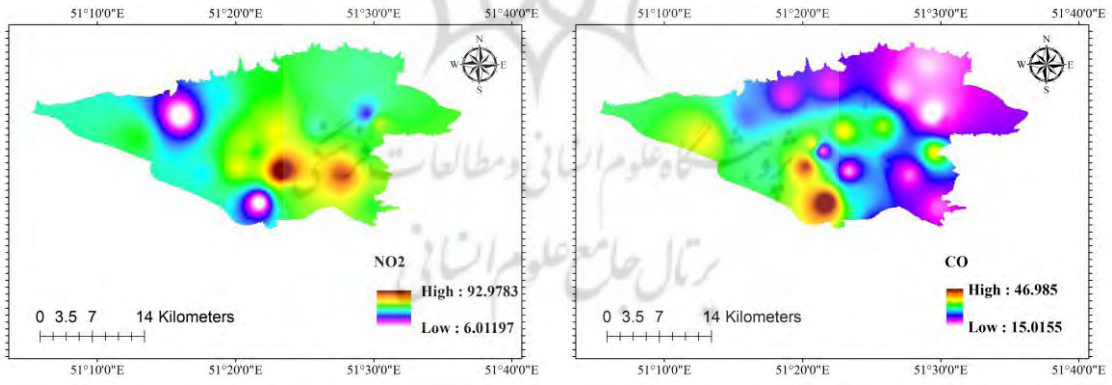
۳-۱- آماده‌سازی داده‌ها

در این تحقیق تأثیر شش معیار مؤثر بر میزان آلودگی هوا در محدوده مطالعاتی مورد بررسی قرار گرفته است که در جدول ۱ نشان داده شده است. این معیارها با توجه به مطالعات گذشته و همچنین محدودیت‌های موجود در دسترسی به داده‌ها انتخاب شده است.

جدول ۱- آلاینده‌های گازی مؤثر بر آلودگی هوا (مأخذ: نویسندگان)

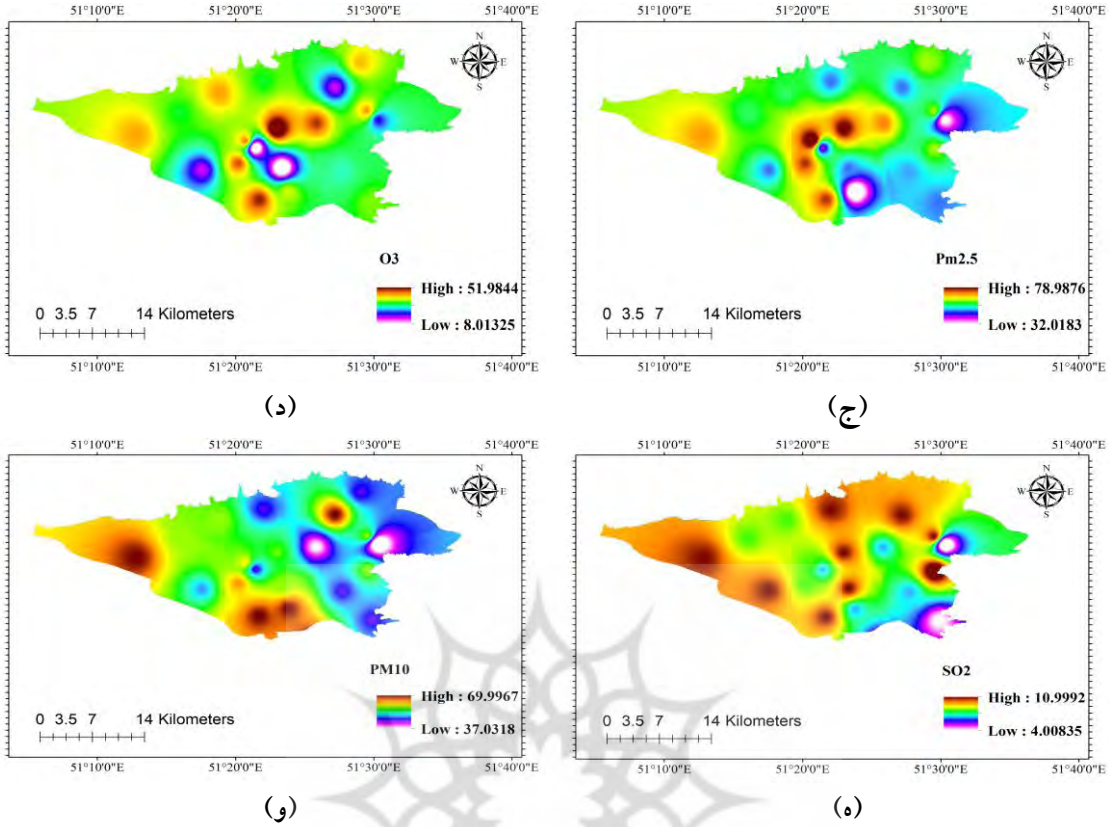
شماره	معیارها	شماره	معیارها
۱	SO2	٤	NO2
۲	CO	٥	PM2.5
۳	O3	٦	PM10

در شکل ۳ نقشه رستری مربوط به میانگین مقادیر ذرات معلق و آلاینده‌های گازی در سال ۱۳۹۹ نشان داده شده است. برای ایجاد لایه مکانی هر کدام از ذرات معلق و آلاینده‌های گازی، میانگین سالیانه آن‌ها برای ۲۱ ایستگاه محاسبه شده و سپس به کمک روش درونیابی کریجینگ، هر کدام از لایه‌ها در قالب نقشه رستری با اندازه پیکسل ۳۰ متر تولید شد (اسلامی‌نژاد و همکاران، ۱٤۰۰ الف، ۱۷۹؛ افتخاری و همکاران[1]، ۲۰۲۱).



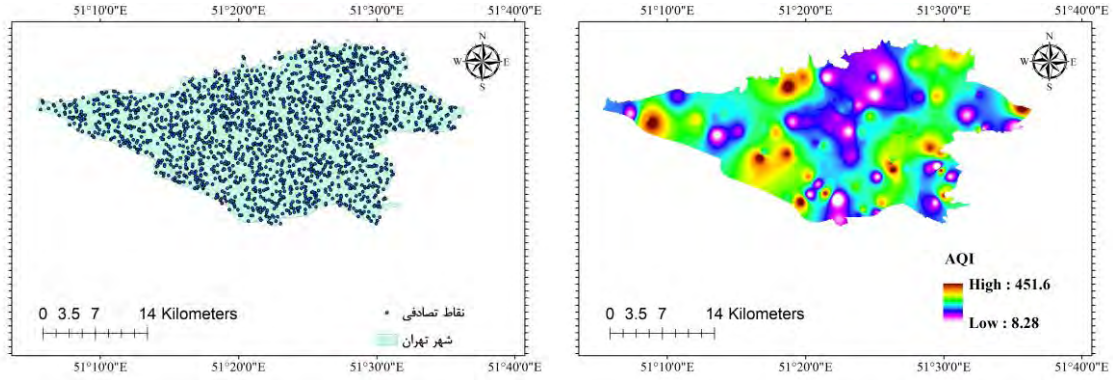
(ب) (الف)

1 Eftekhari et al.



شکل۳- لایه‌های به کار برده شده در منطقه موردمطالعه (الف) CO (ب) NO2 (ج) PM2.5 (د) O3 (ه) SO2 (و) PM10 (مأخذ: نویسندگان)

جهت تولید لایه آلودگی شهر تهران در بازه زمانی مشخص، میانگین مقدار AQI محاسبه شده برای ۲۱ ایستگاه شرکت کنترل ترافیک شهر تهران در سال ۱۳۹۹ محاسبه و سپس به کمک روش درونیابی کریجینگ لایه مکانی آن در قالب نقشه رستری با اندازه پیکسل ۳۰ متر تولید شد (شکل ٤ (الف)). همچنین جهت پیاده‌سازی روش‌های پیشنهادی نیاز به تولید نقاط پراکنده در منطقه موردنظر است. از این رو بر اساس تحلیل Random point در نرم‌افزار Arc Map، ۲۰۰۰ نقطه به صورت تصادفی و یکنواخت تولید شد که این نقاط در شکل ٤ (ب) قابل مشاهده است. سپس مقادیر تمامی لایه‌های اطلاعاتی موجود (نرمال شده متغیرهای مستقل و وابسته) برای این نقاط محاسبه شد. از این تعداد نقاط ۷۰٪ برای آموزش و ۳۰٪ برای آزمایش به صورت تصادفی انتخاب و به صورت یکسان برای تمامی روش‌ها مورد استفاده قرار گرفت.



(ب) (الف)

شکل ٤- (الف) نقشه میزان آلودگی شهر تهران بر اساس میانگین شاخص AQI (ب) نقاط تصادفی تولید شده در منطقه مورد مطالعه (مأخذ: نویسندگان)

نکته مهمی که باید قبل از پیاده‌سازی الگوریتم‌های یادگیری ماشین انجام شود، بررسی مستقل بودن معیارهای موردنظر می‌باشد. به منظور تشخیص هم‌خطی چندگانه در میان عوامل مختلف، آماره تحمل (TOL)[1] و عامل تورم واریانس (VIF)[2] دو پارامتر آماری رایج هستند. طبق جدول ۲، هنگامی‌که مقدار TOL بزرگ‌تر از ۰/۱ و مقدار VIF کوچک‌تر از ۵ باشد، هم‌خطی چندگانه بالایی در میان متغیرهای پیش‌بینی‌کننده وجود ندارد (افتخاری و همکاران، ۱٤۰۰)؛ بنابراین از تمام معیارها در الگوریتم‌های پیشنهادی استفاده گردید.

جدول ۲- آلاینده‌های گازی مؤثر بر آلودگی هوا (مأخذ: نویسندگان)

شماره	معیارها	VIF	TOL
۱	SO2	٤/۲۲	۰/۱٤۸
۲	CO	۳/۹۸	۰/۱٦۹
۳	O3	۲/۲٦	۰/۳٤۵
٤	NO2	۲/۷۸	۰/٤۹۲
۵	PM2.5	۱/۹۵	۰/٦۵۸
٦	PM10	۱/٤٤	۰/٤۵۸

۳-۲- پیش‌بینی آلودگی هوا توسط الگوریتم‌های یادگیری ماشین

با توجه به این‌که یکی از مهم‌ترین پارامترهای ارزیابی روش‌های چندمعیاره، پارامتر RMSE است، از این‌رو تابع برازش الگوریتم ژنتیک، کمینه کردن مقدار RMSE انتخاب شده است تا میزان سنجش توزیع باقیمانده‌های مدل و در

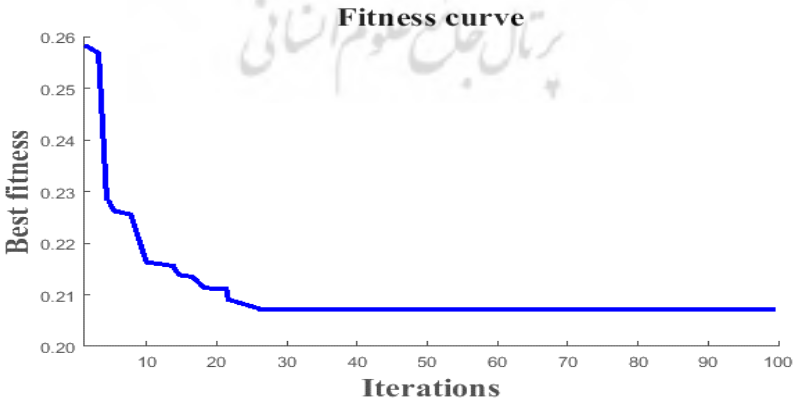
1 Tolerance
2 Variance Inflation Factor

واقع دقت مدل در برآورد میزان آلودگی هوا بررسی شود (اوشان و همکاران، ۲۰۱۹؛ ویلر، ۲۰۱٤). طبق جدول ۳ مقادیر بهینه پارامترهای اولیه الگوریتم ژنتیک، بر اساس روش سعی‌وخطا انتخاب شد. شرط توقف جهت ساده‌سازی روند پیاده‌سازی، تعداد اجرای خاص در نظر گرفته شده است.

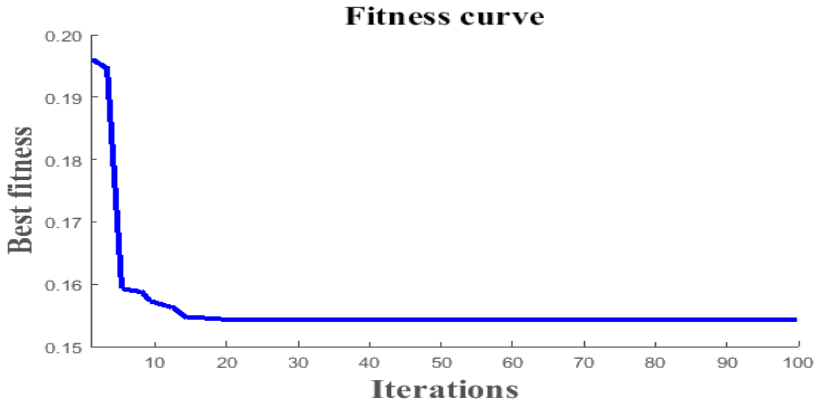
جدول ۳- پارامترهای مورد استفاده در الگوریتم ژنتیک (مأخذ: نویسندگان)

پارامتر	مقدار
اندازه جمعیت	۲۰
تعداد نسل‌ها (تکرار)	۱۰۰
نرخ ترکیب	۰/۸
نوع ترکیب	تک‌نقطه‌ای

معیارها با همان ترتیبی که در جدول ۱ ارائه شده است، کروموزوم‌های الگوریتم ژنتیک را تشکیل داده‌اند. در این تحقیق هر کروموزوم دارای ٦ ژن (معیار) است که به صورت باینری مقداردهی شده است، به گونه‌ای که هر بار تعدادی ژن، جهت ترکیب انتخاب می‌شوند. در این حالت مقدار ۱ به ژن‌های انتخاب شده و مقدار ۰ به ژن‌های انتخاب نشده اختصاص می‌یابد. شکل ۵، نتایج حاصل از ترکیب الگوریتم‌های یادگیری ماشین RF و SVM با الگوریتم GA را نشان می‌دهد. پس از اجرای الگوریتم SVM-GA، بهترین مقدار تابع برازش برابر با ۰/۲۰۷ به دست آمد و بر این اساس، ٤ معیار O3، NO2، PM2.5 و PM10 به عنوان معیارهای مؤثر در برآورد میزان آلودگی هوای شهر تهران شناخته شدند. هم‌چنین برای الگوریتم RF-GA بهترین مقدار تابع برازش برابر با ۰/۱۵۳ به دست آمد و بر این اساس ۵ معیار SO2، O3، NO2، PM2.5 و PM10 به عنوان معیارهای مؤثر در برآورد میزان آلودگی هوای شهر تهران شناخته شدند.



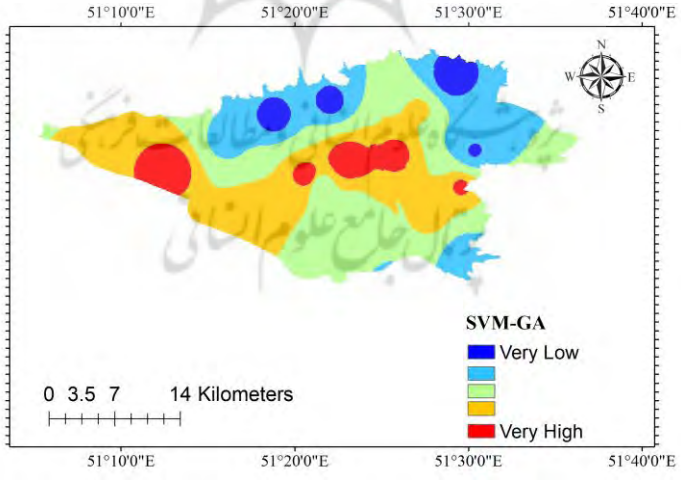
(الف)



(ب)

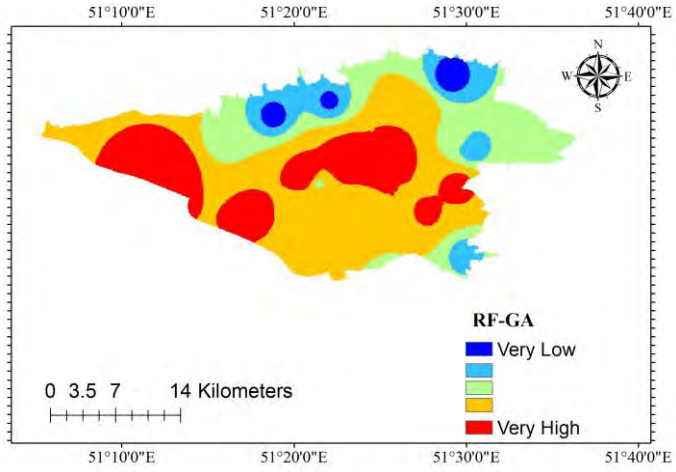
شکل ۵- بهترین مقدار و میانگین مقادیر تابع برازش برای الگوریتم‌های یادگیری ماشین توسعه یافته (الف) SVM-GA (ب) RF-GA (مأخذ: نویسندگان)

جهت ایجاد نقشه‌های برآورد میزان آلودگی شهر تهران، پس از برآورد متغیر وابسته توسط روش‌های یادگیری ماشین توسعه یافته برای نقاط تصادفی تولید شده در محدوده مطالعاتی، سطحی به روش کریجینگ از این نقاط برازش داده می‌شود تا نقشه رستری میزان آلودگی ایجاد گردد. این نقشه‌ها در ۵ کلاس برابر بر اساس روش فاصله مساوی[1] در محدوده [۰،۱] ایجاد شد. شکل ٦ نشان‌دهنده نقشه برآورد میزان آلودگی هوا (AQI) در منطقه مورد مطالعه با استفاده از الگوریتم‌های SVM-GA و RF-GA در ۵ کلاس برابر می‌باشد.



(الف)

1 Equal Interval

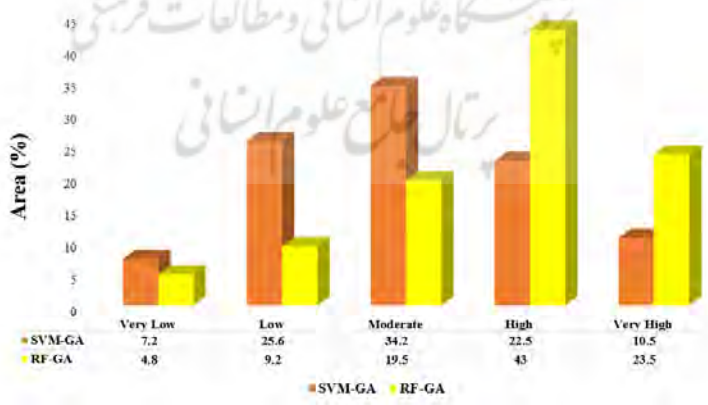


(ب)

شکل ٦- نقشه برآورد میزان آلودگی هوا (AQI) در منطقه مورد مطالعه با استفاده از الگوریتم‌های یادگیری ماشین توسعه یافته (الف) SVM-GA (ب) RF-GA (مأخذ: نویسندگان)

۳-۳- ارزیابی الگوریتم‌های یادگیری ماشین در پیش‌بینی آلودگی هوا

شکل ۷، درصد هر یک از کلاس‌های میزان آلودگی هوای پیش‌بینی شده را توسط مدل‌های RF-GA و SVM-GA نشان می‌دهد. نتایج نشان داد که در مدل RF-GA، ۲ کلاس با آلودگی خیلی زیاد و زیاد درصد بیشتری را در منطقه موردمطالعه نسبت به کلاس‌های مشابه در مدل SVM-GA تحت پوشش قرار می‌دهد. علاوه بر این در مدل SVM-GA، کلاس با آلودگی متوسط، کم و خیلی کم درصد بیشتری را در منطقه مورد مطالعه نسبت به کلاس‌های مشابه در مدل RF-GA دربر گرفته است.



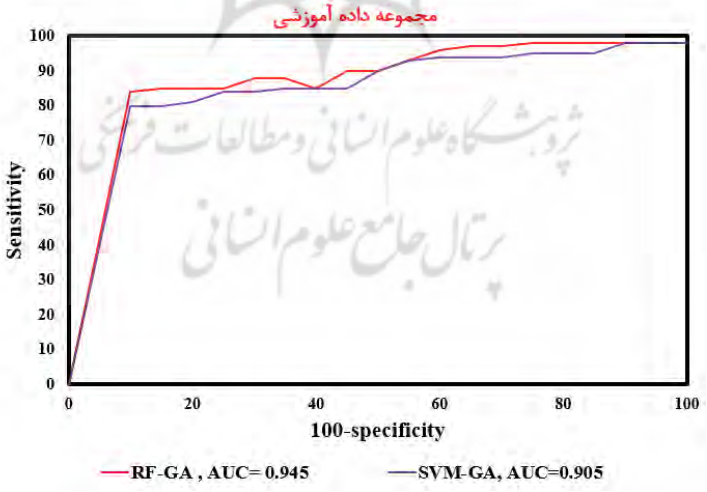
شکل ۷- نمودار میزان درصد کلاس‌های آلودگی هوای پیش‌بینی شده توسط الگوریتم‌های یادگیری ماشین توسعه یافته (مأخذ: نویسندگان)

جدول ٤، مقادیر R^2 و RMSE حاصل از الگوریتم‌های یادگیری ماشین SVM-GA و RF-GA را نشان می‌دهد.

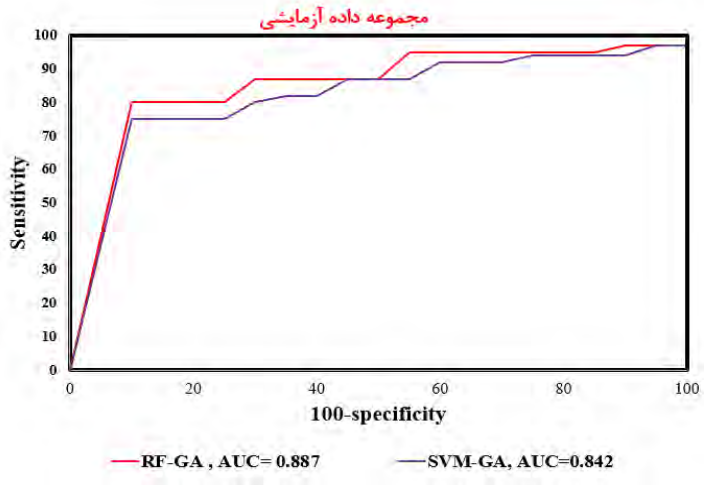
جدول ٤- ارزیابی الگوریتم‌های یادگیری ماشین SVM-GA و RF-GA در برآورد میزان آلودگی هوا (مأخذ: نویسندگان)

نوع روش	R^2	RMSE
SVM-GA	۰/۹۲۵	۰/۲۰۷
RF-GA	۰/۹۹۷	۰/۱۵۳

درنهایت، عملکرد مدل‌های توسعه‌یافته با روش منحنی ROC ارزیابی شد. شکل ۸ منحنی ROC را برای مجموعه داده‌های آموزشی و داده‌های آزمایشی‌مدل‌ها نشان می‌دهد. شکل ۸ (الف) نشان می‌دهد که برای مجموعه داده آموزشی مدل ترکیبی RF-GA دارای بالاترین مقدار AUC (۰/۹٤٥) و پس از آن مدل SVM-GA (AUC= ۰/۹۰٥) قرار دارند. به طور کلی، قابلیت مدل‌ها برای پیش‌بینی آلودگی هوا توسط مجموعه داده‌های آزمایشی ارزیابی می‌شود. بر این اساس شکل ۸ (ب) نشان می‌دهد که مدل ترکیبی RF-GA قوی‌تر بوده و دارای‌بالاترین دقت (AUC= ۰/۸۸۷) نسبت به مدل SVM-GA (AUC= ۰/۸٤۲) می‌باشد. بنابراین، می‌توان نتیجه گرفت که اگر چه تمام مدل‌های ترکیبی به کار رفته قدرت پیش‌بینی خوبی دارند، اما عملکرد مدل RF-GA برای پیش‌بینی آلودگی هوا در منطقه موردمطالعه بهتر است.



(الف)



(ب)

شکل ۸- منحنی ROC و مقدار AUC برای روش‌های پیشنهادی تحقیق (الف) مجموعه داده‌های آموزشی (ب) مجموعه داده‌های آزمایشی (مأخذ: نویسندگان)

٤- بحث و نتیجه‌گیری

با بررسی مطالعات صورت پذیرفته مشخص گردید در تحقیقات پیشین، تحلیل مکانی مناسبی برای تعیین ترکیب بهینه آلاینده‌های گازی مؤثر جهت برآورد میزان آلودگی هوا انجام نگرفته است (ژو و همکاران، ۲۰۱۹؛ لیو و همکاران، ۲۰۱۹). ازآنجایی‌که آلودگی هوا یک مساله پیچیده و چندوجهی می‌باشد و سازوکار مدلسازی آن خود به تنهایی مسأله بسیار پیچیده‌ای است، این تحقیق درنظر دارد بدون درگیر شدن با مفاهیم پیچیده آلودگی هوا و معادلات شیمیایی شکل‌دهنده، به پیش‌بینی آلودگی هوا از نقطه نظر مکانی به مسأله پرداخته و روابط آلودگی هوا را تنها با تکیه بر معادلات مکانی مدلسازی نماید. لذا هدف از این پژوهش تهیه نقشه پهنه‌بندی آلودگی هوا با استفاده از روش‌های نوین یادگیری ماشین مبتنی بر الگوریتم فرااِبتکاری ژنتیک (GA) است؛ بنابراین این مطالعه از مطالعات قبلی متمایز است؛ زیرا در این مطالعه از مدل‌های ترکیبی یادگیری ماشین یعنی RF-GA و SVM-GA جهت برآورد میزان تراکم آلودگی هوای شهر تهران بر مبنای تعیین ترکیب بهینه آلاینده‌های گازی استفاده شده است که نوآوری تحقیق حاضر نیز می‌باشد. مهم‌ترین نتایج این تحقیق به شرح ذیل است:

- ترکیب الگوریتم یادگیری ماشین RF با الگوریتم GA نتایج بهتری را نسبت به ترکیب SVM-GA در اختیار قرار داد.
- مقدار R^2 حاصل از ترکیب الگوریتم‌های RF و SVM با الگوریتم GA به ترتیب برابر ۰/۹۲۵ و ۰/۹۹۷ به دست آمد که نشان‌دهنده سازگاری بالای الگوریتم یادگیری ماشین RF با داده‌های این تحقیق است.

- مقدار RMSE حاصل از ترکیب الگوریتم‌های RF و SVM با الگوریتم GA به ترتیب برابر ۰/۲۰۷ و ۰/۱۵۳ به دست آمد که نشان‌دهنده دقت بالای الگوریتم یادگیری ماشین RF است.
- مدل ترکیبی RF-GA دارای بالاترین دقت (AUC= ۰/۸۸۷) نسبت به مدل SVM-GA (۰/۸۴۲ =AUC) می‌باشد.

بر اساس اطلاعات گرفته شده از شرکت کنترل ترافیک شهر تهران، نتایج حاصل از روش RF بیانگر مناسب بودن انتخاب مدل مذکور جهت برآورد میزان آلودگی هوای شهر تهران بوده است که مطابق با نتایج تحقیق سونگ و همکاران[1] (۲۰۲۱) است. با برآورد مکانی میزان آلودگی هوا پیشنهادهای زیر مطرح می‌گردد:

- نتایج پژوهش حاضر، قابلیت روش‌های ادغام داده‌محور و GIS را در برآورد میزان آلودگی هوای شهر تهران به خوبی نمایان می‌کند. بدین جهت، پیشنهاد می‌گردد که ارگان‌ها، ادارات و سازمان‌های مربوطه با ایجاد بانک‌های اطلاعاتی جامع و به روز از تمام جزئیات و عناصر شهری مبتنی بر GIS، همواره آمادگی لازم برای برای مقابله با پدیده آلودگی هوا را داشته باشند.
- آموزش عمومی، آگاهی و اطلاع‌رسانی دقیق به عموم، در خصوص وجود خطر آلودگی هوا، ابعاد گوناگون آن و هم‌چنین مناطق پرخطر آلودگی هوا.

در این تحقیق از ترکیب الگوریتم ژنتیک با الگوریتم‌های یادگیری ماشین RF و SVM برای شناسایی معیارهای تأثیرگذار در برآورد میزان آلودگی شهر تهران استفاده شد. لذا برای تحقیقات آتی می‌توان توانایی سایر الگوریتم‌های تکاملی مانند الگوریتم انبوه ذرات، کلونی زنبور و ... را در ترکیب با هر یک از الگوریتم‌های یادگیری ماشین جهت شناسایی معیارهای تأثیرگذار مورد ارزیابی قرار داد. هم‌چنین پیشنهاد می‌گردد با تحقیق بیشتر در خصوص آلودگی هوا به عنوان زمینه کاربردی این تحقیق، پارامترهای مؤثر دیگر با شرایط زمانی مناسب که ممکن است در میزان آلودگی هوای یک نقطه اثرگذار باشند، شناسایی گردیده و به پیش بینی زمانی–مکانی شاخص کیفیت آلودگی هوا در منطقه موردمطالعه بپردازند.

کتابنامه

اسلامی‌نژاد، سید احمد؛ افتخاری، مبین؛ محمودی زاده، سعید؛ اکبری، محمد؛ حاجی الیاسی، علی؛ ۱۴۰۰. ارزیابی مدل‌های هوش مصنوعی مبتنی بر درخت به‌منظور پیش‌بینی خطر سیل در بستر GIS. *تحقیقات منابع آب ایران*. ۱۷(۲), ۱۷۴-۱۸۹. https://www.iwrr.ir/article_135317.html

اسلامی‌نژاد، سید احمد؛ افتخاری، مبین؛ اکبری، محمد؛ حاجی الیاسی، علی؛ فرهادیان، هادی؛ ۱۴۰۰. پیش‌بینی مناطق مستعد وقوع سیل با استفاده از مدل‌های پیشرفته یادگیری ماشین (دشت بیرجند). *مدیریت آب و آبیاری*. ۱۱(۴).

1 Song et al.

۸۸۵-۹۰٤. https://doi.org/10.22059/jwim.2022.332875.934

افتخاری، مبین؛ اسلامی نژاد، سید احمد؛ حاجی الیاسی، علی؛ اکبری، محمد؛ ۱٤۰۰. توسعهٔ مدل DRASTIC با استفاده از هوش مصنوعی در پتانسیل آلودگی آبخوان مناطق نیمه خشک. *اکوهیدرولوژی*. (۳) ۸. ٦٥۱-٦٦٥. https://doi.org/10.22059/IJE.2021.323188.1501

حق بیان، سارا؛ تشیع، بهنام؛ ۱۳۹۹. بهبود دقت مدل سازی غلظت ذرات معلق (PM2.5) از طریق ادغام ایستگاه‌های ثابت و همراه سنجش آلودگی هوا. *فصلنامه علمی- پژوهشی اطلاعات جغرافیایی « سپهر»*. ۲۹(۱۱٦). ٤٥-٥۸. https://doi.org/10.22131/sepehr.2021.242859

خزایی، الهه؛ آل شیخ، علی اصغر؛ کریمی، محمد؛ وحیدنیا، محمدحسن؛ ۱۳۹۱. پیش بینی و مدلسازی غلظت آلاینده مونواکسیدکربن با تلفیق شبکه عصبی- فازی تطبیقی و سیستم اطلاعات جغرافیایی. *کاربرد سنجش از دور و GIS در علوم منابع طبیعی*. ۳(۳). ۲۱-۳۳. https://www.sid.ir/paper/189421/fa

رحیمی، جابر؛ رحیمی، علی؛ بذرافشان، جواد؛ ۱۳۹۲. بررسی تداوم روزهای همراه با آلاینده مونوکسیدکربن (CO) در هوای شهر تهران با استفاده از مدل زنجیره مارکف. *نشریه علوم و تکنولوژی محیط زیست*. ۲(۱۵). ۷۹-۹۰. https://www.sid.ir/paper/87572/fa

میری، محمد؛ قانعیان، محمد تقی؛ قلیزاده، عبدالمجید؛ یزدانی، اول محسن؛ نیکونهاد، علی؛ ۱۳۹٤. تحلیل و پهنه بندی آلودگی هوا شهر مشهد با استفاده از مدل‌های مختلف تحلیل فضایی. *مجله مهندسی بهداشت محیط*. (۲) ۳ :-۱٥٤ http://jehe.abzums.ac.ir/article-1-227-fa.html.۱٤

Adams MD, Kanaroglou PS., 2016. Mapping real-time air pollution health risk for environmental management: Combining mobile and stationary air pollution monitoring with neural network models. Journal of Environmental Management .168, 133-141. http://dx.doi.org/10.1016/j.jenvman.2015.12.012

Akbari M, Zahmatkesh H, Eftekhari M., 2021. A GIS-Based System for Real-Time Air Pollution Monitoring and Alerting Based on OGC Sensors Web Enablement Standards. Pollution, 7(1), 25-41. http://dx.doi.org/10.22059/poll.2020.296938.741

Arabgol R, Sartaj M, Asghari K., 2016. Predicting nitrate concentration and its spatial distribution in groundwater resources using support vector machines (SVMs) model. Environmental Modeling & Assessment, 21(1), 71-82. http://dx.doi.org/10.1007/s10666-015-9468-0

de Santana FB, de Souza AM, Poppi RJ., 2018. Visible and near infrared spectroscopy coupled to random forest to quantify some soil quality parameters. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 191, 454-462. http://dx.doi.org/10.1016/j.saa.2017.10.052

Eftekhari M, Eslaminezhad SA, Akbari M, DadrasAjirlou Y, Elyasi AH., 2021. Assessment of the potential of groundwater quality indicators by geostatistical methods in semi-arid regions. Journal of Chinese Soil and Water Conservation, 52(3), 158-67. http://dx.doi.org/10.29417/JCSWC.202109_52(3).0004

Farhadi, R., hadavifar, M., Moeinaddini, M., Amintoosi, M., 2020. Prediction of Air Pollutants Concentration Based on Meteorological Factors in Warm and Cold Season by Artificial

Neural Network and Linear Regression, Case Study: Tehran. Journal of Natural Environment, 73(1), 115-127. http://dx.doi.org/10.22059/JNE.2020.278331.1681

Fawcett T. 2006. An introduction to ROC analysis. Pattern Recognition Letters 27, 861-874. https://doi.org/10.1016/j.patrec.2005.10.010

Ghorbanzadeh O, Blaschke T, Aryal J, Gholaminia K., 2020. A new GIS-based technique using an adaptive neuro-fuzzy inference system for land subsidence susceptibility mapping. Journal of Spatial Science, 65(3), 401–417. https://doi.org/10.1080/14498596.2018.1505564

Gorsevski PV, Gessler PE, Foltz RB, Elliot WJ., 2006. Spatial prediction of landslide hazard using logistic regression and ROC analysis. Transactions in GIS, 10(3), 395-415. https://doi.org/10.1111/j.1467-9671.2006.01004.x

Guevara J, Zadrozny B, Buoro A, Lu L, Tolle J, Limbeck J, Wu M, Hohl D., 2018. A hybrid data-driven and knowledge-driven methodology for estimating the effect of completion parameters on the cumulative production of horizontal wells. In: Proceedings - SPE Annual Technical Conference and Exhibition. Society of Petroleum Engineers (SPE). https://doi.org/10.2118/191446-MS

Kumar D., 2018. Evolving Differential evolution method with random forest for prediction of Air Pollution. Procedia computer science, 132, 824-833. https://doi.org/10.1016/j.procs.2018.05.094

Liu H, Li Q, Yu D, Gu Y., 2019. Air quality index and air pollutant concentration prediction based on machine learning algorithms. Applied Sciences, 9(19), p.4069. https://doi.org/10.3390/app9194069

Masoudi M, Gerami S., 2017. Status of CO as an air pollutant and its prediction, using meteorological parameters in Esfahan, Iran. Pollution. 3 (4), 527-537. https://doi.org/10.22059/poll.2017.62770

McKendry IG., 2015. Evaluation of Artificial Neural Networks for Fine Particulate Pollution (PM10 and PM2.5) Forecasting. Journal of the Air & Waste Management Association 52(9): https://doi.org/1096–1101. 10.1080/10473289.2002.10470836

Mirjalili S., 2019. Genetic algorithm. In Evolutionary algorithms and neural networks (pp. 43-55). Springer, Cham. https://doi.org/10.1007/978-3-319-93025-1

Nejadkoorki F., and Baroutian S., 2012. Forecasting Extreme PM10 Concentrations Using Artificial Neural Networks. Statewide Agricultural Land Use Baseline 2015, 1(1), 277–84. https://doi.org/10.22059/ijer.2011.493

Oshan TM, Li Z, Kang W, Wolf LJ, Fotheringhm AS., 2019. MGWR: A Python implementation of multiscale geographically weighted regression for investigating process spatial heterogeneity and scale, ISPRS International Journal of Geo-Information, 8 (6), p. 269. https://doi.org/10.3390/ijgi8060269

Park S, Kim M, Namgung HG, Kim KT, Cho KH, Kwon SB., 2018. Predicting PM10concentration in Seoul Metropolitan Subway Stations Using Artificial Neural Network (ANN). Journal of Hazardous Materials, 341, 75–82. https://doi.org/10.1016/j.jhazmat.2017.07.050

Quiroz JC, Mariun N, Mehrjou MR, Izadi M, Misron N, Mohd Radzi MA., 2018. Fault detection of broken rotor bar in LS-PMSM using random forests. Measurement, 116, 273-280. https://doi.org/10.1016/j.measurement.2017.11.004

Song XY, Gao Y, Peng Y, Huang S, Liu C, Peng ZR., 2021. A machine learning approach to modelling the spatial variations in the daily fine particulate matter (PM2. 5) and nitrogen

dioxide (NO2) of Shanghai, China. Environment and Planning B: Urban Analytics and City Science, 48(3), 467-483. https://doi.org/10.1177/2399808320975031

Sun Y, Xue B, Zhang M, Yen GG, Lv J., 2020. Automatically designing CNN architectures using the genetic algorithm for image classification. IEEE transactions on cybernetics, 50(9), 3840-3854. https://doi.org/10.1109/TCYB.2020.2983860

Tien Bui D, Shahabi H, Omidvar E, Shirzadi A, Geertsema M, Clague JJ, Lee S., 2019. Shallow landslide prediction using a novel hybrid functional machine learning algorithm. Remote Sensing, 11(8), 931. https://doi.org/10.3390/rs11080931

Wang X, Liu H., 2019. A Knowledge-and Data-Driven Soft Sensor Based on Deep Learning for Predicting the Deformation of an Air Preheater Rotor. IEEE Access 7:159651–159660. https://doi.org/10.1109/ACCESS.2019.2950661

Wheeler DC., 2014. Geographically Weighted Regression. Handbook of Regional Science, Springer: 1435-1459. https://doi.org/10.1007/978-3-642-23430-9_77

Wiemann S, Richter S, Karrasch P, Brauner J, Pech K, Bernard L., 2012. Classification-driven air pollution mapping as for environment and health analysis. 6th International Environmental Modelling and Software Society (iEMSs), 2012, Leipzig, Germany. https://scholarsarchive.byu.edu/iemssconference/2012/Stream-B/353/

Xue J, Xu Y, Zhao L, Wang C, Rasool Z, Ni M, Wang Q, Li D., 2019. Air pollution option pricing model based on AQI. Atmospheric Pollution Research, 10(3), 665-674. https://doi.org/10.1016/j.apr.2018.10.011