نشریه علمی علوم و فنون نقشه‌برداری، دوره ۱۲، شماره ۴، صفحه ۷۷ تا ۸۹، خرداد ماه ۱۴۰۲- مقاله پژوهشی

مدلسازی غلظت ذرات معلق $PM_{2.5}$ بر مبنای جانهی داده‌ها و استفاده از روش‌های یادگیری ماشین

سارا حق‌بیان*[۱]، بهنام تشیع[۲]، مریم حسینی[۳]

[۱] دانشجوی دکتری، دانشکده مهندسی عمران و حمل و نقل، دانشگاه اصفهان
haghbayansara@yahoo.com

[۲] استادیار، دانشکده مهندسی عمران و حمل و نقل، دانشگاه اصفهان
b.tashayo@eng.ui.ac.ir

[۳] دانشجوی کارشناسی ارشد، دانشکده مهندسی عمران و حمل و نقل، دانشگاه اصفهان
maryam.hosseinii1977@gmail.com

(تاریخ دریافت: اسفند ماه ۱۴۰۱، تاریخ تصویب: خرداد ماه ۱۴۰۲)

چکیده

مدلسازی دقیق تغییرات مکانی_زمانی نیازمند روش مناسب و داده‌های کامل و دقیق است. داده‌ها از حسگرهای ایستگاه‌های پایش جمع-آوری می‌شوند. تعداد این ایستگاه‌ها محدود است و به دلیل عوامل اجتناب ناپذیر بخشی از داده‌ها از دست می‌روند. نوآوری مقاله حاضر، غلبه بر محدودیت‌های روش‌های موجود در جانهی مقادیر از دست رفته $PM_{2.5}$ است. محدودیت روش‌های موجود، عدم توجه همزمان به مکانیسم مکانی_زمانی داده‌های از دست رفته‌است. به منظور غلبه بر محدودیت‌های روش‌های موجود، جانهی مقادیر از دست رفته $PM_{2.5}$ همراه با در نظر گرفتن روابط بین متغیرها با حفظ تغییر پذیری و عدم قطعیت طبیعی آن‌ها، با استفاده از مدل‌های درخت اضافی و درخت تصمیم پیاده سازی گردید. نتایج نشان داد که روش درخت اضافی به دلیل کاهش سوگیری با میانگین $R^2=$ ۰.۸۰ دقت بالاتری از روش درخت تصمیم با میانگین $R^2=$ ۰.۶۴ در جانهی مقادیر از دست رفته $PM_{2.5}$ دارد. پس از مدیریت داده‌های از دست رفته با استفاده از روش درخت اضافی، از روش XGBoost به دلیل ارزیابی غیرخطی اهمیت متغیرهای موثر با هدف افزیش دقت و کاهش هزینه محاسباتی برای مدلسازی تغییرات مکانی_زمانی آلاینده $PM_{2.5}$ در بافت‌های مختلف جغرافیایی کلانشهر تهران استفاده گردید. متغیرهای موثر درنظرگرفته شده برای جانهی و مدلسازی شامل داده‌های هواشناسی و سایر آلاینده‌های اصلی نظیر O_3،Pm_{10}،Co،So_2،No_2 است. متغیرهای هواشناسی شامل مجموع بارش، رطوبت نسبی، دما از مدل ECMWF استخراج گردیدند. استفاده از مدل ECMWF علاوه بر افزایش تعداد ایستگاه هواشناسی امکان استفاده از رزولوشن یک ساعتی با تعداد بسیار ناچیز داده از دست رفته را در مقابل تعداد محدود، رزولوشن سه ساعتی با تعداد زیاد داده از دست رفته هواشناسی را فراهم می‌کند.

واژگان کلیدی: $PM_{2.5}$، داده‌های از دست رفته، یادگیری ماشین، ECMWF، درخت تصمیم، XGBoost.

* نویسنده رابط

۱- مقدمه

مدیریت موثر آلاینده‌ها یکی از دغدغه‌های مسولین و مردم است که در بسیاری از کشورها و مناطق قابل حل است [۱]. یکی از راه‌های مدیریت موثر آلاینده‌ها از جمله غلظت ذرات معلق $PM_{2.5}$ استفاده از نتایج مدلسازی دقیق است. با استفاده از نتایج مدلسازی دقیق می‌توان استراتژی-های موثر و مختلفی را برای کاهش آلاینده‌های هوا استفاده کرد [۲]. در حال حاضر متداول‌ترین روش‌های مدلسازی به منظور پیش‌بینی آلاینده‌ها به چهار دسته مدل‌های ریاضی، مدل‌های فیزیکی، روش‌های آماری و روش‌های مبتنی بر یادگیری ماشین قابل تقسیم است [۲]. مدل‌های ریاضی و فیزیکی نیاز به دانش قبلی پیچیده‌ای از حالت انتشار آلاینده و مسیرهای واکنش شیمیایی دارند که غالبا شناخته شده نیست به همین دلیل نسبت به دو روش دیگر کمتر مورد استفاده قرار می‌گیرد. غالبا مطالعات از روش‌های آماری به دلیل سادگی و سهولت استفاده می‌کنند [۳]. روش‌های آماری شامل درون‌یابی، روش تجربی و رگرسیون است. این روش‌ها دارای محدودیت‌هایی از جمله مفروضات محدود کننده‌ای در مورد استقلال مشاهدات و توزیع نتایج، تعریف روابط بین متغیرها از قبل است [۴]. یکی‌دیگر از روش‌های آماری، روش تجربی است که از جمله روش‌های سنتی است که برای پیش‌بینی آلودگی هوا استفاده می‌شود. این روش معمولا براساس اطلاعات کارشناس محیطی توسعه داده می‌شود و دارای محدودیت هستند. عملکرد و دقت این مدل‌ها غالبا به منطقه مورد مطالعه وابسته‌است همچنین استفاده از این مدل‌ها در مقیاس بزرگ دشوار است [۵]. در کنار معایب روش‌های ذکرشده استفاده از روش‌های یادگیری ماشین که به طور مداوم پارامترهای مدل را از طریق تکرار تطبیقی[۱] و تمرین وزن بهینه می‌کند، بسیار مورد توجه محققین قرار گرفته‌است. استفاده از این روش توانایی مدل‌سازی غیرخطی برای پیش بینی غلظت ذرات معلق $PM_{2.5}$ محقق می‌سازد؛ بنابراین طبق نتایج مطالعات، استفاده از روش یادگیری ماشین دقت بالاتری را برای پیش بینی غلظت ذرات معلق $PM_{2.5}$ دارد.

مطالعات متعددی برای مدلسازی غلظت $PM_{2.5}$ در قدرت تفکیک مکانی و زمانی مختلف با روش‌های یادگیری ماشین انجام گرفته‌است. نتایج مطالعات نشان داده که روش غیرخطی و پیشرفته XGBoost، به منظور غلبه بر روش‌های سنتی مانند رگرسیون خطی چندگانه، مدل جمعی تعمیم یافته، ماشین‌های بردار پشتیبان برای تخمین غلظت $PM_{2.5}$ استفاده می‌گردد [۱۳]. در ادامه پاره‌ای از مطالعات انجام گرفته در سال‌های اخیر که حاکی از برتری XGBoost در مدلسازی غلظت $PM_{2.5}$ است؛ بیان می‌گردد. باقری در سال ۲۰۲۲ از روش‌های مختلفی برای پیش بینی غلظت $PM_{2.5}$ استفاده کرد. در این مطالعات ده روش مورد مقایسه قرار گرفت. نتایج نشان داد روش XGBoost بالاترین دقت را در بین روش‌های مورد مقایسه دارد [۶]. زمانی و همکاران در سال ۲۰۱۹ از روش‌های مختلفی برای پیش بینی غلظت $PM_{2.5}$ استفاده کردند. روش‌های مورد مقایسه شامل جنگل‌های تصادفی، XGBoost و روش‌های یادگیری عمیق بود که نتایج نشان داد روش XGBoost با R^2 =۰.۸۰ بالاترین دقت را در بین روش‌های مورد مقایسه دارد [۷]. جینگهویی و همکاران در سال ۲۰۲۰ مطالعه‌ای به منظور پیش بینی غلظت $PM_{2.5}$ انجام دادند. در این مطالعه از روش‌های WRF-Chem، Lasso، XGBoost استفاده کردند. نتایج نشان داد که وقتی مقدار مشاهده شده $PM_{2.5}$ بیشتر از ۵۰ میکروگرم در متر مکعب باشد مدل XGBoost بهترین عملکرد را در بین سه مدل داشت [۸]. گوندوغدو و همکاران در سال ۲۰۲۲ از روش‌های مختلفی برای پیش بینی غلظت $PM_{2.5}$ استفاده کردند. روش‌های مورد مقایسه شامل رگرسیون گام به گام[۲]، جنگل‌های تصادفی، شبکه‌های عصبی، XGBoost بودند که نتایج نشان داد روش XGBoost بهترین عملکرد و رگرسیون گام به گام بدترین عملکرد را داشته‌است [۹].

غالبا روش‌های مدلسازی آلودگی هوا بر مبنای اطلاعات ایستگاه‌های پایش توسعه داده می‌شوند [۲]. حال ایستگاه-های پایش با مشکلات متعددی از جمله قطعی برق، از کار افتادن حسگر، نداشتن برخی از حسگرها در برخی ایستگاه-ها مواجه هستند. همین مشکلات سبب از دست رفتن داده‌های مرتبط با آلاینده‌ها می‌شود؛ لذا داده‌های از دست رفته با روش‌های مناسب و درخور ماهیت پیچیده و دینامیک آلاینده $PM_{2.5}$ باید جانهی شود.

۱ Stepwise Regression

انتخاب یک روش مناسب برای مقابله با داده‌های از دست رفته، به ویژه داده‌های غنی از اطلاعات، چالش برانگیز است و گاهی اوقات می‌تواند منجر به نتایج سوگیرانه‌ای در صورت استفاده از مفروضات نادرست شود [۱۰]. دو روش حذف و جانهی برای مدیریت مقادیر از دست رفته وجود دارد [۱۱]. روش حذف کردن به دلیل سادگی و عدم نیاز به تخمین مقادیر از دست رفته به طور گسترده مورد استفاده قرا می‌گیرد. عیب این روش این است که علاوه بر از دست دادن حجم زیادی داده به ویژه در نمونه‌های کوچک و توزیع غیر تصادفی داده‌ها موجب بایاس[۱] در تجزیه و تحلیل می‌شود [۱۲]. دومین روش، روش جانهی است که خود به دو دسته آماری[۲] و روش‌های یادگیری ماشین[۳] تقسیم می‌شود [۱۳]. روش‌های آماری برای مدیریت داده‌های از دست رفته شامل میانگین، رگرسیون خطی، حداقل مربعات، حداکثرسازی انتظارات[۴] است. در میان این روش‌های آماری، میانگین ساده‌ترین روش مدیریت داده‌های از دست رفته‌است. در این روش مقادیر از دست رفته با مقدار میانگین پر می‌شود. این روش علاوه بر ایجاد بایاس هیچ اطلاعات جدیدی اضافه نمی‌کند؛ تنها حجم نمونه را بالا می‌برد و منجر به خطا می‌شود. این روش مناسب مشاهدات تصادفی هستند که از توزیع نرمال پیروی می‌کنند؛ بنابراین جانهی با روش آماری میانگین عموما پذیرفته شده نیست [۱۳, ۱۴]. در روش رگرسیون خطی روابط بین متغیرهای موجود تخمین زده می‌شود و سپس از ضرایب رگرسیون برای تخمین مقادیر از دست رفته استفاده می‌شود. این روش برخلاف روش حذف کردن، حجم داده را حفظ می‌کند و از تغییر انحراف معیار یا شکل توزیع اجتناب می‌کند [۱۴]. روش حداقل مربعات هم معمولا برای تولید نتیجه پیش بینی نهایی استفاده می‌شود [۱۳]. روش حداکثر سازی انتظارات یک فرآیند تکراری است که تا زمانی ادامه می‌یابد که در برآورد پارامترها همگرایی وجود داشته باشد [۱۱]. لذا استفاده از روش‌های یادگیری ماشین برای غلبه بر روش‌های جانهی آماری در سال‌های اخیر بسیار مورد توجه قرار گرفته‌است.

۱ Bias
۲ Statistical
۳ Machine Learning
۴ Expectation maximization (EM)

روش‌های متداول یادگیری ماشین برای جانهی داده‌های از دست رفته شامل نزدیکترین همسایگی، جنگل تصادفی و درخت تصمیم است [۱۳]. در روش نزدیکترین همسایگی[۵] مقادیر از دست رفته با نزدیکترین همسایگان خود جایگزین می‌شوند. این روش علاوه بر حساسیت به داده‌های پرت، نیاز به انتخاب دقیق پارامتر 'k' دارد. روش جنگل‌های تصادفی به داشتن بایاس شدید در متغیرهای پیوسته شناخته شده‌است [۱۵]. روش‌های جنگل‌های تصادفی و نزدیکترین همسایگی در تعدادی از مطالعات مورد استفاده قرار گرفته‌اند، در حالی که روش درخت تصمیم از سال ۲۰۰۶ تا ۲۰۱۶ به طور مداوم هر سال در مطالعات مورد استفاده قرار گرفته‌است [۱۳]. در روش درخت تصمیم بجای جایگزین کردن یک مقدار واحد برای هر داده از دست رفته، مقادیر از دست رفته با مجموعه‌ای از مقادیر قابل قبول که دارای تغییر پذیری طبیعی و عدم قطعیت مقادیر صحیح هستند، جایگزین شده‌است. این روش، تکرارپذیری را تکرار می‌کند و مجموعه داده‌های جانهی متعددی را ایجاد می‌کند. سپس مجموعه داده‌های تولید شده با استفاده از آمار استاندارد مورد تجزیه و تحلیل قرار می‌گیرد. مزیت این روش بازیابی تغییرپذیری طبیعی مقادیر از دست رفته، عدم قطعیت به دلیل داده‌های از دست رفته را شامل می‌شود که منجر به یک استنتاج آماری معتبر می‌شود. از دیگر مزایای آن نتایج مناسب حتی در حجم نمونه‌های کوچک یا نمونه‌های با تعداد زیاد داده‌های از دست رفته‌است [۱۴]. روش درخت اضافی از الگوریتم‌های رایج یادگیری ماشین نظیر درخت تصمیم، ماشین‌های بردار پشتیبانی و جنگل تصادفی سریع‌تر و دقیق‌تر عمل می‌کند؛ زیرا زمان را برای انتخاب نقطه تقسیم بهینه صرف نمی‌کند و از اطلاعات اضافی در مورد داده‌ها برای بهبود دقت استفاده می‌کند. لذا این روش بایاس و واریانس را کاهش می‌دهد و به طور قابل توجهی از بیش برازش و کمتر برازش جلوگیری می‌کند [۱۶]. علی‌رغم برتری‌های ذکر شده این روش نسبت به سایر روش‌های یادگیری ماشین در جانهی مقادیر از دست رفته کمتر مورد استفاده قرار گرفته‌است.

اهمیت پژوهش حاضر در پنج مورد خلاصه می‌شود. ادغام ایستگاه‌های پایش شهرداری و سازمان حفاظت محیط زیست، استفاده از مدل هواشناسی ECMWF ، انتخاب دو

۵ K-Nearest Neighbours (kNN)

روش برتر یادگیری ماشین در جانهی مقادیر از دست رفته $PM_{2.5}$ ، درنظرگرفتن متغیرهای هواشناسی و سایرآلاینده‌ها در جانهی مقادیر از دست رفته و مدلسازی مکانی_زمانی $PM_{2.5}$، انتخاب روش غیرخطی و پیشرفته XGBoost که حاصل مقایسه ده‌ها روش مختلف مدلسازی مطالعات پیشین است. ادغام ایستگاه‌های پایش به منظور افزایش تعداد ایستگاه با فرض افزایش دقت مدلسازی $PM_{2.5}$ انجام گرفت. طوری‌که در منطقه مورد مطالعه چهار ایستگاه هواشناسی وجود دارد که از طریق وب سایت هواشناسی کشور قابل دسترسی است؛ اما استفاده از مدل ECMWF مستلزم کدنویسی جاوا در محیط Google Earth Engine است. استفاده از مدل ECMWF بجای ایستگاه‌های محدود هواشناسی این امکان را فراهم می‌کند که متغیرهای هواشناسی هر ایستگاه پایش به صورت جداگانه تخمین زده شود؛ بنابراین از چهارده ایستگاه هواشناسی با رزولوشن زمانی یک ساعته بجای چهار ایستگاه هواشناسی با رزولوشن زمانی سه ساعته استفاده گردید. لذا استفاده از این مدل علاوه بر افزایش تعداد ایستگاه‌های هواشناسی و به احتمال زیاد افزایش دقت در تخمین متغیرهای هواشناسی با تعداد بسیار ناچیز داده از دست رفته هواشناسی همراه است. نتایج مطالعات پیشین حاکی از برتری روش‌ها درختی در تخمین مقادیر از دست $PM_{2.5}$ است. استفاده از روش-های درختی علاوه بر این که مشکلات سایر روش‌ها نظیر جایگزین کردن یک مقدار واحد برای هر داده از دست رفته، بایاس، عدم ایجاد اطلاعات اضافه، کاهش حجم نمونه، خطا، تغییر انحراف معیار یا تغییر توزیع شکل داده، حساسیت به داده پرت، نیاز به انتخاب دقیق پارامتر 'k' را نداشته و با درنظرگرفتن ارتباط بین متغیرهای مستقل و وابسته مقادیر از دست رفته را با تکرارپذیری، مجموعه داده‌های جانهی متعددی را ایجاد می‌کند که تغییرپذیری طبیعی مقادیر از دست رفته و عدم قطعیت را شامل می‌شود. درنظرگرفتن ارتباط بین متغیر $PM_{2.5}$ با متغیرهای هواشناسی شامل رطوبت، دما، فشار، میانگین سالانه بارندگی، نقطه شبنم به همراه سایر آلاینده‌ها شامل No_2،So_2،Co،Pm_{10}،O_3 با هدف جانهی و مدلسازی دقیق‌تر تغییرات مکانی‌زمانی $PM_{2.5}$ در بافت‌های مختلف جغرافیایی منطقه مورد مطالعه انجام گرفت. روش XGBoost به دلیل توانایی تخمین ناهمگنی پیچیده مکانی و زمانی داده‌های $PM_{2.5}$ ، حذف متغیرهای کم اهمیت‌تر با هدف افزیش دقت و کاهش هزینه محاسباتی برای پیش بینی ساعتی آلاینده $PM_{2.5}$ در کلانشهر تهران به عنوان منطقه مورد مطالعه انتخاب گردید. گرچه هر روش مدلسازی همواره با خطا همراه است ولی به مدیران و سیستم‌های هوشمند تصمیم‌گیر کمک می‌کند که با هشدار اولیه به مردم مانع فعالیت‌های غیرضروری گردند و با اجرای سیاست و محدودیت فعالیت‌های انسانی سبب کاهش آلودگی هوا شوند.

۲- مبانی نظری

در این پژوهش، داده‌های آلاینده هوا، داده‌های هواشناسی اعم از رطوبت [۱۷]، دما [۱۸, ۱۹]، فشار [۱۹, ۲۰]، بارندگی [۱۷, ۲۱]، دما نقطه شبنم [۱۹, ۲۲] که در مقالات معتبر سالیان گذشته به دلیل همبستگی آن‌ها با آلاینده $PM_{2.5}$ مورد استفاده قرار گرفت. علاوه بر استفاده مطالعات گوناگون از متغیرهای هواشناسی از سایر آلاینده‌های اصلی هوا برای مدلسازی PM2.5 نیز استفاده گردید که می‌توان به مطالعه‌ی که از ترکیب داده‌های هواشناسی با آلاینده‌های CO، CO_2، NO، NO_2، SO_2، و PM_{10} به عنوان پارامترهای کمکی در مدلسازی $PM_{2.5}$ استفاده گردید، اشاره کرد [۲۳]. لذا باتوجه ماهیت پیچیده و ناهمگون $PM_{2.5}$ ارتباط آن با متغیرهای متعددی از جمله داده‌های هواشناسی و سایر آلاینده‌های هوا درنظر گرفته شد.

پس از جمع آوری داده‌ها، پاکسازی داده‌ها مانند حذف نقاط پرت و جانهی مقادیر از دست رفته قبل از آموزش یادگیری ماشین به همان اندازه جمع آوری داده‌ها حیاتی است. پاکسازی داده‌ها به منظور آماده کردن داده‌ها برای یادگیری ماشین اغلب بین ۵۰ تا ۸۰ درصد وقت یک محقق را می‌گیرد [۲۴]. مرحله اول پاکسازی حذف کردن است. مشابه پژوهشی که داده‌های $PM_{2.5}$ ایی که بیش از سه ساعت متوالی داده از دست رفته داشتند را حذف کردند [۲۵]. مرحله دوم پاکسازی جانهی داده‌های از دست رفته‌است که به دو روش آماری و یادگیری ماشین قابل انجام است که در ادامه تشریح می‌گردد.

۲-۱- جانهی با استفاده از روش‌های آماری

جانهی آماری یک روشی برای مقابله با داده‌های از دست رفته‌است. در این روش یک مقدار تخمینی واحد برای

مقادیر از دست رفته به دست می‌آید [۲۶]. روش جانهی آماری مستلزم جایگزینی مقادیر از دست رفته برای هر مقدار جداگانه با استفاده از یک ویژگی کمی یا ویژگی کیفی از همه مقادیر غیر از دست رفته‌است. با جانهی آماری، داده‌های از دست رفته با روش‌های مختلفی مانند میانگین، میانه و درون‌یابی خطی مقادیر موجود مدیریت می‌شوند. در بیشتر مطالعات از روش‌های جانهی آماری به دلیل سادگی، روش مرجع آسان و مورد قبول استفاده می‌شود. با این حال، روش‌های جانهی آماری ممکن است منجر به بایاس یا نتایج غیر واقعی گردد. استفاده از این روش برای داده های با ابعاد بالا ممکن است که عملکرد ضعیفی داشته باشد [۲۷]. استفاده از روش درون‌یابی خطی تأثیر متغیرهای مستقل از دست رفته را در خطای پر کردن وابسته، به شدت کاهش می‌دهد؛ بنابراین استفاده از داده‌ها و دقت را تا حد زیادی بهبود می‌بخشد [۲۸].

۲-۲- جانهی با استفاده از روش‌های یادگیری ماشین

روش جانهی با استفاده از یادگیری ماشین یکی از جذاب ترین روش‌ها برای مدیریت همه منظوره داده‌های از دست رفته‌است [۲۹]. در این روش هر مقدار از دست رفته با مجموعه مقادیر به دست آمده از توزیع پیش بینی شده جایگذاری می‌شود [۳۰]. سپس هر مجموعه داده تجزیه و تحلیل شده و نتایج با هم ترکیب می‌شوند [۳۱]. جانهی بااستفاده از یادگیری ماشین، زمانی‌که توزیع داده‌های مشاهده شده برای تقریب مقادیر متعددی که عدم قطعیت در اطراف مقدار واقعی را منعکس می‌کنند، استفاده می‌شود و این روش بیشتر برای حل محدودیت‌های جانهی آماری اجرا شد [۳۲]. از جمله روش‌های جانهی با استفاده از روش یادگیری ماشین، روش درخت اضافی و درخت تصمیم است.

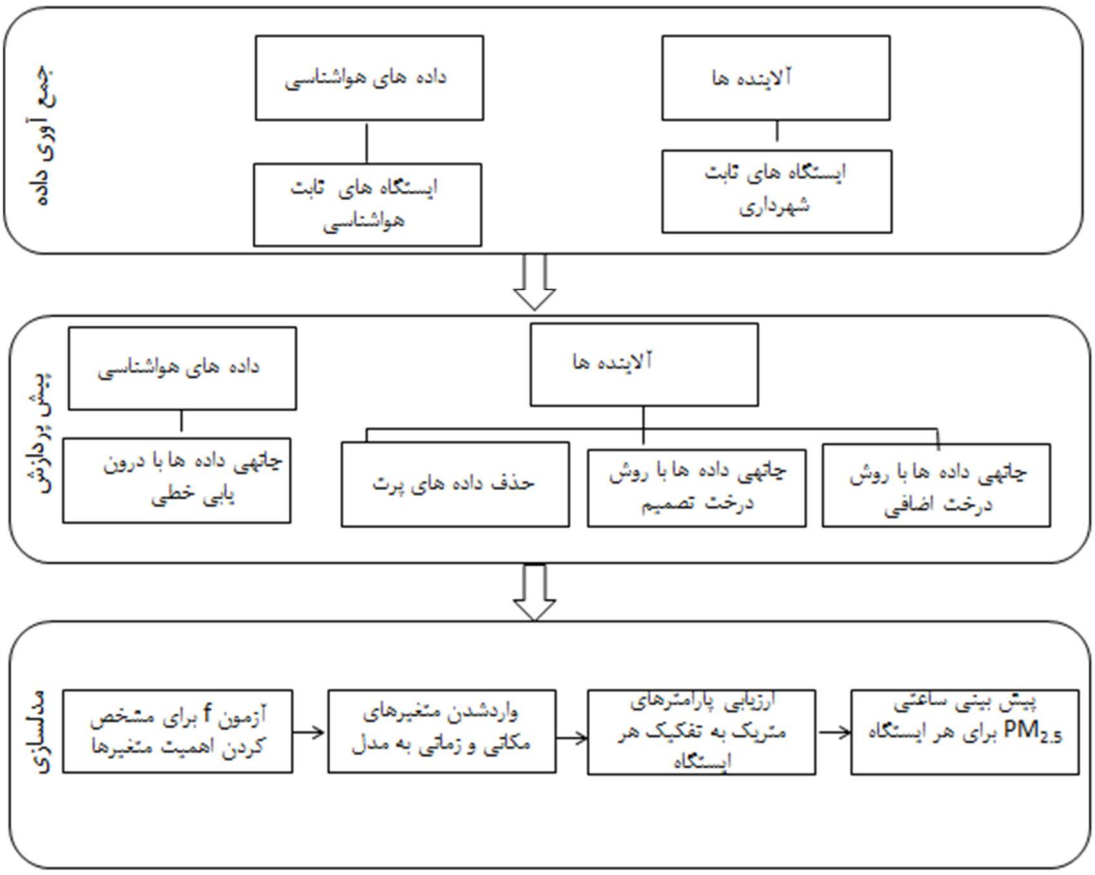
درخت بسیار تصادفی یا درخت اضافی، یک روش نسبتاً جدید یادگیری ماشین است و به عنوان توسعه الگوریتم جنگل تصادفی توسعه یافته‌است و احتمال کمتری برای بیش از حد برازش یک مجموعه داده وجود دارد [۳۳]. درختان اضافی توسط Geurts و همکاران معرفی شدند. آن‌ها لایه تصادفی بودن را به درخت تصمیم اضافه کردند [۳۴]. الگوریتم درخت اضافی از روش کلاسیک بالا به پایین پیروی می‌کند تا یک سری از درختان گرادیان خام یا رگرسیون ایجاد کند. دو تفاوت اصلی بین درخت اضافی و سایر روش‌های خوشه‌بندی مبتنی بر درخت این است که گره‌ها را با انتخاب نقاط برش به صورت تصادفی جدا می‌کند و درخت را با استفاده از کل نمونه یادگیری رشد می‌دهد. به طور کلی درخت اضافی از ویژگی‌های زیرمجموعه تصادفی برای آموزش هر تخمین‌گر پایه استفاده می‌کند [۳۵].

درخت تصمیم یک الگوریتم یادگیری ماشین است که تمام نتایج قابل تصور و مسیرهای منتهی به آن نتایج را در قالب یک ساختار درختی نشان می‌دهد. درخت تصمیم یک الگوریتم کارآمد برای مسائل طبقه بندی و رگرسیون است. ایده اصلی الگوریتم درخت تصمیم این است که یک مسئله پیچیده را به چندین مسئله ساده‌تر تقسیم کند که ممکن است به راه‌حلی منجر شود که تفسیر آن آسان‌تر باشد. یک درخت تصمیم مجموعه‌ای از شرایط را نشان می‌دهد که به صورت سلسله مراتبی سازماندهی شده و به طور متوالی از ریشه تا برگ درخت اعمال می‌شوند. درخت تصمیم یک مدل آموزش‌دیده تولید می‌کند که می‌تواند قوانین منطقی را نشان دهد و سپس می‌تواند برای پیش‌بینی مجموعه داده‌های جدید از طریق فرآیند تکراری تقسیم استفاده شود [۳۳]. محاسبه مقادیر از دست رفته با استفاده از این روش با ساخت درخت‌های تصمیم برای مشاهده انجام می‌شود. مقادیر از دست رفته هر متغیر و سپس مقادیر از دست رفته هر متغیر با استفاده از درخت متناظر آن پر می‌کند؛ سپس پیش‌بینی مقادیر از دست رفته در گره برگ نشان داده می‌شود [۲۷]. پس از اعمال پاکسازی موفق داده‌ها، مجموعه داده‌های نهایی به دست آمده می‌تواند به عنوان یک منبع قابل اعتماد و مناسب برای مدلسازی در نظر گرفته شود [۳۶].

۳- روش‌شناسی

روش پیشنهادی ما برای مدلسازی تغییرات مکانی_زمانی آلاینده $PM_{2.5}$ از سه مرحله تشکیل شده‌است. مرحله نخست، جمع‌آوری داده‌است. مرحله دوم، پیش پردازش و مرحله سوم مدلسازی است. در مرحله نخست جمع آوری داده‌ها شامل داده‌های هواشناسی و داده‌های

حاصل از ایستگاه‌های پایش است. ایستگاه‌های پایش خود از دو طریق سازمان حفاظت محیط زیست و شهرداری جمع‌آوری گردید. مرحله دوم، پیش پردازش است که به منظور افزایش دقت مدلسازی مکانی_زمانی انجام می‌گردد. در این مرحله ابتدا پاکسازی و سپس جانهی مقادیر از دست رفته انجام می‌گردد. پاکسازی و جانهی برای هر دو متغیر هواشناسی و آلاینده‌ها انجام می‌گیرد. در مرحله آخر مدلسازی مکانی_زمانی به منظور پیش بینی ساعتی آلاینده $PM_{2.5}$ با استفاده از روش XGBoost انجام گرفت. شکل ۱ مراحل روش تحقیق را نشان می‌دهد.



شکل ۱. مدلسازی مکانی-زمانی به منظور پیش‌بینی ساعتی $PM_{2.5}$ با استفاده از روش پیشنهادی در منطقه مورد مطالعه

در ادامه هریک از مراحل روش تحقیق که در شکل ۱ آورده شده تشریح می‌گردد.

۳-۱- جمع‌آوری داده‌ها

اولین گام هر پروژه یادگیری ماشین جمع آوری داده‌است. داده‌های اخذ شده شامل آلاینده‌های O_3، So_2، Pm_{10}، Co، $PM_{2.5}$ و متغیرهای هواشناسی است. متغیرهای هواشناسی را می‌توان با استفاده از تعداد محدود و پراکنده ایستگاه‌های هواشناسی و یا با استفاده از مدل‌های معروف آب و هوای جهانی بدست آورد. یکی از مدل‌های معروف آب و هوای جهانی که می‌تواند داده‌های مورد نیاز هواشناسی در پیش بینی غلظت $PM_{2.5}$ را ارائه دهد؛ مدل ECMWF است. در این پژوهش از مدل ECMWF در سامانه Google Earth Engine برای استخراج متغیرهای هواشناسی موردنیاز بجای استفاده از تعداد محدود و پراکنده ایستگاه‌های هواشناسی استفاده گردید. استفاده از این مدل این امکان را می‌دهد که به ازای موقعیت مکانی هریک از ایستگاه‌های پایش، متغیرهای هواشناسی آن‌ها استخراج گردد؛ در نتیجه چهارده ایستگاه هواشناسی بجای چهار ایستگاه هواشناسی موجود در این منطقه در این مطالعه استفاده گردید.

متغیرهای هواشناسی استخراج شده شامل رطوبت، دما، فشار، میانگین سالانه بارندگی، نقطه شبنم است. رطوبت نسبی به طور مستقیم از مدل ECMWF قابل

دستیابی نیست و با استفاده از پارامترهای دما و دمای نقطه شبنم مطابق با فرمول ۱ محاسبه می‌شود:

$$T_a = (112 + 0.9\,T)(RH)^{0.125} + (0.1 - 112) \quad (۱)$$

میانگین سالانه مقادیر پارامتر بارش نسبت به پارامتر مجموع بارش انطباق بیشتری با میانگین سالانه کلانشهر تهران دارد، بنابراین میانگین بارش انتخاب گردید. داده‌های هواشناسی این مدل دارای قدرت تفکیک زمانی ساعتی و مکانی ده کیلومتر است.

آلاینده‌ها نیز با استفاده از سیزده ایستگاه ثابت زمینی از طریق وب سایت کنترل کیفیت هوای تهران و داده‌های غلظت ساعتی آلاینده $PM_{2.5}$ ایستگاه امام خمینی از طریق وب سایت سازمان حفاظت محیط زیست جمع آوری گردید. ایستگاه‌هایی که تمام آلاینده‌ها را داشته باشند و کمتر از ۲۰٪ داده از دست رفته باشند در این مرحله انتخاب شدند. در مرحله بعد، ادغام داده‌های آلاینده و هواشناسی برای هر ایستگاه انجام می‌گردد.

۳-۲- پیش‌پردازش

بعد از اتمام جمع‌آوری داده‌ها، فرآیند پیش پردازش به منظور پاکسازی و جانهی داده‌های از دست رفته انجام می‌شود. در این پژوهش رزولوشن زمانی داده‌های کیفیت هوا ساعتی است و روزهای که بیشتر از هشت ساعت فاقد رکورد داده بودند در مرحله نخست پاکسازی غلظت آلاینده $PM_{2.5}$ حذف شدند. همچنین داده‌های پرت با استفاده از روش IQR[1] بر طبق فرمول (۲) حذف شدند.

$$Q1 - IQR < PM_{2.5} < Q3 + IQR \quad (۲)$$

در این فرمول Q_1 و Q_3 ربع اول و سوم ورودی $PM_{2.5}$ هستند و IQR محدوده بین چارکی مقادیر $PM_{2.5}$ است. مرحله دوم پاکسازی جانهی داده‌های از دست رفته‌است. متغیرهای هواشناسی درنظر گرفته دارای مقادیر از دست رفته در ساعت ۰ تاریخ ۱۴۰۰/۰۶/۱۰ و از ساعت ۴ تا ۱۹ تاریخ ۱۴۰۰/۰۶/۱۲ بود که بدلیل بسیار ناچیز بودن مقادیر از دست رفته از روش درون‌یابی خطی[2] استفاده گردید. معادله درون‌یابی خطی بکاررفته بر طبق فرمول (۳) است.

$$F(x) = f(x_0) + \frac{f(x_1) - f(x_0)}{x_1 - x_0}(x - x_0) \quad (۳)$$

در این معادله X متغیر مستقل است. X_0 و X_1 مقادیر شناخته شده متغیر مستقل و f(x) مقدار متغیر وابسته برای مقدار X از متغیر مستقل است. در این مطالعه متغیر مستقل هواشناسی شامل رطوبت، دما، فشار، بارندگی، مجموع بارش، نقطه شبنم است. بعد از اتمام پاکسازی و جانهی متغیرهای هواشناسی، جانهی مقادیر از دست رفته آلاینده $PM_{2.5}$ آغاز می‌گردد.

از آن‌جایی که آلاینده $PM_{2.5}$ به متغیرهای مختلفی در هر بافت جغرافیایی مرتبط است؛ بنابراین برای تخمین مقادیر از دست رفته $PM_{2.5}$ رابطه آن با هر یک از متغیرهای هواشناسی شامل دما، رطوبت نسبی، فشار سطحی، مجموع بارش و سایر آلاینده‌ها شامل O_3، So_2، Pm_{10} و Co برای هر ایستگاه به صورت جداگانه درنظر گرفته شد. جانهی با استفاده از روش‌های درخت تصمیم و درخت اضافی در بستر Python 3.7.3 با تنظیمات پیش فرض هایپرپارامترها و با استفاده از کتابخانه Sklearn انجام گرفت. در این روش فرآیند جانهی چندین بار تکرار می‌شود تا یک مجموعه داده تکمیل شده ایجاد گردد سپس مجموعه داده تکمیل شده با استفاده از تجزیه و تحلیل آماری برای تولید نتایج مورد تجزیه و تحلیل قرار می‌گیرد و پس از آن، میانگین نتایج گزارش می‌شود. در نظر گرفتن ناهمگونی مکانی-زمانی بسیار مهم است برای تخمین دقیق $PM_{2.5}$ ویژگی‌های زمانی مانند روز و ساعت به مدل نهایی اضافه شد. در زمان‌هایی که این اطلاعات کامل است آموزش و آزمایش را انجام گردید و به اصطلاح یادگیری انجام می‌شود؛ سپس به جانهی مقادیر از دست رفته می‌پردازد.

۳-۳- مدلسازی

پس از اتمام جانهی مقادیر از دست رفته $PM_{2.5}$ به کمک پارامترهای متریک شامل R^2، RMSE، MAE دو روش جانهی درخت اضافی و درخت تصمیم مورد مقایسه قرار گرفتند که نتایج آن در جدول ۲ آورده شده‌است. نتایج نشان داد روش درخت اضافی، روش دقیق‌تری در جانهی مقادیر از دست رفته آلاینده $PM_{2.5}$ است، سپس مدلسازی مکانی_زمانی با استفاده از روش XGBoost انجام گرفت. در صورت نیاز به آگاهی بیشتر از سازوکار روش XGBoost به

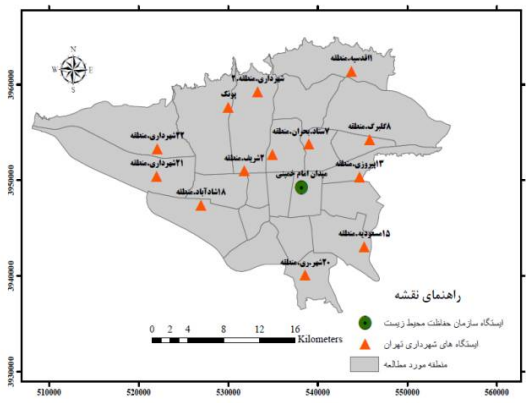
۱ Interquartile Range

۲ Linear Interpolation

[۳۷] مراجعه گردد. در این مطالعه حالات مختلف تقسیم بندی داده‌های آموزش و آزمایشی انجام شد و نتایج نشان داد که نسبت ۸۵ -۱۵٪ نتیجه بهتری نسبت به این تقسیم بندی‌ها در این مطالعه برای این منطقه به همراه دارد. پیاده سازی این الگوریتم نیز در بستر Python 3.7.3 پیاده سازی شد؛ سپس پیش بینی ساعتی $PM_{2.5}$ با استفاده از XGBoost برای چهارده ایستگاه منتخب انجام شد.

۳-۴- منطقه مورد مطالعه

کلانشهر تهران با مختصات جغرافیایی ۳۵ درجه و ۷ دقیقه شمالی و ۵۱ درجه و ۴ دقیقه شرقی است که دارای میانگین ارتفاع ۱۲۰۰ مایل از سطح دریا است. مساحت این شهر بیش از ۷۰۰ کیلومتر مربع است. از لحاظ وضعیت هواشناسی، میانگین دمای سالانه آن ۱۸.۵ درجه سانتی گراد است. از لحاظ بارندگی، میانگین بارندگی این شهر ۱۵۰ میلی متر در سال است که جزو شهرهای کم بارندگی محسوب می‌شود. بادهای غالب از غرب، جنوب و جنوب شرقی می‌وزد که آلودگی کارخانه‌های صنعتی را به مرکز شهر می‌آورد. از لحاظ توپولوژیکی شرایط پایدار و وارونگی دما در این شهر بیشتر در فصل پاییز و زمستان رخ می‌دهد. شکل ۲ موقعیت مکانی ایستگاه پایش آلودگی هوا در منطقه مورد مطالعه را نشان می‌دهد.



شکل ۲. موقعیت مکانی ایستگاه پایش آلودگی هوا در منطقه مورد مطالعه

در این مطالعه ایستگاه‌های اقدسیه، گلبرگ، مسعودیه، پیروزی، پونک، ستاد بحران، شادآباد، شهرداری منطقه۲، شهرداری منطقه ۲۱، شهرداری منطقه ۲۲، شهر ری، شریف، تربیت مدرس که متعلق به شهرداری استان تهران[۱] و ایستگاه امام خمینی متعلق به سازمان حفاظت محیط زیست[۲] منتخب گردید. جدول ۱ مشخصات آماری غلظت آلاینده $PM_{2.5}$ در ایستگاه ثابت زمینی در منطقه مورد مطالعه را آورده‌است.

جدول ۱- مشخصات آماری غلظت آلاینده $PM_{2.5}$ در ایستگاه ثابت زمینی در منطقه مورد مطالعه

ایستگاه‌ها	تعداد داده‌ها		قبل پردازش		بعد پردازش	
	قبل پردازش	بعد پردازش	میانگین	بیشترین	میانگین	بیشترین
امام خمینی	۷۸۴۶	۷۲۹۸	۲۹.۷	۲۵۴.۹	۲۹.۳	۲۵۴.۹
اقدسیه	۸۷۹۴	۸۷۸۳	۲۳.۱	۱۴۰	۲۳.۰	۱۳۹.۶
گلبرگ	۷۹۲۹	۷۸۲۳	۲۲.۹	۱۵۷	۲۲.۷	۱۵۷.۲
مسعودیه	۸۵۲۸	۸۲۷۹	۲۴.۵	۱۴۰	۲۴.۸	۱۴۰.۱
پیروزی	۸۷۶۳	۸۲۵۵	۳۶.۶	۱۹۷	۳۶.۵	۱۹۷.۱
پونک	۸۷۹۴	۷۴۱۵	۱۹.۴	۱۱۰	۱۷.۰	۱۱۰.۳
ستاد بحران	۸۷۸۴	۸۷۸۳	۲۶.۰	۱۲۳	۲۵.۸	۱۲۳.۱
شادآباد	۸۶۰۹	۸۵۴۳	۳۶.۷	۲۶۵	۳۶.۷	۲۵۶.۳
منطقه ۲	۸۷۹۹	۸۸۰۷	۲۵.۸	۱۵۸	۲۵.۹	۱۵۷.۷
منطقه۲۱	۸۸۰۶	۸۸۰۷	۴۲.۱	۳۸۱	۴۱.۹	۳۸۰.۹
منطقه۲۲	۸۷۸۹	۸۷۳۵	۳۰.۳	۱۶۱	۳۰.۲	۱۶۱.۰
شهری	۸۶۹۳	۸۵۱۹	۳۰.۱	۱۳۶	۲۹.۸	۱۳۵.۶
شریف	۸۶۹۱	۸۶۶۳	۴۴.۳	۹۸۰	۴۳.۹	۹۷۹.۹
تربیت مدرس	۸۷۸۹	۸۷۸۳	۳۴.۴	۱۷۸	۳۴.۳	۱۷۸

بازه زمانی داده‌های مورد استفاده در این مطالعه از تاریخ ۱۳۹۹/۰۹/۰۱ تا ۱۴۰۰/۰۹/۰۱ است. دلیل انتخاب این بازه زمانی کمترین میزان داده‌های از دست رفته‌است در چند سال اخیر است. نتایج جدول ۱ نشان می‌دهد که بیشترین داده از دست رفته متعلق به ایستگاه پونک و کمترین تعداد داده از دست رفته متعلق به ایستگاه شهرداری منطقه ۲۱ و ایستگاه ستاد بحران است. بازه داده‌های از دست رفته $PM_{2.5}$ بین ۰.۰۱٪- ۱۵.۶۸٪ است. همچنین نتایج نشان می‌دهد که آلوده‌ترین ایستگاه شریف و پاک ترین ایستگاه پونک است.

۴- بحث و نتایج

در این بخش نتایج مقایسه پارامترهای متریک به منظور مقایسه روش درخت اضافی و درخت تصمیم برای انتخاب بهترین روش جانهی $PM_{2.5}$ بیان می‌گردد. سپس نتایج

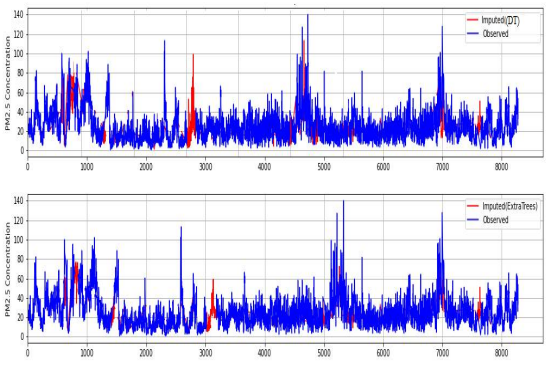
۱ airnow.tehran.ir

۲ aqms.doe.ir

ارزیابی اهمیت متغیرها برای شناسایی و حذف ویژگی‌های کم اهمیت‌تر توسط مقدار F در روش XGBoost آورده می‌شود. در انتها پیش بینی ساعتی آلاینده $PM_{2.5}$ برای تمام ایستگاه‌های منتخب با استفاده از روش XGBoost به نمایش درآمده‌است.

۴-۱- نتایج مقایسه دو روش درخت اضافی و درخت تصمیم در جانهی مقادیر از دست رفته

جانهی مقادیر از دست رفته با هر دو روش یادگیری ماشین برای چهارده ایستگاه منتخب مورد مقایسه قرار گرفت. نتایج مقایسه دو روش درخت تصمیم و درخت اضافی ایستگاه مسعودیه به عنوان یکی از ایستگاه‌های منتخب در شکل ۲ آورده شده‌است.



شکل ۲. مقایسه جانهی دو روش درخت اضافی و تصمیم در ایستگاه مسعودیه

نتایج حاصل از شکل ۲ به لحاظ گرافیکی نشان می‌دهد که روش درخت اضافی نسبت به درخت تصمیم در جانهی مقادیر از دست رفته $PM_{2.5}$ از نظر حفظ توزیع شکل عملکرد بهتری دارد. این برتری نه تنها برای ایستگاه مسعودیه بلکه برای تمامی ایستگاه‌های منتخب هم به لحاظ گرافیکی و هم از مقایسه پارامترهای متریک پرکاربرد مانند R^2، RMSE، MAE به اثبات رسیده‌است. نتایج مقایسه دو روش درخت تصمیم و درخت اضافی با استفاده از پارامترهای متریک برای ایستگاه‌های منتخب در جدول ۲ آورده شده‌است. نتایج حاصل از جدول ۲ نشان داد که روش درخت اضافی با میانگین R^2=۰.۸۰ دقت بالاتری از روش درخت تصمیم با میانگین R^2=۰.۶۴ دارد. بنابراین در مقایسه این دو، مقدار میانگین روش درخت اضافی ۱.۲ برابر دقتش بیشتر از دقت روش درخت تصمیم است.

جدول ۲-مقایسه پارامترهای متریک دو روش درخت اضافی و تصمیم در جانهی مقادیر از دست رفته$PM_{2.5}$

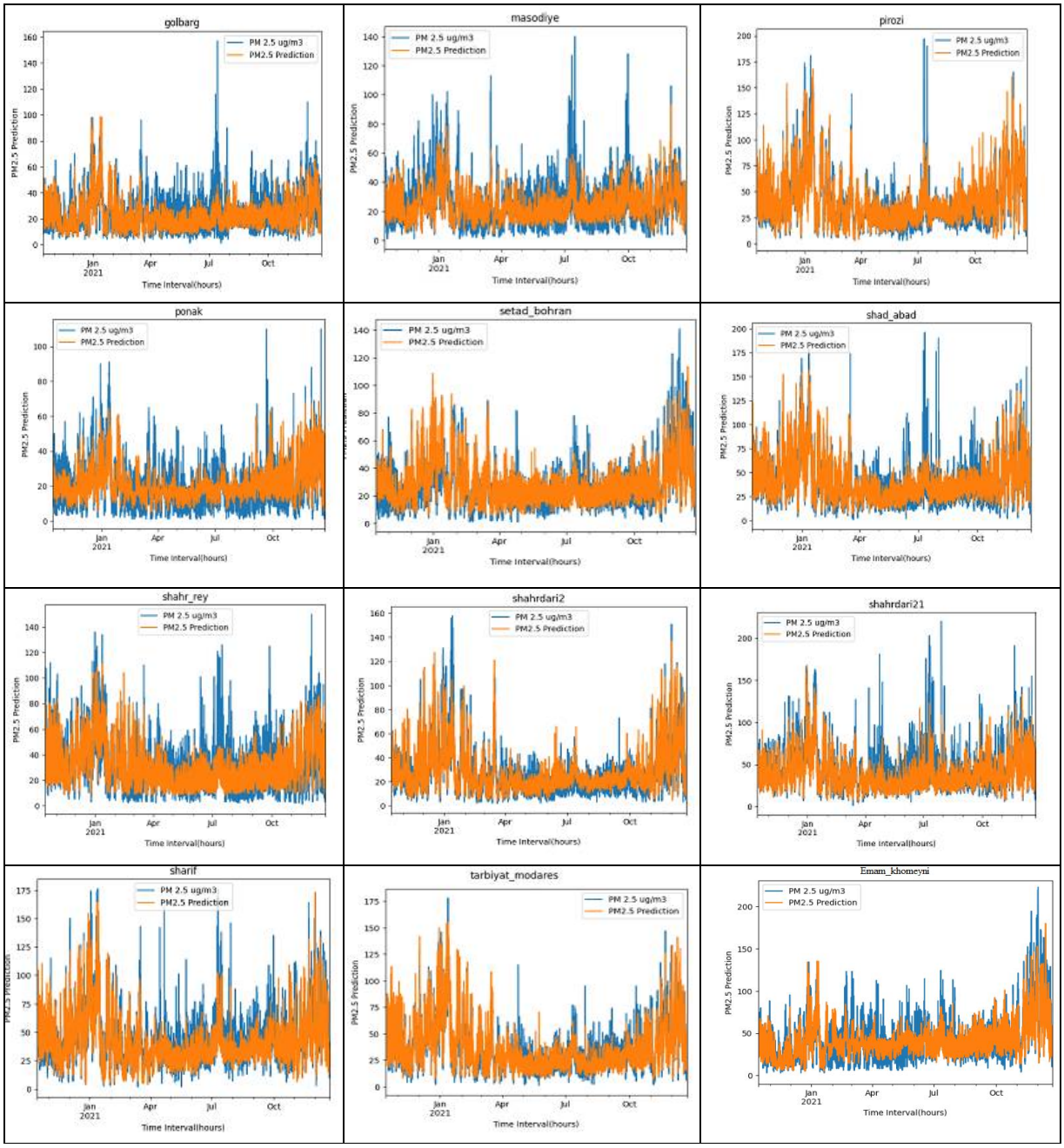
ایستگاه‌ها	درخت اضافی			درخت تصمیم		
	R^2	RMSE	MAE	R^2	RMSE	MAE
امام خمینی	۰.۹۰	۵.۷۶	٤.۲۳	۰.۸۱	۷.۹۲	۵.٦۵
اقدسیه	۰.۸۹	۵.٦۵	۳.٦۳	۰.۸۰	۷.٦٤	٤.۹۹
گلبرگ	۰.٦۰	۸.۵۱	٦.٤۰	۰.۲۵	۱۱.٦۱	۸.٦۱
مسعودیه	۰.۷۲	۰.۷۹	۵.۷۳	۰.٤٦	۱۱.۰۳	۷.٦۰
پیروزی	۰.۹۳	٦.۲۳	٤.۲٦	۰.۸۵	۹.٤۳	٦.۰۱
پونک	۰.٦۱	٦.۸٤	٤.۸۹	۰.۳۲	۹.۰۸	٦.٦۸
ستاد بحران	۰.۷۸	۷.۱۹	۵.٤۲	۰.۵۷	۱۰.۰۹	۷.۵۰
شادآباد	۰.۸۷	۸.۸۰	۵.۸۸	۰.۷٦	۱۱.۸٤	۷.۷۳
منطقه ۲	۰.۸۹	٦.۵۱٦٦	٤.۹۰	۰.۷۵	۹.۹۷	۷.۱۷
منطقه ۲۱	۰.۸۲	۱۰.۱۲	٦.٦٦	۰.٦۸	۱۳.٤۹	۸.۷۵
منطقه۲۲	۰.۸۲	۸.۷۲	٦.۳٤	۰.٦٦	۱۲.۲۱	۸.۵۷
شهرری	۰.۷۹	۸.۸۵	٦.۳۲۱۱	۰.٦۱	۱۲.۲٦۱	۸.۷۹
شریف	۰.۸۰	۹.۸۸	٦.٦٤	۰.۷۱	۱۱.۸۷	۸.۵۰
تربیت مدرس	۰.۹۰	٦.۸۷	٤.۷۸	۰.۸۳	۹.۱۳	٦.۵۹
میانگین کل	۰.۸۰	۷.۱۹	۵.٤۳	۰.٦٤	۱۰.۷٤	۷.۳٦

لذا برآورد نتایج این مقایسه حاکی از برتری روش درخت اضافی در جانهی مقادیر از دست رفته $PM_{2.5}$ در کلیه ایستگاه‌ها منتخب است. مقایسه بین ایستگاه‌ها در هر دو روش نشان می‌دهد که ایستگاه پیروزی بالاترین دقت و ایستگاه گلبرگ پایین‌ترین دقت را در بین ایستگاه‌های منتخب دارد؛ به‌طوریکه ایستگاه پیروزی در روش درخت تصمیم R^2=۰.۸۵۷۱ و در روش درخت اضافی با ۰.۹۳۷ R^2= است. ایستگاه گلبرگ در روش درخت تصمیم با R^2=۰.۲۵۸ و R^2= ۰.۶۰۲ پایین‌ترین مقدار R^2 را در هر دو روش بدست آورد. دلیل بالاترین و پایین‌ترین دقت جانهی ایستگاه پیروزی و گلبرگ به ترتیب داشتن کمترین و بیشترین نقاط از دست رفته $PM_{2.5}$است.

۴-۲- نتایج ارزش F مدلسازی به روش XGBoost

برخی از ویژگی‌ها به افزایش دقت مدل سازی کمک نمی‌کنند و فقط پیچیدگی مدل را افزایش می‌دهند. به همین دلیل، ارزیابی اهمیت ویژگی‌ها برای شناسایی و حذف ویژگی‌های ناکارآمد انجام شده‌است. در روش XGBoost یک عملکرد داخلی وجود دارد که اهمیت ویژگی‌ها را ارزیابی می‌کند. به عبارت دیگر، سهم هر یک از پیش بینی کننده‌ها رادر طول ساخت مدل ارزیابی می‌کند.

در مرحله اول مقدار F معنی داری متغیرهای مستقل را انتخاب می‌کند. پس از آن، متغیرهای مستقل انتخاب می‌شوند سپس این متغیرها به عنوان ورودی ماشین XGBoost استفاده می‌گردند.



شکل ۳. پیش‌بینی ساعتی آلاینده $Pm_{2.5}$ برای تمام ایستگاه‌های منتخب با استفاده از روش XGBoost

این عملکرد ارزش[۱] F نام دارد که در جدول ۳ نتایج این ارزیابی‌ها آورده شده‌است. ارزش F امکان مقایسه واریانس دو مجموعه داده مختلف برای تعیین تفاوت آماری معنی دار را فراهم می‌کند. امتیاز F که امتیاز F1 نیز نامیده می‌شود، اندازه گیری دقت مدل در یک مجموعه داده و برای بسیاری از انواع مدل‌های یادگیری ماشین است. امتیاز F یک معیار پرکاربرد برای عملکرد مدل است. نتایج جدول ۳ اهمیت ویژگی‌ها با استفاده از امتیاز F نشان می‌دهد که نتایج ارزش F برای هر ایستگاه ثابت زمینی متفاوت است.

۱ F Sccore

جدول ۳. ارزیابی اهمیت آلاینده‌های هوا با غلظت $PM_{2.5}$ بر مبانی روش ارزش F

نام ایستگاه‌ها	O_3	Hour	So_2	Pm_{10}	Co	فشارسطحی	دما	رطوبت نسبی	No_2	مجموع بارش	R^2
امام خمینی	۲۷۰	۲۴۳	۳۳۲	۳۴۰	۲۰۸	۲۳۶	۲۱۹	۲۲۷	۲۴۱	۱۰۶	۰.۷۱
اقدسیه	۳۹۲	۳۶۳	۳۴۶	۳۴۵	۳۰۶	۳۰۰	۳۰۰	۲۹۳	۲۸۵	۲۲۸	۰.۸۴
گلبرگ	۳۱۰	۳۴۷	۲۸۴	۳۵۱	۲۷۱	۲۹۱	۲۹۹	۲۵۳	۲۴۲	۱۸۳	۰.۶۰
مسعودیه	۳۳۰	۲۹۴	۲۸۷	۲۸۴	۲۲۱	۲۷۸	۲۳۲	۲۲۴	۱۹۹	۱۳۹	۰.۷۰
پیروزی	۲۳۷	۲۶۲	۲۶۹	۳۱۴	۱۵۴	۲۳۹	۲۱۰	۲۱۷	۲۱۴	۱۳۶	۰.۷۷
پونک	۳۰۰	۲۷۰	۲۱۱	۲۶۴	۲۰۶	۲۳۰	۲۱۲	۲۱۴	۲۲۵	۱۴۶	۰.۷۴
ستاد بحران	۴۲۱	۳۸۹	۳۵۵	۳۶۷	۳۰۸	۳۷۱	۳۲۵	۳۴۵	۳۲۷	۲۱۵	۰.۸۳
شادآباد	۳۴۳	۳۳۸	۳۰۸	۳۸۲	۲۵۷	۳۵۲	۳۰۸	۳۲۴	۲۸۸	۱۷۶	۰.۶۶
منطقه ۲	۳۵۲	۳۲۹	۲۶۴	۳۲۰	۲۲۲	۲۴۴	۲۷۸	۲۳۰	۲۵۱	۱۵۷	۰.۷۶
منطقه ۲۱	۳۱۷	۲۵۸	۲۴۷	۳۳۶	۱۷۱	۲۸۲	۲۶۷	۲۵۶	۲۲۴	۱۴۷	۰.۶۴
منطقه۲۲	۳۱۲	۲۷۹	۲۶۸	۳۵۴	۲۱۲	۲۵۶	۲۳۵	۲۱۴	۲۴۱	۱۶۲	۰.۶۳
شهری	۲۸۰	۲۸۱	۲۸۱	۳۰۴	۲۳۲	۲۲۸	۲۳۴	۲۵۰	۲۱۴	۱۲۹	۰.۶۸
شریف	۳۸۷	۲۹۴	۲۹۰	۳۷۸	۱۱۶	۲۷۱	۲۸۷	۲۷۹	۲۸۴	۱۸۰	۰.۶۷
تربیت مدرس	۴۶۱	۳۰۹	۲۹۶	۳۹۲	۲۳۷	۳۰۵	۳۰۱	۲۸۲	۲۶۲	۲۳۷	۰.۷۳

بیشترین مقدار ارزش F برای تمام ایستگاه O_3 و Pm_{10} است. ارزش اهمیت داده $PM_{2.5}$ با سایر وِیژگی‌ها بدین صورت است که هشت ایستگاه با Pm_{10} و شش ایستگاه با O_3 ارتباط بیشتری نسبت به سایر آلاینده‌ها دارند. همچنین مقدار O_3 بین بازه ۴۲۱-۲۱۴ است. مقدار Pm_{10} نیز بین بازه ۳۹۲-۲۰۷ است. کمترین مقدار ارزش F برای تمام ایستگاه‌ها مربوط به مجموع بارندگی که در بازه ۱۰۶-۲۳۷ قرار دارد. در ادامه پس از کشف ارزش اهمیت ویژگی‌ها پیش بینی ساعتی آلاینده $PM_{2.5}$ برای تمام ایستگاه‌های منتخب با استفاده از روش XGBoost در شکل ۳ نمایش داده شده‌است. ایستگاه اقدسیه که در بافت جغرافیایی شمالی است؛ بالاترین دقت و میانگین $PM_{2.5}$ آن ۲۳.۶ است. قسمت‌های جنوبی کلانشهر تهران شامل ایستگاه شهرری، شادآباد نسبت به سایر مناطق آلوده‌تر است و میزان ذرات معلق $PM_{2.5}$ در آن‌ها بالاتر است. دلیل این افزونگی وجود چندین نیروگاه برق، پالایشگاه نفت در جنوب شهر است.

۵- نتیجه‌گیری

در این مطالعه به منظور افزایش دقت مکانی-زمانی مدلسازی غلظت ذرات معلق $PM_{2.5}$ از روش‌های یادگیری ماشین برای جانهی داده‌های از دست رفته با درنظرگرفتن روابط بین متغیرهای هواشناسی و سایر آلاینده‌ها استفاده گردید. پس از جانهی مقادیر از دست رفته از روش XGBoost به دلیل توانایی در نظر گرفتن ارزش اهمیت ویژگی‌های مستقل با ویژگی‌های وابسته و حذف ویژگی‌های کمتر وابسته که باعث کاهش هزینه‌های محاسباتی می‌شود برای پیش بینی ساعتی آلاینده $PM_{2.5}$ در منطقه شهری تهران استفاده کردیم. در این پژوهش از دو روش درخت تصمیم و درخت اضافی برای جانهی داده‌های از دست رفته PM_2 استفاده گردید. پس از مقایسه این دو روش با استفاده از پارامترهای متریک پرکاربرد مانند R^2، RMSE، MAE نتایج نشان داد که گرچه روش درخت اضافی کمتر در مطالعات برای جانهی مقادیر از دست رفته مورد استفاده قرار گرفته‌است اما دقت بسیار بالاتری از روش درخت تصمیم در منطقه مورد مطالعه دارد.

در این مطالعه داده‌های $PM_{2.5}$ از داده‌های ایستگاه-های ثابت پایش شهرداری تهران و سازمان حفاظت محیط زیست جمع آوری گردید. همچنین از داده‌های هواشناسی حاصل از مدل ECMWF به منظور استخراج داده‌های دما، رطوبت نسبی، مجموع بارندگی، بارندگی، دما نقطه شبنم و فشار برای ایستگاه‌های منتخب استفاده گردید. استفاده از مدل ECMWF این امکان را فراهم کرد که اطلاعات هواشناسی را به صورت ساعتی برای ایستگاه‌های ثابت زمینی منتخب استخراج کرد. درحالی-که تنها چهار ایستگاه هواشناسی در منطقه شهری تهران وجود دارد. نتایج این مطالعه نشان داد که هرچه تعداد

داده‌های از دست رفته کاهش یابد و تعداد ایستگاه‌های ثابت زمینی بیشتر باشد امکان افزایش دقت با روش‌های یادگیری ماشین مانند XGBoost وجود دارد. نتایج امتیاز F حاصل از روش XGBoost نشان می‌دهد که نتایج ارزش F برای هر ایستگاه ثابت زمینی برای همبستگی با داده $PM_{2.5}$ متفاوت است به ترتیب بیشترین مقدار ارزش F برای داده $PM_{2.5}$ در تمام ایستگاه‌های منتخب مربوط به آلاینده O_3 و PM_{10} و سپس داده مربوط به زمان (ساعت، روز) در مرحله بعدی با آلاینده So_2 دارد. کمترین مقدار ارزش F برای تمام ایستگاه‌ها مربوط به مجموع بارندگی است. در یک نتیجه گیری کلی در این مطالعه می‌توان بیان کرد که بیشترین همبستگی بین آلاینده $PM_{2.5}$ با Pm_{10} و کمترین همبستگی بین آلاینده‌های منتخب با No_2 است.

مراجع

[۱] J. Tan, H. Liu, Y. Li, S. Yin, and C. Yu, "A new ensemble spatio-temporal PM2. 5 prediction method based on graph attention recursive networks and reinforcement learning," *Chaos, Solitons & Fractals,* vol. 162, p. 112405, 2022.

[۲] S. Srivastava and I .N. Sinha, "Classification of air pollution dispersion models: a critical review," in *Proceedings of National Seminar on Environmental Engineering with special emphasis on Mining Environment*, 2004.

[۳] X. Xi, Z. Wei, R. Xiaoguang, W. Yijie, B. Xinxin, Y .Wenjun*, et al.*, "A comprehensive evaluation of air pollution prediction improvement by a machine learning method," in *2015 IEEE international conference on service operations and logistics, and informatics (SOLI)*, 2015, pp. 176-181.

[٤] W. Tong, "Machine learning for spatiotemporal big data in air pollution," in *Spatiotemporal Analysis of Air Pollution and Its Application in Public Health*, ed: Elsevier, 2020, pp. 107-134.

[٥] H. Amini, S. M. Taghavi-Shahri, S. B. Henderson, K. Naddafi, R. Nabizadeh, and M .Yunesian, "Land use regression models to estimate the annual and seasonal spatial variability of sulfur dioxide and particulate matter in Tehran, Iran," *Science of the total environment,* vol. 488, pp. 343-353, 2014.

[٦] H. Bagheri, "A machine learning-based framework for high resolution mapping of PM2. 5 in Tehran, Iran, using MAIAC AOD data," *Advances in Space Research,* vol. 69, pp. 3333-3349, 2022.

[۷] M. Zamani Joharestani, C. Cao, X. Ni, B. Bashir, and S. Talebiesfandarani, "PM2. 5 prediction based on random forest, XGBoost, and deep learning using multisource remote sensing data," *Atmosphere,* vol. 10, p. 373, 2019.

[۸] J. Ma, Z. Yu, Y. Qu, J. Xu, and Y. Cao, "Application of the XGBoost machine learning method in PM2. 5 prediction: A case study of Shanghai," *Aerosol and Air Quality Research,* vol. 20, pp. 128-138, 2020.

[۹] S. Gündoğdu, G. Tuna Tuygun, Z. Li, J. Wei, and T. Elbir, "Estimating daily PM2. 5 concentrations using an extreme gradient boosting model based on VIIRS aerosol products over southeastern Europe," *Air Quality, Atmosphere & Health,* vol. 15, pp. 2185-2198, 2022.

[۱۰] Y.-P. Chen, C.-H. Huang, Y.-H. Lo, Y.-Y. Chen, and F. Lai, "Combining attention with spectrum to handle missing values on time series data without imputation," *Information Sciences,* vol. 609, pp. 1271-1287, 2022.

[۱۱] K. J. Nishanth, V. Ravi, N. Ankaiah, and I. Bose, "Soft computing based imputation and hybrid data and text mining: The case of predicting the severity of phishing alerts," *Expert Systems with Applications,* vol. 39, pp. 10583-10589, 2012.

[۱۲] J. Poulos and R. Valle, "Missing data imputation for supervised learning," *Applied Artificial Intelligence,* vol. 32, pp. 186-196, 2018.

[۱۳] W.-C. Lin and C.-F. Tsai, "Missing value imputation: a review and analysis of the literature (2006–2017)," *Artificial Intelligence Review,* vol. 53, pp. 1487-1509, 2020.

[۱٤] H. Kang, "The prevention and handling of the missing data," *Korean journal of anesthesiology,* vol. 64, pp. 402-406, 2013.

[۱٥] S. E. Awan, M. Bennamoun, F .Sohel, F. Sanfilippo, and G. Dwivedi, "A reinforcement learning-based approach for imputing missing data," *Neural Computing and Applications,* vol. 34, pp. 9701-9716, 2022.

[۱٦] I. Belachsen and D. M. Broday, "Imputation of Missing PM2. 5 Observations in a Network of Air Quality Monitoring Stations by a New k NN Method," *Atmosphere,* vol. 13, p. 1934, 2022.

[۱۷] H. Yang, W. Chen, and Z. Liang, "Impact of land use on PM2. 5 pollution in a representative city of middle China," *International Journal of Environmental Research and Public Health,* vol. 14, p. 462, 2017.

[۱۸] S. Z. Shogrkhodaei, S. V. Razavi-Termeh, and A. Fathnia, "Spatio-temporal modeling of PM2. 5 risk mapping using three machine learning algorithms," *Environmental Pollution,* vol. 289, p. 11785.۲۰۲۱ ,۹

[۱۹] X. Xu, "Forecasting air pollution PM2. 5 in Beijing using weather data and multiple kernel learning," *Journal of Forecasting,* vol. 39, pp. 117-125, 2020.

[۲۰] F. Hosseinibalam and A. Hejazi, "Influence of meteorological parameters on air pollution in Isfahan," *IPCBEE,* vol. 46, pp. 7-12, 2012.

[۲۱] Y. Lin, X. Yuan, T. Zhai, and J. Wang, "Effects of land-use patterns on PM2. 5 in China's developed coastal region: Exploration and solutions," *Science of the Total Environment,* vol. 703, p. 135602.۲۰۲۰ ,

[۲۲] Y. Liu, G. Cao, and N. Zhao, "Integrate machine learning and geostatistics for high-resolution mapping of ground-level PM2. 5 concentrations," in *Spatiotemporal Analysis of Air Pollution and Its Application in Public Health*, ed: Elsevier, 202 ,۰pp. 135-151.

[۲۳] M. Faraji, S. Nadi, O. Ghaffarpasand, S. Homayoni, and K. Downey, "An integrated 3D CNN-GRU deep learning method for short-term prediction of PM2. 5 concentration in urban environment," *Science of The Total Environment,* vol. 834, p. 1.۲۰۲۲ ,۵۵۳۲٤

[۲٤] R. J. Chase, D. R. Harrison, A. Burke, G. M. Lackmann, and A. McGovern, "A Machine Learning Tutorial for Operational Meteorology. Part I: Traditional Machine Learning," *Weather and Forecasting,* vol. 37, pp. 1509-1529, 2022.

[۲۵] H. Karimian, Q. Li, C. Wu, Y. Qi, Y. Mo, G. Chen*, et al.*, "Evaluation of different machine learning approaches to forecasting PM2. 5 mass concentrations," *Aerosol and Air Quality Research,* vol. 19, pp. 1400-1410, 2019.

[۲٦] S. Fielding, P. M. Fayers, A. McDonald ,G. McPherson, and M. K. Campbell, "Simple imputation methods were inadequate for missing not at random (MNAR) quality of life data," *Health and Quality of Life Outcomes,* vol. 6, pp. 1-9, 2008.

[۲۷] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, and O. Tabona, "A survey on missing data in machine learning," *Journal of Big Data,* vol. 8, pp. 1-37, 2021.

[۲۸] G. Huang, "Missing data filling method based on linear interpolation and lightgbm," in *Journal of Physics: Conference Series*, 2021, p. 012.۱۸۷

[۲۹] P. D. Allison, "Multiple imputation for missing data: A cautionary tale," *Sociological methods & research,* vol. 28, pp. 301-309, 2000.

[۳۰] D. B. Rubin, *Multiple imputation for nonresponse in surveys* vol. 81: John Wiley & Sons, 2004.

[۳۱] S. Fielding, P. M. Fayers, A. McDonald, G. McPherson, M. K. Campbell, and R. S. Group, "Simple imputation methods were inadequate for missing not at random (MNAR) quality of life data," *Health and Quality of Life Outcomes,* vol. 6, pp. 1-9, 2008.

[۳۲] J. Ma, Z .Shou, A. Zareian, H. Mansour, A. Vetro, and S.-F. Chang, "CDSA: cross-dimensional self-attention for multivariate, geo-tagged time series imputation," *arXiv preprint arXiv:1905.09904,* 2019.

[۳۳] M. W. Ahmad, J. Reynolds, and Y. Rezgui, "Predictive modelling for solar thermal energy systems: A comparison of support vector regression, random forest, extra trees and regression trees," *Journal of cleaner production,* vol. 203, pp. 810-821, 2018.

[۳٤] O. Maier, M. Wilms, J. von der Gablentz, U. M. Krämer, T. F .Münte, and H. Handels, "Extra tree forests for sub-acute ischemic stroke lesion segmentation in MR sequences," *Journal of neuroscience methods,* vol. 240, pp. 89-100, 2015.

[۳۵] E. E. Okoro, T. Obomanu, S. E. Sanni, D. I. Olatunji, and P. Igbinedion, "Application of artificial intelligence in predicting the dynamics of bottom hole pressure for under-balanced drilling: extra tree compared with feed forward neural network model," *Petroleum,* vol. 8, pp. 227-236, 2022.

[۳٦] S. García, S. Ramírez-Gallego, J. Luengo, J. M. Benítez, and F. Herrera, "Big data preprocessing: methods and prospects," *Big Data Analytics,* vol. 1, pp. 1-22, 2016.

[۳۷] H. Dai, G. Huang, H. Zeng, and F. Yang, "PM2. 5 Concentration prediction based on spatiotemporal feature selection using XGBoost-MSCNN-GA-LSTM," *Sustainability,* vol. 13, p. 12071, 2021.