نشریه هواشناسی کشاورزی
(وابسته به انجمن مهندسی آبیاری و آب ایران)
جلد ۱۳، شماره ۲، پاییز و زمستان ۱۴۰۴، صص.۱۴-۳
مقاله علمی پژوهشی

Journal of Agricultural Meteorology
(Iranian Soc. of Water and Irrig. Eng.)
Vol. 13, No. 2, Autumn & Winter 2025, pp. 3-14
DOI: 10.22125/agmj.2025.542575.1188

ارزیابی عملکرد مدل‌های مبتنی بر یادگیری ماشین در برآورد غلظت روزانه منوکسیدکربن زمستانه شهر کرمان

سودابه گلستانی کرمانی[1*]، سید پویا حسینی[2]

تاریخ دریافت: ۱۴۰۳/۱۱/۳۰
تاریخ پذیرش: ۱۴۰۴/۰۶/۳۱

چکیده

مدل‌های مبتنی بر یادگیری ماشین، ابزاری کارآمد جهت تحلیل روابط پیچیده بین پارامترهای کیفی هوا و متغیرهای هواشناسی و تعیین عوامل کلیدی مؤثر بر کیفیت هوا هستند. در تحقیق حاضر با استفاده از اطلاعات روزانه پنج متغیر هواشناسی شامل دمای هوا، بارش، رطوبت نسبی، سرعت باد و فشار هوا و غلظت دو آلاینده هوا (O_3, $PM_{2.5}$) عملکرد ۳ مدل درخت –پایه RF، XGBoost، CatBoost و مدل پرسپترون چند لایه (MLP) ساخته شده بر مبنای شبکه عصبی جهت برآورد غلظت روزانه مونوکسیدکربن فصل زمستان در شهر کرمان ارزیابی شده است. مقایسه مدل‌ها با استفاده از شاخص‌های آماری R^2، RMSE و MAE انجام شد و نتایج بدست آمده نشان داد که مدل CatBoost با R^2 برابر با ۰/۷۷۸، RMSE برابر با (ppb) ۰/۲۸۴ و MAE برابر با (ppb) ۲۰۹۰ در مرحله تست، بالاترین دقت را در تخمین غلظت منوکسیدکربن دارد. نتایج مدل‌های XGBoost و RF تقریباً یکسان بود و R^2 دو مدل در مرحله تست به ترتیب به ۰/۷۴۷ و ۰/۷۲۸ رسید و مدل MLP با R^2 برابر با ۰/۶۹۳، RMSE برابر با (ppb) ۰/۳۰۸ و MAE برابر با (ppb) ۰/۲۳۶کمترین دقت را داشت. این نتایج توانایی مدل‌های درخت-پایه را در مقایسه با مدل ساخته شده بر مبنای شبکه عصبی در برآورد غلظت آلاینده منوکسیدکربن تأیید می‌کند.

واژه‌های کلیدی: آلودگی هوا، کرمان، مونوکسیدکربن، مدل یادگیری ماشین

مقدمه

امروزه یکی از نگرانی‌های عمده جامعه بشری، حفاظت از محیط زیست و رعایت معیارهای زیست محیطی به منظور تداوم حیات روی کره زمین است (Najafpoor et al., 2014; Alizadeh Dakhel et al., 2009). در این میان آلودگی هوا به دلیل تأثیر بر سلامت انسان، اکولوژی و محیط زیست مورد توجه ویژه قرار گرفته و در کلان شهرها به یک چالش اساسی تبدیل شده است. آلودگی هوا در حقیقت وجود یک یا چند آلاینده از قبیل گرد و غبار، گازها، بو، دود و... در هوا است که تحت تأثیر فعالیت انسانی یا طبیعی تولید شده و به واسطه کمیت، ویژگی و زمان ماندی که دارند برای زندگی انسان و سایر موجودات زنده و حتی آثار و ابنیه خطرناک و مضر هستند (Wark, 1998). شناخته شده‌ترین آلاینده‌های هوا شامل ذرات معلق (PM)، اکسیدهای نیتروژن (NO_X)، ازن (O_3)، مونوکسیدکربن (CO) و دی اکسید گوگرد (SO_2) هستند که تحت تأثیر عواملی مانند پیشرفت صنایع و فناوری، توسعه شهری، افزایش و تراکم جمعیت، افزایش وسایل تقلیه موتوری و احتراق ناقص سوخت، مصرف سوخت‌های فسیلی و در برخی موارد شرایط خاص اقلیمی و توپوگرافی شهرها تولید و تجمع پیدا می‌کنند و باعث کاهش کیفیت هوا و افزایش

[1] استادیار و عضو هیات علمی، گروه علوم و مهندسی آب، دانشکده کشاورزی، دانشگاه شهید باهنر کرمان، کرمان، ایران
(*نویسنده مسئول: s.golestani@uk.ac.ir)

[2] دانش آموخته مقطع کارشناسی ارشد، گروه علوم و مهندسی آب، دانشکده کشاورزی، دانشگاه شهید باهنر کرمان، کرمان، ایران

نحوه ارجاع مقاله:
گلستانی کرمانی، س.، حسینی، س. پ. ۱۴۰۴. ارزیابی عملکرد مدل‌های مبتنی بر یادگیری ماشین در برآورد غلظت روزانه منوکسیدکربن زمستانه شهر کرمان. نشریه هواشناسی کشاورزی، ۱۳(۲)، ۳-۱۴. DOI: 10.22125/agmj.2025.542575.1188

Golestani Kermani, S., Hosseini, S. P. 2025. Evaluation of machine learning-based models for estimating winter season carbon monoxide concentration in Kerman. Journal of Agricultural Meteorology, 13(2), 3-14. DOI: 10.22125/agmj.2025.542575.1188

DOI: 10.22125/agmj.2025.542575.1188

بیماری و مرگ و میر در مناطق مختلف می‌شوند (Khorasani et al., 2002; Wright, 2002). به طوری‌که بر اساس گزارش WHO آلودگی هوا سالانه موجب مرگ بیش از ۷ میلیون نفر در جهان می‌شود که نزدیک به ۶۰۰ هزار نفر از آنها کودک هستند و پیش‌بینی می‌شود که مجموع این رقم تا سال ۲۰۳۰ به حدود ۹ میلیون نفر برسد (Tipton, 2022). از این رو برنامه‌ریزی و ارائه راهکار جهت کاهش اثر آلودگی هوا جز الزامات زندگی شهری بوده و اولین قدم در این مسیر، اطلاع از وضع کیفیت هوا است. در سال‌های اخیر، پژوهش‌های متعددی در زمینه پیش‌بینی آلودگی هوا با استفاده از مدل‌های عددی و هوشمند انجام شده و مقایسه نتایج به‌دست آمده نشان می‌دهد که مدل‌های هوشمند علیرغم عدم نیاز به داشتن اطلاعاتی مانند ضرایب انتشار آلاینده که اغلب دسترسی به آن با مشکلاتی همراه است، دارای ساختار ساده‌تر و انعطاف‌پذیری بیشتر در شرایط کمبود داده هستند و توانایی پیش‌بینی پدیده‌های غیر خطی و پیچیده را با دقت قابل قبول دارند (Noori et al., 2013; Moazami et al., 2017). الگوریتم‌های یادگیری ماشین از جمله جنگل تصادفی (RF[۱])، ماشین بردار پشتیبان (SVM)[۲] و غیره که در واقع زیرمجموعه مدل‌های هوشمند محسوب می‌شوند، توانایی یادگیری روابط بین یک مجموعه داده را دارند و از تشخیص الگو برای توصیف روابط بین متغیرهای مستقل و وابسته استفاده می‌کنند و به دلیل آنکه محدود به مفروضات سنتی در مورد ویژگی‌های داده نیستند، قدرت بیشتری برای حل روابط پیچیده و چند وجهی دارند (Hojjati et al., 2022). از این رو کاربرد آنها در مطالعات آلودگی هوا مورد توجه محققین قرار گرفته است از جمله Masoudi and Gerami, (2017) دقت قابل قبول شبکه عصبی مصنوعی پرسپترون سه لایه در پیش‌بینی غلظت CO هوای شهر اصفهان را تأیید کردند و این مطلب را ناشی از توانایی مدل در لحاظ کردن روابط غیرخطی بین آلاینده‌ها دانستند. Akbarzadeh et al., (2020) دقت بالای مدل‌های ساخته شده بر مبنای SVM را در تخمین میزان CO شهر تهران در مقایسه با مدل‌های ANN و ANFIS تأیید کردند. Farhadi et al., (2020) از شبکه عصبی پرسپترون سه لایه برای پیش‌بینی غلظت CO در هوای شهر تهران استفاده کردند و بیشترین مقدار R^2 را در فصل سرد سال و به مقدار ۰/۷۶ گزارش کردند. Lei et al., (2022) مدل RF را به عنوان مدل برتر برای تخمین غلظت $PM_{2.5}$ و PM_{10} در ماکائو معرفی کردند. Juarez and Petersen, (2022) دقت مدل XGBoost[۳] در پیش‌بینی ازن هوای شهر دهلی را در مقایسه با سایر مدل‌ها تأیید کردند. Mohammadi et al., (2024) دقت مدل‌های مختلف را در پیش‌بینی غلظت $PM_{2.5}$ در هوای شهر اصفهان مورد بررسی قرار دادند و به ترتیب مدل‌های ANN[۴]، RF، SVM و KNN[۵] را به عنوان دقیق‌ترین مدل‌ها معرفی کردند. Rahman et al., (2024) نیز توانایی برخی از مدل‌های یادگیری ماشین را در پیش‌بینی غلظت آلاینده‌های هوا مورد بررسی قرار دادند و در نهایت دقت بیشتر مدل جنگل تصادفی و درخت تصمیم را تأیید کردند. اگرچه تخمین کیفیت هوا به دلیل تعاملات پیچیده بین عوامل مختلف چالش برانگیز است، اما جمع‌بندی مطالعات انجام شده نشان می‌دهد که الگوریتم‌های یادگیری ماشین در مدیریت و پردازش مجموعه داده‌های گسترده، دقت قابل قبولی داشته و امکان درک عمیق از روابط پیچیده بین آلاینده‌ها و پارامترهای هواشناسی را فراهم می‌کنند که نتایج آن می‌تواند در اتخاذ تصمیمات مدیریتی جهت کاهش آلودگی هوا موثر باشد. مونوکسیدکربن به عنوان یکی از مهم‌ترین آلاینده‌های هوا که عمدتاً از احتراق ناقص سوخت های حاوی کربن مانند بنزین تولید می‌شود، با اتصال به هموگلوبین موجود در گلبول‌های قرمز خون، توانایی حمل اکسیژن را کاهش می‌دهد که منجر به برزو علائمی مانند سردرد و سرگیجه شده و در غلظت‌های بالا، می‌تواند منجر به مسمومیت و خفگی شود. این آلاینده اغلب به عنوان یک شاخص مهم برای بررسی وضعیت احتراق مورد مطالعه قرار می‌گیرد و غلظت آن اغلب در فصل سرد افزایش می‌یابد (Chadalavada et al., 2025). لذا در پژوهش حاضر به ارزیابی دقت برخی از مدل‌های یادگیری ماشین جهت تخمین غلظت

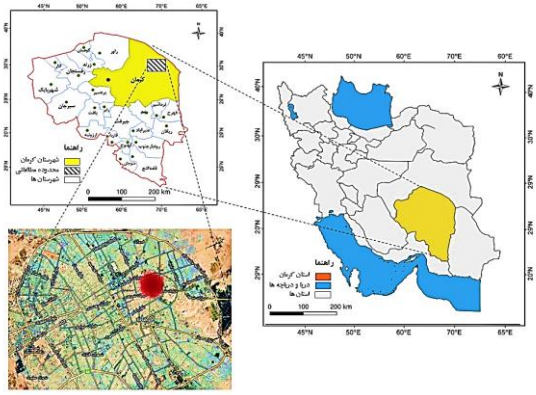
[۱] Random Forest
[۲] Support Vector Machine
[۳] Extreme Gradient Boosting
[۴] Artificial Neural Networks
[۵] K-Nearest Neighbors

CO هوای شهر کرمان در فصل زمستان پرداخته شده که در تحقیقات پیشین، کمتر مورد توجه قرار گرفته است.

مواد و روش‌ها

محدوده مورد مطالعه

استان کرمان به عنوان پهناورترین استان کشور در گستره‌ای به مساحت ۱۸۳۲۸۵ کیلومتر مربع (تقریب به ۱۱ درصد مساحت کشور) در جنوب شرق کشور و در امتداد فلات مرکزی ایران قرار گرفته و شهر کرمان با مختصات عرض جغرافیایی ۳۰/۲۸ درجه شمالی و طول جغرافیایی ۵۷/۰۸ درجه شرقی و ارتفاع ۱۷۵۵ متر از سطح دریا، مساحتی معادل ۲۴۰ کیلومتر مربع از این استان را شامل می‌شود که خلاصه برخی از اطلاعات هواشناسی آن در جدول ۱ و موقعیت جغرافیایی محدوده مطالعاتی مورد نظر در شکل ۱ ارائه شده است. در سال‌های اخیر بروز عواملی مانند رشد جمعیت و توسعه زیر ساخت‌های شهری، افزایش وسائل نقلیه و مصرف بالای سوخت در کنار اقلیم خشک منطقه و کاهش ریزش‌های جوی، شرایط مساعدی جهت کاهش کیفیت هوا و افزایش غلظت آلاینده‌هایی مانند مونوکسیدکربن به خصوص در زمستان‌های کم‌بارش فراهم نموده است.



شکل۱- موقعیت جغرافیایی محدوده مورد مطالعه

Figure 1- Location of the study area

جدول ۱- میانگین اطلاعات هواشناسی شهر کرمان مربوط به دوره آماری ۳۰ ساله (دریافت شده از سازمان هواشناسی شهرستان کرمان ، ۱۴۰۴)

Table 1 – Average meteorological data of Kerman city for the 30-year statistical period (obtained from the Kerman county meteorological organization, 2025)

Sunshine duration ($hr.y^{-1}$)	Relative humidity (%)	Precipitation ($mm.y^{-1}$)	Wind direction	Wind speed ($m.s^{-1}$)	Temperature (◦C)
3309	33.2	116.5	Northern	3.2	16.7

داده‌های مورداستفاده

در پژوهش حاضر برای مدل‌سازی غلظت CO از اطلاعات روزانه متغیرهای هواشناسی و همچنین غلظت آلاینده‌هایی مانند ($PM_{2.5}$ ($\mu g.m^{-3}$ و O3(ppb) در فصل زمستان سال های (۱۴۰۳-۱۳۹۶) شهر کرمان استفاده شد. غلظت آلاینده‌ها در بازه زمانی مورد نظر توسط سامانه پایش کیفی هوا متعلق به سازمان محیط زیست کشور که در میدان شهدا شهر کرمان نصب شده، اندازه‌گیری گردید و اطلاعات هواشناسی روزانه نیز از سایت سازمان هواشناسی دریافت شد. ابتدا پایش اولیه اطلاعات انجام شد و پارامترهایی که دارای داده گمشده بودند، حذف گردید و در مجموع از اطلاعات ۵۴۳ روز برای مدل‌سازی استفاده شد. سپس جهت بررسی همبستگی بین متغیرها و انتخاب پارامترهای تأثیر گذار بر CO از محاسبه ضریب همبستگی R^2 و جهت تعیین اهمیت متغیرها از مدل جنگل تصادفی استفاده شد و در نهایت ۷ متغیر متوسط سرعت باد (WS)، متوسط دما (T)، متوسط فشارهوا (P)، متوسط رطوبت نسبی هوا (RH)، تابش خورشیدی (R)، ازن (O_3)، ذرات معلق ($PM_{2.5}$) به عنوان مهم‌ترین پارامترها انتخاب شدند که خلاصه ویژگی های آماری آنها در جدول ۲ و میزان اهمیت آنها در شکل ۲ ارائه شده است. به علت تعدد متغیرهای ورودی و واحدهای اندازه‌گیری و جهت تسهیل در مقایسه داده‌ها، نرمال کردن داده‌ها قبل از آموزش مدل با استفاده از معادله ۱ انجام شد.

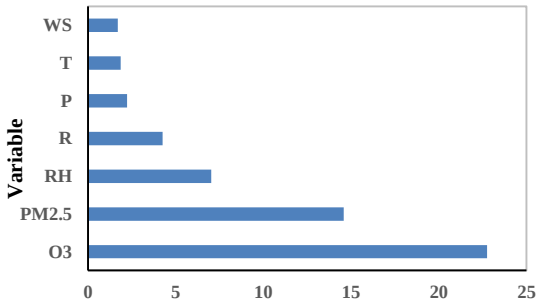
$$X_n = \frac{X_i - X_{min}}{X_{max} - X_{min}} \quad (۱)$$

که در آن X_n داده استاندارد شده، X_i داده مشاهداتی و X_{max} و X_{min} به ترتیب بیشترین و کمترین داده‌ها می‌باشد.

جدول ۲- مشخصات آماری اطلاعات مورد استفاده

Table 2- Statistical characteristics of the used data

parameters	WS ($m.se^{-1}$)	T (◦C)	P (mbar)	RH (%)	R ($kj.m^{-2}$)	O_3 (ppb)	$PM_{2.5}$ ($\mu g.m^{-3}$)
Max	10	21.3	1033.46	96.37	2992	49.12	431.45
Min	0.5	-3.1	997.22	8.87	50	1.44	0
Mean	3.1	8.83	1016.4	38.9	1581.27	23.9	23.58
Median	2.62	8.8	1015.82	34.75	1582.5	19.61	19.71
Standard deviation	1.54	4.46	5.99	18.62	526.93	12.03	23.06
Skewness	1.4	-0.02	0.36	0.91	-0.37	0.55	12.05



شکل۲- میزان اهمیت متغیرهای انتخاب شده

Figure 2- Importance levels of the selected variables

معرفی مدل‌ها

در تحقیق حاضر از ۴ مدل یادگیری ماشین برای شبیه سازی غلظت CO استفاده گردید که در ادامه به شرح مختصر ویژگی‌های هر یک پرداخته شده است.

جنگل تصادفی (RF)

جنگل تصادفی یک نوع از مدل‌های درختی (درخت-پایه) است که شامل انبوهی از روش‌های کلاس‌بندی و رگرسیونی می‌باشد. در این مدل یک مجموعه از درختان تصمیم ایجاد می‌شود که هر کدام روی یک زیرمجموعه تصادفی از داده‌های آموزشی، ساخته و آموزش داده می‌شود (Zhou et al., 2023). در نتیجه درختان از کل مجموعه داده استفاده نمی‌کنند و در هر گره، ویژگی بهینه از مجموعه‌ای از ویژگی‌های موجود انتخاب می‌شود. مدل جنگل تصادفی از دو فراپارامتر شامل تعداد درختان تصمیم برای رشد و تعداد ویژگی‌هایی که به طور تصادفی در هر تقسیم نمونه برداری شده‌اند، استفاده می‌کند و پیش‌بینی نهایی بعد از آموزش همه درختان و با میانگین‌گیری از نتایج بدست می آید (Dong et al., 2020; Azizi Mobaser et al., 2025) . اتخاذ این رویکرد در مدل جنگل تصادفی باعث شده که الگوهای پیچیده داده به طور مؤثر مدیریت شود و در عین حال تعمیم‌پذیری مدل حفظ شود که در نتیجه آن، جنگل تصادفی به گزینه‌ای محبوب در یادگیری ماشین تبدیل شده است (Rahman et al., 2024). جنگل تصادفی در برابر بیش‌برازش مقاوم است، داده‌های با ابعاد بالا را به خوبی مدیریت می‌کند و با ویژگی‌های دسته‌ای و عددی به خوبی کار می‌کند (Chadalavada et al., 2025). در این مدل، با وجود اینکه پیش‌بینی تک‌تک درختان به نویز در مجموعه آموزشی حساس است، استفاده از چندین درخت به شرط آنکه همبستگی بین درختان وجود نداشته باشد باعث کاهش حساسیت می‌شود. هنگامی که مدل آموزش داده شد، پیش‌بینی با استفاده از معادله ۲ محاسبه می شود.

$$\hat{Y}_i = \frac{1}{M}\sum_{m=1}^{M} T_m(f_t) \qquad (۲)$$

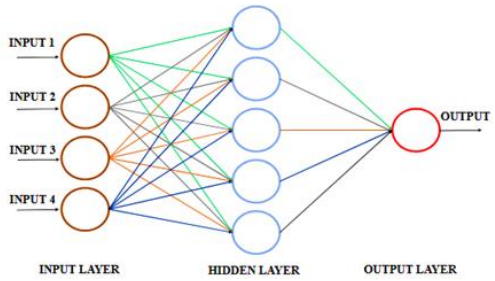
که در آن M تعداد درخت‌ها، T_m درخت تصمیم منفرد و f_t بردار پیش‌بینی کننده است. (Ruiz-AÍvarez et al., 2021).

شبکه عصبی پرسپترون چند لایه (MLP[1])

شبکه عصبی پرسپترون چند لایه یکی از محبوب‌ترین مدل‌های یادگیری ماشین است که بر اساس شباهت با ساختار مغز انسان ساخته شده است (Zounemat-Kermani et al., 2020). این مدل در واقع یک شبکه عصبی پیشرو با پس انتشار خطا است که از سه لایه ورودی، پنهان و خروجی تشکیل شده (شکل ۳) و هدف از آموزش این مدل رسیدن به قابلیت پذیری یادگیری و تعمیم‌پذیری است. بدین معنا که

[1] Multi-Layer Perceptrom

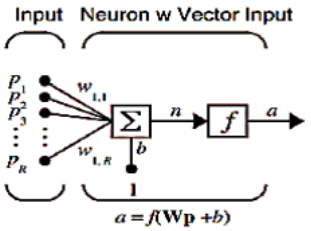
مدل قادر باشد هم الگوهای آموزش دیده و هم الگوهای آموزش ندیده را به درستی تشخیص دهد (Samii et al., 2023).



شکل ۳ - یک شبکه عصبی مصنوعی با سه لایه (Samii et al., 2023)

Figure 3 – An artificial neural network with three layers (Samii et al., 2023)

در این مدل، لایه ورودی وظیفه توزیع مقادیر ورودی به لایه بعدی را دارد و لایه خروجی نیز پاسخ تولید شده را ارائه می‌دهد. تعداد نرون‌ها در لایه ورودی و خروجی به ترتیب با تعداد متغیرهای ورودی و خروجی برابر است. لایه یا لایه‌های پنهان وظیفه پردازش و ارتباط بین لایه ورودی و خروجی را بر عهده دارند و تعداد نرون‌ها و لایه‌ها در این قسمت به پیچیدگی مساله بستگی دارد. در واقع هر نرون یک عنصر محاسبه‌گر با یک یا چند ورودی و یک خروجی است. شکل ۴ ساختار یک نرون مصنوعی را نشان می‌دهد که شامل سه بخش وزن‌ها، ورودی و تابع محرک (تابع فعال ساز) است که این تابع برای انتقال خروجی‌های حاصل شده از هر لایه به لایه بعدی مورد استفاده قرار می‌گیرد.



شکل ۴- مدل چند ورودی یک نرون (Beyranvand and Sahraian, 2018)

Figure 4 – Multi-input model of a neuron (Beyranvand and Sahraian, 2018)

در مدل MLP اتصال نرون‌ها بدین صورت است که تمام نرون‌های لایه ورودی به نرون‌های لایه پنهان و تمام نرون‌های لایه پنهان به لایه خروجی متصل است و برای آموزش از الگوریتم پس انتشار خطا استفاده می‌شود. در واقع خروجی محاسبه شده توسط مدل با خروجی واقعی مقایسه شده و مقدار خطا محاسبه می‌شود و این خطا به منظور تنظیم وزن‌ها و بایاس به عقب انتشار می‌یابد تا خطا در دفعات بعد کمتر شود (Shokati and Kaffash Charandabi, 2021).

الگوریتم تقویت گرادیانی پیشرفته (XGBoost)

این الگوریتم مبتنی بر تقویت درخت و در واقع یک نسخه بهبود یافته از روش تقویت گرادیان درخت تصمیم است که با ترکیب دو روش تقویت ومنظم سازی، یک مدل پیش‌بینی قوی و دقیق ایجاد می‌کند (Chen and Guestrin, 2016). در روش تقویت، چندین مدل ضعیف (معمولاً درخت تصمیم) به طور متوالی آموزش داده می‌شود و هر مدل جدید، اشتباهات مدل قبلی را اصلاح می‌کند (Suhendra et al., 2023). در حالی که منظم سازی از طریق کاهش نرخ یادگیری و تنظیم اهمیت ویژگی‌ها انجام می‌شود که این امر باعث جلوگیری از بیش‌برازش و بهبود توانایی تعمیم مدل می‌شود (Maulana et al., 2023). بدین منظور، بسط تیلور تابع هدف را تقریب زده و بهینه‌سازی می‌کند تا توابع ضرر پیچیده و غیر یکپارچه به صورت کارامد بهینه شوند (Chen and Guestrin, 2016). در واقع در هر مرحله از فرآیند آموزش، الگوریتم XGBoost یک درخت تصمیم ایجاد می‌کند تا مجموع تابع ضرر و عبارت منظم سازی را به حداقل برساند. عبارت منظم سازی جهت کنترل پیچیدگی مدل مورد استفاده قرار می‌گیرد. با تنظیم وزن مربوط به عبارت منظم سازی، این الگوریتم قادر است ضمن کاهش پیچیدگی و حفظ عملکرد مدل، توانایی تعمیم‌دهی آن را بهبود بخشد (Luo et al., 2024). در الگوریتم XGBoost تابع ضرر معمولاً به صورت جمع کل اختلاف بین پیش‌بینی مدل و مقادیر واقعی و با معادلات ۳ و ۴ محاسبه می‌شود.

$$(y_i,\tilde{y}_i)=(y_i-\tilde{y}_i)^2 \quad (۳)$$

$$\Omega(f)=\gamma\times T+\frac{1}{2}\lambda\times\sum_{j=1}^{T}w_j^2 \quad (۴)$$

درروابط مذکور y_i ماتریس مقادیر واقعی، $\tilde{y}_i$ ماتریس مقادیر پیش‌بینی شده، T تعداد گره‌ها، wj وزن گره‌ها، γ و λ پارامترهای منظم سازی هستند که پیچیدگی و میزان جریمه

DOI: 10.22125/agmj.2025.542575.1188

برای وزن‌های گره‌های برگ را کنترل می‌کنند. در این مدل در هر مرحله از آموزش از الگوریتم نزول گرادیان برای به روز رسانی پارامترهای مدل و کاهش تابع ضرر و از بسط دوم تیلور برای تقریب تابع ضرر و محاسبه پارامترهای به روز رسانی شده استفاده می‌شود. به علاوه از استراتژی های مختلف مانند زیر نمونه‌گیری ستون‌ها، زیر نمونه‌گیری ردیف‌ها و ارزیابی اهمیت ویژگی‌ها برای بهبود عملکرد و پایداری مدل به خصوص در شرایط مواجهه با داده‌های بالا و پراکنده استفاده می‌شود (Dhaliwal et al., 2022).

الگوریتم CatBoost[۱]

CatBoost یک الگوریتم قدرتمند گرادیان بوستینگ است که در آن مجموعه‌ای از درختان تصمیم ضعیف به صورت متوالی ساخته می‌شوند. این الگوریتم به طور تکراری مدل را با برازش درختان جدید روی باقی‌مانده‌های تکرارهای قبلی برای به حداقل رساندن تابع ضرر، بهبود می‌بخشد. برخلاف روش‌های سنتی گرادیان بوستینگ، CatBoost از تکنیک order boosting استفاده می‌کند که توانایی بهتر در درک روابط بین ویژگی‌های دسته‌ای با در نظر گرفتن وابستگی‌های ترتیبی را دارد که منجر به پیش‌بینی‌های دقیق‌تر می‌شود. همچنین به کاهش مشکل بیش‌برازش که در دیگر الگوریتم‌های گرادیان بوستینگ به‌ویژه هنگام کار با مجموعه داده‌های کوچک یا نامتوازن دیده می‌شود، کمک می‌کند و به دلیل پایداری، عملکرد بالا و مدیریت کارآمد متغیرهای دسته‌ای محبوبیت زیادی پیدا کرده است (Bentéjac et al., 2021; Luo et al., 2021).

ویژگی مدل‌های مورد استفاده

پیاده سازی و اجرا کلیه مدل‌های یادگیری ماشین ذکر شده در محیط نرم افزار Python 3.8 انجام شد. در این پژوهش از اطلاعات روزانه ۵ متغیر هواشناسی و ۲ آلاینده هوا به عنوان ورودی مدل استفاده شد و غلظت CO شبیه سازی شد. در همه الگوریتم‌ها ۷۰٪ داده‌ها برای آموزش و ۳۰٪ داده‌ها برای تست استفاده شد. برای افزایش دقت مدل‌ها و رسیدن به نتیجه مطلوب، روش‌های مختلف بهینه سازی و همچنین تنظیم پارامترهای مختلف برای هر مدل تست و نتایج بررسی شد و در نهایت، از روش RandomizedSearchCV برای تنظیم پارامتردر مدل‌های درخت-پایه استفاده شد که ویژگی‌های هر مدل در جدول ۳ ارائه شده است.

جدول ۳- ویژگی‌های هر مدل

Table 3 – Features of each model

Models ML	Optimal Parameters
Random Forest	Bootstrap:True, max_depth:20, max_features:0.5, min_samples_leaf:2, min_samples_split: 4, n_estimators: 463
MLP	Hidden Layer Structure: three hidden layer (100,100,100), Activation Function: Rectified Linear Unit (ReLU) Regularization Parameter (Alpha)=0.0061 Initial Learning Rate= 0.0046 Solver=Adam Maximum Number of Iterations (Epochs)=500 Random State=42
XGBoost	subsample: 0.8, reg_lambda:5, reg_alpha:0.1, n_estimators:300, max_depth:3, learning_rate: 0.03, colsample_bytree: 0.8
CatBoost	subsample:0.7, learning_rate:0.03, l2_ leaf_reg:3, iterations:1000, depth:6, border_count: 64, bagging_temperature: 5

ارزیابی عملکرد و دقت مدل‌ها

به منظور ارزیابی عملکرد مدل‌های مذکور از شاخص‌های آماری ضریب تعیین (R^2)، جذر میانگین مربعات خطا (RMSE[۲]) و میانگین خطای مطلق (MAE[۳]) به شرح معادله‌های ۵ تا ۷ استفاده شد (Segovia et al., 2023).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(O_i - P_i)^2}{\sum_{i=1}^{n}(O_i - \bar{O})^2} \quad (۵)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(P_i - O_i)^2}{n}} \quad (۶)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|P_i - O_i| \quad (۷)$$

[۱] Categorical Boosting
[۲] Root Mean Square Error
[۳] Mean Absolute Error

در معادلات بالا P_i مقدار پیش‌بینی شده توسط مدل، O_i مقدار مشاهداتی، $\overline{O}$ میانگین مقادیر مشاهداتی و n تعداد داده است. هر چه مقدار RMSE و MAE کمتر و به صفر نزدیک و مقدار R^2 به ۱ نزدیک باشد، نشان‌دهنده دقیق‌تر بودن شبیه‌سازی و کارایی بهتر مدل‌ها در هر مرحله است.

نتایج و بحث

در جدول ۴ نتایج حاصل از ارزیابی عملکرد مدل‌های RF،MLP، XGBoost، CatBoost در برآورد غلظت CO زمستانه در هوای شهر کرمان در مراحل آموزش و تست ارائه شده است. در مجموع، بررسی شاخص‌های آماری موید دقت بیشتر مدل‌های درخت-پایه در مقایسه با مدل ساخته شده بر مبنای شبکه عصبی است و نمودارهای رسم شده در شکل ۵ نیز نشان می‌دهد که تجمع نقاط مشاهداتی و شبیه سازی شده در اطراف خط نیمساز (Y=X) در مدل‌های درخت-پایه نسبت به مدل MLP بیشتر است. همچنین نمودار رسم شده در شکل ۶ و نمودار تجمیعی رسم شده در شکل ۷ نیز دقت بیشتر مدل‌های درخت –پایه را تأیید می‌کند. محققان دیگری نظیر Shahriar et al., (2021); Karami et al., (2023)، Rahman et al., (2024); Kothandaraman et al.,(2022) نیز در تحقیقات مشابهی برتری مدل‌های درخت پایه در تخمین غلظت آلاینده‌های هوا را تأیید کردند. بیشترین مقدار R^2 در مرحله آموزش و تست در مدل CatBoost و به ترتیب به مقدار ۰/۹۰۱ و ۰/۷۷۸ مشاهده شد. همچنین کمترین مقدار RMSE مرحله آموزش و تست برابر با (ppb) ۰/۱۱۸ و (ppb) ۰/۲۸۴ و کمترین مقدار MAE دو مرحله برابر با (ppb) ۰/۱۱۳ و (ppb) ۰/۲۰۹ در مدل CatBoost مشاهده شد که نشان دهنده توانایی بهتر این مدل در فهم روابط و آموزش بهتر است. در تحقیق حاضر از برخی اطلاعات هواشناسی و اطلاعات مربوط به غلظت دو آلاینده هوا به عنوان متغیرهای ورودی به مدل استفاده شد که به صورت روزانه ثبت شدند و بررسی آماری انجام شده نشان داد که داده‌ها توزیع نرمال ندارند. همچنین دارای نویز بالایی هستند که ممکن است به دلیل بی‌دقتی دستگاه یا اپراتور ثبت کننده باشد و یا اینکه حتی ممکن است یک مقدار واقعی اندازه‌گیری شده باشد.

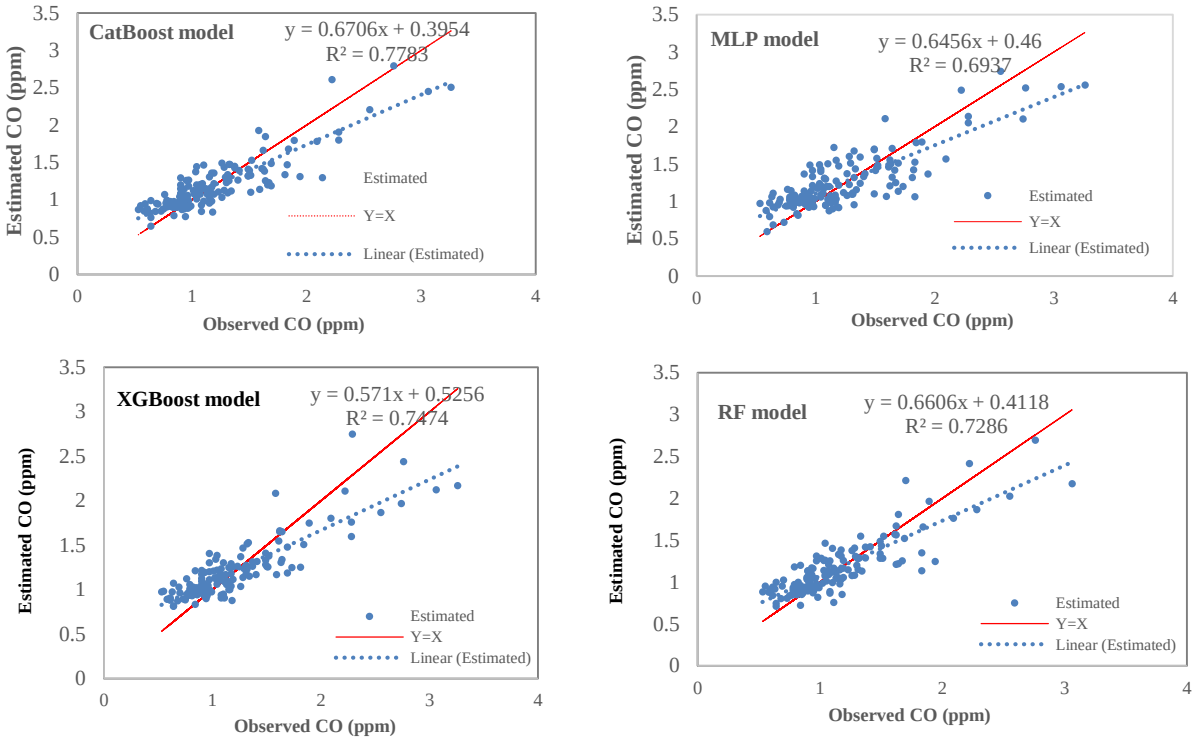
جدول۴- شاخص‌های آماری محاسبه شده مدل های مذکور در مرحله آموزش و تست

Table 4 – Statistical indices calculated for the mentioned models during training and testing phases

	Train			Test		
ML Models	**R^2**	**RMSE (ppb)**	**MAE (ppb)**	**R^2**	**RMSE (ppb)**	**MAE (ppb)**
CatBoost	**0.901**	**0.118**	**0.113**	**0.778**	**0.284**	**0.209**
XGBoost	**0.888**	**0.153**	**0.11**	**0.747**	**0.291**	**0.217**
RF	**0.856**	**0.176**	**0.132**	**0.728**	**0.294**	**0.212**
MLP	**0.774**	**0.248**	**0.19**	**0.693**	**0.308**	**0.236**

در تحقیق حاضر سعی شد تا از داده‌های واقعی برای ارزیابی دقت مدل‌ها استفاده شود و تا حد ممکن از حذف مقادیر پرت (به جز چند نقطه) صرفنظر گردید تا عملکرد مدل‌ها در شرایط نزدیک به وضعیت واقعی بررسی گردد. بنابراین مشاهده نتایج دقیق‌تر در مدل CatBoost با توجه به ویژگی‌های این مدل قابل انتظار است. زیرا این مدل از ترکیب مناسب درخت تصمیم و الگوریتم تقویتی برای افزایش توانایی جهت شناسایی الگوهای پیچیده و مقاومتی در برابر داده‌های با توزیع غیرمتوازن استفاده می‌کند. همچنین با استفاده از روش order boosting مشکل بیش برازش کاهش یافته و در نتیجه با تنظیم پارامترها، دقت مدل افزایش می‌یابد (Prokhorenkiva et al., 2018). مقایسه نتایج حاصل از مدل‌های XGBoost و RF نیز نشان می‌دهد که هر دو مدل دقت نسبتاً یکسانی در برآورد غلظت CO دارند. به طوری‌که نسبت R^2 مدل XGBoost به RF در مرحله آموزش و تست به ترتیب برابر با ۱/۰۳۷ و ۱/۰۲۶ محاسبه شد که برتری اندک مدل XGBoost را تأیید می‌کند. مقدار R^2 در مرحله آموزش و تست در مدل XGBoost به ترتیب ۰/۸۸۸ و ۰/۷۴۷ و در مدل RF به ترتیب برابر با ۰/۸۵۶ و ۰/۷۲۸ محاسبه شد. مقدار RMSE نیز در دو مرحله آموزش و تست در مدل

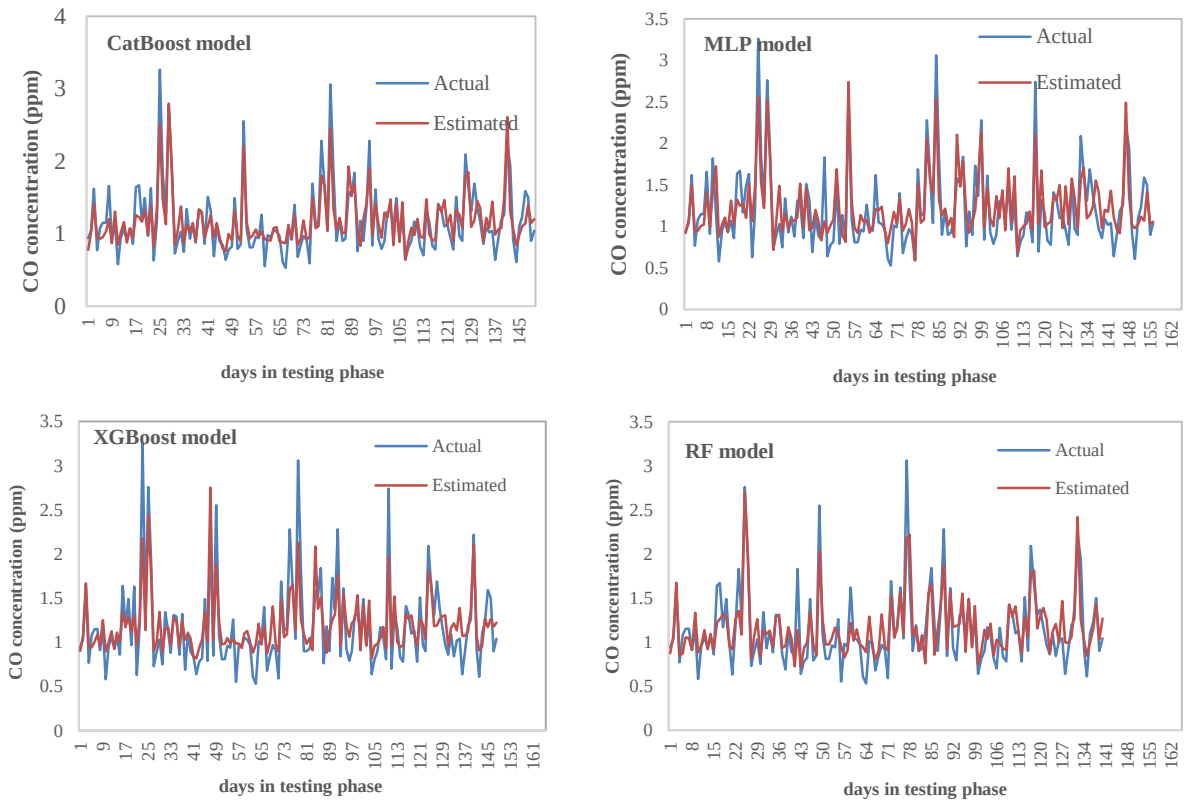
XGBoost برابر با (ppb) ۰/۱۵۳ و (ppb) ۰/۲۹۱ و در مدل RF برابر با (ppb) ۰/۱۷۶ و (ppb) ۰/۲۹۴ محاسبه شد. این نتایج با نتایج ارائه شده توسط Juarez and Peterson, (2024) در مورد تأیید برابری نتایج حاصل از مدل XGBoost و RF در تخمین غلظت ازن همخوانی دارد.



شکل ۵- نمودار مقایسه مقادیر مشاهداتی و پیش‌بینی شده با خط نیمساز در مرحله تست

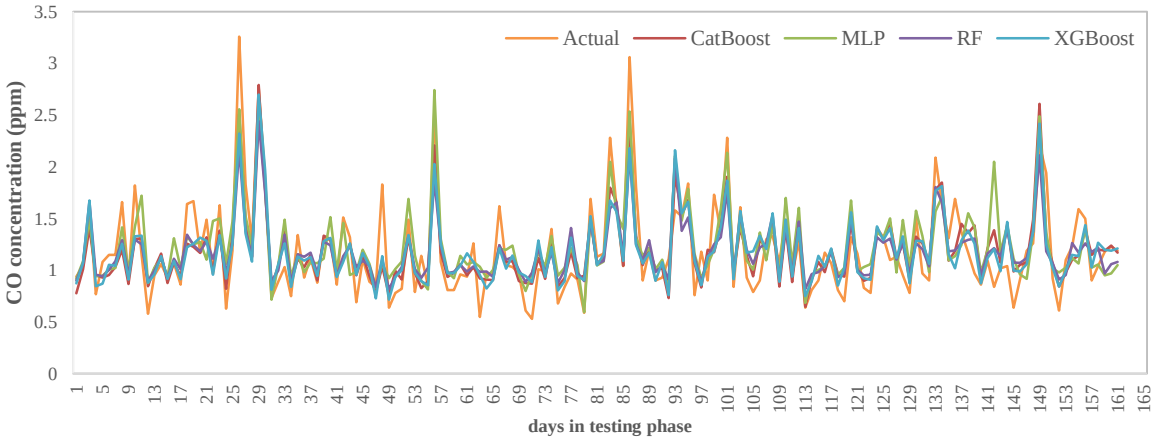
Figure 5 – Comparison chart of observed and estimated values with the bisector line in the testing phase

XGBoost و CatBoost مدل‌هایی بر پایه الگوریتم گرادیان بوستینگ هستند که این الگوریتم سعی می‌کند با ترکیب یادگیرنده‌های ضعیف، یعنی مدل‌هایی که عملکرد بهتر از حد تصادفی دارند، یک یادگیرنده قوی بسازد. این مدل‌ها در واقع نسخه‌های اصلاح شده‌ای از الگوریتم گرادیان بوستینگ هستند که در آنها سرعت آموزش و توانایی تعمیم‌پذیری مدل بهبود یافته است. XGBoost از درختان تصمیم به صورت مرحله‌ای استفاده می‌کند. در واقع هر درخت روی خطاهای درخت قبلی تمرکز می‌کند و از گرادیان نزولی برای بهینه‌سازی خطاها استفاده می‌شود. مهم‌ترین بهبود رخ داده در این مدل این است که از آنجا که در این مدل فقط از درختان تصمیم به عنوان مدل پایه استفاده می‌شود، یک نسخه تغییر یافته از تابع ضرر به کار برده می‌شود تا بتواند پیچیدگی درخت‌ها را کنترل کند که این موضوع نقش مهمی در کاهش بیش‌برازش و افزایش تعمیم پذیری مدل و افزایش سرعت آموزش دارد. در CatBoost هدف اصلی کاهش جابه‌جایی پیش‌بینی است که در طول آموزش رخ می‌دهد. این جابه‌جایی در واقع اختلاف بین مقدار تابع $F(X_i)$ برای نمونه آموزشی X_i و مقدار تابع $F(X)$ برای نمونه تست X است و به این دلیل رخ می‌دهد که در فرآیند آموزش، گرادیان بوستینگ از یک سری نمونه برای براورد گرادیان‌ها و همچنین آموزش مدل‌هایی که این گرادیان‌ها را حداقل می‌کنند، استفاده می‌کند.



شکل۶- نمودار مقایسه مقادیر مشاهداتی و شبیه سازی شده در مرحله تست

Figure 6 – Comparison chart of observed and simulated values in the testing phase



شکل۷- نمودار تجمیعی مقایسه مقادیر مشاهداتی و شبیه سازی شده در مرحله تست

Figure 7 – Cumulative comparison chart of observed and simulated values in the testing phase

اما در CatBoost برای حل این مشکل، گرادیان‌ها با استفاده از یک توالی از مدل‌های پایه که آن نمونه خاص را در مجموعه اموزش خود ندارند، تخمین زده می‌شود (Bentéjac et al., 2021). هر چند مدل CatBoost از نظر عملکرد بر مدل‌هایی مانند XGBoost برتری دارد، اما این اختلاف در پژوهش انجام شده توسط (Bentéjac et al., 2021) از نظر آماری معنی‌دار نبود که در تحقیق حاضر نیز نزدیکی نتایج نهایی دو مدل تأیید شد و با نتایج ایشان همخوانی دارد. اما

RF یک الگوریتم مبتنی بر Bagging است که از تعداد زیادی درخت تصمیم مستقل تشکیل شده و هر درخت روی یک زیر مجموعه تصادفی از داده‌ها آموزش می‌بیند. سپس نتایج تمام درخت‌ها ترکیب می‌شوند و این روش بر کاهش واریانس تمرکز دارد. ویژگی قابل توجه این مدل این است که بدون تنظیم پارامتر معمولاً نتیجه مطلوب را ارائه می‌دهد و در غیر این صورت، تنظیم پارامترها اغلب تأثیر اندکی بر بهبود عملکرد دارد. بررسی نتایج حاصل از مدل MLP نشان داد که این مدل کمترین دقت را در تخمین غلظت CO در مقایسه با سایر مدل‌ها داشته است. مقدار R2 در مرحله آموزش و تست به ترتیب ۰/۷۴۵ و۰/۶۹۳ و مقدار RMSE دراین دو مرحله به ترتیب به ۰/۲۴۸ و ۰/۳۰۸ و مقدار MAE نیز به ترتیب به ۰/۱۹۰ و ۰/۲۳۶ (ppb) رسید که نشان دهنده دقت پایین مدل در مقایسه با مدل‌های درخت-پایه است. در تحقیقات انجام شده توسط (Choi et al., 2023; Li et al., 2024) نیز برتری مدل‌های درخت-پایه در تخمین آلاینده‌های هوا در مقایسه با مدل‌های مبتنی بر شبکه عصبی گزارش شده است که این برتری به دلیل استفاده از تکنیک‌هایی مانند بوستینگ است که از بیش برازش در مدل‌های درخت-پایه جلوگیری نموده و نتایج دقیق‌تری را به ویژه در شرایط استفاده از داده‌های واقعی و دارای نویز ارائه می‌کند.

نتیجه‌گیری

یکی از چالش‌های جامعه مدرن امروزی، کاهش کیفیت هواست که شدت آن به واسطه صنعتی شدن و افزایش جمعیت شهرها، روند افزایشی دارد. بنابراین برنامه‌ریزی جهت مقابله با پیامدهای منفی آلودگی هوا بر سلامت انسان و محیط زیست بر اساس داشتن دانش دقیق از آلاینده‌ها و عوامل مؤثر بر آن‌ها ضروری است. داده‌های مربوط به کیفیت هوا دارای ویژگی‌های تصادفی، نامنظم و ناپایدار هستند و همین امر پیش‌بینی آلاینده‌ها و کیفیت هوا را دشوار می‌کند. از این رو استفاده از چارچوب‌های ریاضی در قالب مدل‌های یادگیری ماشین روشی بهینه و مقرون‌به‌صرفه برای مدل‌سازی آلودگی هوا به شمار می‌رود که نتایج حاصل از آن می‌تواند در اعمال مدیریت جهت کاهش اثرات الودگی هوا موثر باشد. در تحقیق حاضر از ۴ مدل یادگیری ماشین MLP، RF، XGBoost و CatBoost جهت تخمین غلظت مونوکسیدکربن زمستانه در شهر کرمان استفاده شد. از اطلاعات روزانه ۵ پارامتر هواشناسی و ۲ آلاینده هوا به عنوان ورودی به مدل‌ها استفاده شد و نتایج بدست آمده نشان داد که مدل‌های درخت-پایه دقت قابل قبولی در تخمین غلظت مونوکسید کربن در مقایسه با مدل ساخته شده بر مبنای شبکه عصبی دارند. بیشترین مقدار R^2 و حداقل مقدار RMSE و MAE در مدل CatBoost مشاهده شد. نتایج حاصل از مدل XGBoost و RF تقریباً یکسان بود، هر چند برتری اندک مدل XGBOOST مشاهده شد. کمترین دقت نیز در مدل MLP مشاهده شد.

سپاسگزاری

نویسندگان از سازمان هواشناسی و اداره محیط زیست شهرستان کرمان به خاطر در اختیار قرار دادن اطلاعات تشکر می‌نمایند. همچنین از همکاری و مشاوره آقای دکتر امیرحسین نجف آبادی پور قدردانی می‌شود.

منابع

Akbarzadeh, A., Vesali Naseh, M. R., and NodeFarahani, M. 2020. Carbon monoxide prediction in the atmosphere of Tehran using developed support vector machine. Pollution, 6(1), 43-57, DOI:10.22059/poll.2019.279412.618

Alizadeh Dakhel, A., Ghavidel, A., and Panahande, M. 2009. Kerman cement suspended particles distribution modeling using computational fluid dynamics. Journal of Environmental Health Research Forum, 67-74.

Azizi Mobaser, J., Rasoulzadeh, A., and Akbari Majd, A. 2025. Assessment of machine learning and remote sensing in quantifying reference evapotranspiration. Journal of Water and Irrigation Management, 15(1), 0180-205. https://doi.org/10.22059/jwim.2025.383457.1182

Bentéjac, C., Csörgő A., and Martínez-Muñoz, G. 2021. A comparative analysis of gradient boosting algorithms. Artificial Intelligence Review, Springer Nature, 54, 1937–1967. https://doi.org/10.1007/s10462-020-09896-5

Beyranvand, S.H., and Sahraian, K. 2018. The types of neural networks and their application. 3th international conference on management and humanistic science research. 5 July, Tehran.

Chadalavada, S., Faust, O., Salvi, M., Seoni, S. Raj, N., Raghavendra, U., Gudigar, A., Barua, P.D., Molinari, F., and Acharya, R. 2025. Application of

Iranian Journal of Natural Resources, 4(55), 8 [In Farsi].

Kothandaraman, D., Praveena, N., Varadarajkumar, K., Madhav Rao, B., Dhabliya, D., Satla, S., and Abera, W. 2022. Intelligent forecasting of air quality and pollution prediction using machine learn-ing. Adsorption Science and Technology, https:// doi.org/ 10.1155/ 2022/ 50866 22

Lei, T.M.T., Siu, S.W.I., Monjardino, J., Mendes, L., and Ferreira, F. 2022. Using machine learning methods to forecast air quality: Atmosphere, 13, 1412. https://doi.org/10.3390/atmos13091412

Li, Y., Zhang, M., Ma, G., Ren, H., Yu, E. 2024. Analysis of primary air pollutants' spatiotemporal distributions based on satellite imagery and machine-learning techniques. Atmosphere, 15(3), 1-21. https://doi.org/10.3390/atmos15030287

Luo, M., Wang, Y., Xie, Y., Zhou, L., Qiao, J., Qiu, S., and Sun, Y. 2021. Combination of featureselection and catboost for prediction: The firstapplication to the estimation of aboveground biomass. Forests, 12(2), 216. https://doi.org/10.3390/f12020216

Luo, S., El, X., and Li, X. 2024. Data preprocessing method for landslide displacement prediction based on XG Boost. 13th Data Driven Control and Learning Systems Conference (DDCLS), 745-750.

Masoudi, M., and 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

Maulana, A., Noviandy, T. R., Suhendra, R., Earlia, N., Sofyan, H., Subianto, M., and Idroes, R. 2023. Performance analysis and feature extraction for classifying the severity of atopic dermatitis diseases. 2nd International Conference on Computer System, Information Technology, and Electrical Engineering (COSITE), 226–231. https://doi.org/10.1109/COSITE60233.2023.10249760

Moazami, S., Noori, R., Mohammad Salimian, M., Momeni, M., and Vesali Naseh, M.R. 2017. Evaluation of support vector machine performance for carbon monoxide prediction. Modares Civil Engineering Journal, 17(3), 195-202. [In Farsi].

Mohammadi, F., Teiri, H., Yaghoub Hajizadeh, Y., Abdolahnejad, A., and Ebrahimi, A. 2024. Prediction of atmospheric PM2.5 level by machine learning techniques in Isfahan, Iran. Scientific Report, 14, 2109. https://doi.org/10.1038/s41598-024-52617-

Najafpoor, A.A., Allahyari, S., Javid, A.B., and Esmaily, H. 2014. Modeling of air pollution (carbon monoxide and nitrogen oxides) from artificial intelligence in air pollution monitoring and forecasting: A systematic review. Environmental Modelling and Software, 185: 106312. https://doi.org/10.1016/j.envsoft.2024.106312

Chen, T., and Guestrin, C. 2016. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 785-794.

Choi, H., Park, S., Kang, Y., Im, J., and Song, S. 2023. Retrieval of hourly PM2.5 using top-of-atmosphere reflectance from geostationary ocean color imagers I and II. Environmental Pollution, 15; 323, 121169. DOI: 10.1016/j.envpol.2023.121169. Epub 2023 Feb 9.

Dhaliwal, J. K., Panday, D., Saha, D., Lee, J., Jagadamma, S., Schaeffer, S., Mengistu, A. 2022. Predicting and Interpreting cotton yield and its determinants under long-term conservation management practices using machine learning. Computers and Electronics in Agriculture, 199, DOI:10.1016/j.compag.2022.107107

Dong, X., Yu, Z., Cao, W., Shi, Y., and Ma, Q. 2020. A survey on ensemble learning. Frontiers of Computer Science, 14(2), 241-258. DOI:https://10.1007/s11704-019-8208-z

ıdi, R., hadavifar, M., Moeinaddini, M., and 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

Hojjati, S.M., Tafazoli, M., Asadiyan, M., and Baluee, A. 2022. Estimation of carbon sequestration and forest soil respiration using machine learning models in Eastern Forests of Mazandaran Province. Journal of Forest Research and Development, 8(4), 371-387. [In Farsi]. 10.30466/jfrd.2022.54304.1613

Juarez, E.K., and Petersen, M.R.A. 2022. Comparison of machine learning methods to forecast tropospheric ozone levels in Delhi. Atmosphere, 13, 46.

Karami, P., Eslaminezhad, S. A., Eftekhari, M., Boroumand, F., and 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

Khorasani, N., Cheraghi, M., Nadafi, K., and Karami, M. 2002. Study of air quality in Tehran and Isfahan in 1377 and provide solutions for improving that.

Earlia, N., Niode, N.J., and Idroes, R. 2023. Evaluation of gradient boosted classifier in atopic dermatitis severity score classification. Heca Journal of Applied Sciences, 1(2), 54–61. https://doi.org/10.60084/hjas.v1i2.85.

Shahriar, S.A., Kayes, I., Hasan, K., Hasan, M., Islam, R., Awang, N.R., Hamzah, Z., Rak, A.E., and Salam, M.A. 2021. Potential of ARIMA-ANN, ARIMA-SVM, DT and CatBoost for atmospheric $PM_{2.5}$ forecasting in Bangladesh. Atmosphere, 12(1), 100.

Shokati, H., and Kaffash Charandabi, N. 2021. Evaluation and comparison of FFB, CFB, and MLP artificial neural networks for the identification of potential sites for the construction of photovoltaic solar power plants in East Azarbaijan province. Journal of Geography and Environmental Planning, 31(4), 1-28.

Tipton, J. 2022. Leaving the city: health and happiness in the other America. John Hunt Publishing.

Wark, K.W.D. 1998. Air Pollution, Its Origin and Control. Third Ed. New York.

Wright, J. 2002. Chronic and occult carbon monoxide poisoning: we don't know what we're missing. **J**ournal of Emergency Medicine, 19 (5), 386–390. http://dx.doi.org/10.1136/emj.19.5.386

Zhou, X., Guo, M., Li, Z., Yu, X., Huang, G., Li, Z., Zhang, X and Liu, L. 2023. Associations between air pollutant and pneumonia and asthma requiring hospitaliza-tion among children aged under 5 years in Ningbo 2015–2017. Frontires in Public Health. https://doi.org/10.3389/fpubh.2022.1017105

Zounemat-Kermani, M., Ramezani-Charmahineh, A., Razavi, R., Alizamir, M, and Ouarda, T. 2020. Machine learning and water economy: a new approach to predicting dam's water sales revenue. Water Resources Management, 34, 1893–1911. https://doi.org/10.1007/s11269-020-02529-0

vehicles in the city of Mashhad in 2010. Journal of North Khorasan University of Medical Sciences, 6(2), 258. [In Farsi].

Noori, R., Hoshyaripour, G., Ashrafi, K., and Rasti, O. 2013 .Introducing an appropriate model using support vector machine for predicting carbon monoxide daily concentration in Tehran atmosphere. Iranian Journal of Health and Environment, 6(1):1-10. [In Farsi].

Prokhorenkiva, P., Gusev, G., Vorobev, A., Dorogush, A.V., and Gulin, A. 2018. CatBoost: unbiased boosting with categorical features. NIPS'18: Proceedings of the 32nd International Conference on Neural Information Processing Systems, 6639 - 6649

Rahman, M., Hussain Nayeem, E., Ahmed, S., Akther Tanha, K., Alam Sakib, S., Mohiuddin, K.H., and Hasan Babu, H. 2024. AirNet: predictive machine learning model for air quality forecasting using web interface. Environmental Systems Research, 13, 44. https://doi.org/10.1186/s40068-024-00378-z

Ruiz-AÍvarez, M., Gomariz-Castillo, F., and Alonso-Sarría, F. 2021. Evapotranspiration response to climate change in semi-arid areas: using random forest as multi-model ensemble method. Water, 13(2), https://doi.org/10.3390/w13020222

Samii, A., Karami, H., Ghazvinian, H., Safari, A., and Dadras Ajirlou, Y. 2023. Comparison of DEEP-LSTM and MLP models in estimation of evaporation pan for arid regions. Journal of Soft Computing in Civil Engineering, 7(2), 155-175. https://doi.org/10.22115/scce.2023.367948.1550

Segovia, J.A., Toaquiza, J., Llanos, J., Rivas, D.R. 2023. Meteorological Variables Forecasting System Using Machine Learning and Open-Source Software. Electronics, 12(4), 1007. https://doi.org/10.3390/electronics12041007

Suhendra, R., Suryadi, S., Husdayanti, N., Maulana, A., Noviandy, T.R., Sasmita, N.R., Subianto, M.,

Journal of Agricultural Meteorology
(Iranian Soc. of Water and Irrig. Eng.)
Vol. 13, No. 2, Autumn & Winter 2025, pp. 3-14
DOI: 10.22125/agmj.2025.542575.1188

نشریه هواشناسی کشاورزی
(وابسته به انجمن مهندسی آبیاری و آب ایران)
جلد ۱۳، شماره ۲، پاییز و زمستان ۱۴۰۴، صص.۱۴-۳
مقاله علمی پژوهشی

Evaluation of machine learning-based models for estimating winter season carbon monoxide concentration in Kerman

S. Golestani Kermani[1]*, S. P. Hosseini[2]

Received: 18/02/2025
Accepted: 22/09/2025

Abstract
Machine learning-based models are effective and practical tools for analyzing complex relationships between air quality parameters and meteorological variables and identifying key factors influencing air quality. The aim of this study is evaluation of three tree-based models' performance namely —Random Forest (RF), XGBoost, and CatBoost—and a multilayer perceptron (MLP) neural network model, using daily data of five meteorological varibles, including wind speed, temperature, air pressure, relative humidity, and rainfall and two air pollutants (O3 and PM2.5), to estimate the daily concentration of carbon monoxide during winter season in Kerman city, south of Iran.The models' performance was assessed using statistical indices including the coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE). Results indicated that the CatBoost model had the highest accuracy in estimating CO concentrations, with a R^2 of 0.778, RMSE of 0.284 (ppb), and MAE of 0.209 (ppb) during the test phase. The performance of the XGBoost and RF models was relatively similar, with R^2 values of 0.747 and 0.728, respectively. The MLP model showed the lowest accuracy, with R^2, RMSE and MAE of 0.693, 0.308 and 0.236 pbb, respectively. These results confirm the superior skill of tree-based models in comparison to the neural network-based model for estimating CO concentration in study region.

Keywords: Air pollution, Carbon monoxide, Machine learning, Kerman

[1] Assistant Professor, Department of Water Engineering, Faculty of Agriculture, Shahid Bahonar University of Kerman, Kerman, Iran.
(*Corresponding Author Email Address: s.golestani@uk.ac.ir)
[2] M.Sc. Graduated, Department of Water Engineering, Faculty of Agriculture, Shahid Bahonar University if Kerman, Kerman, Iran

نحوه ارجاع مقاله:
گلستانی کرمانی، س.، حسینی، س. پ. ۱۴۰۴. ارزیابی عملکرد مدل‌های مبتنی بر یادگیری ماشین در برآورد غلظت روزانه منوکسیدکربن زمستانه شهر کرمان. نشریه هواشناسی کشاورزی، ۱۳(۲)، ۳-۱۴. DOI: 10.22125/agmj.2025.542575.1188
Golestani Kermani, S., Hosseini, S. P. 2025. Evaluation of machine learning-based models for estimating winter season carbon monoxide concentration in Kerman. Journal of Agricultural Meteorology, 13(2), 3-14. DOI: 10.22125/agmj.2025.542575.1188