Research Paper

Relative Humidity Prediction using XGBoost Machine Learning Model, Case Study: Bajgah Climatological Station, Iran

Reza Piraei[1], Ali Mohammadi[2], Seied Hosein Afzali[3]*
1. PhD Student of Water Recourses Management, Department of Civil and Environmental Engineering, Shiraz University, Shiraz, Iran
2. MSc Student of Water Recourses Management, School of Civil and Environmental Engineering, Tarbiat Modares University, Tehran, Iran
3. Associate Prof. of Civil Engineering, Department of Civil and Environmental Engineering, Shiraz University, Shiraz, Iran

Received: 2023/12/06
Revised: 2023/12/10
Accepted: 2024/01/19

Use your device to scan and read the article online

DOI:
10.30495/wej.2024.32823.2403

Keywords:
Bajgah, Machine Learning, Relative Humidity, XGBoost

Abstract

Introduction: Relative humidity is one of the most important hydrological parameters that significantly influences evapotranspiration water resource management, plant growth and even concrete settings. Hence, accurate prediction and estimation of relative humidity paramount importance.

Methods: In this study, since two parameters relative humidity and the minimum and maximum temperatures of preceding days, have the most significant impact on predicting future relative humidity, and given the prevalence of available data for only these two parameters in many parts of the country, various scenarios involving these parameters were studied. The best scenario for predicting relative humidity was obtained using the XGBoost model. To assess the accuracy of the model, the Bajgah region in Fars Province was chosen as a case study, and the accuracy of different scenarios was compared using data from the past 30 years (1993 to 2023). In this regard, missing data were estimated using the KNN Imputer model. The correlation between mean relative humidity of one to ten days before and the target variable (predicted relative humidity on day t) was calculated using Pearson correlation. Based on the results indicating the insignificance of data from the fourth day and earlier, data from one to three days before were utilized.

Findings and Conclusion: Finally, by comparing the results based on six statistical criteria (RMSE, MAE, MARE, MXARE, NSE, and R^2), it was determined the scenario based on relative humidity and the maximum and minimum temperatures of the preceding 3 days provides the best estimation.

Citation: Piraei R, Mohammadi A, Afzali SH. Relative Humidity Prediction using XGBoost Machine Learning Model, Case Study: Bajgah Climatological Station, Iran. Water Resources Engineering Journal. 2024; 17 (62): 40- 53.

⁎**Corresponding author:** Seied Hosein Afzali
Address: Dept. of Civil and Environmental Engineering, Shiraz University, Shiraz, Iran
Tell: +989171112935
Email: afzali@shirazu.ac.ir

Extended Abstract

Introduction

Meteorological variables, particularly relative humidity, exert significant influence on both natural ecosystems and human activities. This impact extends to economic aspects, notably affecting agricultural and water systems, as well as renewable and solar energy management. In hydrology, relative humidity plays a crucial role in the water cycle, influencing transpiration rates from various surfaces. Its importance in agriculture is highlighted in irrigation management and crop productivity due to its effect on transpiration. Additionally, relative humidity contributes to water resources management, influencing the Penman-Montite evapotranspiration calculation method. Measurement of relative humidity is commonly done using hygrometers, but challenges arise in certain situations, leading to the exploration of statistical and machine learning methods for estimation. While recent years have seen increased use of artificial intelligence for climate variable estimation, research specifically focusing on relative humidity remains limited. Using various combination of input variables, this study utilizes XGBoost for the first time to predict relative humidity at the Bajgah meteorological station in Iran, considering 30 years of data and evaluating model performance using six statistical criteria.

Materials and Methods

In this study, meteorological data spanning September 1993 to September 2023, sourced from the Bajgah Fars meteorological station, was employed. Various modeling scenarios were devised, integrating diverse input variables such as relative humidity, minimum and maximum temperatures from one to three days prior. The selection of this 3-day timeframe is grounded in Pearson correlation findings derived from an analysis of relative humidity data over the preceding one to ten days. Nine distinct scenarios were implemented, and hyperparameter optimization was employed during model training, facilitated by grid search. The dataset was partitioned, allocating 80% for training purposes and 20% for testing. Model performance evaluation encompassed six statistical criteria, and to comprehensively gauge models across all criteria, a ranking scheme from literature was adopted. Finally, XGBoost feature importance analyses were conducted on the models, elucidating the significance of each feature in predicting relative humidity.

Findings

After optimizing model hyperparameters, the study assessed their performance on test data. Graphical analysis revealed that models trained solely on minimum and maximum temperatures exhibited wide dispersion and low correlation between observed and predicted relative humidity. Conversely, models relying solely on relative humidity demonstrated significantly improved correlation. Notably, Model I exhibited step-wise predictions, indicating reduced performance despite reasonable correlation. Incorporating both temperature and humidity variables enhanced correlation, with Model VII showing the best test data performance (RMSE=6.73, MARE=0.11, NSE=0.75, R2=0.75, MXARE=1, MAE=4.82). Models exclusively relying on temperature variables performed weakest. The ranking scheme, based on a comprehensive assessment of 12 criteria, places Models XI and VIII jointly at the top, followed by Models VII, II, III, and I, with Models V, VI, and IV placing at the bottom positions. Relative humidity of the previous day emerged as the most important variable in XGBoost feature importance analysis, emphasizing its significance in accurate predictions.

Discussion

Based on the results of this study, models relying solely on temperature variables performed less effectively. On the other hand, models utilizing historical relative humidity showed a significance enhancement in performance. The addition of temperature variables to previous days' relative humidity slightly improved performance. Feature importance analysis indicated that relative humidity, especially

from the previous day, had higher importance than temperature variables. However, the study emphasized that variable importance results were specific to the data and models used. While the findings aligned with previous studies, the research recommended further exploration of the relationship between current and past relative humidity for deeper insights. Overall, the study highlighted the potential for improved climate variable predictions through machine learning model optimization and careful selection of input variables.

Conclusion

Accurate prediction of relative humidity holds significant importance due to its pivotal role in the hydrological cycle and subsequent impacts on agriculture and human well-being. Utilizing the XGBoost machine learning model with 30 years of data, this study demonstrates superior performance when combining historical relative humidity with temperature variables compared to models relying solely on temperature. Models IX and VII, consistently outperformed others, highlighting the crucial role of historical relative humidity data. While these findings guide variable selection, future research should explore additional climatic factors. Additionally, given the ongoing progress in machine learning models, future research could delve into further comparative analyses of alternative models for enhanced insights.

Ethical Considerations compliance with ethical guidelines

The cooperation of the participants in the present study was voluntary and accompanied by their consent.

Funding

No funding.

Authors' contributions

Design and conceptualization: Seied Hosein Afzali, Reza Piraei, Ali Mohammadi.

Methodology and data analysis: Seied Hosein Afzali, Reza Piraei, Ali Mohammadi.

Supervision and final writing: Seied Hosein Afzali.

Conflicts of interest

The authors declared no conflict of interest.

مقاله پژوهشی

پیش‌بینی رطوبت نسبی به وسیله مدل یادگیری ماشین XGBoost، مطالعه موردی باجگاه، ایران

رضا پیرایی[۱]، علی محمدی[۲]، سید حسین افضلی[۳*]
۱. دانشجو دکتری رشته مهندسی و مدیریت منابع آب، بخش عمران و محیط زیست، دانشکده مهندسی، دانشگاه شیراز، شیراز، ایران
۲. دانشجو کارشناسی ارشد رشته مهندسی و مدیریت منابع آب، دانشکده مهندسی عمران و محیط زیست، دانشگاه تربیت مدرس، تهران، ایران
۳. دانشیار مهندسی عمران، بخش عمران و محیط زیست، دانشکده مهندسی، دانشگاه شیراز، شیراز، ایران

تاریخ دریافت: ۱۴۰۲/۰۹/۱۵
تاریخ داوری: ۱۴۰۲/۰۹/۱۹
تاریخ پذیرش: ۱۴۰۲/۱۰/۲۹

از دستگاه خود برای اسکن و خواندن مقاله به صورت آنلاین استفاده کنید

DOI: 10.30495/wej.2024.32823.2403

واژه‌های کلیدی:
رطوبت نسبی، مدل XGBoost، یادگیری ماشین، باجگاه

چکیده

مقدمه: رطوبت نسبی هوا یکی از مهمترین پارامترهای هیدرولوژیکی است که در مدیریت منابع آب، رشد گیاهان و حتی گیرش بتن تاثیر زیادی دارد. لذا پیش بینی و تخمین آن از اهمیت بسزایی برخوردار است.

روش: در این پژوهش از آنجا که پارامترهای رطوبت نسبی و میزان دمای حداقل و حداکثر روزهای قبل، بیشترین تاثیر را در تخمین رطوبت نسبی روز آینده دارند و همچنین وجود آمار تنها این پارامترها در برخی از نقاط کشور، سناریوهای مختلفی مشتمل بر این دو پارامتر مورد مطالعه قرار گرفته است و بهترین سناریو برای پیش بینی رطوبت نسبی با استفاده از مدل XGBoost بدست آمده است. جهت بررسی کارآیی مدل مذکور، منطقه باجگاه در استان فارس مورد تحلیل قرار گرفته و با استفاده از آمار مربوط به سی سال گذشته (۱۳۷۲تا ۱۴۰۲) صحت و دقت سناریوهای مختلف مورد مقایسه قرار گرفته اند. در این راستا ابتدا مقادیری برای داده‌های گمشده به کمک KNN Imputer تخمین زده شده است. سپس میزان ارتباط داده‌های پیشین به کمک همبستگی پیرسون بین متغیر هدف (رطوبت نسبی روز t) و میانگین رطوبت روزانه در بازه یک تا ده روز قبل، محاسبه شده و با توجه به نتایج حاصله مبنی بر کم تاثیر بودن آمار روز چهارم به قبل، آمار مربوط به سه روز قبل مورد استفاده قرار گرفته است.

یافته‌ها و نتیجه‌گیری: در نهایت بر اساس مقایسه نتایج حاصل از ۶ شاخص آماری RMSE, MAE, MARE, MXARE, NSE و R^2، مشخص گردید که در بین سناریوهای مختف، سناریو مبتنی بر رطوبت نسبی و دمای حداکثر و حداقل ۳ روز قبل بهترین تخمین را ارایه می دهد.

* **نویسنده مسئول:** سید حسین افضلی
نشانی: بخش عمران و محیط زیست، دانشکده مهندسی، دانشگاه شیراز، شیراز، ایران.
تلفن: ۰۹۱۷۱۱۱۲۹۳۵
پست الکترونیکی: afzali@shirazu.ac.ir

مقدمه

متغیرهای هواشناسی از جمله رطوبت نسبی به طور مداوم بر زندگی انسان تأثیر می‌گذارند و تأثیرات مستقیمی بر محیطهای طبیعی و انسانی دارند. از جمله تأثیرات این متغیرها بر امنیت اقتصادی یک منطقه می‌توان به تاثیر آن‌ها بر سیستم‌های کشاورزی و آبی و همچنین بر مدیریت سیستم‌های انرژی تجدیدپذیر و خورشیدی اشاره کرد (۱). در هیدرولوژی، رطوبت نسبی یک عامل کلیدی در چرخه آب است که بر نرخ تبخیرتعرق از سطوح مختلف آبی و خاکی تأثیر می‌گذارد. در کشاورزی، رطوبت نسبی به دلیل تاثیری که بر تبخیرتعرق دارد یک متغیر حیاتی برای مدیریت آبیاری و بهره‌وری محصولات کشاورزی است. یکی از کاربردهای مهم رطوبت نسبی در مدیریت منابع آب تاثیر آن در روش محاسبه تبخیرتعرق مرجع استاندارد پنمن–مونتیت است (۲). در نتیجه، به طور کلی اهمیت اندازه‌گیری و پیش بینی رطوبت نسبی در هیدرولوژی و کشاورزی، در تاثیر آن بر چرخه آب، فیزیولوژی گیاهان و پایداری منابع آب و تولید مواد غذایی است. رطوبت نسبی عبارت است از نسبت مقدار آبی که جو در یک دما نگه می‌دارد (فشار بخار آب در هوا) به حداکثر مقدار آبی که جو می‌تواند در همان دما نگه دارد (فشار بخار اشباع). رطوبت نسبی بدون بعد است و معمولاً به صورت درصد ارائه می‌شود. اگرچه فشار واقعی بخار آب ممکن است طی روز نسبتاً ثابت باشد، لیکن رطوبت نسبی بین مقدار حداکثر خود (در نزدیکی طلوع خورشید) و حداقل خود (در اوایل بعدازظهر) نوسان می‌کند. این تغییرات در مقدار رطوبت نسبی نتیجه تغییرات فشار بخار اشباع توسط دمای هوا در طول روز است. با تغییر دما در طول روز، رطوبت نسبی نیز به طور قابل توجهی تغییر می‌کند.

رطوبت نسبی در محل معمولا به وسیله نم‌سنج[۱] اندازه‌گیری می‌شود (۳). در سال‌های اخیر روش‌های آماری مختلفی برای تخمین رطوبت نسبی ارائه شده‌اند و پژوهشگران برای تخمین متغیرهای آب و هوایی به استفاده از مدل‌های هوش مصنوعی و یادگیری ماشین روی آورده‌اند (۴–۷). با این حال، پژوهش‌های محدودی بر روی تخمین رطوبت نسبی به کمک مدل‌های یادگیری ماشین ارائه شده‌اند (۱, ۲, ۸, ۹). در این پژوهش به کمک مدل نوین درختی eXtreme Gradient Boosting (XGBoost)، سناریوهای مختلفی بر اساس ترکیب های گوناگونی از متغیرهای آب و هوایی چون دمای بیشینه، دمای کمینه و رطوبت نسبی روزهای قبل، رطوبت نسبی روز آینده تخمین زده شده و دقیق ترین سناریو به دست می آید. این کار برای ایستگاه هواشناسی باجگاه واقع در نزدیکی شهر شیراز استان فارس ایران به کمک ۳۰ سال داده انجام می‌شود. در این راستا، عملکرد سناریوها به کمک ۶ معیار آماری مختلف ارزیابی می‌شود.

پیشینه پژوهش

همان طور که در مقدمه اشاره شد، مطالعات محدودی روی تخمین رطوبت نسبی با استفاده از مدل‌های یادگیری ماشین انجام گردیده است. در این راستا خطیبی و همکاران (۱) در پژوهش خود به بررسی عملکرد دو مدل Gene expression programming (GEP) و ANN برای تخمین رطوبت نسبی روزانه در سانتا کلاریتا، کالیفرنیا پرداخته‌اند و نشان داده اند مدل شبکه‌های عصبی عملکرد بهتری دارند. ایشان برای تخمین رطوبت نسبی، از متغیرهای هواشناسی مانند دما، سرعت باد و رطوبت نسبی روز پیشین استفاده کرده اند و نشان داده اند ترکیب متغیرهای هواشناسی و متغیر رطوبت نسبی روز پیشین بهترین نتیجه را ارائه می‌دهد. بیاتوارکشی و همکاران (۸) برای پیش‌بینی روزانه و ماهانه رطوبت نسبی در ۳۰ ایستگاه هواشناسی در ایران از چند مدل یادگیری ماشین استفاده کرده اند و نشان داده اند که مدل WPCA–ANN عملکرد بهتری نسبت به سایر مدل‌ها دارد. آنها در پژوهش خود از مقادیر روزانه رطوبت نسبی چهار روز گذشته برای ورودی مدل‌های یادگیری ماشین خود استفاده کرده اند و دلیل انتخاب این متغیرها را ضریب همبستگی بالای آن‌ها با رطوبت نسبی هدف عنوان کرده اند. تائو و همکاران (۲) در پژوهش خود به پیش بینی ماهانه رطوبت نسبی با استفاده از داده‌های هواشناسی در دو ایستگاه کوت و موصل در کشور عراق با استفاده از مدل‌های SVR، RF و Multivariate Adaptive Regression Spline (MARS) پرداخته اند. ایشان دمای بیشینه و کمینه، ساعات آفتابی روزانه، بارش، سرعت باد و میزان تبخیر را به عنوان متغیرهای ورودی در نظر گرفته اند. در پژوهش آنها از مدل XGBoost تنها برای انتخاب بهترین ترکیب متغیرهای ورودی استفاده شده است. تحقیق ایشان نشان داد که در ایستگاه کوت مدل RF و در ایستگاه موصل مدل MARS بهترین عملکرد را دارد. مرابت و هدم (۹) نیز در پژوهش خود برای تخمین رطوبت نسبی در دو ایستگاه هواشناسی در الجزایر از سه مدل یادگیری ماشین Extreme learning Machine (ELM)، ANN و RF و ترکیب آنها با سه مدل تجزیه سیگنال (Empirical Mode Decomposition، Variational Mode Decomposition و Empirical Wavelet Transform) استفاده کرده اند و نهایتا به این نتیجه رسیده اند که مدل‌های ترکیبی بر اساس Empirical Wavelet Transform عملکرد بسیار خوبی از خود نشان می دهد. آنها در پژوهش خود بحث روشنی را در مورد عوامل مؤثر بر رطوبت نسبی ارائه می‌دهند و ترکیبات متغیر ورودی قابل توجهی را مورد بررسی قرار می‌دهند. تحقیق آنها نشان می دهد دمای کمینه و دمای بیشینه مهمترین عامل اصلی تأثیرگذار بر رطوبت نسبی می باشد. با توجه به موارد فوق، از آنجا که رطوبت نسبی و دمای کمینه و بیشینه از عوامل اصلی در میزان رطوبت نسبی می باشند و از طرفی بدلیل اینکه در برخی از ایستگاه های موجود در کشور تنها این دو متغیر اندازه گیری شده اند و داده مربوط به دیگر متغیرها وجود ندارد، در این تحقیق از این دو پارامتر در سناریوهای مختلف به عنوان متغیر ورودی مدل یادگیری ماشین برای پیش بینی رطوبت نسبی استفاده شده است. از طرفی علی رغم اینکه تائو و همکاران (۲) برای انتخاب بهترین ترکیب ورودی‌های خود از مدل XGBoost بهره گرفتند، اما از این مدل تا کنون برای تخمین رطوبت نسبی روزانه استفاده نشده است. لذا در این پژوهش به کمک ۳۰ سال داده ی بدست آمده از ایستگاه باجگاه واقع در ایران، برای نخستین بار از این مدل برای پیش بینی رطوبت

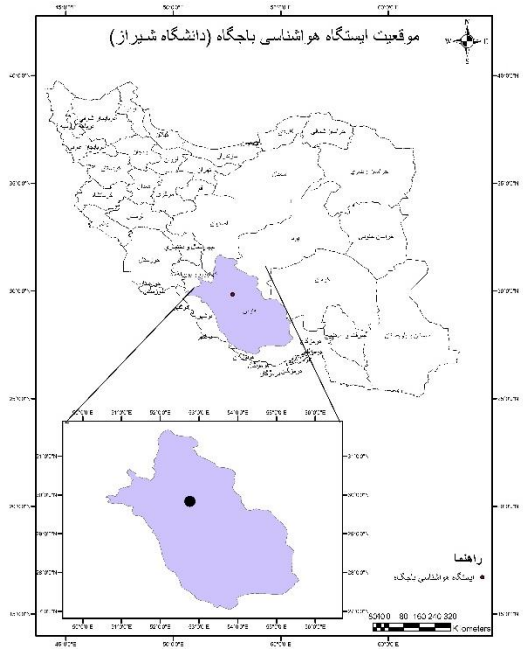
[۱] Hygrometer

نسبی استفاده می شود.

مواد و روش‌ها

منطقه مورد مطالعه

داده‌های مورد استفاده در این پژوهش از ایستگاه هواشناسی دانشکده کشاورزی دانشگاه شیراز واقع در باجگاه، استان فارس، دریافت شده است. این ایستگاه در طول جغرافیایی ´۴۶ °۵۲، عرض جغرافیایی، ´۵۰ °۲۹ و ارتفاع ۱۸۱۰ متری از سطح دریا واقع شده است، موقعیت تقریبی این ایستگاه در شکل ۱ نشان داده شده است. میانگین بارش سالانه در این ایستگاه ۳۵۲/۱۳ میلیمتر در سال و میانگین روزانه دما ۱۴/۴۶ درجه سلسیوس می‌باشد. داده‌های استفاده شده مربوط به بازه زمانی ۱۰ شهریور ۱۳۷۲ تا ۹ شهریور ۱۴۰۲ به صورت روزانه شامل متغیرهای دمای بیشینه، دمای کمینه، رطوبت نسبی بیشینه و رطوبت نسبی کمینه می‌باشد. در این پژوهش تلاش برآن شده است تا میانگین رطوبت نسبی روزانه در افق یک روزه به وسیله مدل‌های یادگیری ماشین پیش‌بینی شود. در این راستا، میانگین رطوبت نسبی روزانه با میانگین‌گرفتن از رطوبت نسبی بیشینه و کمینه هر روز به دست آمده است. با توجه به اینکه داده‌ها به صورت سری زمانی می‌باشند و متغیرهای ورودی داده‌های روزهای پیشین می‌باشند، با در نظر گرفتن ترتیب داده‌ها ۸۰٪ به عنوان داده‌های آموزش و ۲۰٪ به عنوان داده‌های آزمون به صورت تصادفی انتخاب شده اند.



شکل ۱- موقعیت تقریبی ایستگاه هواشناسی باجگاه

مراحل پیش پردازش

در این پژوهش، پیش از آموزش مدل برای تخمین رطوبت نسبی، ابتدا مقادیری برای داده‌های گمشده به کمک مدل KNN[۱] Imputer تخمین زده می‌شوند. این کار با استفاده از مفهوم نزدیکی بین داده‌های مشاهده شده انجام می‌شود (۱۰). داده‌های گمشده بر اساس توابع فاصله که مجاورت داده‌های مشاهده شده با داده گمشده را می‌سنجند تولید می‌شوند. یکی از توابع فاصله مشهور، تابع اقلیدسی می‌باشد که به صورت زیر نزدیکی بین مقدار هدف، x^* و مقادیر در همسایگی آن، x_i، با توجه به انتخاب تعدادی از همسایگان، K، محاسبه می‌شود:

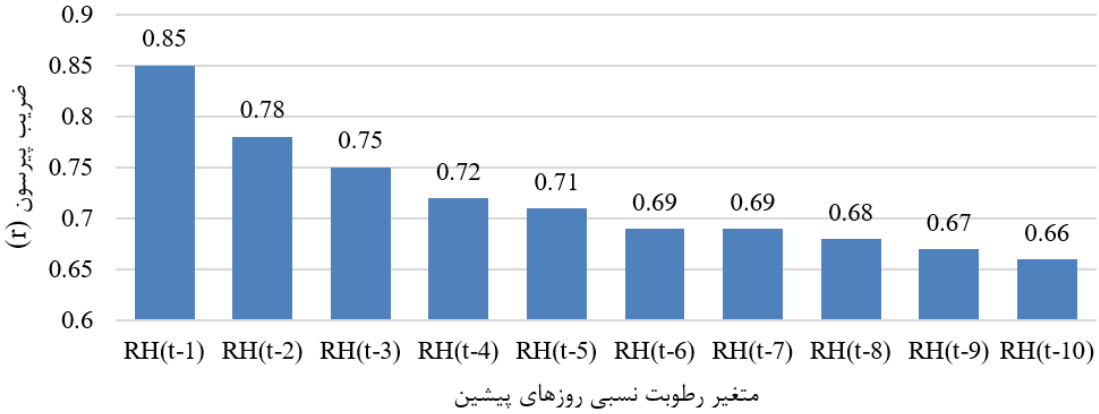
$$d(x^*, x_i) = \sqrt{\sum_{i=1}^{K}(x^* - x_i)^2} \qquad (۱)$$

پس از حل مشکل داده‌های گمشده، با توجه به اینکه در این پژوهش از داده‌های پیشین جهت تخمین روز جدید استفاده می‌شود، میزان ارتباط داده‌های پیشین به کمک همبستگی پیرسون بین متغیر هدف (پیش‌بینی میانگین رطوبت نسبی روز t) و میانگین رطوبت نسبی روزانه در بازه یک تا ده روز قبل، محاسبه شده اند و نتایج حاصله در شکل ۲ ارائه شده است. با توجه به انتخاب آستانه ۰٫۷۵، بازه یک تا سه روز قبل، با بیشترین مقادیر همبستگی برای داده های ورودی انتخاب شده و برای بررسی تاثیر دما بر دقت مدل، سناریوهای دیگری با متغیرهای دمای بیشینه و کمینه نیز در این بازه تعریف شده اند. در نتیجه، با انتخاب متغیرهای ورودی مختلف، ۹ سناریو مختلف تعریف شده که به طور خلاصه در جدول ۱ شرح داده شده‌اند.

مدل یادگیری ماشین

برای تدوین معادله‌ای برای تخمین رطوبت نسبی روزانه، لازم است رابطه‌ای بین متغیرهای مختلف آب و هوایی استخراج شود. در سوی دیگر، الگوریتم‌های یادگیری ماشین بدون توجه به روابط بنیادین میان متغیرهای ورودی و متغیر خروجی و یا هیچ گونه درکی از مسئله مورد نظر، به راحتی رابطه‌ای میان این متغیرها ایجاد می‌کنند. با این حال، با توجه به همین موضوع یعنی عدم درک صحیح مدل از مسئله، پیش از آموزش یا برازش یک مدل یادگیری ماشین به یک مجموعه داده‌ها، باید یک سری فرایند تبدیل حیاتی روی داده‌ها انجام شوند که می‌توانند تأثیر قابل‌توجهی بر عملکرد مدل داشته باشند (۱۱). برای این منظور، در این پژوهش برای کاهش نوسانات و تدقیق نتایج حاصله از الگوریتم تبدیل MinMaxScaler از کتابخانه Scikit-learn برای نرمالیزه کردن داده‌ها استفاده شده است. در این تبدیل، برای هر متغیر مقدار حداقل از هر یک از داده‌ها کم شده و نتیجه آن بر اختلاف بین مقادیر حداکثر و حداقل تقسیم می‌شود. از طریق این تبدیل، هر متغیر بین ۰ و ۱ مقیاس دهی می‌شود. علاوه بر این، اگر در هر یک از سناریوها، مدل، رطوبت نسبی را به عنوان یک مقدار منفی پیش‌بینی کند، با توجه به منطقی نبودن این موضوع آن مقدار با صفر باید جایگزین شود. مدل استفاده شده برای تخمین رطوبت نسبی در این پژوهش مدل XGBoost می‌باشد که در ادامه به شرح مختصر این مدل پرداخته می‌شود.

[۱] K-Nearest Neighbor



شکل ۲- نتایج همبستگی پیرسون

جدول ۱- ترکیب‌های مختلف متغیرهای ورودی برای سناریوهای مختلف

متغیرهای اساسی	مدل	متغیرهای ورودی	متغیر خروجی
RH	I	RH_{t-1}	RH_t
	II	RH_{t-1}, RH_{t-2}	RH_t
	III	$RH_{t-1}, RH_{t-2}, RH_{t-3}$	RH_t
T_{max} & T_{min}	IV	$T_{\min_{t-1}}, T_{\max_{t-1}}$	RH_t
	V	$T_{\min_{t-1}}, T_{\max_{t-1}}, T_{\min_{t-2}}, T_{\max_{t-2}}$	RH_t
	VI	$T_{\min_{t-1}}, T_{\max_{t-1}}, T_{\min_{t-2}}, T_{\max_{t-2}}, T_{\min_{t-3}}, T_{\max_{t-3}}$	RH_t
RH & T_{max} & T_{min}	VII	$RH_{t-1}, T_{\min_{t-1}}, T_{\max_{t-1}}$	RH_t
	VIII	$RH_{t-1}, RH_{t-2}, T_{\min_{t-1}}, T_{\max_{t-1}}, T_{\min_{t-2}}, T_{\max_{t-2}}$	RH_t
	IX	$RH_{t-1}, RH_{t-2}, RH_{t-3}, T_{\min_{t-1}}, T_{\max_{t-1}}, T_{\min_{t-2}}, T_{\max_{t-2}}, T_{\min_{t-3}}, T_{\max_{t-3}}$	RH_t

XGBoost در واقع نسخه بهینه شده الگوریتم Gradient Boosting است که شامل ویژگی‌های اضافی مانند رگولاریزاسیون و هرس درختان[۱] برای جلوگیری از مشکل برازش بیش از حد است (۱۲). XGBoost، یک مدل یادگیری ماشین پرکاربرد برای هر دو مسائل طبقه بندی و رگرسیون است که در فرایند انتخاب توابع ضرر برای ارزیابی مدل انعطاف پذیری را فراهم می کند. این مدل به دلیل پردازش سریع و کارآمد مجموعه داده‌های گسترده، به فناوری بلوک[۲] و موازی‌سازی فرایند آموزش به کمک پردازشگر چند رشته‌ای معروف است. همچنین این مدل به طور مداوم الگوریتم خود را برای دقت بهتر بهبود می‌بخشد (۱۳). به طور کلی، این مدل تابع هدف منحصر به فردی دارد که از دو جزء اصلی تشکیل شده است: (الف) جزء اول که با کاهش پیچیدگی مدل منجر به کاهش مشکل برازش بیش از حد می‌شود و (ب) جزء دوم که از عبارت رگولاریزاسیون و تابع ضرر برای تعیین باقیمانده‌ها (تفاوت بین مقادیر مشاهده شده و پیش‌بینی شده) استفاده می کند (۱۴). با توجه به اینکه این مدل از ترکیب چند درخت تصمیم‌گیری تشکیل شده است، کاربرد باقیمانده‌ها اساساً در جهت اصلاح خطاهای درخت تصمیم‌گیری پیشین در طول هر تکرار الگوریتم می‌باشد. همان طور که اشاره شد، مدل XGBoost به طور مکرر چندین یادگیرنده ضعیف (درخت تصمیم‌گیری) را برای ایجاد یک یادگیرنده قوی ترکیب می‌کند. پیش‌بینی نهایی مدل در واقع جمع وزنی خروجی هر یک از درختان می باشد. این وزن‌ها بر اساس مشتقات مقادیر باقی‌مانده به دست می آیند (۱۵). در این پژوهش، مدل یادگیری ماشین XGBoost در پایتون با استفاده از کتابخانه xgboost پیاده‌سازی شده است. برای بکارگیری این مدل، ابرپارامترهای اولیه مدل که در جدول ۲ معرفی شده‌اند، با استفاده از فرآیند جست و جو شبکه[۳] تنظیم شده اند. همچنین برای سایر ابرپارامترهای مدل مقادیر پیش‌فرض انتخاب شده اند.

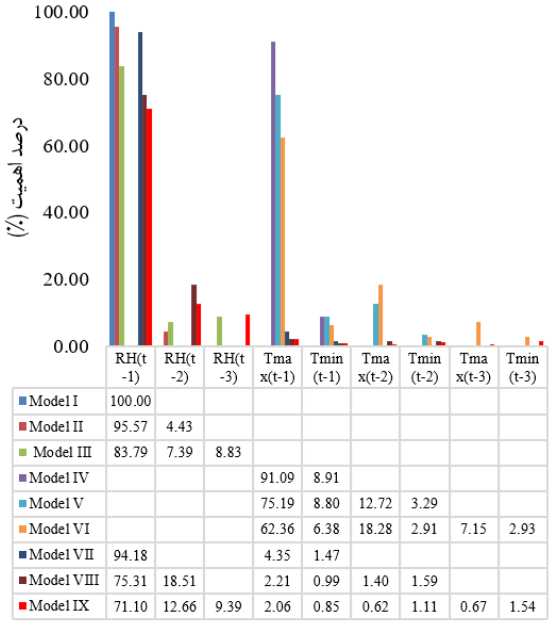
[۱] Tree pruning
[۲] Block Technology
[۳] Grid Search

جدول ۲- ابرپارامترهای مدل XGBoost

ابرپارامتر	توضیحات
n_estimators	تعداد درختان تصمیم گیری که با هم ترکیب می‌شوند.
max_depth	حداکثر عمق هر یک از درختان را مشخص می‌کند، که مقداری مثبت و یا "None" (عدم محدودیت) می‌تواند باشد.
learning_rate	وزنی که به هر یک از درختان در طول هر تکرار الگوریتم انتصاب می‌گردد، که مقداری بین ۰ تا ۱ می‌تواند باشد.
reg_alpha	عبارت رگولاریزاسیون L1
reg_lambda	عبارت رگولاریزاسیون L2
min_split_loss	حداقل کاهش تلفات مورد نیاز برای تقسیم یک گره در هر درخت، که مقادیر صفر تا بینهایت را می‌توان برای آن قرار داد.
min_child_weight	حداقل مجموع وزن مورد نیاز در هر گره، که اگر مجموع وزن داده‌ها در آن گره کمتر از این آستانه باشد، تقسیم‌بندی متوقف می‌شود.

تحلیل اهمیت ویژگی

در الگوریتم XGBoost یک خصوصیت تحت عنوان تحلیل اهمیت ویژگی وجود دارد که می تواند به عنوان نوعی تحلیل حساسیت برای ارزیابی اهمیت نسبی هریک از متغیرهای ورودی مورد استفاده قرار گیرد. مقادیر اهمیت ویژگی را می‌توان با دو معیار تعیین کرد: (الف) معیار وزن که تعداد دفعات استفاده از هر متغیر برای تقسیم داده‌ها در تمام درخت‌های تصمیم‌گیری مدل را مد نظر قرار می دهد و یا (ب) معیار بهره[۱] که میانگین بهبود یافته‌ی بدست آمده در مدل بر اساس استفاده از هر یک از متغیرها را برای تقسیم داده‌ها در نظر می گیرد (۱۶). مقادیر اهمیت بالاتر نشان دهنده تأثیر بیشتر آن متغیر بر پیش‌بینی مدل می باشد. نتایج اهمیت ویژگی می تواند به عنوان یک راهنما برای تجزیه و تحلیل بیشتر و انتخاب بهترین ترکیب از متغیرها استفاده شود. در این پژوهش، مقادیر اهمیت ویژگی با استفاده از دستور پایتون model.feature_importances_ و معیار بهره تعیین می‌شوند. در شکل ۳ نتایج حاصل از تحلیل اهمیت ویژگی XGBoost نشان داده شده است. تحلیل اهمیت ویژگی XGBoost نه تنها می‌تواند اهمیت نسبی هر متغیر ورودی را در یک مدل اندازه‌گیری کند، بلکه می تواند به شناسایی موثرترین متغیرها برای تخمین‌های دقیق‌تر نیز کمک کند. همان طور که در شکل ۳ مشخص است، به طور کلی در تمام سناریوها بیشترین اهمیت را متغیر رطوبت نسبی یک روز قبل دارد، که این نشان دهنده اهمیت زیاد وجود این متغیر به عنوان یکی از متغیرهای ورودی در هر سناریو می باشد. به طور کلی در تمام سناریوهایی که از رطوبت نسبی روزهای پیشین استفاده شده، رطوبت نسبی روز قبل بیشترین درصد اهمیت را دارد. با این حال، با مقایسه سناریوهای III و IX نمی‌توان به طور حتمی گفت که پس از رطوبت نسبی روز گذشته، کدام یک از رطوبت‌های نسبی دو و سه روز گذشته اهمیت بیشتری دارند. پس از رطوبت نسبی، دمای بیشینه یک روز قبل مهم‌ترین متغیر میان متغیرهای دمایی می باشد. پس از آن، در اکثر سناریوها متغیر دمای بیشینه دو روز قبل از دمای کمینه یک روز قبل اهمیت بیشتری دارد. با این حال، این موضوع در رابطه با سناریو IX صدق نمی‌کرد و به طور کلی با مقایسه سناریوهایی که از متغیر دما استفاده می‌کردند، نمیتوان نتیجه قطعی به جز اهمیت بسیار بالای دمای بیشینه روز قبل گرفت و برای نتیجه‌گیری های دقیق‌تر لازم است که تحلیل‌های بیشتری بر روی تاثیر این متغیرها انجام شود.



	RH(t-1)	RH(t-2)	RH(t-3)	Tmax(t-1)	Tmin(t-1)	Tmax(t-2)	Tmin(t-2)	Tmax(t-3)	Tmin(t-3)
Model I	100.00								
Model II	95.57	4.43							
Model III	83.79	7.39	8.83						
Model IV				91.09	8.91				
Model V				75.19	8.80	12.72	3.29		
Model VI				62.36	6.38	18.28	2.91	7.15	2.93
Model VII	94.18			4.35	1.47				
Model VIII	75.31	18.51		2.21	0.99	1.40	1.59		
Model IX	71.10	12.66	9.39	2.06	0.85	0.62	1.11	0.67	1.54

شکل ۳- نتایج تحلیل اهمیت ویژگی برای تمام سناریوهای مختلف

معیارهای آماری

برای ارزیابی عملکرد مدل‌ها از شش معیار آماری مختلف استفاده شده است. این معیارها شامل: (۱) جذر میانگین مربعات خطا (RMSE)، (۲) میانگین قدرمطلق خطا (MAE)، (۳) میانگین قدر مطلق خطای نسبی (MARE)، (۴) حداکثر قدر مطلق خطای نسبی (MXARE)، (۵) ضریب کارآیی نش–ساتکلیف (NSE) و (۶) ضریب تعیین (R^2) (۱۷) می باشند. معادلات ریاضی این معیارها به صورت زیر می‌باشند:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(O_i - P_i)^2}{n}} \qquad (۲)$$

[۱] Gain

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|O_i - P_i| \quad (۳)$$

$$MARE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{O_i - P_i}{O_i}\right| \quad (۴)$$

$$MXARE = max\left(\left|\frac{O_i - P_i}{O_i}\right|\right) \; for\; i = 1, \dots, n \quad (۵)$$

$$NSE = 1 - \frac{\sum_{i=1}^{n}(O_i - P_i)^2}{\sum_{i=1}^{n}\left(O_i - \frac{\sum_{i=1}^{n} O_i}{n}\right)^2} \quad (۶)$$

$$R^2 = \left\{\sum_{i=1}^{n}\left[\left(O_i - \frac{\sum_{i=1}^{n} O_i}{n}\right)\left(P_i - \frac{\sum_{i=1}^{n} P_i}{n}\right)\right] \Big/ \sqrt{\sum_{i=1}^{n}\left(O_i - \frac{\sum_{i=1}^{n} O_i}{n}\right)^2 \sum_{i=1}^{n}\left(P_i - \frac{\sum_{i=1}^{n} P_i}{n}\right)^2}\right\}^2 \quad (۷)$$

که در این معیارها، نمادهای n، O و P به ترتیب نشان‌دهنده تعداد داده‌ها، مقدار مشاهده‌شده رطوبت نسبی و مقدار پیش‌بینی شده رطوبت نسبی می‌باشند. با توجه به تعاریف ارائه شده برای هر معیار، هرچه مقادیر R^2 و NSE بزرگتر (نزدیک به یک) و همچنین هرچه مقادیر RMSE، MAE، MARE و MXARE کمتر (نزدیک به صفر) باشند، نشان دهنده عملکرد بهتر مدل می‌باشد.

نتایج

نتایج بهینه سازی ابرپارامترها

پیش از اینکه مدل برای هر یک از سناریوها آموزش داده شود لازم است ابرپارامترهای مدل XGBoost برای هر یک از این سناریوها به طور بهینه، تنظیم شوند. در این پژوهش از روش جستجوی شبکه برای تنظیم ابرپارمترها استفاده شده و نتایج حاصله در جدول ۳ آورده شده است. پس از آموزش مدل برای هر سناریو براساس ابرپارامترهای بهینه آن مدل، می‌توان نتایج حاصل از هر سناریو را برای داده‌های آزمون بررسی کرد، که در ادامه به آن پرداخته می‌شود.

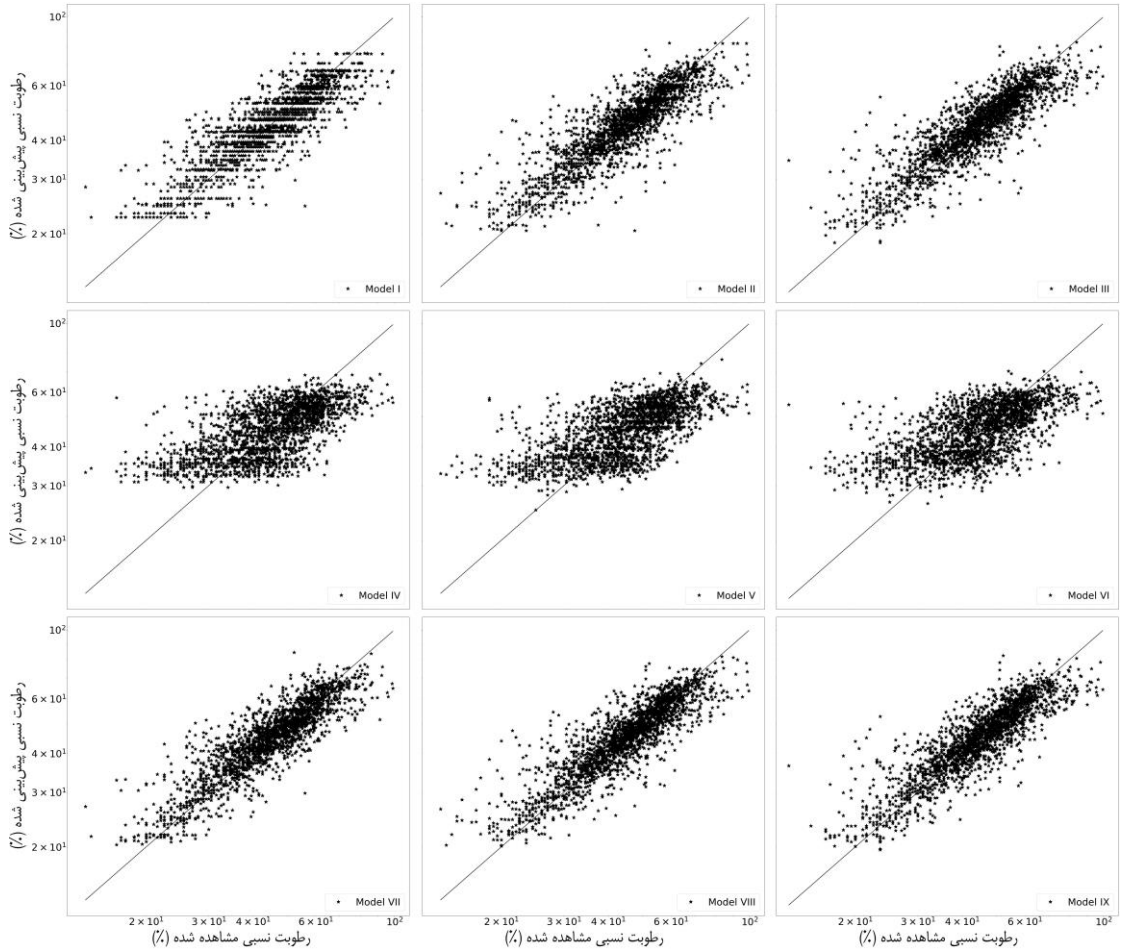
جدول ۳- مقادیر بهینه ابرپارامترها در هر سناریو

مدل	n_estimators	max_depth	learning_rate	reg_alpha	reg_lambda	min_split_loss	min_child_weight
I	۱۰۰	۲	۰,۱	۲	۲	۰	۱
II	۱۰۰	۴	۰,۱	۰,۹	۱,۸	۰	۵
III	۱۰۰	۳	۰,۱	۰,۷	۱,۴	۰	۶
IV	۱۰۰	۲	۰,۲	۰,۰	۲	۰	۱۱
V	۱۰۰	۲	۰,۳	۰,۴	۱,۹	۰	۳
VI	۱۰۰	۳	۰,۲	۱,۲	۱,۷	۰	۳
VII	۱۰۰	۳	۰,۱	۱,۸	۱,۷	۰	۹
VIII	۱۰۰	۳	۰,۱	۰,۹	۱,۶	۰	۱۲
IX	۱۰۰	۳	۰,۱	۱,۳	۱,۵	۰	۶

تحلیل دقت مدل

مقایسه میان رطوبت نسبی مشاهده شده و پیش‌بینی شده توسط مدل یادگیری ماشین برای سناریوهای مختلف در شکل ۴ نمایش داده شده است. محور افقی نمایانگر رطوبت نسبی مشاهده شده به صورت لگاریتمی و محور عمودی نمایانگر رطوبت نسبی پیش‌بینی شده به صورت لگاریتمی توسط مدل می‌باشد. با توجه به شکل ۴، سناریوهایی که فقط برپایه دمای کمینه و بیشینه آموزش داده شده‌اند، پراکندگی زیادی دیده می شود و رطوبت نسبی مشاهده شده با مقادیر پیش‌بینی شده اختلاف زیادی دارد. در سوی دیگر، در سناریوهای مبتنی بر رطوبت نسبی، مقادیر محاسبه شده به مقادیر مشاهده شده نزدیک تر است. این نشان می‌دهد که استفاده از رطوبت نسبی روزهای گذشته برای پیش‌بینی رطوبت نسبی روز آینده، باعث افزایش دقت سناریو شده است.

البته سناریو I نوعی روند پله پله‌ای را نشان می‌دهد که منجر به بروز خطا شده است. به عبارت دیگر، در این سناریو به ازای یک رطوبت نسبی مشاهده شده، چندین رطوبت نسبی مختلف پیش‌بینی شده که این امر نشان دهنده عملکرد ضعیف‌تر این سناریو نسبت به سایر سناریوها می‌باشد. با مقایسه نمودارهای مربوط به سناریوهای بر پایه دما و رطوبت نسبی با نمودارهایی که فقط بر پایه رطوبت نسبی می باشند، تا حدودی می‌توان نتیجه گرفت که اضافه کردن متغیرهای دمای کمینه و بیشینه باعث بهبود همبستگی بیشتر میان رطوبت نسبی مشاهده شده و پیش‌بینی شده توسط سناریو می‌شود. اگر چه نمودارهای موجود در شکل ۴ دید کلی نسبت به عملکرد سناریوها می‌دهند، اما برای بررسی دقیق‌تر عملکرد هر سناریو و ارزیابی و نتیجه گیری جامع‌تر آنها نیاز به استفاده از معیارهای آماری می‌باشد که در بخش بعد به آن پرداخته شده است.



شکل ۴- نتایج همبستگی میان رطوبت نسبی مشاهده شده و پیش‌بینی شده در هر یک از مدل‌ها برای داده‌های آزمون

نتایج معیارهای عملکرد

همانگونه که ذکر گردید در این پژوهش برای ارزیابی میزان دقت سناریوهای مختلف از معیارهای آماری RMSE، MAE، NSE، R^2، MXARE و MARE استفاده شده است. با توجه به تعریف هر یک از معیارها، مقادیر معیارهایRMSE ، MAE، MXARE و MARE هرچه به صفر نزدیک‌تر باشند و در معیارهای R^2 و NSE هرچه به یک نزدیک تر باشند، نشان دهنده دقت بالاتر سناریو می‌باشند. نتایج معیارهای آماری سناریوهای مختلف در جدول ۴ نمایش داده شده است. سناریو VI ضعیف‌ترین عملکرد نسبت به سایر سناریوها در داده های آزمون را داشته اند ($RMSE = 10.41$، $MARE = 0.20$، $NSE = 0.41$، $R^2 = 0.41$، $MXARE = 3.21$ و $MARE = 0.20$). به طور کلی، سناریوهای فقط بر پایه دمای بیشینه و کمینه ضعیف‌ترین عملکرد را داشتند. در سوی دیگر، بهترین عملکرد برای داده‌های آزمون را سناریو VII، با مقادیر RMSE، MAE، NSE، R^2، MXARE و MARE به ترتیب ۶,۷۳، ۴,۸۲، ۰,۷۵، ۰,۷۵، ۱ و ۰,۱۱، از خود نشان داد. در همین راستا، به طور کلی عملکرد سناریوهایی که در آنها هم از رطوبت نسبی و هم از دمای بیشینه و کمینه استفاده شده است بهتر از سایر سناریوها بودند. با این حال علاوه بر داده‌های آزمون، عملکرد هر سناریو نسبت به داده‌های آموزش نیز حائز اهمیت است و برای ارزیابی عملکرد هر سناریو، باید عملکرد آن سناریو در هر دو سری داده‌های آموزش و آزمون در کنار هم بررسی شود. با توجه به اینکه در این پژوهش از ۶ معیار آماری مختلف برای ارزیابی عملکرد سناریوها استفاده شده است، برای ارزیابی جامع‌تر عملکرد هر سناریو، نیازمند یک روش و سیستم رتبه بندی جامع مشتمل بر هر۶ معیار می باشد که در ادامه به شرح آن پرداخته می‌شود.

جدول ۴- نتایج معیارهای آماری مختلف برای سناریوهای مختلف

مدل	RMSE		MAE		NSE		R^2		MXARE		MARE	
	آموزش	آزمون	آموزش	آزمون	آموزش	آزمون	آموزش	آزمون	آموزش	آزمون	آموزش	آزمون
I	7.02	6.93	4.98	4.94	0.73	0.74	0.73	0.74	2.14	1.10	0.12	0.11
II	6.57	7.02	4.70	4.89	0.76	0.73	0.76	0.73	2.05	1.16	0.11	0.11
III	6.54	7.05	4.67	4.96	0.77	0.73	0.77	0.73	2.11	1.65	0.11	0.11
IV	10.24	10.22	8.05	8.05	0.43	0.42	0.43	0.43	3.19	2.49	0.20	0.19
V	10.03	10.27	7.90	8.00	0.45	0.43	0.45	0.43	3.14	2.10	0.20	0.20
VI	9.72	10.41	7.65	8.08	0.48	0.41	0.48	0.41	2.45	3.21	0.19	0.20
VII	6.69	**6.73**	4.77	**4.82**	0.76	**0.75**	0.76	**0.75**	**2.03**	**1.00**	0.11	**0.11**
VIII	6.45	6.90	4.63	4.84	0.77	0.74	0.77	0.74	2.04	1.20	0.11	0.11
IX	**6.33**	6.89	**4.54**	4.88	**0.78**	0.74	**0.78**	0.74	2.03	1.81	**0.11**	0.11

نتایج رتبه بندی سناریوهای مختلف

با توجه به اینکه در این پژوهش از ۶ معیار آماری مختلف در دو سری داده آموزش و آزمون برای ارزیابی عملکرد هر سناریو استفاده شده است، لذا ۱۲ معیار برای سنجش هر سناریو وجود دارد، که برای بررسی عملکرد و انتخاب بهترین سناریو نیاز به یک سیستم رتبه بندی جامع می‌باشد که تمام ۱۲ معیار را در کنار یکدیگر بررسی کند. در این پژوهش، ابتدا هر یک از معیارها در هر سری داده، از ۱ (بهترین) تا ۹ (بدترین)، رتبه بندی می‌شوند. به عنوان مثال، سناریویی که کمترین مقدار RMSE برای داده‌های آموزش را دارد رتبه ۱ و به همین ترتیب سناریویی که بیشترین مقدار RMSE را در آن سری داده دارد رتبه ۹ را کسب می‌کند. سپس، رتبه‌های بدست آمده برای هر ۶ معیار را برای سری‌های آموزش و آزمون بدست آورده و به طور جداگانه با یکدیگر جمع می شوند. سپس به نتایج حاصله به صورت صعودی مرتب می‌شوند. کوچکترین مقدار حاصل رتبه ۱ و بزرگترین مقدار رتبه ۹ را دریافت می‌کند. به این صورت ۱۲ معیار مختلف اکنون در ۲ ستون رتبه سناریوها نسبت به سری داده آموزش و آزمون خلاصه می‌شود. در نهایت مجددا مقادیر این دو ستون با هم جمع، به طور صعودی مرتب و رتبه بندی می‌شوند. بدین صورت رتبه نهایی هر یک از سناریوها بدست می‌آید. نتایج رتبه بندی سناریو ها با توجه به شیوه ذکر شده در جدول ۵ نمایش داده شده است.

با توجه به نتایج جدول ۵، سناریو VIII هم در داده های آزمون و هم در داده‌های آموزش رتبه دوم را کسب نموده که این عملکرد با ثبات و بسیار خوب این سناریو منجر به این شد که در کنار سناریو IX به طور مشترک در رتبه اول قرار گیرند. به عبارتی سناریوهایی که مبتنی بر رطوبت نسبی و دمای بیشینه و دمای کمینه مربوط به دو روز یا سه روز قبل باشند، بهترین عملکرد را نشان می دهند و رطوبت نسبی روز آینده را دقیق تر پیش بینی می کنند. به طور کلی سناریوهای مبتنی بر رطوبت نسبی و دما بهترین عملکرد، سناریوهای مبتنی بر تنها رطوبت نسبی عملکرد قابل قبول و سناریوهای مبتنی بر تنها دما ضعیف‌ترین عملکرد را نشان دادند.

جدول ۵- نتایج رتبه‌بندی عملکرد سناریوهای مختلف

مدل	RMSE		MAE		NSE		R^2		MXARE		MARE		مجموع رتبه ها		رتبه در مجموعه		مجموع رتبه ها	رتبه کل
	آموزش	آزمون	آموزش	آزمون	آموزش	آزمون	آموزش	آزمون	آموزش	آزمون	آموزش	آزمون	آموزش	آزمون	آموزش	آزمون		
I	6	4	6	5	6	4	6	4	6	2	6	4	36	23	6	4	10	4
II	4	5	4	4	4	5	4	5	4	3	4	6	24	28	4	5	9	3
III	3	6	3	6	3	6	3	6	5	5	3	5	20	34	3	6	9	3
IV	9	7	9	8	9	8	9	8	9	8	9	7	54	46	9	8	17	7
V	8	8	8	7	8	7	8	7	8	7	8	8	48	44	8	7	15	5
VI	7	9	7	9	7	9	7	9	7	9	7	9	42	54	7	9	16	6
VII	5	1	5	1	5	1	5	1	1	1	5	1	26	6	5	1	6	2
VIII	2	3	2	2	2	2	2	2	3	4	2	3	13	16	2	2	4	1
IX	1	2	1	3	1	3	1	3	2	6	1	2	**7**	19	1	3	4	1

بحث

متغیرهای آب و هوایی تاثیر زیادی بر انسان و طبیعت دارند. این متغیرها در فرآیندهای طبیعی همچون چرخه هیدرولوژیکی نقش بسزایی دارند و بر بخش‌های مهمی از زندگی انسان، همچون کشاورزی و امنیت

غذایی، تاثیر می‌گذارند. رطوبت نسبی یک عامل مهم در چرخه هیدرولوژیکی است و با تاثیرگذاری بر تبخیرتعرق نقش بسزایی در بخش کشاورزی و زندگی انسان دارد، به همین دلیل پیش‌بینی آن برای انسان بسیار مهم است. در سال‌های اخیر پژوهشگران برای تخمین برخی از متغیرهای آب و هوایی از مدل‌های یادگیری ماشین استفاده کرده‌اند و نتایج بسیار امیدوار کننده‌ای از آن گرفته‌اند (۴-۷). یکی از مدل‌های نوین یادگیری ماشین که در مطالعات پیشین برای تخمین تبخیرتعرق مرجع، نتایج بسیار خوبی از خود نشان داده است، مدل XGBoost می‌باشد. در این پژوهش برای اولین بار از این مدل جهت تخمین رطوبت نسبی استفاده شده است. با توجه به نتیجه تحقیق سایر محققین مبنی بر اینکه دو پارامتر رطوبت نسبی و میزان دمای حداقل و حداکثر روزهای قبل، بیشترین تاثیر را در تخمین رطوبت نسبی روز آینده دارد و همچنین وجود آمار تنها این دو پارامتر در بسیاری از نقاط کشور، سناریوهای مختلفی مشتمل بر این دو پارامتر با استفاده از مدل مذکور مورد مطالعه قرار گرفته اند. با توجه به آزمون همبستگی پیرسون بر اساس ۳۰ سال داده ایستگاه هواشناسی باجگاه فارس، مبنی بر اینکه داده های مربوط به ۳ روز قبل بیشترین تاثیر را در میزان رطوبت روز آینده دارند، رطوبت نسبی، دمای کمینه و دمای بیشینه یک تا سه روز قبل به عنوان داده‌های ورودی مدل استفاده شده اند. همچنین بر اساس اینکه فقط از داده‌های رطوبت نسبی استفاده شود، یا فقط از داده‌های مربوط به دما استفاده شود و یا از هر سه رطوبت نسبی و دمای بیشینه و کمینه استفاده شود، به طور کلی ۹ سناریو مختلف بدست می آید که بر اساس اعداد لاتین از I تا IX نام گذاری شده اند. پس از بهینه سازی ابرپارامترهای مدل، نتایج حاصل از هر یک از ۹ سناریو در جدول‌های ۴ و ۵ و شکل‌های ۳ و ۴ نشان داده شده اند. به طور کلی نشان داده شد که استفاده از رطوبت نسبی روزهای پیشین در سناریوها به عنوان متغیر ورودی مدل می‌توان دقت مدل را بسیار افزایش دهد. همان طور که در شکل ۳ نشان داده شد، سناریوهایی که تنها از دمای بیشینه و دمای کمینه روزهای پیشین برای تخمین رطوبت نسبی استفاده می‌کردند، پراکندگی زیادی را نشان دادند. این نتیجه در جدول ۴ نیز بر اساس معیارهای آماری تایید شد و با توجه به سیستم رتبه بندی ارائه شده در جدول ۵ این سناریوها ضعیف ترین عملکرد را نشان دادند و رتبه‌های انتهایی را میان سایر مدل‌ها کسب کردند. مدل‌هایی که تنها از رطوبت نسبی روزهای پیشین استفاده کردند، بهبود عملکرد قابل توجهی را از خود نشان دادند. به طوری که مقادیر R2 برای داده‌های آزمون در این مدل ها بین ۰,۷۳ تا ۰,۷۴ بود، در حالی که این مقدار برای مدل‌هایی که تنها از متغیرهای دما بهره می‌بردند بین ۰,۴۱ تا ۰,۴۳ بود. اضافه کردن متغیرهای دما به رطوبت نسبی روزهای پیشین منجر به بهبود نسبی عملکرد بیشتر در سناریوها شد به طوری که مقادیر R2 برای داده‌های آزمون در این سناریوها به ۰,۷۵ رسید. شکل ۳ نیز همبستگی بسیار بهتر این سناریوها نسبت به سناریوهایی که فقط از متغیرهای دما استفاده کرده بودند را نشان می‌دهد. با توجه به سیستم رتبه بندی، مدل IX در کنار مدل VIII بهترین عملکرد را داشتند. لازم به ذکر است که مدل VIII نسبت به هر دو سری داده آموزش و آزمون نتیجه یکسان از خود نشان داد و در هر دو سری داده رتبه دوم را کسب نمود. در حالی که سناریو IX نسبت به داده‌های آموزش عملکرد بسیار بهتری را نشان داد و جایگاه نخست را در آن سری داده کسب نمود. عملکرد ضعیف‌تر این سناریو نسبت به داده‌های آزمون، با اینکه تا حدودی می‌توان نشان دهنده موضوع برازش بیش از حد باشد، به حدی نبود که مانع کسب رتبه نخست توسط این سناریو شود. پس از این دو، سناریو VII در جایگاه سوم قرار گرفت. این مدل نسبت به داده‌های آزمون عملکرد بسیار خوبی را از خود نشان داد و جایگاه نخست را در این سری داده کسب کرد. اما عملکرد ضعیف‌تر آن در داده‌های آموزش می‌تواند نشان‌دهنده این باشد که این سناریو علی رغم نتایج بسیار خوبی که برای داده‌های آزمون نشان می‌دهد، نمی‌توان به اندازه دو مدل پیشین قابل اتکا باشد. پس از این سناریو، سناریوهای II و III مشترکا در رتبه سوم قرار گرفتند در حالی که سناریو I رتبه چهارم را کسب کرد. همان طور که در شکل ۳ می‌توان مشاهده کرد، سناریو I به ازای هر رطوبت نسبی مشاهده شده، ممکن است چند مقدار مختلف را به عنوان پیش‌بینی خود ارائه دهد که این مشکل منجر به قرارگیری این سناریو در جایگاه پایین‌تر شده است. خصوصیت ویژه مدل XGBoost تحت عنوان تحلیل اهمیت ویژگی که نتایج آن در شکل ۴ نشان داده شده است، اولویت انتخاب متغیرهای ورودی مدل را بر اساس میزان اهمیت و تاثیر متغیر بر عملکرد مدل را میسر می سازد. بر اساس این تحلیل، در تمام سناریوها اهمیت متغیر رطوبت نسبی به خصوص رطوبت نسبی یک روز قبل بسیار بیشتر از دما بوده است که این با نتایج آماری بدست آمده همخوانی دارد. پس از رطوبت‌های نسبی، دمای بیشینه یک روز قبل مهم‌ترین متغیر بر اساس نتایج بود. در رابطه با اینکه بین متغیر رطوبت نسبی دو و سه روز قبل و همچنین میان سایر متغیرهای دما، کدام یک ارجحیت بیشتری بر دیگری دارد، با توجه به نتایج متفاوتی که هر مدل نشان داده نمی‌توان با قطعیت نظر داد. به طور کلی تحلیل اهمیت ویژگی نشان دهنده این است که هر متغیر به چه مقدار در تقسیم داده‌ها در برگ‌های درختان تصمیم‌گیری نقش داشته‌اند. به عبارت دیگر، اهمیت بالای رطوبت نسبی یک روز گذشته نشان دهنده این است که در یک مجموعه داده ورودی، وجود این متغیر و همچنین اندازه‌گیری دقیق آن جهت استفاده از آن برای مدل آموزش داده شده در این پژوهش بسیار حائز اهمیت است. لازم به ذکر است که نتایج بدست آمده تنها مختص به داده‌ها و سناریوهای به کار رفته در این پژوهش بوده و از آنجایی که تحلیل اهمیت ویژگی اهمیت نسبی رطوبت نسبی یک روز قبل را نشان می‌دهد، می‌توان توصیه کرد که در مطالعات آینده علت این میزان تاثیر زیاد رطوبت نسبی روز پیشین مورد بررسی قرار گیرد تا درک عمیق‌تری از چنین تأثیری به دست آید. به طور کلی این نتیجه با نتایج مطالعات پیشین نیز همخوانی زیادی داشت.

نتیجه‌گیری و پیشنهادها

با توجه به نقش محوری رطوبت نسبی در چرخه هیدرولوژیکی و اثرات متعاقب آن بر کشاورزی و رفاه انسان، پیش‌بینی دقیق رطوبت نسبی از اهمیت ویژه‌ای برخوردار است. در این پژوهش، با استفاده از مدل یادگیری ماشین XGBoost و ۳۰ سال داده از ایستگاه هواشناسی

دانشکده کشاورزی دانشگاه شیراز واقع در باجگاه، رطوبت نسبی با استفاده از ترکیب‌های مختلفی از متغیرهای رطوبت نسبی و دمای بیشینه و کمینه روزهای پیشین تخمین زده شده است. نتایج بدست آمده برتری سناریوهایی که رطوبت نسبی روزهای قبل را در نظر می گیرند، در مقایسه با سناریوهایی که صرفاً به متغیرهای دما متکی هستند، نشان می‌دهد. با توجه به نتایج بدست آمده، سناریوهایی که مبتنی بر رطوبت نسبی و دمای بیشینه و دمای کمینه مربوط به دو روز یا سه روز قبل باشند، بهترین عملکرد را نشان دادند و مشترکا رتبه نخست را کسب کردند. پس از آن، به ترتیب سناریوهای مبتنی بر تنها رطوبت نسبی روزهای قبل و سناریوهایی که فقط بر پایه دما قرار گرفته اند بهترین عملکرد را دارند. به عبارتی دیگر با توجه به نتایج حاصله می‌توان نتیجه گرفت که رطوبت نسبی روزهای پیشین به عنوان متغیر ورودی نسبت به دما تاثیر بیشتری در پیش بینی رطوبت نسبی دارد اما در نظر گرفتن هر دو متغیر رطوبت و دما به عنوان متغیرهای ورودی مدل، دقت پیش بینی رطوبت نسبی روز آینده را افزایش می دهد. تحلیل اهمیت ویژگی، اهمیت متغیرهای رطوبت نسبی، به خصوص رطوبت نسبی یک روز قبل، را نشان می‌دهد. با این حال نتایج این تحلیل مختص به شرایط خاص این پژوهش بوده و به طور کلی به عنوان راهنمایی بیشتر جهت انتخاب متغیر ورودی می‌توان از آن بهره برد. همانگونه که قبلا ذکر گردید با توجه به نتیجه تحقیق سایر محققین و عدم دسترسی به دیگر متغیرها در بسیاری از ایستگاه های هواشناسی در این تحقیق صرفا از دو متغیر رطوبت نسبی و دما استفاده شده است. لذا در پژوهش‌های آتی پژوهشگران می‌توانند اثر سایر متغیرهای آب و هوایی را بر رطوبت نسبی بررسی کنند. همچنین به کمک سایر مدل‌های یادگیری ماشین می‌توانند رطوبت نسبی را تخمین بزنند و عملکرد آن مدل ها را با مدل XGBoost مقایسه کنند.

ملاحظات اخلاقی پیروی از اصول اخلاق پژوهش

همکاری مشارکت‌کنندگان در تحقیق حاضر به صورت داوطلبانه و با رضایت آنان بوده است.

حامی مالی

هزینه تحقیق حاضر توسط نویسندگان مقاله تامین شده است.

تعارض منافع

بنابر اظهار نویسندگان، مقاله حاضر فاقد هرگونه تعارض منافع بوده است.

References

1. Khatibi, R., L. Naghipour, M.A. Ghorbani, and M.T. Aalami, Predictability of relative humidity by two artificial intelligence techniques using noisy data from two Californian gauging stations. Neural Computing and Applications, 2013. **23**(7): p. 2241-2252.
2. Tao, H., S.M. Awadh, S.Q. Salih, S.S. Shafik, and Z.M. Yaseen, Integration of extreme gradient boosting feature selection approach with machine learning models: application of weather relative humidity prediction. Neural Computing and Applications, 2022. **34**(1): p. 515-533.
3. Allen, R.G., L.S. Pereira, D. Raes, and M. Smith, Crop evapotranspiration-Guidelines for computing crop water requirements-FAO Irrigation and drainage paper 56. Fao, Rome, 1998. **300**(9): p. D05109.
4. Fan, J., et al., Evaluation of SVM, ELM and four tree-based ensemble models for predicting daily reference evapotranspiration using limited meteorological data in different climates of China. Agricultural and Forest Meteorology, 2018. **263**: p. 225-241.
5. Ferreira, L.B. and F.F. da Cunha, New approach to estimate daily reference evapotranspiration based on hourly temperature and relative humidity using machine learning and deep learning. Agricultural Water Management, 2020. **234**: p. 106113.
6. Bellido-Jiménez, J.A., J. Estévez, and A.P. García-Marín, New machine learning approaches to improve reference evapotranspiration estimates using intra-daily temperature-based variables in a semi-arid region of Spain. Agricultural Water Management, 2021. **245**: p. 106558.
7. Abdallah, M., et al., Reference evapotranspiration estimation in hyper-arid regions via D-vine copula based-quantile regression and comparison with empirical approaches and machine learning models. Journal of Hydrology: Regional Studies, 2022. **44**: p. 101259.
8. Bayatvarkeshi, M., K. Mohammadi, O. Kisi, and R. Fasihi, A new wavelet

conjunction approach for estimation of relative humidity: wavelet principal component analysis combined with ANN. Neural Computing and Applications, 2020. **32**(9): p. 4989-5000.
9. Merabet, K. and S. Heddam, Improving the accuracy of air relative humidity prediction using hybrid machine learning based on empirical mode decomposition: a comparative study. Environmental Science and Pollution Research, 2023. **30**(21): p. 60868-60889.
10. Gezgen, D., Comparison of missing data imputation methods applied to daily temperature and precipitation data in Turkey. 2023, Middle East Technical University.
11. Bisong, E., Building machine learning and deep learning models on Google cloud platform. 2019: Springer.
12. Bandara, A., et al. A generalized ensemble machine learning approach for landslide susceptibility modeling. in Data Management, Analytics and Innovation: Proceedings of ICDMAI 2019, Volume 2. 2020. Springer.
13. Lu, H. and X. Ma, Hybrid decision tree-based machine learning models for short-term water quality prediction. Chemosphere, 2020. **249**: p. 126169.
14. Katipoğlu, O.M. and M. Sarıgöl, Prediction of flood routing results in the Central Anatolian region of Türkiye with various machine learning models. Stochastic Environmental Research and Risk Assessment, 2023: p. 1-20.
15. Han, Y., et al., Coupling a bat algorithm with xgboost to estimate reference evapotranspiration in the arid and semiarid regions of china. Advances in Meteorology, 2019. **2019**: p. 1-16.
16. Piraei, R., S.H. Afzali, and M. Niazkar, Assessment of XGBoost to Estimate Total Sediment Loads in Rivers. Water Resources Management, 2023.
17. Piraei, R., M. Niazkar, and S.H. Afzali, Assessment of data-driven models for estimating total sediment discharge. Earth Science Informatics, 2023. **16**(3): p. 2795-2812.