نشریه علمی علوم و فنون نقشه‌برداری، دوره ۱۶، شماره ۱، صفحه ۱۷ تا ۳۴، شهریور ماه ۱۴۰۵ - مقاله پژوهشی

# مدل‌سازی مکانی - زمانی $PM_{2.5}$ در شهر تهران با استفاده از داده‌های چندمنبعی و یادگیری ماشین توضیح‌پذیر

**حسین احمدی ریابی[۱]، پرستو پیله فروش‌ها[*۲]**

[۱] دانشجوی کارشناسی ارشد سیستم اطلاعات مکانی – دانشکده مهندسی نقشه‌برداری – دانشگاه صنعتی خواجه نصیر الدین طوسی
hossein.ahmadiriabi@email.kntu.ac.ir

[۲] استادیار دانشکده مهندسی نقشه‌برداری – دانشگاه صنعتی خواجه نصیر الدین طوسی
pilehforoosh@kntu.ac.ir

(دریافت: فروردین ۱۴۰۵، تصویب: مرداد ۱۴۰۵)

## چکیده

آلودگی هوا ناشی از ذرات معلق کمتر از ۲/۵ میکرون ($PM_{2.5}$) یکی از مهم‌ترین چالش‌های زیست‌محیطی و سلامت عمومی در کلان‌شهرها است. با وجود کاربرد گسترده مدل‌های یادگیری ماشین در پیش‌بینی $PM_{2.5}$ بیشتر مطالعات اثرات مکانی و زمانی را به‌صورت مجزا بررسی کرده و کمتر به ارزیابی هم‌زمان مقیاس‌های مکانی–زمانی، برهم‌کنش متغیرهای مؤثر و تبیین‌پذیری مدل‌ها پرداخته‌اند. در این پژوهش، یک چارچوب یادگیری ماشین توضیح‌پذیر برای مدل‌سازی مکانی–زمانی $PM_{2.5}$ در شهر تهران ارائه شد. بدین منظور، داده‌های ایستگاه‌های پایش، متغیرهای هواشناسی، ویژگی‌های زمانی، بستر و پوشش زمین، مورفولوژی شهری، ویژگی‌های همسایگی مکانی–زمانی و ویژگی‌های تعاملی تلفیق و عملکرد سه الگوریتم درخت تصمیم، جنگل تصادفی و XGBoost در دو مقیاس مکانی (۵۰۰ و ۱۰۰۰ متر) و دو بازه زمانی (هفتگی و ماهانه) ارزیابی شد. همچنین، از روش SHAP برای تبیین رفتار مدل و تحلیل اهمیت متغیرها در سطوح سراسری و محلی استفاده شد. نتایج نشان داد که مدل XGBoost در مقیاس مکانی 1000متر و بازه زمانی هفتگی بهترین عملکرد را با $R^2$ برابر با ۰/۵۵۷ ارائه داد. همچنین، متغیرهای هواشناسی، ویژگی‌های همسایگی مکانی–زمانی و ویژگی‌های تعاملی از مهم‌ترین عوامل مؤثر بر تغییرات $PM_{2.5}$ بودند و تحلیل‌های SHAP ناهمگنی فضایی اثر این عوامل را آشکار کرد. نتایج این پژوهش نشان می‌دهد که ارزیابی هم‌زمان مقیاس‌های مکانی–زمانی، همراه با بهره‌گیری از ویژگی‌های تعاملی و یادگیری ماشین توضیح‌پذیر، دقت پیش‌بینی و شناخت عوامل مؤثر بر آلودگی هوا را بهبود داده و می‌تواند از تصمیم‌گیری در مدیریت کیفیت هوا و برنامه‌ریزی شهری پشتیبانی کند.

**واژگان کلیدی:** $PM_{2.5}$، مدل‌سازی مکانی-زمانی، داده‌های چندمنبعی، یادگیری ماشین، SHAP.

[*] نویسنده رابط

## ۱- مقدمه

آلودگی هوا در دهه‌های اخیر به یکی از مهم‌ترین چالش‌های زیست‌محیطی جهان تبدیل شده و آثار نامطلوبی بر سلامت انسان و محیط‌زیست بر جای گذاشته است. بر اساس گزارش سازمان جهانی بهداشت، نزدیک به ۹۹ درصد جمعیت جهان در معرض هوایی قرار دارند که غلظت آلاینده‌های آن از حدود توصیه شده فراتر است [۱]. این مسئله در کلان‌شهرهای آسیایی، از جمله تهران شدت بیشتری یافته‌است [۲].

آلودگی هوای تهران تحت‌تأثیر منابع انسان‌ساخت، به‌ویژه حمل‌ونقل و فعالیت‌های صنعتی، و نیز شرایط توپوگرافی و اقلیمی شهر شکل می‌گیرد [۳]. در این میان، غلظت ذرات معلق با قطر کمتر از ۲/۵ میکرومتر ($PM_{2.5}$) یکی از مهم‌ترین آلاینده‌های شهر است که غلظت آن، به ویژه در نیمه دوم سال، در اثر پایداری جو و وارونگی دما افزایش می‌یابد [۴]. ازاین‌رو، مدل‌سازی دقیق $PM_{2.5}$ برای پایش کیفیت هوا و پشتیبانی از سیاست‌گذاری‌های شهری در تهران اهمیت زیادی دارد.

در سال‌های اخیر، روش‌های متنوعی از مدل‌های آماری تا یادگیری ماشین و یادگیری عمیق برای مدل‌سازی آلودگی هوا توسعه یافته‌اند [۳، ۵-۱۰]. بااین‌حال، روش‌های آماری در برابر روابط غیرخطی میان $PM_{2.5}$ و عوامل موثر بر آن با محدودیت‌هایی روبه‌رو هستند و روش‌های یادگیری عمیق نیز معمولا به داده و توان محاسباتی بیشتری نیاز داشته و از توضیح‌پذیری کمتری برخوردارند. از این رو، الگوریتم‌های یادگیری ماشین به دلیل توازن مناسب میان دقت و توضیح‌پذیری، گزینه‌ای مناسب برای مدل‌سازی $PM_{2.5}$ در مقیاس شهری محسوب می‌شوند [۱۱].

برخی مطالعات انجام شده در زمینه مدل‌سازی آلودگی هوا بر تحلیل الگوهای فضایی آلودگی هوا تمرکز داشته‌اند. برای نمونه، Habibi و همکاران (۲۰۱۷) نشان دادند که توزیع $PM_{2.5}$ در تهران یکنواخت نیست و تمایز فضایی مشخصی میان نیمه شمالی و جنوبی شهر وجود دارد [۱۰]. همچنین، Navabi و همکاران (۲۰۱۹) با استفاده از داده‌های AOD، متغیرهای هواشناسی و الگوریتم‌های یادگیری ماشین به برآورد فضایی $PM_{2.5}$ در تهران پرداختند و نشان دادند که مدل جنگل تصادفی (Random Forest, RF) عملکرد بهتری نسبت به سایر الگوریتم‌ها دارد. بااین‌حال، ویژگی‌های شهری و شاخص‌های مکانی-زمانی در این مطالعات به صورت مستقیم وارد مدل‌سازی نشده‌است [۱۲].

Delavar و همکاران (۲۰۱۹) از روش‌های یادگیری ماشین برای مدل‌سازی آلودگی هوا در تهران استفاده کردند و نشان دادند که ترکیب داده‌های چند منبعی شامل متغیرهای هواشناسی، توپوگرافی، زمانی و مقادیر ایستگاه‌های همسایه، می‌تواند عملکرد مدل را در مقیاس شهری بهبود دهد [۸]. به طور مشابه، Kaviani Rad و همکاران (۲۰۲۲) با استفاده از الگوریتم گرادیان تقویتی شدید ( Extreme Gradient Boosting, XGBoost) در سه شهر ایران نشان دادند که عملکرد مدل به شدت تحت تاثیر تنوع متغیرهای ورودی قرار دارد. بااین‌حال، در این مطالعات نقش ویژگی‌های مکانی و شهری به صورت محدود بررسی شده است [۱۳].

در مطالعات دیگر نیز، الگوریتم‌های یادگیری ماشین برای مدل‌سازی آلاینده‌های جوی در محیط‌های شهری به کار گرفته شده‌اند. به‌عنوان نمونه، Tella و Balogun (۲۰۲۲) در برآورد $PM_{10}$ در مالزی نشان دادند که مدل‌های مبتنی بر درخت، از جمله RF و XGBoost عملکرد مطلوب‌تری نسبت به سایر روش‌ها دارند [۶]. همچنین، Mohammadi و همکاران (۲۰۲۴) با استفاده از الگوریتم‌های RF و ماشین بردار پشتیبان (Support Vector Machine, SVM) نشان دادند که ترکیب متغیرهای هواشناسی و زمانی می‌تواند عملکرد مدل‌های یادگیری ماشین را نسبت به روش‌های آماری سنتی بهبود دهد [۴]. در مطالعه‌ای دیگر، Makhdoomi و همکاران (۲۰۲۵) با استفاده از چند الگوریتم یادگیری ماشین نشان دادند که مدل‌های مبتنی بر درخت تصمیم، به ویژه مدل‌های تقویتی، توان بالایی در برآورد $PM_{2.5}$ دارند [۱۴]. همچنین، Zhaledoost و Taleai (۲۰۲۵) نشان دادند که افزودن ویژگی همسایگی زمانی و انتخاب مقیاس مکانی مناسب می‌تواند عملکرد مدل‌سازی آلودگی هوای شهری، به ویژه برای $PM_{2.5}$ را بهبود دهد [۵]. بااین‌حال، تمرکز این مطالعات بیشتر بر بهبود دقت مدل و ارزیابی اثر مقیاس و ویژگی‌های زمانی بوده و تحلیل نقش نسبی متغیرهای موثر در شکل‌گیری الگوهای آلودگی مورد توجه قرار نگرفته است.

با وجود پیشرفت‌های قابل‌توجه در پیش‌بینی غلظت PM2.5، همچنان چندین محدودیت اساسی در مطالعات پیشین وجود دارد. نخست، بخش عمده‌ای از پژوهش‌ها عمدتاً بر متغیرهای هواشناسی و زمانی یا مقایسه عملکرد

الگوریتم‌های یادگیری ماشین متمرکز بوده‌اند و نقش هم‌زمان ویژگی‌های محیطی و شهری، از جمله کاربری اراضی، توپوگرافی، تراکم شبکه راه‌ها و ساختار فضایی شهر، به‌صورت یکپارچه مورد بررسی قرار نگرفته است. دوم، در بسیاری از مطالعات، وابستگی‌های مکانی و زمانی غلظت آلاینده‌ها به‌طور صریح در ساختار مدل لحاظ نشده و هر ایستگاه به‌عنوان یک واحد مستقل در نظر گرفته شده است؛ در حالی که غلظت PM2.5 به‌شدت تحت تأثیر شرایط ایستگاه‌های مجاور و وضعیت زمانی گذشته قرار دارد. سوم، اغلب پژوهش‌ها بر بهبود شاخص‌های عملکرد مدل تمرکز داشته‌اند و توجه کمتری به تفسیرپذیری مدل، تبیین نحوه تصمیم‌گیری آن و بررسی تغییرپذیری فضایی اثر متغیرهای مؤثر در نواحی مختلف شهر داشته‌اند. ازاین‌رو، توسعه یک چارچوب یکپارچه که بتواند داده‌های چندمنبعی را با وابستگی‌های مکانی–زمانی ترکیب کرده و علاوه بر دقت مناسب، تفسیرپذیری قابل قبولی نیز ارائه دهد، همچنان یک نیاز پژوهشی محسوب می‌شود.

برای غلبه بر مسئله عدم شفافیت مدل‌های پیچیده، در سال‌های اخیر رویکرد هوش مصنوعی توضیح‌پذیر (Expalinable AI, XAI) موردتوجه قرار گرفته است. در این زمینه Ribeiro و همکاران (۲۰۱۶) و Lundberg و Lee (۲۰۱۷) به ترتیب با معرفی روش‌های تبیین‌های محلی مدل ناوابسته ( Local Interpretable Model-agnostic Explanations, LIME) و تبیین افزایشی شاپلی ( Shapley Additive exPlanations, SHAP)، ابزارهایی برای توضیح نتایج مدل‌های یادگیری ماشین ارائه کردند [۱۵, ۱۶]. در حوزه کیفیت هوا نیز این ابزارها امکان تحلیل روابط غیرخطی و پیچیده میان متغیرهای محیطی و غلظت آلاینده‌ها را فراهم کرده‌اند. برای نمونه، Liu و Su (۲۰۲۵) با استفاده از مدل درخت تقویت گرادیانی ( Gradient Boosting Decision Tree, GBDT) و روش SHAP، روابط غیرخطی میان کاربری اراضی شهری و شش آلاینده اصلی را در ۲۶۵ شهر چین بررسی کردند و نشان دادند که به‌کارگیری SHAP می‌تواند سهم نسبی متغیرها در خروجی مدل به صورت کمی تبیین کند [۱۷]. با وجود این پیشرفت‌ها، استفاده از رویکردهای مبتنی بر توضیح‌پذیری در مدل‌سازی آلودگی کلان‌شهرهای ایران، به ویژه در شهرهایی با ویژگی‌های خاص مورفولوژیکی و اقلیمی مانند تهران، همچنان محدود است.

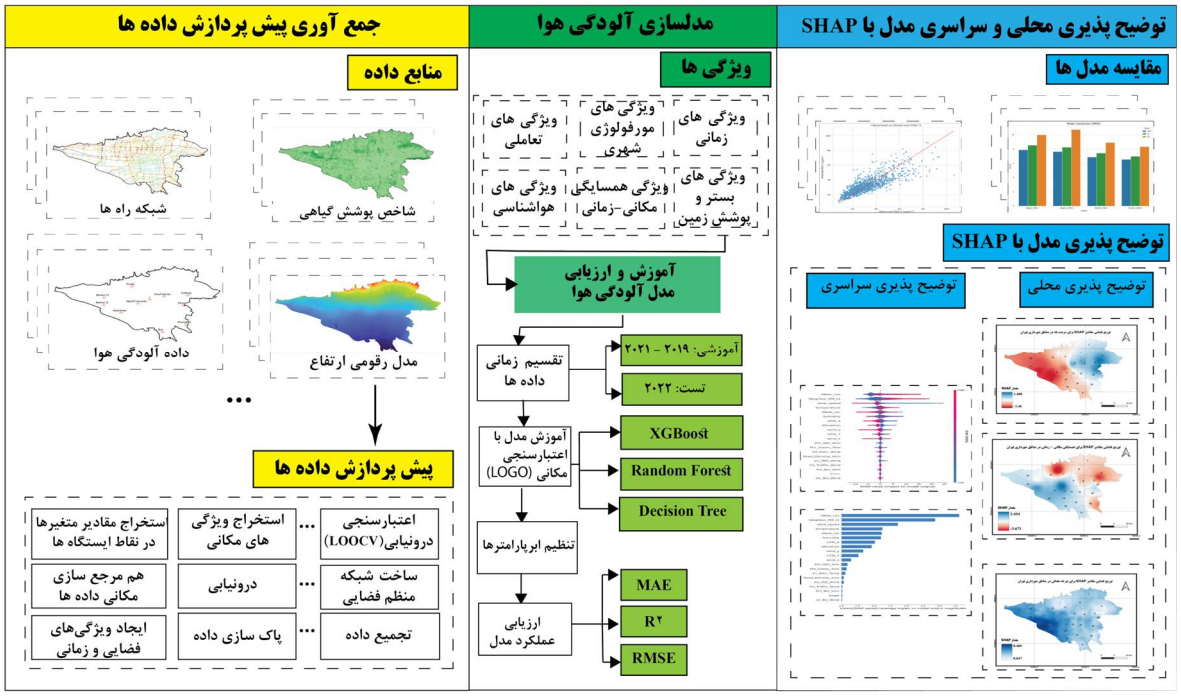
با توجه به چالش‌های مطرح‌شده، این پژوهش با هدف ارائه یک چارچوب یکپارچه، چندمنبعی و توضیح‌پذیر برای مدلسازی مکانی–زمانی غلظت PM2.5 در شهر تهران انجام شده است. در این چارچوب، داده‌های آلودگی هوا، متغیرهای هواشناسی، ویژگی‌های زمانی، ویژگی‌های بستر و پوشش زمین و شاخص‌های مورفولوژی شهری در دو مقیاس مکانی ۵۰۰ و ۱۰۰۰ متر و دو بازه زمانی هفتگی و ماهانه تلفیق شدند. همچنین، به‌منظور بازنمایی بهتر وابستگی‌های مکانی–زمانی و روابط غیرخطی میان متغیرها، ویژگی‌های همسایگی مکانی–زمانی و مجموعه‌ای از ویژگی‌های تعاملی طراحی و به مدل افزوده شد. سپس عملکرد سه الگوریتم درخت تصمیم، جنگل تصادفی و XGBoost با استفاده از یک چارچوب اعتبارسنجی مکانی–زمانی مورد ارزیابی قرار گرفت و در نهایت، روش SHAP برای تبیین رفتار مدل در دو سطح سراسری و محلی به کار گرفته شد.

نوآوری‌های اصلی این پژوهش را می‌توان در سه محور خلاصه کرد. نخست، توسعه یک پایگاه داده چندمنبعی و چندمقیاسی که اطلاعات آلودگی هوا، متغیرهای هواشناسی، ویژگی‌های محیطی و شهری، ویژگی‌های همسایگی مکانی–زمانی و متغیرهای تعاملی را در یک چارچوب واحد ادغام می‌کند. دوم، ارزیابی عملکرد مدل‌های درخت‌مبنا در مقیاس‌های مختلف مکانی و زمانی به‌منظور بررسی اثر مقیاس تحلیل بر دقت مدل. سوم، به‌کارگیری روش SHAP در دو سطح سراسری و محلی که علاوه بر شناسایی عوامل مؤثر بر پیش‌بینی مدل، امکان تحلیل ناهمگنی فضایی اثر متغیرها در بخش‌های مختلف شهر تهران را فراهم می‌سازد. این ویژگی‌ها موجب می‌شود که چارچوب پیشنهادی علاوه بر دقت مناسب، از قابلیت تفسیر و پشتیبانی از تصمیم‌گیری نیز برخوردار باشد.

در ادامه، در بخش ۲، مواد و روش‌های پژوهش شامل معرفی محدوده مطالعاتی، داده‌های مورد استفاده، مراحل پیش‌پردازش، مهندسی ویژگی‌ها، الگوریتم‌های یادگیری ماشین و روش توضیح‌پذیری SHAP تشریح شده است. در بخش ۳، نتایج حاصل از مدل‌سازی مکانی–زمانی، ارزیابی عملکرد مدل‌ها و تحلیل‌های سراسری و محلی SHAP ارائه و مورد بحث قرار گرفته است. در نهایت، در بخش ۴، مهم‌ترین یافته‌های پژوهش جمع‌بندی شده، محدودیت‌های مطالعه بیان گردیده و پیشنهادهایی برای پژوهش‌های آینده ارائه شده است.

## ۲- مواد و روش‌ها

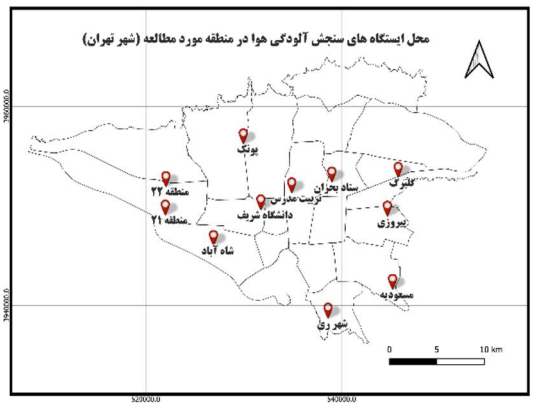
در این پژوهش، فرایند مدل‌سازی مکانی – زمانی غلظت $PM_{2.5}$ در شهر تهران مطابق فلوچارت شکل (۱) انجام شد. ابتدا داده‌های آلودگی هوا، متغیرهای هواشناسی و لایه‌های محیطی و شهری گردآوری، پیش‌پردازش و در قالب یک پایگاه داده مکانی–زمانی یکپارچه سازمان‌دهی شدند. سپس، داده‌ها در دو بازه زمانی هفتگی و ماهانه و دو مقیاس مکانی ۵۰۰ و ۱۰۰۰ متر آماده شدند تا اثر مقیاس زمانی و مکانی بر عملکرد مدل بررسی شود. در ادامه، ویژگی‌های زمانی، هواشناسی، بستر و پوشش زمین، مورفولوژی شهری، همسایگی مکانی–زمانی $PM_{2.5}$ و ویژگی‌های تعاملی استخراج شدند. در نهایت، عملکرد الگوریتم‌های درخت تصمیم، جنگل تصادفی و XGBoostمقایسه و با استفاده از روش SHAP ، رفتار مدل و نقش متغیرهای مؤثر در دو سطح سراسری و محلی تبیین شد. جزئیات محدوده مطالعاتی، داده‌ها، پیش‌پردازش، استخراج ویژگی‌ها، مدل‌سازی و تحلیل توضیح‌پذیری به‌ترتیب در بخش‌های ۲-۱ تا ۲-۶ ارائه شده‌اند.




شکل ۱- فلوچارت مدل پیشنهادی تحقیق

### ۲-۱- محدوده مطالعاتی

محدوده مطالعاتی این پژوهش، شهر تهران در دامنه جنوبی رشته‌کوه البرز و در بازه تقریبی طول جغرافیایی ۵۱/۰۸ تا ۵۱/۶۷ درجه شرقی و عرض جغرافیایی ۳۵/۵۶ تا ۳۵/۸۳ درجه شمالی قرار دارد. تراکم جمعیت، گسترش شبکه حمل‌ونقل و تمرکز کاربری‌های شهری و صنعتی، تهران را به یکی از مهم‌ترین کانون‌های آلودگی هوا در کشور تبدیل کرده است. همچنین، شیب ارتفاعی شمال به جنوب، محصورشدگی نسبی ناشی از ارتفاعات البرز و شرایط جوی پایدار، زمینه انباشت آلاینده‌ها را فراهم می‌کنند. ازاین‌رو، تهران به‌عنوان محدوده‌ای مناسب برای بررسی تغییرات مکانی–زمانی $PM_{2.5}$ و اثر عوامل هواشناسی، محیطی و شهری انتخاب شد. موقعیت محدوده مطالعاتی و ایستگاه‌های پایش در شکل (۲) نشان داده شده است.



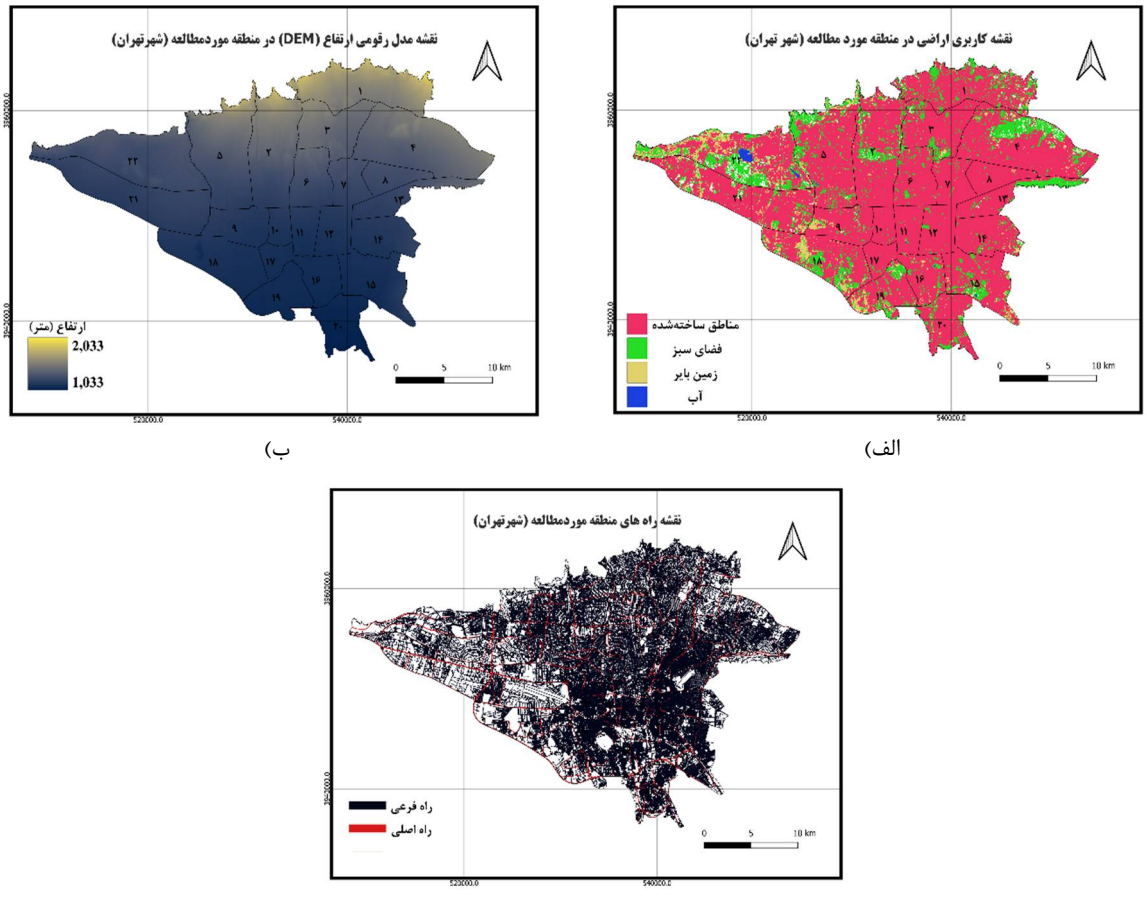

شکل ۲- محدوده مطالعاتی تحقیق

## ۲-۲- داده‌های مورد استفاده

در این پژوهش، غلظت $PM_{2.5}$ به‌عنوان متغیر وابسته با واحد $\mu g/m^3$ در نظر گرفته شد و مجموعه‌ای از داده‌های هواشناسی، محیطی و مکانی به عنوان ورودی مدل مورد استفاده قرار گرفت. داده‌های $PM_{2.5}$و متغیرهای هواشناسی از مجموعه داده ارائه شده در مطالعه Abbasi و همکاران (۲۰۲۵) اخذ شد [۱۸]. این داده‌ها شامل مشاهدات ایستگاه‌های پایش کیفیت هوا و داده‌های هواشناسی متناظر در بازه زمانی ۲۰۱۹ تا ۲۰۲۲ بودند و به عنوان داده پایه در تحلیل حاضر مورد استفاده قرار گرفتند. متغیرهای هواشناسی شامل دما، رطوبت، فشار جو، سرعت باد، نقطه شبنم و مولفه‌های برداری باد بودند. این متغیرها به دلیل نقش مستقیم در تولید، انتقال، تجمع و پراکنش آلاینده‌ها، به عنوان بخشی از ورودی‌های اصلی مدل در نظر گرفته شدند.

به‌منظور توصیف ویژگی‌های محیطی و کالبدی شهر، از چندلایه کمکی نیز استفاده شد. در این راستا، نقشه کاربری اراضی شهر تهران با استفاده از داده‌های ESA WorldCover باقدرت تفکیک مکانی ۱۰ متر تهیه شد. کلاس‌های اولیه این پایگاه داده برای کاهش پیچیدگی تحلیل، در چهار دسته اصلی شامل مناطق ساخته‌شده، پوشش گیاهی، پهنه‌های آبی و اراضی بایر بازطبقه‌بندی شدند. همچنین، داده‌های ارتفاعی SRTM با قدرت تفکیک مکانی ۳۰ متر برای بازنمایی شرایط توپوگرافی، و داده‌های شبکه راه‌ها از OpenStreetMap (OSM) برای استخراج ویژگی مرتبط با مورفولوژی شهری مورد استفاده قرار گرفتند.

در مجموع، داده‌های به‌کاررفته در این پژوهش شامل داده‌های آلودگی هوا و هواشناسی با منشأ ایستگاهی، و لایه‌های محیطی و شهری با منشأ سنجش از دوری و برداری بودند که مبنای تولید پایگاه داده مکانی - زمانی و استخراج ویژگی‌های مدل را تشکیل دادند. نمونه‌ای از لایه‌های ورودی مورد استفاده در پژوهش، شامل کاربری اراضی، مدل رقومی ارتفاع و شبکه راه‌ها در شکل (۳) نمایش داده شده است. جزئیات ویژگی‌های نهایی استخراج شده و دسته‌بندی آن‌ها در جدول ۲ ارائه شده است.




شکل ۳- نمونه‌ای از لایه‌های ورودی مورد استفاده در پژوهش شامل (الف) کاربری اراضی، (ب) مدل رقومی ارتفاع و (ج) شبکه راه‌ها

## ۲-۳- پیش‌پردازش داده‌ها

پس از گردآوری داده‌های پایه، مراحل پیش‌پردازش به‌منظور آماده‌سازی آن‌ها برای فرایند مدل‌سازی انجام شد. در گام نخست، داده‌های $PM_{2.5}$ و متغیرهای هواشناسی از نظر کامل بودن و سازگاری بررسی شدند و مشاهدات ناقص یا ناسازگار حذف شدند. سپس داده‌ها در دو مقیاس زمانی هفتگی و ماهانه تجمیع شدند تا ضمن کاهش نوسانات کوتاه مدت، امکان استخراج الگوهای پایدارتر زمانی و مقایسه عملکرد مدل‌ها در دو سطح زمانی متفاوت فراهم شود.

در ادامه، برای ایجاد یک چارچوب منظم جهت تحلیل فضایی، محدوده مطالعاتی به دو شبکه منظم با ابعاد ۵۰۰ × ۵۰۰ متر و ۱۰۰۰ × ۱۰۰۰ متر تقسیم شد. در مقیاس ۵۰۰ متری، مجموعا ۲۷۳۹ سلول و در مقیاس ۱۰۰۰ متری، ۶۸۵ سلول کل محدوده شهر تهران را پوشش دادند. شبکه ۵۰۰ متری با هدف ایجاد توازن میان نمایش دقیق تغییرات محلی $PM_{2.5}$ و حفظ کارایی محاسباتی انتخاب شد. این مقیاس، امکان بازنمایی مناسب تغییرات خردمقیاس را فراهم می‌کند. همچنین، شبکه ۱۰۰۰ متری برای ارزیابی تأثیر مقیاس فضایی بر عملکرد مدل‌ها در مقیاس‌های مختلف مورد استفاده قرار گرفت.

برای آماده‌سازی لایه‌های مکانی، تمامی داده‌ها به یک سیستم مختصات مشترک (EPSG:32639) منتقل و از نظر محدوده مکانی، قالب داده‌ای و تفکیک‌پذیری همسان‌سازی شدند. سپس لایه‌های محیطی و شهری متناسب با محدوده شهر تهران برش داده شدند و برای استفاده در چارچوب گریدی پژوهش آماده‌سازی شدند.

از آنجا که بخش مهمی از داده‌های مورد استفاده ماهیت نقطه‌ای داشتند، برای تبدیل آن‌ها به سطوح پیوسته مکانی، دو روش درون‌یابی وزن‌دهی معکوس فاصله ( Inverse Distance Weighting, IDW) و کریجینگ (Kriging) بررسی شدند. در روش IDW ، مقدار هر متغیر در هر مکان بر اساس میانگین وزنی مقادیر نقاط مجاور و به‌صورت معکوس با فاصله محاسبه می‌شود (رابطه ۱) [۱۹]. روش Kriging نیز یک روش درون‌یابی زمین‌آماری است که بر پایه ساختار همبستگی فضایی داده‌ها عمل می‌کند و با استفاده از واریوگرام، مقدار متغیر را در مکان‌های ناشناخته برآورد می‌کند [۲۰]. برای مقایسه عملکرد این دو روش، از اعتبارسنجی متقابل (Cross-Validation) و شاخص ریشه میانگین مربعات خطا ( Roost Mean Square Error, RMSE) استفاده شد (رابطه ۲). نتایج این مقایسه در جدول (۱) نشان داد شده است. اگرچه نتایج اعتبارسنجی متقابل نشان داد که روش kriging در برخی متغیرها از خطای کمتری برخوردار است، ارزیابی عملکرد مدل‌های یادگیری ماشین نشان داد که لایه‌های تولید شده با این روش، در مقایسه با IDW، توان تعمیم‌پذیری پایین‌تری ایجاد می‌کنند. بنابراین، با توجه به هدف پژوهش در حفظ تغییرات محلی و تولید ورودی‌های مناسب برای مدل‌سازی مکانی – زمانی، از روش IDW برای تولید لایه‌های نهایی استفاده شد. در ادامه، مقادیر حاصل از سطوح درون‌یابی شده برای سلول‌های ۵۰۰ و ۱۰۰۰ متری استخراج و در پایگاه داده نهایی ذخیره شد.

$$Z(x_0) = \frac{(\sum_{i=1}^{n} w_i Z(x_i))}{(\sum_{i=1}^{n} w_i)}, \quad w_i = \frac{1}{d(x_0, x_i)^p} \qquad (۱)$$

$$RMSE = \sqrt{\frac{1}{n}\Sigma_i^n (y_i - \hat{y_i})^2} \qquad (۲)$$

در روابط فوق، $Z(x_0)$ مقدار برآورد شده در نقطه $x_0$، $Z(x_i)$ مقدار مشاهده شده در نقطه i، $d(x_0, x_i)$ فاصله بین نقطه هدف و نقطه مشاهده‌ای و $p$ توان فاصله، $y_i$ مقدار مشاهده شده $PM_{2.5}$ و $\hat{y_i}$ مقدار برآورد شده توسط مدل می‌باشد.

جدول ۱- نتایج درونیابی متغیرها با استفاده از دو روش درونیابی IDW و Kriging

| نام ویژگی | ماهانه | | هفتگی | |
|---|---|---|---|---|
| | IDW | Kriging | IDW | Kriging |
| نقطه شبنم | ۰/۲۲ | ۰/۱۳ | ۰/۲۳ | ۰/۱۴ |
| رطوبت | ۰/۹۸ | ۰/۵۹ | ۱/۰۰ | ۰/۶۱ |
| فشارهوا | ۰/۱۰ | ۰/۰۶ | ۰/۱۰ | ۰/۰۶ |
| دما | ۰/۳۵ | ۰/۲۱ | ۰/۳۵ | ۰/۲۲ |
| سرعت باد | ۰/۳۳ | ۰/۲۰ | ۰/۳۳ | ۰/۲۱ |
| مؤلفه افقی جهت باد | ۰/۲۳ | ۰/۲۵ | ۰/۲۴ | ۰/۲۶ |
| مؤلفه عمودی جهت باد | ۰/۲۶ | ۰/۱۹ | ۰/۲۹ | ۰/۲۱ |

نتایج جدول ۱ نشان داد که با وجود خطای کمتر Kriging در برخی متغیرها، لایه‌های تولید شده با IDW عملکرد بهتری در مدل‌سازی $PM_{2.5}$ داشتند که ناشی از حفظ بهتر تغییرات محلی است. بنابراین، لایه‌های تولید شده با روش

IDW به‌عنوان ورودی مدل‌های یادگیری ماشین مورد استفاده قرار گرفتند.

## ۲-۴- استخراج و مهندسی ویژگی‌ها

پس از آماده‌سازی پایگاه داده مکانی–زمانی، مجموعه‌ای از ویژگی‌های مستقل برای مدل‌سازی $PM_{2.5}$ استخراج شد. این مرحله، با هدف بازنمایی الگوهای زمانی، شرایط هواشناسی، ویژگی‌های بستر و پوشش زمین، ساختار مورفولوژی شهر، وابستگی‌های مکانی – زمانی و اثرات تعاملی میان متغیرها انجام شد. بر این اساس، ویژگی‌های مورد استفاده در شش گروه اصلی شامل ویژگی‌های زمانی، هواشناسی، بستر و پوشش زمین، مورفولوژی شهری، همسایگی مکانی-زمانی و ویژگی‌های تعاملی سازمان‌دهی شدند. مشخصات این ویژگی‌ها در جدول (۲) ارائه شده و جزئیات آن‌ها در بخش ۲-۴-۱ الی ۲-۴-۶ آمده است.

جدول ۲- مشخصات ویژگی‌های مورد استفاده در مدل و منابع استخراج آن‌ها

| دسته اصلی | نام ویژگی | نام اختصاری | منبع داده | فرمت | دقت تفکیک |
|---|---|---|---|---|---|
| زمانی | مؤلفه کسینوسی ماه | Month_Cos | استخراج از سری زمانی | عددی | ماهانه/ هفتگی |
| | مؤلفه سینوسی ماه | Month_Sin | | | |
| | مؤلفه سینوسی هفته | Week_Sin | | | |
| | مؤلفه کسینوسی هفته | Week_Cos | | | |
| هواشناسی | دما | Temperature | Abbasi و همکاران (۲۰۲۵) | رستری | ۱۰۰۰/۵۰۰متری- ماهانه/هفتگی |
| | رطوبت | Humidity | | | |
| | فشار جو | Pressure | | | |
| | سرعت باد | Wind_Speed | | | |
| | نقطه شبنم | Dew_Point | | | |
| | مؤلفه‌های برداری باد | Wind_x/ Wind_y | | | |
| بستر و پوشش زمین | کاربری اراضی | LULC | ESA World Cover | رستری | ۱۰ متر |
| | درصد کاربری ساخته‌شده در شعاع ۱کیلومتری | PCT_BU_1km | محاسبات تحقیق | رستری | ۱۰۰۰/۵۰۰متری |
| | درصد کاربری فضای سبز در شعاع ۳ کیلومتری | PCT_Green_3km | | | |
| | درصد کاربری صنعتی در شعاع ۳ کیلومتری | PCT_IND_3km | | | |
| مورفولوژی شهری | مدل رقومی ارتفاع | DEM | SRTM | رستری | ۳۰ متر |
| | شیب | Slope | محاسبات تحقیق | رستری | ۳۰ متر |
| | تراکم راه‌ها در شعاع ۱ کیلومتری | Road_Density_1km | OpenStreetMap (OSM) | برداری | گرید ۵۰۰/۱۰۰۰ متر |
| مکانی - زمانی | همسایگی مکانی -زمانی | Neighbor_PM_t1 | محاسبات تحقیق | ماتریسی | همسایگی چهار ایستگاه مجاور در گام زمانی پیشین |
| تعاملات فیزیکی | تعامل ارتفاع و دما | Int_Elev_Temp | محاسبات تحقیق | عددی | ۱۰۰۰/۵۰۰متری- ماهانه / هفتگی |
| | نسبت تراکم راه‌ها به‌سرعت باد | Int_Traffic_Wind | | | |
| | نسبت درصد کاربری صنعتی به‌سرعت باد | Int_IND_Wind | | | |
| | تعامل پوشش سبز و رطوبت | Int_Green_Humidity | | | |
| | نسبت درصد مناطق ساخته‌شده به‌سرعت باد | Int_BU_Wind | | | |

### ۲-۴-۱-ویژگی‌های زمانی

به‌منظور بازنمایی رفتار چرخه‌ای زمان در مدل، از کدگذاری سینوسی و کسینوسی برای ماه و هفته استفاده شد. این نوع کدگذاری از ایجاد ناپیوستگی مصنوعی میان ابتدا و انتهای چرخه زمانی جلوگیری می‌کند، به‌گونه‌ای که برای مثال، ماه دوازدهم و ماه اول به‌عنوان دو مقدار نزدیک در چرخه زمانی در نظر گرفته می‌شوند. بر این اساس، متغیرهای $Month_{sin}$ و $Month_{cos}$ برای مجموعه داده‌های ماهانه و متغیرهای $Week_{sin}$ و $Week_{cos}$ برای مجموعه داده‌های هفتگی مطابق روابط (۳) و (۴) محاسبه شدند که در آنکه در آن $Month$ و$Week$ به‌ترتیب شماره ماه و هفته در چرخه زمانی هستند. این ویژگی امکان بازنمایی مناسب‌تر تغییرات فصلی و هفتگی $PM_{2.5}$ را در مدل فراهم می‌کند.

$$Month_{sin} = sin\left(\frac{2\pi \times \text{Month}}{12}\right) \qquad (۳)$$
$$Month_{cos} = cos\left(\frac{2\pi \times \text{Month}}{12}\right)$$

$$Week_{sin} = sin\left(\frac{2\pi \times \text{Week}}{52}\right) \qquad (۴)$$
$$Week_{cos} = cos\left(\frac{2\pi \times \text{Week}}{52}\right)$$

### ۲-۴-۲- ویژگی‌های هواشناسی

ویژگی‌های هواشناسی به‌عنوان یکی از اصلی‌ترین گروه‌های ورودی مدل در نظر گرفته شدند، زیرا این متغیرها نقش مستقیمی در تولید، انتقال، انباشت و پراکنش آلاینده‌ها در محیط شهری دارند. در این پژوهش، دما، رطوبت، فشار جو، سرعت باد، نقطه شبنم و مؤلفه‌های برداری باد به‌عنوان ویژگی‌های هواشناسی مورد استفاده قرار گرفتند. این متغیرها امکان بازنمایی شرایط جوی مؤثر بر پایداری یا تهویه جو و در نتیجه تغییرات مکانی – زمانی $PM_{2.5}$ را فراهم می‌کنند.

### ۲-۴-۳-ویژگی‌های بستر و پوشش زمین

برای بازنمایی ویژگی‌های طبیعی و محیطی مؤثر بر تغییرات$PM_{2.5}$ ، از متغیرهای بستر و پوشش زمین استفاده شد. در این راستا، شیب و ارتفاع از مدل رقومی ارتفاع استخراج شدند تا اثر توپوگرافی بر جریان هوا و انباشت آلاینده‌ها لحاظ شود. همچنین، لایه کاربری اراضی برای نمایش تفاوت‌های فضایی ناشی از استقرار انواع کاربری‌ها به کار گرفته شد. شعاع جستجوی هر متغیر متناسب با دامنه اثرگذاری آن بر کیفیت هوا تعیین شد؛ به‌طوری‌که برای درصد سطح ساخته‌شده از شعاع ۱ کیلومتر به‌منظور بازنمایی ویژگی‌های محیط نزدیک و برای درصد کاربری صنعتی و فضای سبز از شعاع ۳ کیلومتر به‌دلیل گستره اثرگذاری بیشتر آن‌ها استفاده شد. انتخاب این شعاع‌ها بر اساس مطالعات پیشین و ویژگی‌های فضایی شهر تهران انجام شد.

### ۲-۴-۴-ویژگی مورفولوژی شهری

برای بازنمایی اثر مورفولوژی شهری بر تغییرات $PM_{2.5}$، از شاخص تراکم راه‌ها استفاده شد. این شاخص از نسبت مجموع طول شبکه راه‌ها به مساحت هر سلول در شعاع ۱ کیلومتری محاسبه و به‌عنوان نماینده‌ای از شدت توسعه شهری، فعالیت‌های ترافیکی و نزدیکی به منابع انتشار آلاینده‌های ناشی از حمل‌ونقل وارد مدل شد. انتخاب شعاع ۱ کیلومتری با هدف بازنمایی اثر محیط پیرامون هر سلول و بر اساس مطالعات پیشین در زمینه مدل‌سازی آلودگی هوای شهری انجام شد.

### ۲-۴-۵- ویژگی همسایگی مکانی–زمانی $PM_{2.5}$

با توجه به ماهیت پویا و پیوسته $PM_{2.5}$ در بعد زمان و نیز وابستگی مکانی غلظت این آلاینده، در این پژوهش یک ویژگی همسایگی مکانی– زمانی برای $PM_{2.5}$ تعریف شد. این ویژگی بر اساس میانگین مقادیر $PM_{2.5}$ در چهار ایستگاه مجاور و در گام زمانی پیشین محاسبه شد و هدف از آن، وارد کردن اطلاعات مربوط به محیط پیرامونی و رفتار گذشته آلاینده به فرایند مدل‌سازی بود. در نتیجه، مدل علاوه بر شرایط محلی هر ایستگاه، از تغییرات اخیر $PM_{2.5}$ در ایستگاه‌های نزدیک نیز بهره می‌گیرد. این شاخص مطابق رابطه شماره (۵) محاسبه شد.

$$STN_i(t) = \frac{1}{4}\sum_{j=1}^{k} PM_j(t-1) \qquad (۵)$$

که در آن $STN_i(t)$ شاخص همسایگی مکانی – زمانی برای سلول i در زمان t، $PM_j(t-1)$ مقدار $PM_{2.5}$ در همسایه j در گام زمانی قبلی و k تعداد همسایه‌های در نظر گرفته شده است. استفاده از غلظت آلاینده در گام زمانی قبل، ضمن در نظر گرفتن وابستگی فضایی-زمانی، از بروز

بیش‌برازش و نشت اطلاعات جلوگیری می‌کند؛ زیرا در این حالت اطلاعات مربوط به زمان هدف در فرایند آموزش مدل وارد نمی‌شود.

### ۲-۴-۶- ویژگی‌های تعاملی

به‌منظور بازنمایی اثرات مشترک و غیرخطی میان متغیرهای محیطی و هواشناسی، مجموعه‌ای از ویژگی‌های تعاملی شامل برهم‌کنش ارتفاع و دما، تراکم راه‌ها و سرعت باد، کاربری صنعتی و سرعت باد، فضای سبز و رطوبت نسبی، و سطح ساخته‌شده و سرعت باد مطابق روابط (۶) تا (۱۰) تعریف شدند. این ویژگی‌ها با الهام از سازوکارهای شناخته‌شده انتشار، انتقال و تجمع $PM_{2.5}$ طراحی شده‌اند تا اثر هم‌زمان چند متغیر به‌صورت صریح در فرآیند یادگیری مدل لحاظ شود.

$$Int_{Elev_{Temp}} = Elevation \times Temperature \quad (۶)$$

$$Int_{Traffic_{Wind}} = \frac{Road\ Density_1km}{Wind\ Spreed + 1} \quad (۷)$$

$$Int_{IND_{Wind}} = \frac{PCT_IND_3km}{Wind\ Speed + 1} \quad (۸)$$

$$Int_{Green_{Humidity}} = PCT_Green_3km \times Humidity \quad (۹)$$

$$Int_{BU_{Wind}} = \frac{PCT_BU_1km}{Wind\ Speed + 1} \quad (۱۰)$$

روابط (۶) تا (۱۰) هر یک جنبه‌ای از برهم‌کنش میان عوامل مؤثر بر تغییرات $\boldsymbol{PM_{2.5}}$ را بازنمایی می‌کنند. تعامل ارتفاع و دما برای نمایش اثر هم‌زمان توپوگرافی و شرایط حرارتی، تعامل تراکم راه‌ها، کاربری صنعتی و سطح ساخته‌شده با سرعت باد برای بازتاب وابستگی اثر منابع انتشار به ظرفیت تهویه جو، و تعامل فضای سبز و رطوبت نسبی برای نمایش اثر هم‌زمان پوشش گیاهی و شرایط رطوبتی در نظر گرفته شدند. همچنین، در روابط مبتنی بر سرعت باد، مقدار ثابت ۱ به مخرج افزوده شد تا از بروز خطای تقسیم بر صفر در شرایط سکون باد جلوگیری شود.

## ۲-۵- مدل‌سازی یادگیری ماشین

پس از آماده‌سازی مجموعه‌داده‌ها و استخراج ویژگی‌های توضیحی، فرایند مدل‌سازی $PM_{2.5}$ با استفاده از الگوریتم‌های یادگیری ماشین انجام شد. با توجه به ماهیت غیرخطی روابط میان $PM_{2.5}$ و متغیرهای موثر بر آن، و نیز با هدف مقایسه عملکرد مدل‌های درخت مبنا در بازنمایی این روابط، سه الگوریتم درخت تصمیم، جنگل تصادفی و XGBoost برای این پژوهش انتخاب شدند. مدل‌سازی برای هریک از چهار مجموعه داده حاصل از ترکیب دو مقیاس مکانی و دو مقیاس زمانی به صورت مستقل انجام شد تا اثر مقیاس تحلیل بر عملکرد مدل‌ها نیز بررسی شود.

انتخاب این الگوریتم‌ها به دلیل توانایی آن‌ها در مدل‌سازی روابط غیرخطی، حساسیت کمتر به نرمال بودن داده‌ها و قابلیت استفاده در مسائل رگرسیونی با مجموعه‌ای متنوع از ویژگی‌های ورودی صورت گرفت. مدل درخت تصمیم به‌عنوان ساده‌ترین مدل درخت مبنا، امکان تقسیم‌بندی داده‌ها بر اساس قواعد تصمیم‌گیری را فراهم می‌کند و به‌عنوان مبنایی برای مقایسه عملکرد سایر مدل‌ها مورد استفاده قرار گرفت. جنگل تصادفی با ترکیب چندین درخت تصمیم و استفاده از نمونه‌گیری تصادفی، از توان تعمیم‌پذیری بیشتری برخوردار است و معمولاً نسبت به‌پیش برازش مقاومت بیشتری دارد. الگوریتم XGBoost نیز به‌عنوان یک روش تقویتی مبتنی بر درخت، به دلیل قدرت بالا در مدل‌سازی روابط پیچیده و عملکرد مناسب در مسائل پیش‌بینی و برآورد، در این پژوهش به کار گرفته شد.

### ۲-۵-۱- آموزش، تنظیم و اعتبارسنجی مدل‌ها

برای آموزش مدل‌ها، مجموعه ویژگی‌های استخراج شده به‌عنوان ورودی و غلظت $PM_{2.5}$ به عنوان خروجی مدل در نظر گرفته شد. فرایند آموزش و ارزیابی برای هریک از چهار مجموعه داده ۵۰۰ متری-هفتگی، ۵۰۰ متری-ماهانه، ۱۰۰۰ متری-هفتگی و ۱۰۰۰ متری-ماهانه به صورت مستقل انجام شد. به منظور دستیابی به تنظیم مناسب مدل‌ها، پارامترهای اصلی هر الگوریتم در فرایند آموزش مورد تنظیم قرار گرفت تا از بیش برازش جلوگیری شده و توان تعمیم‌پذیری مدل افزایش یابد.

با توجه به ماهیت مکانی–زمانی داده‌ها، عملکرد مدل‌ها در دو بعد زمانی و مکانی ارزیابی شد. در ارزیابی زمانی، داده‌ها بر اساس توالی زمانی به مجموعه‌های آموزش و اعتبارسنجی تفکیک شدند. به‌طوری که داده‌های سال‌های ۲۰۱۹ تا ۲۰۲۱ برای آموزش مدل و داده‌های سال ۲۰۲۲ برای اعتبارسنجی مورد استفاده قرار گرفتند.

برای ارزیابی توانایی تعمیم‌پذیری مکانی مدل بهینه، از رویکرد اعتبارسنجی متقاطع مکانی از نوع Leave-One-Group-Out (LOGO) استفاده شد. در این روش، داده‌های هر ایستگاه پایش آلودگی هوا به‌عنوان یک گروه مستقل در نظر گرفته شد و در هر تکرار، داده‌های یک ایستگاه برای آزمون و داده‌های سایر ایستگاه‌ها برای آموزش مدل مورداستفاده قرار گرفت.

### ۲-۵- معیارهای ارزیابی مدل‌ها

برای ارزیابی عملکرد مدل‌های یادگیری ماشین در برآورد $PM_{2.5}$، از سه معیار ریشه میانگین مربعات خطا (RMSE)، میانگین قدرمطلق خطا (MAE) و ضریب تعیین ($R^2$) استفاده شد. معیار RMSE که رابطه آن در بخش پیش‌پردازش داده‌ها ارائه شد، در این مرحله نیز برای سنجش میزان انحراف کلی مقادیر برآورد شده از مقادیر مشاهده شده به کار رفت. علاوه بر این، از معیار MAE برای اندازه‌گیری میانگین قدرمطلق خطاها و از ضریب تعیین برای سنجش توان مدل در تبیین تغییرات $PM_{2.5}$ استفاده شد. روابط این دو معیار به ترتیب در روابط (۱۱) و (۱۲) ارائه شده‌اند.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y_i}| \qquad (۱۱)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y_i})^2}{\sum_{i=1}^{n}(y_i - \bar{y_i})^2} \qquad (۱۲)$$

که در آن $y_i$ مقدار مشاهده شده $PM_{2.5}$، $\hat{y_i}$ مقدار برآورد شده توسط مدل، $\bar{y_i}$ میانگین مقادیر مشاهده شده و n تعداد نمونه است. در این پژوهش، مدل‌ها براساس کمینه بودن RMSE و MAE و بیشینه بودن $R^2$ با یکدیگر مقایسه شدند و از این سه معیار برای ارزیابی عملکرد مدل‌ها در سناریوهای مختلف مکانی–زمانی استفاده شد.

### ۲-۶- توضیح‌پذیری مدل

اگرچه مدل‌های یادگیری درخت مبنا، به‌ویژه XGBoost توان بالایی در مدل‌سازی روابط پیچیده و غیرخطی دارند، اما تفسیر مستقیم نحوه اثرگذاری متغیرهای ورودی بر خروجی مدل در آن‌ها به‌سادگی امکان‌پذیر نیست. ازاین‌رو، در این پژوهش برای تحلیل و توضیح رفتار مدل از رویکرد هوش مصنوعی توضیح‌پذیر (Explainable Artificial Intelligence, XAI) استفاده شد. در این راستا، روش SHAP به‌عنوان یکی از معتبرترین ابزارهای توضیح‌پذیری مدل‌های یادگیری ماشین به کار گرفته شد.

روش SHAP بر پایه نظریه بازی‌های تعاونی توسعه یافته و سهم هر متغیر را در پیش‌بینی مدل به‌صورت کمی برآورد می‌کند [۱۶]. در این روش، خروجی مدل برای هر نمونه به مجموع مقدار پایه مدل و سهم هریک از ویژگی‌ها تجزیه می‌شود. رابطه کلی SHAP مطابق رابطه (۱۳) بیان می‌شود.

$$f(x) = \phi_0 + \sum_{i=1}^{M}\phi_i \qquad (۱۳)$$

که در آن $f(x)$ خروجی مدل برای نمونه$x$ ، $\phi_0$ مقدار پایه مدل، $M$ تعداد ویژگی‌ها، و $\phi_i$ سهم ویژگی i در خروجی مدل است. در این پژوهش، برای محاسبه مقادیر SHAP از الگوریتم TreeExplainer بر روی مدل XGBoost استفاده شد. این روش امکان تحلیل رفتار مدل را در دو سطح سراسری و محلی فراهم می‌کند.

#### ۲-۶-۱- توضیح‌پذیری مدل در سطح سراسری

در سطح سراسری، از نمودار خلاصه SHAP ( SHAP Summary Plot) برای ارزیابی اهمیت کلی متغیرها و جهت اثر آن‌ها بر مدل‌سازی $PM_{2.5}$ استفاده شد. این نمودار نشان می‌دهد که کدام متغیرها در عملکرد کلی مدل بیشترین سهم را دارند و افزایش یا کاهش هر متغیر چگونه با تغییر مقدار متغیر هدف مرتبط است.

#### ۲-۶-۲- توضیح‌پذیری مدل در سطح محلی

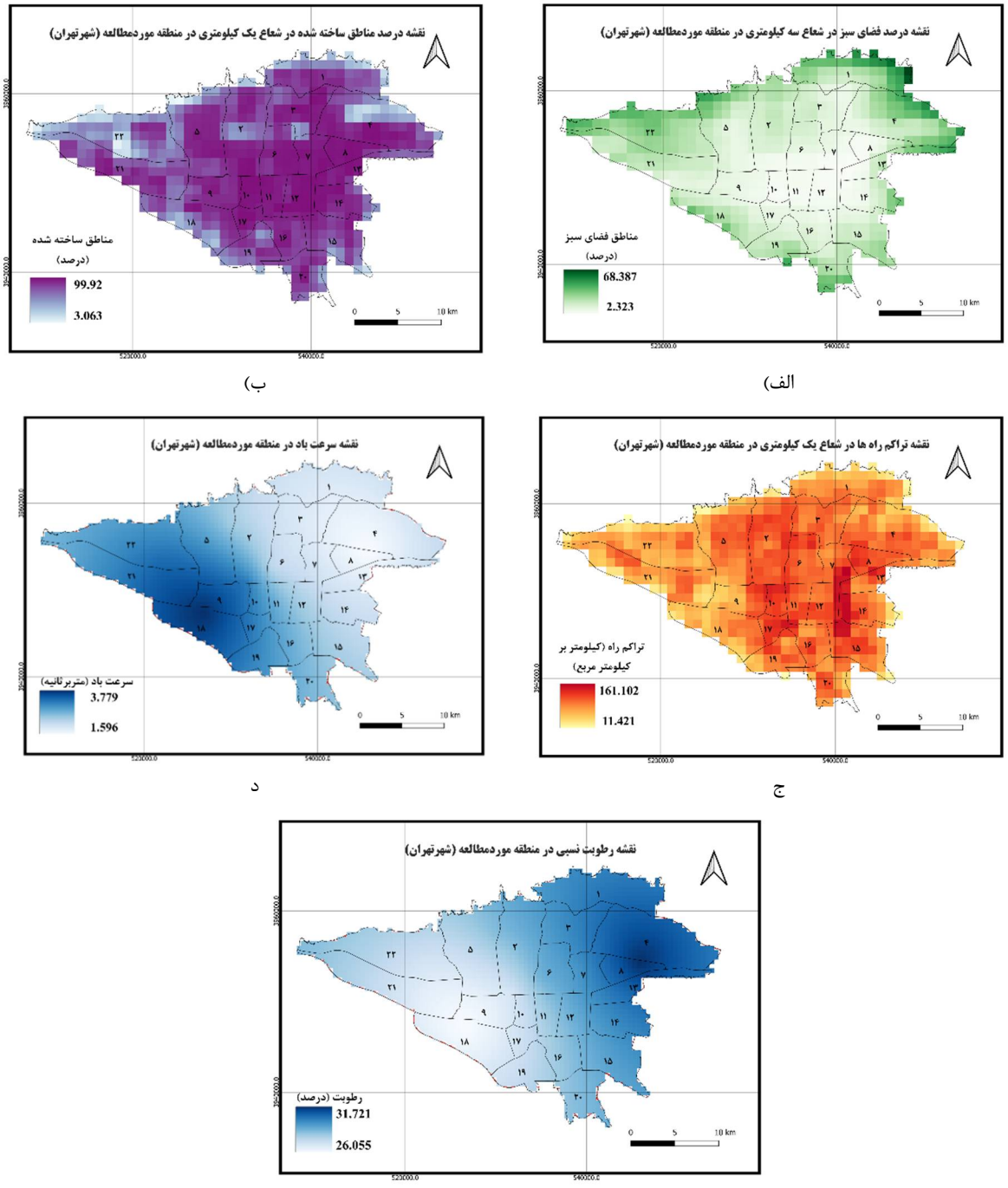
در سطح محلی، مقادیر SHAP برای هر واحد مکانی محاسبه و با اتصال به مختصات گریدها، در محیط GIS با استفاده از روش Kriging درون‌یابی شدند. لایه‌های حاصل به‌صورت نقشه‌های پیوسته نمایش داده شدند تا الگوی فضایی اثر متغیرها و ناهمگنی مکانی آن‌ها در سطح شهر تحلیل شود. این تحلیل نشان می‌دهد که سهم هر متغیر در پیش‌بینی غلظت $PM_{2.5}$ در بخش‌های مختلف شهر یکسان نیست.

شایان ذکر است که استفاده از Kriging در این مرحله صرفا با هدف پهنه‌بندی و بصری‌سازی مقادیر SHAP انجام شده است و به معنای برتری این روش برای تولید داده‌های

ورودی مدل نیست. در حالی که روش درون‌یابی متغیرهای ورودی بر اساس ارزیابی عملکرد مدل انتخاب شد، Kriging به دلیل در نظر گرفتن خودهمبستگی مکانی و تولید سطوح پیوسته، برای نمایش و تفسیر فضایی نتایج SHAP به کار گرفته شد.

## ۳- پیاده‌سازی و نتایج

در این بخش، نتایج حاصل از پیاده‌سازی مدل پیشنهادی مورد بررسی قرار می‌گیرد. بدین منظور، در بخش ۳-۱ به بررسی توزیع مکانی متغیرهای ورودی پرداخته شده‌است.

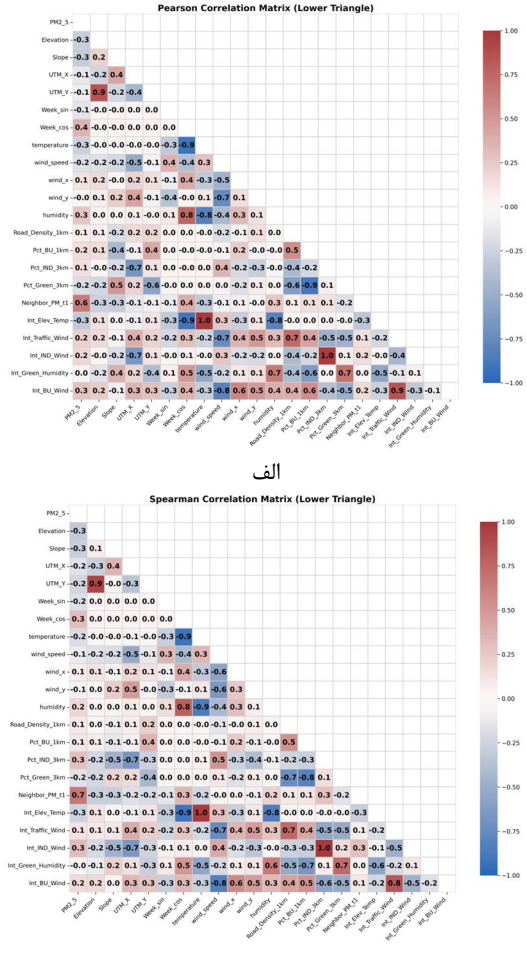



شکل ۴- نمونه‌ای از متغیرهای مستقل مورد استفاده در مدل شامل (الف) درصد فضای سبز در شعاع سه کیلومتری، (ب) درصد مناطق ساخته شده در شعاع یک کیلومتری، (ج) تراکم راه‌ها در شعاع یک کیلومتری، د) سرعت باد و ه) رطوبت نسبی

سپس، عملکرد مدل‌های یادگیری ماشین در بخش ۳-۲ ارزیابی شده و در نهایت، نتایج حاصل از توضیح‌پذیری مدل‌ها در بخش ۳-۳ تشریح گردیده‌اند.

## ۳-۱- تحلیل فضایی و همبستگی متغیرها

در این بخش، توزیع فضایی متغیرهای کلیدی و ساختار همبستگی میان متغیرهای مستقل و متغیر هدف بررسی شده است. شکل (۴) نمونه‌ای از توزیع فضایی متغیرهای ورودی را در محدوده مطالعاتی نشان می‌دهد.




الف

ب

شکل ۵- بررسی همبستگی بین پارامترهای مستقل و وابسته: الف) ماتریس همبستگی پیرسون و ب) ماتریس همبستگی اسپیرمن

نتایج حاکی از ناهمگنی فضایی متغیرهای محیطی و شهری در تهران است؛ به‌طوری که تراکم راه‌ها و درصد مناطق ساخته‌شده عمدتا در نواحی مرکزی و متراکم شهر و سهم فضای سبز در بخش‌های شمالی و پیرامونی بیشتر است. این ناهمگنی فضایی اهمیت لحاظ کردن ویژگی‌های مکانی در مدل‌سازی غلظت $PM_{2.5}$ را نشان می‌دهد.

علاوه بر تحلیل فضایی، برای بررسی روابط میان متغیرها، ضرایب همبستگی پیرسون و اسپیرمن برای سناریوی منتخب محاسبه و نتایج آن در شکل (۵) ارائه شد. این تحلیل با ارزیابی هم‌زمان روابط خطی و یکنواخت، تصویری اولیه از شدت و جهت ارتباط متغیرهای مستقل با غلظت $PM_{2.5}$ و نیز همبستگی میان ویژگی‌های ورودی فراهم می‌کند. نتایج نشان داد که برخی متغیرهای مرتبط با محیط ساخته شده و شرایط هواشناسی همبستگی قوی‌تری با $PM_{2.5}$ دارند که بیانگر نقش توامان عوامل شهری و جوی در تغییرات این آلاینده است. همچنین، بررسی همبستگی میان متغیرهای مستقل در شناخت ساختار داده‌ها و تفسیر رفتار مدل‌های یادگیری ماشین مفید است. با این حال، از آنجا که تحلیل همبستگی تنها روابط خطی یا یکنواخت را نشان می‌دهد، تفسیر نهایی اهمیت و نحوه اثرگذاری متغیرها بر اساس نتایج مدل‌سازی و تحلیل SHAP در بخش‌های بعدی ارائه می‌شود.

## ۳-۲- نتایج مدل‌سازی مکانی – زمانی $PM_{2.5}$

به‌منظور مدل‌سازی مکانی–زمانی $PM_{2.5}$، عملکرد سه الگوریتم یادگیری ماشین شامل درخت تصمیم، جنگل تصادفی و XGBoost در دو مقیاس زمانی (هفتگی و ماهانه) و دو مقیاس مکانی (۵۰۰ و ۱۰۰۰ متر) ارزیابی شد. نتایج ارزیابی زمانی مدل‌ها، با استفاده از داده‌های سال‌های ۲۰۱۹ تا ۲۰۲۱ برای آموزش و داده‌های سال ۲۰۲۲ برای اعتبارسنجی، در جدول (۳) ارائه شده است. عملکرد مدل‌ها بر اساس معیارهایMAE ، RMSE و $R^2$ مقایسه شد.

نتایج نشان داد که مدل XGBoost در تمامی سناریوهای مورد بررسی بهترین عملکرد را با کمترین مقادیر MAE و RMSE و بیشترین مقدار $R^2$ ارائه کرد و توانایی بیشتری در مدل‌سازی روابط پیچیده میان متغیرهای ورودی و غلظت $PM_{2.5}$ داشت. پس از آن، جنگل تصادفی عملکرد بهتری نسبت به درخت تصمیم نشان داد، در حالی که درخت تصمیم ضعیف‌ترین نتایج را در تمامی سناریوها به دست آورد.

افزون بر مقایسه الگوریتم‌ها، نتایج نشان داد که سناریوی مبتنی بر مقیاس مکانی ۱۰۰۰ متر و بازه زمانی هفتگی بهترین عملکرد را در میان تمامی سناریوهای مورد بررسی ارائه می‌دهد. در این سناریو، مدل XGBoost با مقادیر

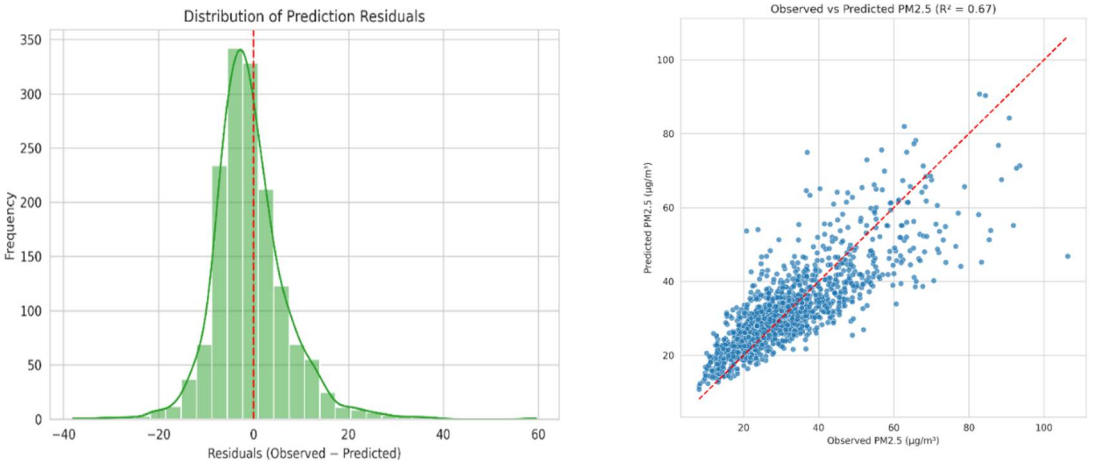
MAE=5.743، RMSE=7.617 و $R^2$=0.557 بالاترین دقت را کسب کرد و به عنوان مدل نهایی پژوهش انتخاب شد. این یافته نشان می‌دهد که انتخاب مناسب مقیاس مکانی و زمانی می‌تواند بر عملکرد مدل‌های یادگیری ماشین در پیش‌بینی غلظت $PM_{2.5}$ تاثیر قابل‌توجهی داشته باشد.

## ۳-۳- ارزیابی مدل بهینه و تعمیم‌پذیری مکانی

به منظور ارزیابی دقیق‌تر عملکرد مدل بهینه، نمودارهای اعتبارسنجی شامل مقایسه مقادیر پیش‌بینی شده در برابر مقادیر مشاهده شده و نیز توزیع باقی‌مانده‌های مدل بررسی شد. همان‌طور که در شکل (۶-الف) مشاهده می‌شود، بخش عمده‌ای از نقاط در اطراف خط برابری ($y = x$) قرار گرفته‌اند که بیانگر توان مناسب مدل در بازتولید مقادیر واقعی $PM_{2.5}$ است. مقدار ضریب تعیین به‌دست آمده نیز نشان می‌دهد که مدل بخش قابل توجهی از واریانس داده‌ها را تبیین می‌کند. بااین‌حال، با افزایش غلظت $PM_{2.5}$، پراکندگی نقاط بیشتر می‌شود که بیانگر کاهش دقت مدل در پیش‌بینی مقادیر بالا است.

جدول ۳- نتایج حاصل از مدل‌سازی آلاینده PM2.5 در شبکه با رزولوشن‌های مکانی و زمانی مختلف

| رزولوشن مکانی (متر) | رزولوشن زمانی | XGBoost | | | Random Forest | | | Decision Tree | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | MAE | R2 | RMSE | MAE | R2 | RMSE | MAE | R2 | RMSE |
| ۵۰۰ | ماهانه | ۵/۵۳۳ | ۰/۲۲۷ | ۶/۸۵۹ | ۵/۸۷۱ | ۰/۰۶۱ | ۷/۴۲۳ | ۷/۰۲۷ | -۰/۱۵۱ | ۸/۸۹۶ |
| | هفتگی | ۶/۰۴۴ | ۰/۴۸۹ | ۷/۸۶۱ | ۶/۴۵۱ | ۰/۳۹۲ | ۸/۵۲۶ | ۷/۲۷۲ | ۰/۱۸۹ | ۹/۹۶۵ |
| ۱۰۰۰ | ماهانه | ۵/۲۰۰ | ۰/۳۷۱ | ۶/۵۲۹ | ۵/۴۶۸ | ۰/۲۹۹ | ۶/۹۵۱ | ۶/۷۸۴ | -۰/۰۷۲ | ۸/۳۲۳ |
| | هفتگی | ۵/۷۴۳ | **۰/۵۵۷** | ۷/۶۱۷ | ۶/۱۴۲ | ۰/۴۶۶ | ۸/۲۳۴ | ۷/۹۱۰ | ۰/۱۳۵ | ۱۰/۷۰۴ |




الف ب

شکل ۶- نمودارهای عملکرد مدل بهینه در مدل‌سازی $PM_{2.5}$. الف) مقایسه مقادیر پیش‌بینی شده در برابر مقادیر مشاهده شده، ب) هیستوگرام باقی‌مانده‌های مدل

همچنین، هیستوگرام باقی‌مانده‌های مدل در شکل (۶-ب) نشان می‌دهد که خطاهای پیش‌بینی عمدتا حول مقدار صفر متمرکز هشتند و توزیع آن‌ها از الگوی نسبتا متقارنی برخوردار است. این ویژگی حاکی از آن است که مدل به‌طورکلی فاقد سوگیری سیستماتیک آشکار در پیش‌بینی مقادیر $PM_{2.5}$ است. با این حال، وجود تعدادی باقی‌مانده دورافتاده در دو انتهای توزیع نشان می‌دهد که مدل در بازنمایی برخی رخدادهای شدید آلودگی با محدودیت‌هایی مواجه است.

برای ارزیابی توان تعمیم‌پذیری مکانی مدل بهینه، از روش اعتبارسنجی متقاطع مکانی Leave-One-Group-Out (LOGO) استفاده شد. در این ارزیابی، هر ایستگاه پایش به عنوان یک گروه مستقل در نظر گرفته شد و مدل در هر تکرار با حذف کامل داده‌های یک ایستگاه از فرایند آموزش، برای

پیش‌بینی در آن مکان آزمون شد (جدول ۴). نتایج نشان می‌دهد که مدل بهینه، در مجموع، از پایداری و تعمیم‌پذیری مکانی قابل قبولی برخوردار است، به طوری که میانگین شاخص‌های عملکرد در سطح ایستگاه‌ها بیانگر توان مناسب مدل در پیش‌بینی $PM_{2.5}$ در مکان‌های مشاهده نشده است.

بااین‌حال، عملکرد مدل در میان ایستگاه‌های مختلف یکسان نبوده و ناهمگنی قابل توجهی در دقت پیش‌بینی مشاهده شد. در حالی که برخی ایستگاه‌ها از مقادیر بالای $R^2$ برخوردار بودند، در تعدادی دیگر عملکرد مدل ضعیف‌تر بود. این موضوع نشان می‌دهد که اگرچه مدل در مقیاس کلی عملکرد موفقی دارد، دقت آن در برخی موقعیت‌های خاص می‌تواند تحت تاثیر ویژگی‌های محلی، منابع انتشار موضعی با پیچیدگی‌های توپوگرافی قرار بگیرد.

جدول ۴- نتایج LOGO مدل بهینه

| ایستگاه | $R^2$ | MAE | RMSE |
|---|---|---|---|
| ۱ | ۰/۷۰ | ۵/۲۹۰ | ۷/۴۳۵ |
| ۲ | ۰/۴۵۳ | ۷/۴۶۹ | ۹/۹۹۱ |
| ۳ | ۰/۲۵۲ | ۴/۹۴۶ | ۶/۲۱۲ |
| ۴ | ۰/۸۷۳ | ۳/۶۳۲ | ۴/۵۱۵ |
| ۵ | ۰/۵۵۳ | ۶/۲۷۸ | ۷/۴۰۴ |
| ۶ | ۰/۶۳۱ | ۵/۰۶۳ | ۷/۴۰۴ |
| ۷ | ۰/۲۹۰ | ۵/۸۰۱ | ۶/۹۰۵ |
| ۸ | ۰/۶۰۳ | ۶/۹۱۴ | ۸/۵۲۵ |
| ۹ | ۰/۷۳۹ | ۴/۸۵۵ | ۷/۶۲۴ |
| ۱۰ | -۰/۰۲۴ | ۵/۱۷۵ | ۶/۹۶۹ |
| ۱۱ | ۰/۷۴۸ | ۳/۷۳۶ | ۵/۵۳۳ |

## ۳-۴- اهمیت متغیرها و توضیح‌پذیری مدل

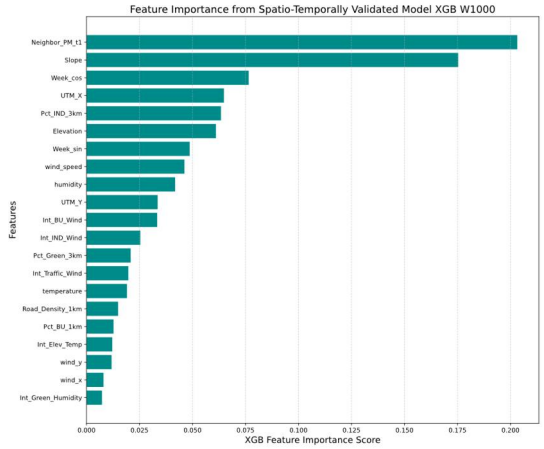
در این بخش، برای تفسیر رفتار مدل بهینه، ابتدا در بخش ۳-۴-۱ اهمیت نسبی متغیرهای ورودی بر اساس معیارهای درونی مدل بررسی می‌شود و سپس نتایج توضیح‌پذیری مدل با استفاده از روش SHAP در دو سطح سراسری و محلی در بخش ۳-۵ ارائه می‌گردد.

### ۳-۴-۱- تحلیل اهمیت متغیرها در مدل بهینه

به‌منظور بررسی سهم نسبی متغیرهای ورودی در عملکرد مدل بهینه، اهمیت ویژگی‌ها بر اساس معیار بهره اطلاعاتی (Gain) محاسبه شد و نتایج آن در شکل (۷) ارائه شده است. نتایج نشان داد که ویژگی همسایگی مکانی-زمانی بیشترین اهمیت را دارد که بیانگر نقش برجسته خودهمبستگی مکانی-زمانی در تغییرات $PM_{2.5}$ است. پس از آن، شیب، مختصات مکانی و برخی ویژگی‌های زمانی در رتبه‌های بعدی قرار گرفتند، در حالی که متغیرهای بستر و پوشش زمین، مورفولوژی شهری و شرایط هواشناسی نیز در بهبود عملکرد مدل سهم قابل توجهی داشتند.

## ۳-۵- نتایج توضیح‌پذیری مدل با SHAP

اگرچه بررسی اهمیت ویژگی‌ها بر اساس معیار آماری بهره اطلاعاتی شناخت مناسبی از قدرت پیش‌بینی‌کنندگی مدل‌ها ارائه می‌دهد، اما این روش سنتی قادر به توضیح جهت تأثیر (افزایش یا کاهشی بودن اثر آلودگی) و رفتار غیرخطی متغیرها در مقادیر مختلف نیست. به‌منظور غلبه بر این محدودیت، در این تحقیق از روش SHAP در دو سطح سراسری و محلی استفاده شد که نتایج آن به ترتیب در بخش‌های ۳-۵-۱ و ۳-۵-۲ ارائه شده‌اند.




شکل ۷- نمودار اهمیت متغیرها در مدل بهینه

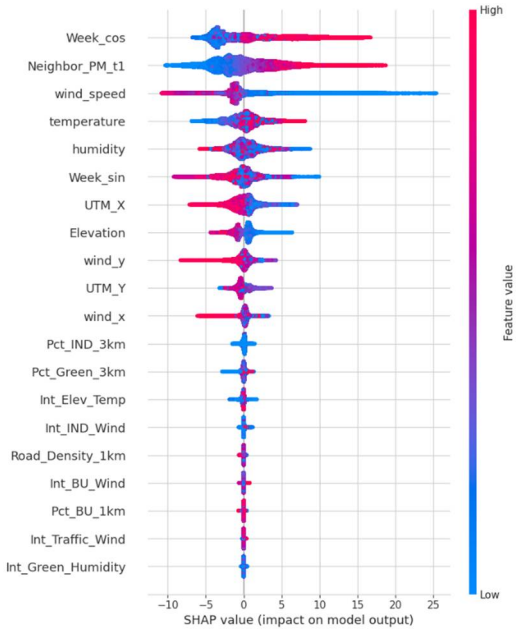
### ۳-۵-۱- نتایج توضیح‌پذیری در سطح سراسری

به‌منظور تفسیر رفتار مدل بهینه، تحلیل سراسری SHAP برای بررسی اهمیت و جهت اثر متغیرهای ورودی انجام شد. نمودار خلاصه SHAP (شکل ۸) نشان داد که رفتار مدل عمدتاً تحت تأثیر پنج متغیر کلیدی شامل ویژگی همسایگی مکانی–زمانی $PM_{2.5}$، چرخه زمانی، دما، سرعت باد و رطوبت قرار دارد. در میان آن‌ها، ویژگی همسایگی مکانی–زمانی بیشترین اهمیت را داشت؛ به‌طوری که مقادیر بالای آن عمدتاً با مقادیر مثبت SHAP همراه بود و نقش خودهمبستگی مکانی–زمانی در تغییرات $PM_{2.5}$ را نشان می‌دهد.

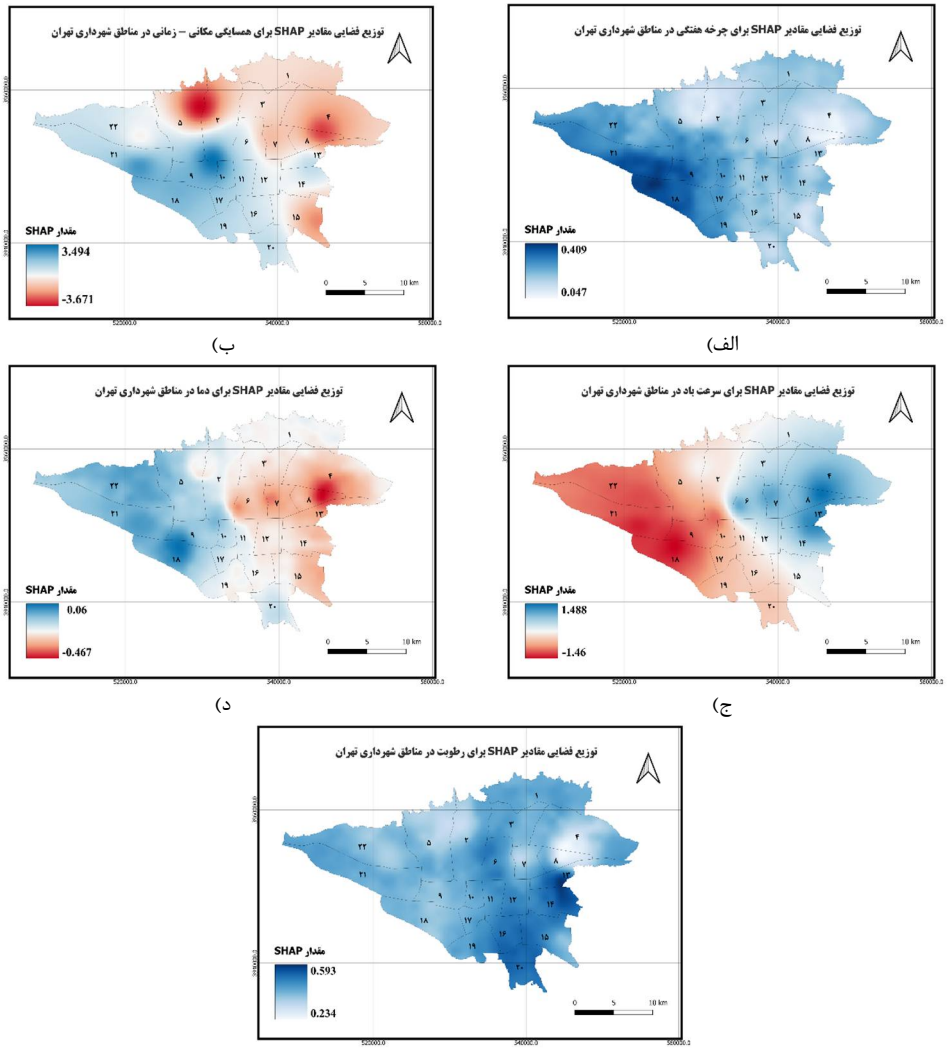
همچنین، اهمیت بالای ویژگی چرخه زمانی بیانگر تأثیر الگوهای زمانی بر تغییرات $PM_{2.5}$ است. در میان متغیرهای هواشناسی، افزایش سرعت باد عمدتاً اثر کاهنده، رطوبت اثر افزایشی و دما رفتاری غیرخطی بر خروجی مدل نشان داد. به‌طور کلی، نتایج SHAP نشان می‌دهد که مدل بهینه تغییرات $PM_{2.5}$ را بر پایه تعامل عوامل مکانی–زمانی و شرایط هواشناسی بازنمایی می‌کند.

### ۳-۵-۲- نتایج توضیح‌پذیری در سطح محلی

به‌منظور بررسی ناهمگنی فضایی رفتار مدل، توزیع مکانی مقادیر محلی SHAP برای پنج متغیر کلیدی شامل چرخه زمانی، ویژگی همسایگی مکان - زمانی $PM_{2.5}$، سرعت باد، دما و رطوبت در قالب نقشه‌های پیوسته ارائه شد (شکل ۹).




شکل ۸- نمودار SHAP در سطح سراسری برای مدل بهینه




شکل ۹- توزیع فضایی مقادیر SHAP محلی برای ویژگی‌های مهم در مدل بهینه در شهر تهران: الف) چرخه هفتگی، ب) همسایگی مکانی – زمانی، ج) سرعت باد، د) دما و ه) رطوبت

این نقشه‌ها نشان می‌دهند که اگرچه این متغیرها در سطح سراسری از مهم‌ترین عوامل مثر بر خروجی مدل هستند، شدت و حتی جهت آن‌ها در بخش‌های مختلف شهر یکسان نیست. به عبارت دیگر، سهم هر متغیر در مدل‌سازی $PM_{2.5}$ دارای ناهمگنی فضایی مشخصی است و بسته به شرایط محلی تغییر می‌کند.

به‌منظور بررسی ناهمگنی فضایی رفتار مدل، توزیع مکانی مقادیر محلی SHAP برای پنج متغیر کلیدی شامل چرخه زمانی، ویژگی همسایگی مکانی–زمانی $PM_{2.5}$، سرعت باد، دما و رطوبت در قالب نقشه‌های پیوسته ارائه شد (شکل ۹). نتایج نشان داد که اگرچه این متغیرها در سطح سراسری از مهم‌ترین عوامل مؤثر بر خروجی مدل هستند، شدت و جهت اثر آن‌ها در بخش‌های مختلف شهر متفاوت است.

ویژگی همسایگی مکانی–زمانی آشکارترین ناهمگنی فضایی را نشان داد؛ به‌طوری که اثر آن در نواحی مرکزی و جنوب‌غربی شهر قوی‌تر و در بخش‌های شمالی و شمال‌شرقی ضعیف‌تر بود. همچنین، متغیرهای هواشناسی نیز رفتار مکانی متفاوتی داشتند؛ به‌گونه‌ای که سرعت باد در بخش‌های غربی و شمال‌غربی عمدتاً اثر کاهنده، دما رفتاری غیرخطی و وابسته به موقعیت مکانی، و رطوبت در بیشتر مناطق اثر افزایشی بر خروجی مدل نشان داد. در مقابل، چرخه زمانی نسبت به سایر متغیرها الگوی فضایی یکنواخت‌تری داشت، هرچند شدت اثر آن نیز در سطح شهر ثابت نبود.

به‌طور کلی، نتایج تحلیل محلی SHAP نشان می‌دهد که اثر متغیرهای کلیدی در سطح شهر یکنواخت نیست و تفسیرپذیری محلی می‌تواند درک دقیق‌تری از تفاوت‌های فضایی عوامل مؤثر بر تغییرات $PM_{2.5}$ فراهم کند.

## ۴- بحث و نتیجه‌گیری

در این پژوهش، یک چارچوب توضیح‌پذیر برای مدل‌سازی مکانی–زمانی $PM_{2.5}$ در شهر تهران ارائه شد. برای این منظور، داده‌های آلودگی هوا با متغیرهای هواشناسی، ویژگی‌های بستر و پوشش زمین، مورفولوژی شهری، ویژگی همسایگی مکانی–زمانی و ویژگی‌های تعاملی تلفیق شدند و عملکرد سه الگوریتم درخت تصمیم، جنگل تصادفی و XGBoostدر دو مقیاس مکانی (۵۰۰ و ۱۰۰۰ متر) و دو مقیاس زمانی (هفتگی و ماهانه) مقایسه شد.

نتایج نشان داد که مدل XGBoost در تمامی سناریوها بهترین عملکرد را داشت و در مقایسه با جنگل تصادفی و درخت تصمیم، مقادیر کمتر MAE و RMSE و مقدار بیشتر $R^2$ را به دست آورد. همچنین، مقایسه سناریوهای مختلف نشان داد که ترکیب تفکیک مکانی ۱۰۰۰ متر و بازه زمانی هفتگی مناسب‌ترین ساختار را برای مدل‌سازی $PM_{2.5}$ فراهم می‌کند. این یافته نشان می‌دهد که انتخاب مقیاس مکانی و زمانی مناسب می‌تواند نقش مهمی در بهبود عملکرد مدل‌های یادگیری ماشین داشته باشد.

تحلیل اهمیت ویژگی‌ها و نتایج روش SHAP نشان داد که ویژگی همسایگی مکانی–زمانی، چرخه زمانی و متغیرهای هواشناسی، به‌ویژه دما، سرعت باد و رطوبت، مهم‌ترین عوامل مؤثر بر پیش‌بینی $PM_{2.5}$ هستند. اهمیت بالای ویژگی همسایگی مکانی–زمانی بیانگر نقش خودهمبستگی مکانی–زمانی در تغییرات این آلاینده است، در حالی که متغیرهای هواشناسی و زمانی نشان می‌دهند تغییرات $PM_{2.5}$ حاصل برهم‌کنش شرایط جوی، الگوهای زمانی و ساختار فضایی شهر است.

یکی از نوآوری‌های اصلی این پژوهش، استفاده از روش SHAP در دو سطح سراسری و محلی برای تفسیر رفتار مدل بود. تحلیل سراسری متغیرهای کلیدی و جهت کلی اثر آن‌ها را مشخص کرد، در حالی که تحلیل محلی نشان داد شدت و حتی جهت اثر این متغیرها در بخش‌های مختلف شهر یکسان نیست. این نتایج بیانگر وجود ناهمگنی فضایی در عوامل مؤثر بر $PM_{2.5}$ است و نشان می‌دهد که تحلیل‌های محلی می‌توانند اطلاعاتی را آشکار کنند که در تحلیل‌های سراسری قابل مشاهده نیست.

از نظر کاربردی، چارچوب پیشنهادی می‌تواند به‌عنوان ابزاری برای پشتیبانی از مدیریت کیفیت هوا مورد استفاده قرار گیرد. نقشه‌های پیش‌بینی $PM_{2.5}$ و تحلیل‌های توضیح‌پذیر SHAPامکان شناسایی مناطق با ریسک بالاتر، تعیین عوامل مؤثر در هر ناحیه و اولویت‌بندی اقدامات مدیریتی مانند کنترل ترافیک، توسعه فضای سبز و بهینه‌سازی شبکه پایش کیفیت هوا را فراهم می‌کنند. همچنین، نتایج این پژوهش نشان می‌دهد که استفاده از گرید ۱۰۰۰ متری می‌تواند مبنایی برای طراحی شبکه‌های پایش و استقرار حسگرهای کم‌هزینه در مطالعات آینده باشد.

با وجود عملکرد مناسب الگوریتم XGBoost، مقدار $R^2$ به‌دست‌آمده نشان می‌دهد که بخشی از تغییرات $PM_{2.5}$

همچنان توسط متغیرهای مورد استفاده تبیین نشده است. این موضوع به محدودیت تعداد و پراکندگی ایستگاه‌های پایش، نبود داده‌های مستقیم ترافیک، منابع انتشار و برخی متغیرهای محیطی مؤثر مربوط می‌شود. ازاین‌رو، نتایج این پژوهش باید با در نظر گرفتن این محدودیت‌ها تفسیر و بررسی شوند.

مهم‌ترین محدودیت‌های این پژوهش شامل تعداد محدود ایستگاه‌های پایش کیفیت هوا و عدم دسترسی به داده‌های زمانی و مکانی ترافیک، نظیر حجم، سرعت و شمار تردد وسایل نقلیه، بود. به همین دلیل، از تراکم شبکه معابر به‌عنوان نماینده‌ای از منابع متحرک انتشار استفاده شد که اگرچه الگوی فضایی ترافیک را بازنمایی می‌کند، اما قادر به نمایش تغییرات زمانی آن نیست. همچنین، عدم استفاده از برخی متغیرهای محیطی مؤثر، مانند دمای سطح زمین (LST) و اثرات جزیره حرارتی شهری، می‌تواند در تبیین نشدن بخشی از تغییرات $PM_{2.5}$ مؤثر بوده باشد. ازاین‌رو، استفاده از داده‌های واقعی ترافیک، محصولات سنجش‌ازدور مانند AOD وLST، توسعه شبکه‌های پایش و به‌کارگیری مدل‌های پیشرفته‌تری نظیر شبکه‌های عصبی گراف، می‌تواند در مطالعات آینده به بهبود دقت پیش‌بینی و درک بهتر پویایی مکانی–زمانی آلودگی هوا کمک کند.

## سپاسگزاری

این مقاله با حمایت دانشگاه صنعتی خواجه نصیرالدین طوسی (بنیاد علم ایران (INSF) با کد گرنت ۴۰۴۰۹۵۰۱) انجام شده است.

## مراجع


[۱] World Health Organization. (2022, April 4). Billions of people still breathe unhealthy air: new WHO data. https://www.who.int/news/item/04-04-2022-billions-of-people-still-breathe-unhealthy-air-new-who-data

[۲] Naddafi, K., Hassanvand, M. S., Yunesian, M., Momeniha, F., Nabizadeh, R., Faridi, S., & Gholampour, A. (2012). Health impact assessment of air pollution in megacity of Tehran, Iran. Iranian journal of environmental health science & engineering, 9(1), 28.

[۳] Shahbazi, H., Reyhanian, M., Hosseini, V., & Afshin, H. (2016). The relative contributions of mobile sources to air pollutant emissions in Tehran, Iran: an emission inventory approach. Emission control science and technology, 2(1), 44–56.

[۴] Mohammadi, F., Teiri, H., Hajizadeh, Y., Abdolahnejad, A., & Ebrahimi, A. (2024). Prediction of atmospheric PM2. 5 level by machine learning techniques in Isfahan, Iran. Scientific Reports, 14(1), 2109.

[۵] Zhalehdoost A, Taleai M. (2025). Unravelling the importance of spatial and temporal resolutions in modeling urban air pollution using a machine learning approach. Scientific Reports. Jul 29;15(1):27708.

[۶] Tella, A., & Balogun, A.-L. (2022). GIS-based air quality modelling: Spatial prediction of PM10 for Selangor State, Malaysia using machine learning algorithms. Environmental Science and Pollution Research, 29(57), 86109–86125.

[۷] Sayeed, A., Choi, Y., Eslami, E., Lops, Y., Roy, A., & Jung, J. (2020). Using a deep convolutional neural network to predict 2017 ozone concentrations, 24 hours in advance. Neural Networks, 121, 396–408.

[۸] Delavar, M.R., Gholami, A., Shiran, G.R., Rashidi, Y., Nakhaeizadeh, G.R., Fedra, K. and Hatefi Afshar, S. (2019). A novel method for improving air pollution prediction based on machine learning approaches: a case study applied to the capital city of Tehran. ISPRS International Journal of Geo-Information, 8(2), p.99.

[۹] Huang, C.-J., & Kuo, P.-H. (2018). A deep CNN-LSTM model for particulate matter (PM2. 5) forecasting in smart cities. Sensors, 18(7), 2220.

[۱۰] Habibi, R., Alesheikh, A.A., Mohammadinia, A. and Sharif, M. (2017). An assessment of spatial pattern characterization of air pollution: A case study of CO and PM2. 5 in Tehran, Iran. ISPRS international journal of Geo-information, 6(9), p.270.

[۱۱] Rybarczyk, Y., & Zalakeviciute, R. (2018). Machine learning approaches for outdoor air quality modelling: A systematic review. Applied Sciences, 8(12), 2570.

[۱۲] Nabavi, S.O., Haimberger, L. and Abbasi, E. (2019). Assessing PM2. 5 concentrations in Tehran, Iran, from space using MAIAC, deep blue, and dark target AOD and machine learning algorithms. Atmospheric Pollution Research, 10(3), pp.889-903.

[۱۳] Kaviani Rad, A. K., Shamshiri, R. R., Naghipour, A., Razmi, S.-O., Shariati, M., Golkar, F., & Balasundram, S. K. (2022). Machine learning for determining interactions between air pollutants and environmental parameters in three cities of Iran. Sustainability, 14(13), 8027.

[۱۴] Makhdoomi, A., Sarkhosh, M. and Ziaei, S. (2025). PM2. 5 concentration prediction using machine learning algorithms: an approach to virtual monitoring stations. Scientific Reports, 15(1), p.8076.

[۱۵] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). " Why should i trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining.

[۱۶] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.

[۱۷] Liu, X., & Su, H. (2025). Unraveling the complex relationships between urban land use pattern and air quality: Evidence from GBDT-SHAP analysis. Urban Climate, 64, 102711.

[۱۸] Abbasi, M. T., Alesheikh, A. A., & Rezaie, F. (2025). A Lightweight Spatiotemporal Graph Framework Leveraging Clustered Monitoring Networks and Copula-Based Pollutant Dependency for PM2. 5 Forecasting. Land, 14(8), 1589.

[۱۹] Burrough, P. A., & McDonnell, R. A. (1998). Principles of Geographical Information Systems. Oxford University Press.

[۲۰] Li, J., & Heap, A. D. (2014). A review of spatial interpolation methods for environmental scientists. Geoscience Australia.

[۲۱] Xu, W., Jin, X., Liu, M., Ma, Z. and Zhou, Y. (2021). Analysis of spatiotemporal variation of PM2. 5 and its relationship to land use in China. Atmospheric Pollution Research, 12(9), p.101151.