Physical Geography Research Quarterly

Online ISSN: 2423-7760

Journal Homepage: jphgr.ut.ac.ir

Institute of Geography

Application of Entropy Theory and Principal Component Analysis to Determine Input Variables for Estimating Solar Radiation using Machine Learning Algorithms

Somayeh Soltani-Gerdefaramarzi [1], Mozhgan Askarizadeh [2]

1. (Corresponding Author) *Department of Water Sciences and Engineering, Faculty of Agriculture and Natural Resources, Ardakan University, Ardakan, Iran*
Email: ssoltani@ardakan.ac.ir
2. *Department of Computer Engineering, Faculty of Engineering, Ardakan University, Ardakan, Iran*
Email: maskari@ardakan.ac.ir

Article Info

Article type:
Research Article

Article History:
Received:
4 August 2024
Received in revised form:
18 October 2024
Accepted:
24 November 2024
Available online:
25 December 2024

Keywords:
Geometric Specifications,
Machine Learning,
Solar Zenith Angle,
Radiation,
Yazd.

ABSTRACT

Solar radiation is crucial in energy balance models and plant growth simulations. This research investigates the performance of Principal Component Analysis (PCA) and Shannon Entropy Theory (ENT) in determining the input for machine learning models – Random Forest (RF), Linear Regression (LR), Support Vector Regression (SVR), K-Nearest Neighbors (KNN), Decision Tree (DT), and XGBoost (XGB) – for estimating solar radiation at the Yazd synoptic station between 2006 and 2023. Daily data for average temperature, minimum temperature, maximum temperature, sunshine hours, relative humidity, and solar radiation were obtained from the Meteorological Organization. Extraterrestrial radiation, the relative Earth-Sun distance, solar declination angle, and maximum sunshine hours were calculated using existing formulas and selected as inputs for the pre-processing methods. The results of machine learning algorithms indicated their acceptable accuracy in estimating solar radiation. By reducing the dimensionality of the input data to the machine learning algorithms, the results showed that the Principal Component Analysis (PCA) method increased the model's accuracy. Among the models used, the PCA-SVR model showed the best result at the Yazd station with a coefficient of determination of 0.923 and an accuracy of 92.84%. It is worth mentioning that the Shannon entropy theory method failed to improve the modeling results compared to the method without initial pre-processing. This analysis shows that using dimensionality reduction techniques and selecting appropriate models can lead to greater accuracy and less computational complexity in prediction problems. However, sufficient care should be taken when selecting a pre-processing model for the initial data.

Cite this article: Soltani-Gerdefaramarzi, S., & Askarizadeh, M. (2024). Application of Entropy Theory and Principal Component Analysis to Determine Input Variables for Estimating Solar Radiation using Machine Learning Algorithms. *Physical Geography Research Quarterly, 56* (4), 73-87.
http://doi.org/10.22059/JPHGR.2025.386916.1007858

Publisher: University of Tehran Press

Extended Abstract

Introduction

In terms of selecting all influential parameters and the lack of statistical information, the complexity of meteorological and hydrological systems makes complete modeling of these systems impossible. Using system modeling based on mathematical relationships is of interest in such conditions. Solar radiation is one of the important and effective meteorological variables in estimating evapotranspiration and the water needs of plants, and it is the energy source for all atmospheric and surface processes. Although the measurement of this variable has a relatively long history in Iran, due to the high costs of measuring instruments, many existing stations in the country lack a radiometer or pyranometer, or face issues such as calibration problems and the accumulation of water and dust on the sensor. Even at weather stations that measure radiation, there are days when radiation data is not recorded, or unrealistic values outside the expected range are observed due to equipment malfunctions or other issues. On the other hand, due to the many factors affecting solar radiation studies, it is impossible to include all elements in the relevant equations. As a result, only a limited number of these variables are applicable for estimating solar radiation using empirical and semi-empirical equations. In recent years, many researchers have focused their studies on using data mining methods and mathematical modeling to estimate solar radiation.

Methodology

The data used in this research are daily climatic variables measured at the Yazd synoptic station from 2006 to 2023. The Yazd station is located at 31.8974° North latitude and 54.3569° East longitude, at an altitude of 1216 meters above sea level. The average solar and extraterrestrial radiation at the Yazd synoptic station are 19.35 and 32 megajoules per square meter per day. The ratio of sunshine hours to maximum possible sunshine hours is 0.75, the average relative humidity is 27%, and the average temperature is 28°C. Data from 2006 to 2014 were used for calibrating the equations, and data from 2015 to 2023 were used for evaluating the results. Extraterrestrial radiation and maximum daily sunshine hours, which depend on the geographical latitude and day number based on the Gregorian calendar, were calculated using the relationships presented by Duffie and Beckman (1991). This research investigates the performance of Principal Component Analysis (PCA) and Shannon Entropy (ENT) for determining the input variables of Random Forest (RF), Linear Regression (LR), Support Vector Regression (SVR), K-Nearest Neighbors (KNN), Decision Tree (DT), and XGBoost (XGB) machine learning models in estimating solar radiation. Daily data for mean temperature, minimum temperature, maximum temperature, sunshine hours, relative humidity, and solar radiation were obtained from the Meteorological Organization. Extraterrestrial radiation, relative earth-sun distance, solar declination angle, and maximum sunshine hours were calculated using existing relationships and selected as inputs for the preprocessing methods.

Results and discussion

Results showed that, in the training phase, the employed models were well-trained and exhibited acceptable results. In the testing phase, the modeling results for the raw input data (without pre-processing) also yielded satisfactory results for all models. The coefficient of determination varied between 0.790 for the KNN model and 0.893 for the SVR model, depending on the algorithms used. In other words, regarding R-squared values, all the algorithms used showed good results for solar radiation prediction. Considering all evaluation metrics, the Support Vector Regression (SVR) algorithm performed better than other models to predict solar radiation with RMSE = 1.732, MSE = 0.003, MAE = 0.826, R^2 = 0.893, and an accuracy of 90.75%. Results showed that using Principal Component Analysis (PCA) for dimensionality reduction, the first principal component accounted for approximately 49% of the variance, and the second principal component accounted for approximately 36%. The first two principal components comprised over 85% of the original data's variability; therefore, these

two components were considered as inputs for the predictive models to estimate solar radiation. Based on the training results, the PCA-DT and ENT-DT models exhibited the best performance in solar radiation estimation and model training at the Yazd station, achieving zero mean squared error and mean absolute percentage error, and a coefficient of determination of 1.00 compared to other models. The results of the model testing section indicate that the PCA-SVR model outperforms other methods. As can be seen, the PCA-SVR model, with a coefficient of determination of 0.923 and an accuracy of 92.84%, achieved the best results among the mentioned models at Yazd station, exhibiting the lowest error metrics. The ENT-DT model, with a coefficient of determination of 0.535 and an accuracy of 79.34%, showed weaker results among the models used at Yazd station.

Conclusion

Given the importance of accurate solar radiation estimation in hydrological phenomena and the need for advanced methods in its estimation, this research utilized Principal Component Analysis (PCA) and entropy theory for data pre-processing. Model inputs for the estimation models were identified using these two methods. Modeling was performed using Random Forest (RF), Linear Regression (LR), Support Vector Regression (SVR), K-Nearest Neighbors (KNN), Decision Tree (DT), and XGBoost (XGB) models. Entropy theory results indicated that at the Yazd station, solar declination angle, minimum temperature, minimum relative humidity, and average relative humidity were effective variables in estimating solar radiation. Furthermore, PCA reduced the number of input variables to two principal components, and modeling was performed using these two derived input variables. Overall, the modeling results showed that the PCA-SVR model outperformed other models in estimating solar radiation. In general, PCA pre-processing demonstrated that this method determines better inputs for the estimation models. It is worth noting that Shannon's theoretical method did not improve the modeling results compared to the method without pre-processing. This analysis shows that using dimensionality reduction techniques and selecting appropriate models can lead to higher accuracy and lower computational complexity in prediction problems. However, care must be taken when selecting the pre-processing model for the initial data. Similar research using new data or in different geographical conditions could also help further validate the results.

Funding

There is no funding support.

Authors' Contribution

In this study, the authors' contributions are as follows: Somayeh Soltani-Gardfaramarzi was responsible for the study design, data collection, analysis, writing the initial draft, and final editing of the article, and Mojgan Askarizadeh was responsible for modeling and results.

Conflict of Interest

Authors declared no conflict of interest.

Acknowledgments

We are grateful to all the scientific consultants of this paper.

فصلنامه پژوهش‌های جغرافیای طبیعی
شاپا الکترونیکی: 2423-7760
Journal Homepage: jphgr.ut.ac.ir

کاربرد تئوری آنتروپی و تحلیل مؤلفه اصلی جهت تعیین متغیرهای ورودی تخمین تابش خورشیدی با الگوریتم‌های یادگیری ماشین

سمیه سلطانی گردفرامرزی [۱]، مژگان عسکری زاده [۲]

۱- نویسنده مسئول، گروه علوم و مهندسی آب، دانشکده کشاورزی و منابع طبیعی، دانشگاه اردکان، اردکان، ایران. رایانامه: ssoltani@ardakan.ac.ir
۲- گروه مهندسی کامپیوتر، دانشکده فنی و مهندسی، دانشگاه اردکان، اردکان، ایران. رایانامه: maskari@ardakan.ac.ir

اطلاعات مقاله

نوع مقاله:
مقاله پژوهشی

تاریخ دریافت:
۱۴۰۳/۰۵/۱۴
تاریخ بازنگری:
۱۴۰۳/۰۷/۲۷
تاریخ پذیرش:
۱۴۰۳/۰۹/۰۴
تاریخ چاپ:
۱۴۰۳/۱۰/۰۷

واژگان کلیدی:
مشخصات هندسی،
یادگیری ماشین،
زاویه میل خورشیدی،
تابش،
یزد.

چکیده

تابش خورشیدی به‌عنوان یکی از متغیرهای مهم در مدل‌های بیلان انرژی و شبیه‌سازی رشد گیاهان اهمیت زیادی دارد. در این پژوهش عملکرد روش تحلیل مؤلفه اصلی (PCA) و تئوری آنتروپی شانون (ENT) برای تعیین ورودی مدل‌های یادگیری ماشین جنگل تصادفی (RF)، رگرسیون خطی (LR)، ماشین بردار پشتیبان (SVR)، نزدیک‌ترین همسایه (KNN)، درخت تصمیم (DT) و XGBoost(XGB) در برآورد تابش خورشیدی در ایستگاه سینوپتیک یزد در حد فاصل سال‌های ۲۰۰۶ تا ۲۰۲۳ موردبررسی قرار گرفت. متغیرهای میانگین دما، دمای کمینه، دمای بیشینه، ساعات آفتابی، رطوبت نسبی و تابش خورشیدی به‌صورت روزانه از سازمان هواشناسی دریافت و متغیرهای تابش فرازمینی، فاصله نسبی زمین تا خورشید، زاویه میل خورشیدی و حداکثر ساعات آفتابی با روابط موجود محاسبه و به‌عنوان ورودی روش‌های پیش‌پردازش انتخاب شدند. نتایج الگوریتم‌های یادگیری ماشین حاکی از دقت قابل‌قبول آن‌ها در تخمین تابش خورشیدی بود. با کاهش بعد داده‌های ورودی به الگوریتم‌های یادگیری ماشین، نتایج نشان داد که روش تحلیل مؤلفه اصلی دقت مدل را افزایش داد و در بین مدل‌های به‌کاررفته، مدل PCA-SVR با ضریب تبیین ۰/۹۲۳ و دقت ۹۲/۸۴٪ بهترین نتیجه را در ایستگاه یزد نشان داد. لازم به ذکر است که روش تئوری آنتروپی شانون نتوانست نتایج مدل‌سازی را نسبت به روش بدون پیش‌پردازش اولیه بهبود بخشد. این تحلیل نشان می‌دهد که استفاده از تکنیک‌های کاهش ابعاد و انتخاب مدل‌های مناسب می‌تواند منجر به‌دقت بیشتر و پیچیدگی محاسباتی کمتر در مسائل پیش‌بینی شود، هرچند در انتخاب مدل پیش‌پردازش داده‌های اولیه باید دقت کافی داشت.

استناد: سلطانی گردفرامرزی، سمیه و عسکری زاده، مژگان. (۱۴۰۳). کاربرد تئوری آنتروپی و تحلیل مؤلفه اصلی جهت تعیین متغیرهای ورودی تخمین تابش خورشیدی با الگوریتم‌های یادگیری ماشین. *مجله پژوهش‌های جغرافیای طبیعی*، ۵۶ (۳)، ۸۷-۷۳.
http://doi.org/10.22059/JPHGR.2025.386916.1007858

ناشر: مؤسسه انتشارات دانشگاه تهران

مقدمه

پیچیدگی سیستم‌های هواشناسی و هیدرولوژیکی از نظر انتخاب تمام پارامترهای تأثیرگذار و نقص اطلاعات آماری، امکان مدل‌سازی کامل این سیستم‌ها را غیرممکن می‌سازد. در چنین شرایطی استفاده از مدل‌سازی سیستمی که مبنی بر روابط ریاضی باشد، موردتوجه می‌باشد (Abdelhafidi et al., 2021: 207). از مطالعات انجام‌شده در این زمینه می‌توان به مدل‌سازی تبخیر و تعرق گیاه مرجع (شیخ‌الاسلامی و همکاران، ۱۳۹۳: ۴۲۲)، بارش (محمدی و امامقلی زاده، ۱۳۹۵: ۷۷)، ضریب پخشیدگی آلودگی (سلطانی گردفرامرزی و همکاران، ۱۳۹۴: ۸۷)، فرسایش خاک (Boroughani et al., 2022: 25) و.... با روش‌های مختلف یادگیری ماشین و مدل‌سازی ریاضی اشاره کرد. تابش خورشیدی یکی از متغیرهای مهم و مؤثر هواشناسی در برآورد تبخیر و تعرق و نیاز آبی گیاهان است و منشأ انرژی برای همه تحولات جو سطح زمین می‌باشد. اندازه‌گیری این متغیر اگرچه در ایران سابقه نسبتاً طولانی دارد ولی به دلیل هزینه‌های زیاد وسایل اندازه‌گیری در بسیاری از ایستگاه‌های موجود کشور دستگاه تابش‌سنج یا پیرانومتر وجود ندارد و یا مشکلاتی همچون واسنجی آن، تجمع آب و گردوغبار بر روی سنسور آن وجود دارد (Rahimikhoob, 2010: 2132). حتی در ایستگاه‌های هواشناسی هم که تابش را اندازه می‌گیرند، روزهایی وجود دارد که داده‌های تابش ثبت نمی‌شود یا مقادیر غیرواقعی و خارج از بازه مورد انتظار به دلیل نقص دستگاه و یا مشکلات دیگر مشاهده می‌شود (Hunt et al., 1998: 295). هرچند در اغلب این ایستگاه‌ها ساعات آفتابی به‌طور روزانه اندازه گرفته می‌شود. تحقیقات مختلفی برای تخمین تابش خورشیدی با استفاده از داده‌های هواشناسی انجام‌شده و روش‌های زیادی توسعه‌یافته است. محققین به دنبال راهی برای تخمین بهتر و دقیق‌تر تابش خورشیدی می‌باشند. در نتیجه روابط تجربی و رگرسیونی، فنون سنجش‌ازدور و میانیابی خطی توسعه یافتند (Sabziparvar & Shetaee, 2007: 650). از طرف دیگر به دلیل کثرت عناصر مؤثر در مطالعات تابش خورشیدی نمی‌توان تمام عناصر را در معادلات مربوطه وارد کرد. در نتیجه برای تخمین تابش خورشیدی توسط معادلات تجربی و نیمه تجربی تنها تعداد محدودی از این متغیرها کاربرد دارد. در سال‌های اخیر پژوهشگران زیادی مطالعات خود را بر مبنای استفاده از روش‌های داده‌کاوی و مدل‌سازی ریاضی برای تخمین تابش خورشیدی معطوف داشته‌اند (عوض پور و همکاران، ۱۳۹۸: ۱۸۶۱؛ Yadav& Chandel, 2015: 681؛ Olalekan et al., 2018: 9؛ محمدی و همکاران، ۱۳۹۸: ۶۳۲؛ Meenal& Selvakumar, 2018: 331، Radosevic et al., 2020: 104780؛ Abdelhafidi et al., 2021: 209؛ سلطانی گردفرامرزی، ۱۴۰۲: ۸؛ سلطانی گردفرامرزی و مؤمنی، ۱۴۰۲: ۳۰). با توجه به اینکه در بسیاری از ایستگاه‌ها امکان اندازه‌گیری دقیق تابش خورشیدی وجود ندارد و از طرف دیگر لازم است تا متغیرهای هواشناسی کمتری در معادلات مورداستفاده قرار گیرد، لذا بررسی روابط غیرخطی موجود بین متغیرهای هواشناسی مؤثر بر تابش خورشیدی ضروری به نظر می‌رسد. از طرف دیگر قابلیت یک مدل هیدرولوژیکی هم باید از نظر تئوری بکار رفته در آن و هم از نظر میزان داده موردنیاز و موجودیت آن داده‌ها ارزیابی شود. پس اینکه چه تعداد داده و چه تعداد متغیر ورودی از یک پدیده مانند تابش خورشیدی در اختیار داشته باشیم نیز حائز اهمیت است (محمدی و همکاران، ۱۳۹۸: ۶۳۰؛ سلطانی گردفرامرزی، ۱۴۰۲: ۱۰). از طرف دیگر با توجه به عدم دسترسی به همه داده‌های مؤثر بر تابش خورشیدی در اکثر ایستگاه‌های هواشناسی، باید بتوان بر روی متغیرهای موجود تحلیل مناسبی انجام داد تا بتوان تأثیر نسبی آن‌ها را بر تابش خورشیدی بررسی کرد. روش‌های پیش‌پردازش داده‌ها با شناسایی ورودی‌های مؤثر بر کاهش ابعاد ورودی‌های مدل و از بین بردن روند ناایستائی موجود در ورودی‌ها، موجب افزایش دقت و کارایی مدل برای پیش‌بینی پدیده‌ها در آینده خواهد شد (محمدی و همکاران، ۱۳۹۸: ۶۳۲). از جمله روش‌های پیش‌پردازش داده‌ها که در این پژوهش موردبررسی قرارگرفته است، روش تئوری آنتروپی و تحلیل مؤلفه‌های اصلی می‌باشد. لازم به ذکر است که مطالعات زیادی در خصوص برآورد تابش خورشیدی در برخی

ایستگاه‌های ایران انجام‌شده است که تنها از داده‌های هواشناسی استفاده‌شده است. تنها مطالعه موجود که تابش خورشیدی در کرمان را با استفاده از مشخصات هندسی، نجومی، جغرافیایی و هواشناسی برآورد کرده است مربوط به مطالعه صفاریان و محرابی‌پور (۱۳۸۸) می‌باشد. نتایج این پژوهش نشان داد که علاوه بر داده‌های هواشناسی داده‌های هندسی و نجومی نیز در برآورد تابش خورشیدی اهمیت دارند. با توجه به اهمیت تابش رسیده به سطح زمین، هدف از این پژوهش تخمین تابش خورشیدی در ایستگاه یزد با یک اقلیم خشک و گرم است که با استفاده از مدل‌های یادگیری ماشین جنگل تصادفی (RF)، رگرسیون خطی (LR)، ماشین بردار پشتیبان (SVR)، نزدیک‌ترین همسایه (KNN)، درخت تصمیم (DT) و XGB انجام‌شده و سعی گردیده تا تخمین تابش همراه با کاهش پیچیدگی مدل‌سازی به‌وسیله پیش‌پردازش داده‌ها با روش‌های تحلیل مؤلفه اصلی و تئوری آنتروپی شانون و ترکیبی از داده‌های هواشناسی، هندسی، نجومی و جغرافیایی صورت گیرد و نتایج به‌دست‌آمده با نتایج بدون کاهش بعد داده‌های ورودی به مدل‌ها، مقایسه گردد. کاربرد داده‌های هواشناسی، هندسی و نجومی به‌طور هم‌زمان و استفاده از روش‌های کاهش بعد داده‌های ورودی در تخمین تابش خورشیدی از نوآوری‌های این پژوهش می‌باشد.

روش پژوهش

روش آنالیز مؤلفه‌های اصلی

در روش آنالیز مؤلفه‌های اصلی P متغیر اصلی همبسته به p مؤلفه غیرهمبسته یا متعامد تبدیل می‌شوند. با اعمال PCA متغیرهای ورودی اصلی به متغیرهای جدید که بدون همبستگی می‌باشند، تبدیل می‌شوند. مؤلفه‌های ایجادشده ترکیبی خطی متغیرهای اصلی می‌باشند. اگر در P متغیر اصلی تنها مقداری همبستگی وجود داشته باشد، انجام روش PCA می‌تواند مفید باشد. با انتخاب چند مؤلفه اصلی اول، سایر مؤلفه‌ها از محاسبات بعدی حذف می‌شوند. از نمودار واریزه‌ای (اسکری پلات) که در آن مقادیر ویژه در مقابل شماره مؤلفه‌ها رسم می‌شود، برای تشخیص آستانه حذف استفاده می‌شود. در این روش مرز بین مؤلفه‌های اصلی و غیر اصلی محلی است که نمودار میل به خطی شدن به‌صورت افقی می‌نماید (Liu et al., 2003: 80).

تئوری آنتروپی شانون

شانون در سال ۱۹۴۸ نشان داد که وقایع با احتمال وقوع زیاد، اطلاعات واضح و مشهودی در اختیار می‌گذارند و برعکس هر چقدر احتمال وقوع یک رخداد کمتر باشد، اطلاعات حاصل از آن جدیدتر و برای محققین مفیدتر است. به‌عبارت‌دیگر، از تئوری آنتروپی می‌توان به‌عنوان شاخصی برای کمی کردن میزان عدم آگاهی و دانش نسبت به مشخصات یک سامانه، استفاده نمود (Xu et al., 2015: 142). برای به دست آوردن وزن آنتروپی (wj) از رابطه (۱) استفاده می‌شود:

رابطه (۱) $$w_j = \frac{1-e_j}{\sum_{i=1}^{n} 1-e_j}$$

متغیر ej مقدار آنتروپی انتقال اطلاعات را نشان می‌دهد. هر چه قدر مقدار آنتروپی کمتر باشد، تأثیر j بیشتر خواهد بود. آنتروپی انتقال اطلاعات بین دو متغیر i و j به‌صورت زیر محاسبه می‌شود (Shannon, 1948: 640):

رابطه (۲) $$e_j = -\frac{1}{\ln m}\sum_{i=1}^{m} P_{ij} \ln P_{ij}$$

مدل‌های هوش مصنوعی استفاده‌شده

۱- جنگل تصادفی[1]

جنگل تصادفی یک الگوریتم یادگیری ماشین مبتنی بر درخت تصمیم است که از مجموعه‌ای از درخت‌های تصمیم استفاده می‌کند (Breiman, 2001). هر درخت در این الگوریتم روی یک زیرمجموعه تصادفی از داده‌ها ویژگی‌ها آموزش می‌بینند. نتیجه نهایی مدل با رأی‌گیری (برای طبقه‌بندی) یا میانگین‌گیری (برای رگرسیون) از خروجی‌های این درخت‌ها به دست می‌آید. این روش از بیش برازش جلوگیری کرده و معمولاً عملکرد بسیار خوبی روی‌داده‌های پیچیده دارد. ویژگی های این مدل شامل مقاومت در برابر نویز و داده‌های پرت و مناسب برای مسائل طبقه‌بندی و رگرسیون می‌باشد.

۲- رگرسیون خطی[2]

رگرسیون خطی یکی از ساده‌ترین و پرکاربردترین روش‌های یادگیری ماشین است که برای پیش‌بینی متغیر عددی به کار می‌رود (Draper& Smith, 1998). در این الگوریتم، هدف یافتن خطی است که بهترین تطابق را با داده‌ها داشته باشد. این خط به‌گونه‌ای تعریف می‌شود که مجموع مربعات خطاهای پیش‌بینی را به حداقل برساند. ویژگی‌های این مدل ساده و قابل تفسیر بودن و عملکرد خوب در داده‌های خطی یا نزدیک به خطی است.

۳- رگرسیون بردار پشتیبان[3]

رگرسیون بردار پشتیبان نسخه‌ای از ماشین بردار پشتیبان است که برای مسائل رگرسیون استفاده می‌شود (Drucker et al., 1997). این الگوریتم یک مدل پیش‌بینی خطی یا غیرخطی را با استفاده از مفهوم هسته‌ها ایجاد می‌کند. هدف SVRاین است که یک خط یا منحنی پیش‌بینی پیدا کند که بیشترین تعداد نقاط داده‌ها در یک محدوده خطا قرار گیرند. ویژگی‌های این مدل عملکرد خوب در مسائل غیرخطی و حساس به پارامترها و مقیاس داده‌ها است.

۴- نزدیک‌ترین همسایه[4]

نزدیک‌ترین همسایه یک روش یادگیری مبتنی بر نمونه است که پیش‌بینی را بر اساس نزدیک‌ترین نمونه‌ها انجام می دهد (Cover& Hart, 1967). در این روش،K همسایه نزدیک‌ترین داده‌ها بر اساس فاصله (مانند اقلیدسی) مشخص می شوند و خروجی پیش‌بینی با رأی‌گیری (برای طبقه‌بندی) یا میانگین‌گیری (برای رگرسیون) انجام می‌شود. ویژگی‌های این مدل ساده و قابل‌فهم بودن و عملکرد خوب در داده‌های کوچک اما حساس به نویز و مقیاس داده‌هاست.

۵- درخت تصمیم[5]

درخت تصمیم یک الگوریتم یادگیری ماشین است که از ساختار درختی برای تقسیم‌بندی داده‌ها استفاده می‌کند (Quinlan, 1986). این الگوریتم داده‌ها را با استفاده از ویژگی‌هایی که بیشترین اطلاعات را ارائه می‌دهند، به شاخه‌های مختلف تقسیم می‌کند. ساده و قابل تفسیر بودن از مزایای این مدل است ولی ممکن است دچار بیش برازش شود، مگر اینکه محدودیت‌هایی اعمال شود.

۶- XGBoost[6]

یک الگوریتم قدرتمند یادگیری ماشین است که به‌طور خاص برای مسائل رده‌بندی و پیش‌بینی طراحی‌شده است. این الگوریتم یکی از محبوب‌ترین و مؤثرترین روش‌ها در مسابقات یادگیری ماشین و کاربردهای واقعی است. این الگوریتم

1. Random Forest (RF)
2. Linear Regression (LR)
3. Support Vector Regression (SVR)
4. k-nearest neighbors (KNN)
5. Decision Tree (DT)
6. Extreme Gradient Boosting

به‌طور خودکار می‌تواند با مقادیر گمشده تعامل داشته باشد و این داده‌ها را در هنگام آموزش مدل مدیریت کند. همچنین به‌خوبی می‌تواند با داده‌های بزرگ و پیچیده کار کند و معمولاً در یادگیری ماشین عملکرد بسیار خوبی دارد (Saraswat et al., 2024).

محدوده موردمطالعه

داده‌های مورداستفاده در این پژوهش داده‌های اندازه‌گیری شده در ایستگاه سینوپتیک یزد طی سال‌های ۲۰۰۶ تا ۲۰۲۳ در مقیاس روزانه می‌باشد. استان یزد به‌عنوان یکی از استان‌های با تابش خورشیدی بالا در ایران شناخته می‌شود. این ویژگی باعث می‌شود که داده‌های مربوط به تابش خورشیدی در این منطقه به‌عنوان نمونه‌ای مناسب برای ارزیابی تکنیک‌ها و مدل‌های پیش‌بینی تابش مورداستفاده قرار گیرد. ایستگاه یزد در موقعیت طول و عرض جغرافیایی به ترتیب ۳۱/۸۹۷۴ درجه شمالی و ۵۴/۳۵۶۹ شرقی در ارتفاع ۱۲۱۶ متری از سطح دریا قرارگرفته است. روند تغییرات مقادیر اندازه‌گیری شده تابش خورشیدی در ایستگاه یزد طی سال‌های ۲۰۰۶ تا ۲۰۲۳ در شکل (۱) آورده شده است. میانگین تابش خورشیدی و تابش فرازمینی در ایستگاه سینوپتیک یزد به ترتیب ۱۹/۳۵ و ۳۲ مگاژول بر مترمربع در روز، نسبت ساعات آفتابی به حداکثر ساعات آفتابی ۰/۷۵، رطوبت نسبی میانگین ۲۷ درصد، دمای میانگین ۲۸ درجه سانتی‌گراد می‌باشد. از آمار سال‌های ۲۰۰۶ تا ۲۰۱۴ برای واسنجی معادله‌ها و از آمار ۲۰۱۵ تا ۲۰۲۳ برای ارزیابی نتایج استفاده گردید. هم‌چنین مقادیر تابش فرازمینی و حداکثر ساعات روشنایی روزانه که وابسته به عرض جغرافیایی محل و شماره روز سال بر مبنای تقویم میلادی می‌باشند، از روابط ارائه‌شده در زیر توسط Duffie & Beckman (1991) محاسبه شدند.

$$R_a = \frac{24\times3600}{\pi} I_{gs} d_r \left[\cos\varphi \cos\delta \sin w_s + \frac{\pi}{180} \sin\varphi \sin\delta w_s\right] \quad (۱)$$

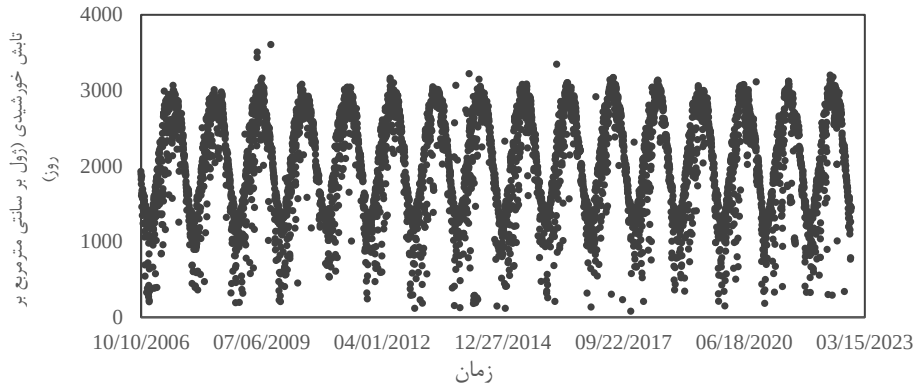
$$d_r = 1 + 0.033 \left[\cos\frac{360J}{365}\right] \quad (۲)$$

$$\delta = 23.45\sin(\frac{360(J+284)}{365}) \quad (۳)$$

$$w_s = \cos^{-1}(-\tan\delta \tan\varphi) \quad (۴)$$

$$N = \frac{2}{15} w_s \quad (۵)$$

جایی که Ra تابش فرازمینی بر حسب مگاژول بر مترمربع بر روز، I_{gs} ثابت خورشیدی و برابر با ۱۳۶۷ وات بر مترمربع، J شماره روز از سال، d_r فاصله نسبی زمین تا خورشید، φ عرض جغرافیایی بر حسب درجه، δ زاویه میل خورشیدی بر حسب درجه، w_s زاویه ساعتی خورشیدی بر حسب درجه و N حداکثر ساعات روشنایی روز است.



شکل ۱. سری زمانی مقادیر تابش خورشیدی در ایستگاه سینوپتیک یزد

معیارهای ارزیابی مدل

با استفاده از آماره‌های مختلفی می‌توان عملکرد مدل‌ها را مورد ارزیابی قرار داد. یکی از این آماره‌ها استفاده از معیارهای ارزیابی است. از جمله معیارهای ارزیابی پرکاربرد ریشه میانگین مربعات خطا[1]، قدر مطلق میانگین خطا[2]، خطای میانگین مربعات[3] و ضریب تبیین (R^2) می‌باشند که به ترتیب در روابط (۳) تا (۵) آورده شده است. دقیق‌ترین مدل با توجه به این معیارها، مدلی خواهد بود که مقدار این چهار معیار به ترتیب نزدیک به صفر، صفر، صفر و یک باشد. در این روابط N تعداد مشاهدات، O_i و P_i به ترتیب مقادیر مشاهده‌شده تابش خورشیدی و مقادیر برآورد شده تابش خورشیدی و P_{ave} و O_{ave} به ترتیب میانگین مقادیر برآورد شده تابش خورشیدی و مشاهده‌شده تابش خورشیدی می‌باشد.

$$RMSE = \sqrt{\frac{\int_{i=1}^{N}(O_i-P_i)^2}{N}}$$ رابطه (۳)

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|O_i - P_i|$$ رابطه (۴)

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(O_i - P_i)^2$$ رابطه (۵)

$$R^2 = \left(\frac{\sum_{i=1}^{N}(O_i-O_{ave})(P_i-P_{ave})}{\sqrt{(\sum_{i=1}^{N}(O_i-O_{ave})^2)(\sum_{i=1}^{N}(P_i-P_{ave})^2)}}\right)^2$$ رابطه (۶)

یافته‌ها

مدل‌سازی بدون پیش‌پردازش داده‌های ورودی

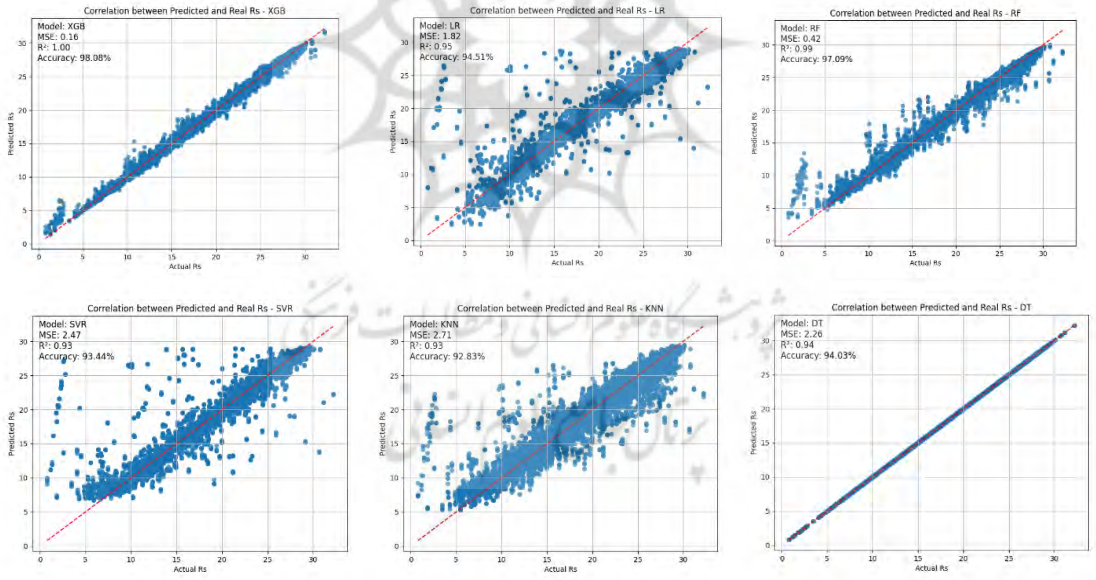
در جدول (۱) نتایج معیارهای ارزیابی الگوریتم‌های بکار رفته در این مطالعه در مرحله آموزش و آزمون ارائه‌شده است. همان‌طور که نتایج نشان می‌دهد، در مرحله آموزش مدل‌های بکار رفته به‌خوبی آموزش‌دیده و نتایج قابل‌قبولی نشان دادند. در مرحله آزمون همچنین نتایج مدل‌سازی داده‌های ورودی بدون پیش‌پردازش اولیه نتایج مناسبی برای همه مدل‌ها ارائه داد. ضریب تعیین بسته به الگوریتم‌های بکار رفته بین ۰/۷۹۰ در مدل KNN و ۰/۸۹۳ برای مدل SVR متغیر است. به‌عبارت‌دیگر از نظر مقادیر ضریب تبیین، همه الگوریتم‌های مورداستفاده نتایج خوبی برای پیش‌بینی تابش خورشیدی نشان دادند. با توجه به نتایج همه معیارها مشاهده می‌شود که الگوریتم رگرسیون بردار پشتیبان با RMSE = ۱/۷۳۲، ۳/۰۰

1. Root Mean Square Error (RMSE)
2. Mean Absolute Error (MAE)
3. Mean Squared Error (MSE)

MSE=۰/۸۲۶،MAE= ۰/۸۹۳،R^2= و دقت ۹۰/۷۵٪ بهتر از سایر مدل‌ها عمل کرده و تابش خورشیدی را با دقت بالاتری پیش‌بینی کرده است.

جدول ۱. مدل‌سازی بدون پیش‌پردازش اولیه بر روی داده‌ها در مرحله آموزش و آزمون مدل

Model	RMSE	MSE	MAE	R^2	Accuracy (%)
مرحله آموزش					
XGB	۰/۴۰۰	۰/۱۶۰	۰/۲۶۶	۰/۹۹۶	۹۸/۰۸
RF	۰/۸۲۹	۰/۶۷۵	۰/۴۰۳	۰/۹۸۲	۹۶/۱۲
LR	۲/۱۴۵	۴/۶۰۸	۱/۰۸۷	۰/۸۷۶	۸۹/۳۴
SVR	۲/۱۰۳	۴/۴۳۲	۰/۸۹۴	۰/۸۸۱	۹۰/۲۳
KNN	۱/۹۱۵	۳/۶۷۴	۱/۰۴۱	۰/۹۰۱	۹۰/۳۹
DT	۰/۰۰	۰/۰۰	۰/۰۰	۱/۰۰	۱۰۰/۰
مرحله آزمون					
XGB	۱/۹۷۲	۳/۸۸۹	۱/۱۲۰	۰/۸۷۲	۹۰/۱۲
RF	۱/۸۰۷	۳/۲۶۵	۰/۹۶۸	۰/۸۱۵	۸۹/۳۶
LR	۱/۷۷۰	۳/۱۳۳	۱/۰۰	۰/۸۲۶	۸۹/۶۴
SVR	۱/۷۳۲	۳/۰۰۱	۰/۸۲۶	۰/۸۹۳	۹۰/۷۵
KNN	۲/۰۲۶	۴/۱۰۳	۱/۱۷۵	۰/۷۹۰	۸۸/۳۴
DT	۲/۷۲۷	۷/۴۳۶	۱/۳۷۳	۰/۷۹۶	۸۸/۵۷



شکل ۲. نمودار پراکنش نتایج بخش آموزش مدل‌های مورداستفاده بدون پیش‌پردازش اولیه

تعیین تعداد مؤلفه اصلی ورودی

پس از تعیین ساختارهای ورودی هر یک از مدل‌ها و تعیین متغیرهای ورودی مؤثر با روش‌های پیش‌پردازش تحلیل مؤلفه اصلی و تئوری آنتروپی، هر یک از مدل‌های جنگل تصادفی (RF)، رگرسیون خطی (LR)، ماشین بردار پشتیبان (SVR)، نزدیک‌ترین همسایه (KNN)، درخت تصمیم (DT) و XGB به ازای داده‌های آموزش مورد واسنجی قرارگرفته و سپس عملکرد مدل‌های آموزش‌دیده به ازای داده‌های بخش صحت سنجی ارزیابی گردیده است. در ادامه خلاصه نتایج

مربوط به هر یک از مدل‌ها در تخمین تابش خورشیدی ارائه‌شده است. مقادیر ویژه حاصل از پارامترها و درصد واریانس هر مؤلفه در آنالیز مؤلفه‌های اصلی در جدول (۲) نمایش داده شده است. همان‌طور که نتایج نشان می‌دهد مؤلفه اول حدود ۴۹ درصد از واریانس و مؤلفه دوم حدود ۳۶ درصد واریانس را به خود اختصاص داده و به‌صورت تجمعی دو مؤلفه اول بیش از ۸۵ درصد پراکندگی داده‌های اصلی را شامل می‌شوند. در نتیجه این دو مؤلفه به‌عنوان ورودی مدل‌های تخمین‌گر برای تخمین تابش خورشیدی در نظر گرفته می‌شود. برای تشکیل هر مؤلفه باید مقادیر متغیر پارامترها را در بردارهای ویژه مربوط به هر متغیر (جدول ۳) ضرب و حاصل را با هم جمع کرد. در نتیجه، مؤلفه‌هایی حاصل می‌شود که می‌توان از آن‌ها به‌جای متغیرهای اولیه به‌عنوان ورودی به مدل‌های تخمین‌گر استفاده کرد (روابط ۷ و ۸):

$$PC1=(0.910*T_{min})+(0.963*T_{max})+(0.949*T_{mean})+(-0.801*RH)+(0.576*n/N)+(-0.147*dr)+(349*\delta)+(0.343*Ra)$$ رابطه (۷)

$$PC2=(-0.201*T_{min})+(-0.190*T_{max})+(-0.198*T_{mean})+(0.237*RH)+(0.015*n/N)+(-0.986*dr)+(0.936*\delta)+(0.939*Ra)$$ رابطه (۸)

جدول ۲. مشخصات مقادیر ویژه حاصل از پارامترها و درصد واریانس

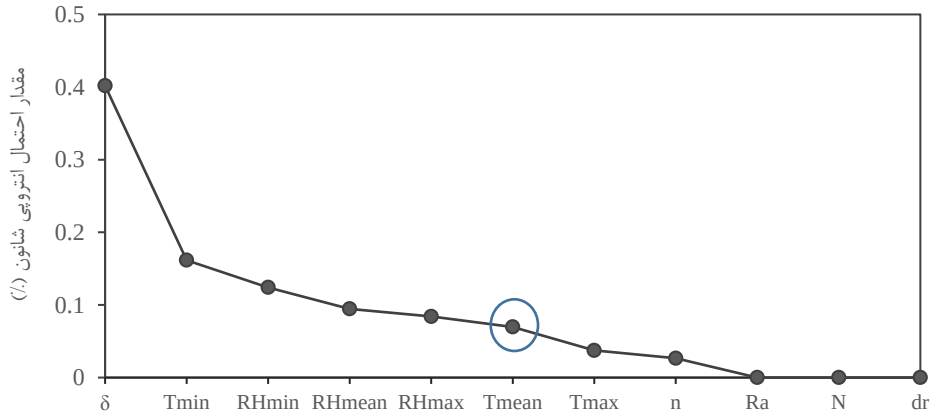
مؤلفه	مقادیر ویژه		
	کل	درصد از واریانس	درصد تجمعی
۱	۳/۹۲۰	۴۹/۰۰۳	۴۹/۰۳
۲	۲/۹۰۹	۳۶/۳۶۷	۸۵/۳۷۰
۳	۰/۷۹۳	۹/۹۱۳	۹۵/۲۸۳
۴	۰/۳۴۳	۴/۲۹۲	۹۹/۵۷۵
۵	۰/۰۲۸	۰/۳۵۱	۹۹/۹۲۶
۶	۰/۰۰۶	۰/۰۷۴	۱۰۰
۷	$۳/۲۹۵\times۱۰^{-۱۰}$	$۴/۱۱۸\times۱۰^{-۹}$	۱۰۰
۸	$۷/۴۲۳\times۱۰^{-۱۸}$	$۹/۲۷۹\times۱۰^{-۱۷}$	۱۰۰

جدول ۳. مقادیر بردار ویژه مربوط به پارامترهای موردبررسی

پارامترها	مؤلفه اول	مؤلفه دوم
دمای کمینه (T_{min})	۰/۹۱۰	−۰/۲۰۱
دمای بیشینه (T_{max})	۰/۹۶۳	−۰/۱۹۰
دمای میانگین (T_{mean})	۰/۹۴۹	−۰/۱۹۸
رطوبت نسبی (RH)	−۰/۸۰۱	۰/۲۳۷
نسبت ساعات آفتابی (n/N)	۰/۵۷۶	۰/۰۱۵
فاصله نسبی زمین تا خورشید (dr)	−۰/۱۴۷	−۰/۹۸۶
زاویه میل خورشیدی (δ)	۰/۳۴۹	۰/۹۳۶
تابش فرازمینی (Ra)	۰/۳۴۳	۰/۹۳۹

در این پژوهش تئوری شانون نیز برای تک‌تک متغیرهای ورودی محاسبه گردید که نتایج آن در شکل (۳) نشان داده شده است. بر اساس تئوری آنتروپی شانون متغیری که دارای بیشترین احتمال وقوع باشد، ارتباط بیشتر و معنادارتری با تابش خورشیدی دارد. با توجه به شکل (۳) متغیر زاویه میل خورشیدی با اختلاف نسبت به دمای حداقل، رطوبت نسبی حداقل، میانگین و حداکثر، دمای میانگین، مؤثرترین متغیر برای تخمین تابش خورشیدی شناسایی شد. همچنین متغیرهای فاصله نسبی زمین تا خورشید، تابش فرازمینی و حداکثر ساعات آفتابی کمترین تأثیر را در ارتباط با تابش خورشیدی در ایستگاه یزد دارند. برای تعیین ترکیب بهینه با استفاده از تئوری آنتروپی نمودار مقادیر احتمال آنتروپی در برابر متغیرهای

ورودی رسم شد. نقطه‌ای که نمودار دچار تغییر شیب شدید می‌شود، به‌عنوان نقطه عطف انتخاب می‌شود. پس با توجه به اینکه متغیرهایی که مقدار آنتروپی آن‌ها بیشتر است در تخمین تابش خورشیدی مؤثرتر هستند، در این مدل متغیرهای زاویه میل خورشیدی، دمای حداقل، رطوبت نسبی حداقل، میانگین رطوبت نسبی و دمای میانگین به‌عنوان ورودی مدل پیش‌بینی انتخاب می‌شوند.



شکل ۳. انتخاب ترکیب مؤثر برای تخمین تابش خورشیدی با تئوری شانون

مدل‌سازی همراه با پیش‌پردازش توسط تحلیل مؤلفه اصلی و تئوری آنتروپی

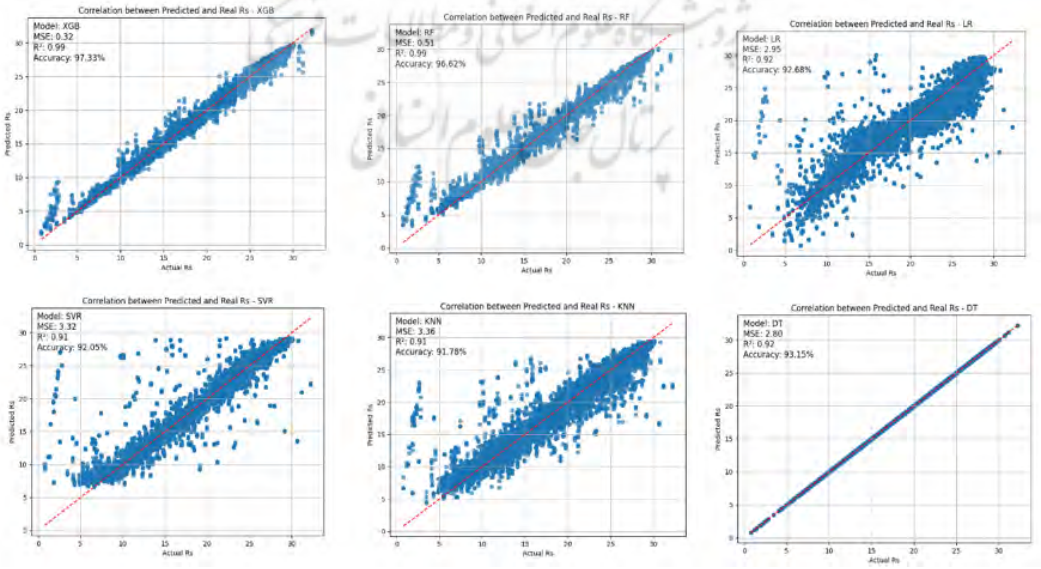
نتایج جدول (۴) و (۵) به ترتیب نشان‌دهنده مدل‌سازی با استفاده از داده‌های حاصل از پیش‌پردازش PCA و تئوری آنتروپی در مرحله آموزش و آزمون مدل می‌باشد. یکی از دلایل عمده کاهش دقت مدل در مرحله آزمون، پدیده بیش تناسبی مسئله است. این مسئله زمانی رخ می‌دهد که مدل به‌قدری بر روی‌داده‌های آموزشی تنظیم‌شده باشد که نه‌تنها الگوهای واقعی را یاد می‌گیرد، بلکه همچنین نویز و جزئیات خاصی که فقط در مجموعه داده آموزشی وجود دارد را نیز حفظ می‌کند. در نتیجه، وقتی داده‌های جدید (داده‌های آزمون) وارد مدل می‌شوند، توانایی مدل در تعمیم و پیش‌بینی صحیح کاهش می‌یابد. به‌طورکلی نتایج حاصل از کاهش بعد به روش PCA بهتر از نتایج کاهش بعد به روش تئوری شانون در هر دو بخش آموزش و آزمون به دست آمد. با توجه به نتایج قسمت آموزش (جدول ۴) مدل PCA-DT و ENT-DT با مجذور میانگین مربعات خطا و میانگین قدر مطلق خطای نسبی صفر و ضریب تبیین ۱/۰۰ نسبت به سایر مدل‌ها در ایستگاه یزد بهترین عملکرد در تخمین تابش خورشیدی و آموزش مدل‌ها را داشته است. شکل (۴) و (۵) نیز به ترتیب نمودار پراکنش نتایج بخش آموزش مدل‌سازی مدل‌های مورداستفاده را همراه با پیش‌پردازش روش تئوری آنتروپی و تحلیل مؤلفه اصلی نشان می‌دهد. نتایج بخش آزمون مدل (جدول ۵) حاکی از برتری مدل PCA-SVR نسبت به سایر روش‌ها است. همان‌طور که مشاهده می‌شود مدل PCA-SVR با ضریب تبیین ۰/۹۲۳ و دقت ۹۲/۸۴٪ با کمترین مقدار معیارهای خطا بهترین نتیجه را در بین مدل‌های مذکور در ایستگاه یزد دارد. مدل ENT-DT با ضریب تبیین ۰/۵۳۵ و دقت ۷۹/۳۴٪ نتایج ضعیف‌تری در بین مدل‌های استفاده‌شده در ایستگاه یزد نشان داد.

جدول ٤. مدل‌سازی با استفاده از داده‌های حاصل از پیش‌پردازش PCA و تئوری شانون در مرحله آموزش مدل

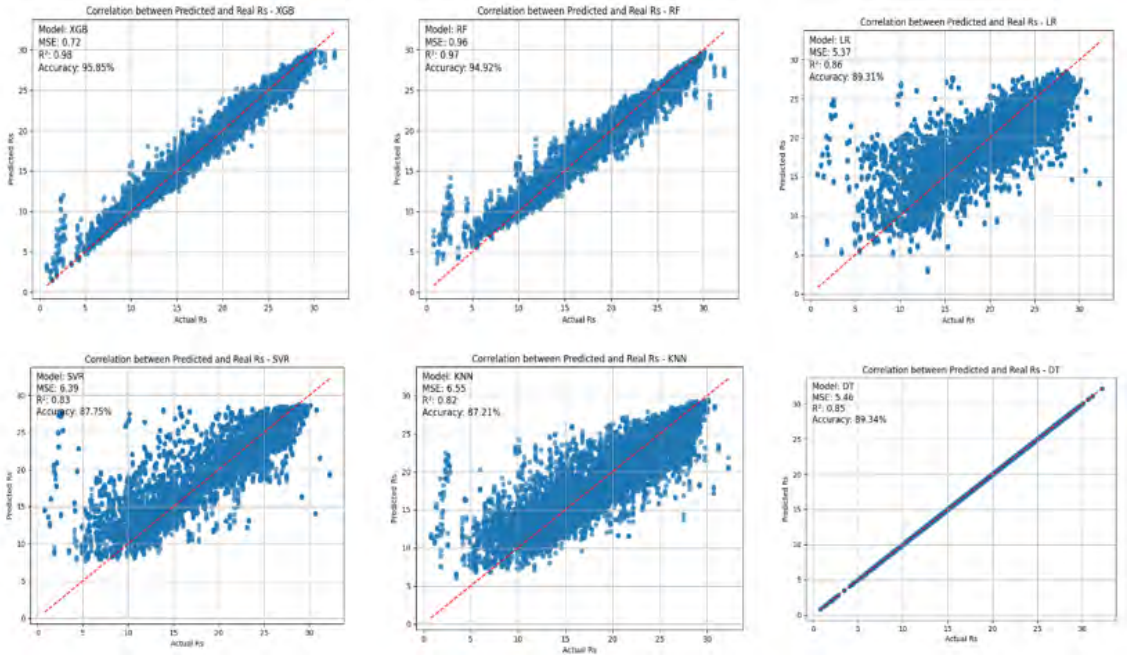
Accuracy (%)	R^2	MAE	MSE	RMSE	Model
۹۷/۳۳	۰/۹۹۱	۰/۳۵۸	۰/۳۲۲	۰/۵۶۷	PCA-XGB
۹۵/۹۱	۰/۹۸۱	۰/۴۲۶	۰/۷۰۴	۰/۸۳۹	PCA-RF
۸۴/۸۰	۰/۷۹۰	۱/۸۶۹	۷/۸۳۰	۲/۷۹۸	PCA-LR
۹۰/۱۶	۰/۸۸۱	۰/۹۱۱	۴/۴۳۵	۲/۱۰۴	PCA-SVR
۹۰/۷۲	۰/۹۰۵	۰/۹۹۸	۳/۵۳۱	۱/۱۸۸	PCA-KNN
۱۰۰	۱/۰۰	۰/۰۰	۰/۰۰	۰/۰۰	PCA-DT
۹۵/۸۴	۰/۹۸۱	۰/۵۶۱	۰/۷۱۸	۰/۸۴۶	ENT-XGB
۹۳/۹۹	۰/۹۶۷	۰/۶۸۵	۱/۱۹۷	۱/۰۹۴	ENT-RF
۷۸/۰۹۹	۰/۶۱۹	۲/۷۲۷	۱۴/۱۸۶	۳/۷۶۶	ENT-LR
۸۳/۰۵	۰/۷۴۶	۱/۸۴۳	۹/۴۵۵	۳/۰۷۴	ENT-SVR
۸۵/۰۴	۰/۸۰۶	۱/۷۸۱	۷/۲۱۶	۲/۳۸۵	ENT-KNN
۱۰۰	۱/۰۰	۰/۰۰	۰/۰۰	۰/۰۰	ENT-DT

جدول ٥. مدل‌سازی با استفاده از داده‌های حاصل از پیش‌پردازش PCA و تئوری شانون در مرحله آزمون مدل

Accuracy (%)	R^2	MAE	MSE	RMSE	Model
۹۱/۱۰	۰/۹۰	۱/۱۶۰	۳/۸۹۶	۱/۹۷۴	PCA-XGB
۹۱/۹۶	۰/۹۱۳	۱/۰۱	۳/۳۲۳	۱/۸۲۳	PCA-RF
۸۷/۶۵	۰/۸۳۰	۱/۷۸۲	۶/۴۹۳	۲/۵۴۸	PCA-LR
۹۲/۸۴	۰/۹۲۳	۰/۸۱۷	۲/۹۵۱	۱/۷۱۸	PCA-SVR
۹۱/۰۹	۰/۸۹۳	۱/۱۴	۴/۰۸۵	۲/۰۲	PCA-KNN
۸۹/۴۱	۰/۸۴۶	۱/۳۹	۵/۸۹۷	۲/۴۳	PCA-DT
۸۲/۱۹	۰/۷۲۳	۲/۱۳۴	۱۰/۲۷۹	۳/۲۰۱	ENT-XGB
۸۲/۷۹	۰/۷۴۸	۲/۰۰۵	۹/۴۱۴	۳/۰۶۰	ENT-RF
۷۷/۳۸	۰/۵۹۴	۲/۸۱۶	۱۵/۰۹۹	۳/۸۸۱	ENT-LR
۸۲/۵۶	۰/۷۳۴	۱/۹۰۵	۹/۸۹۲	۳/۱۳۶	ENT-SVR
۸۱/۱۹	۰/۶۹۸	۲/۲۴۳	۱۱/۲۲۸	۳/۳۴۴	ENT-KNN
۷۹/۳۴	۰/۵۳۵	۲/۶۸۲	۱۷/۲۸۹	۴/۱۵۲	ENT-DT



شکل ٤. نمودار پراکنش نتایج بخش آزمون مدل‌های مورداستفاده را همراه با پیش‌پردازش روش تئوری آنتروپی



شکل ۵. نمودار پراکنش نتایج بخش آزمون مدل‌های مورداستفاده را همراه با پیش‌پردازش روش تحلیل مؤلفه اصلی

بحث

کاهش بعد یکی از جنبه‌های مهم در تحلیل داده‌ها و مدل‌سازی به‌ویژه در زمینه‌های یادگیری ماشین، آمار و داده‌کاوی است. این فرآیند شامل کاهش تعداد متغیرهای تصادفی موردبررسی است و به‌سادگی مجموعه داده‌ها کمک می‌کند، درحالی‌که ویژگی‌های اصلی آن حفظ می‌شود. این روند برای بهبود عملکرد مدل، کاهش هزینه‌های محاسباتی و افزایش قابلیت تفسیر بسیار حیاتی است. در این پژوهش دو روش کاهش بعد شامل روش تحلیل مؤلفه اصلی (PCA) و تئوری آنتروپی شانون (ENT) برای تعیین ورودی مدل‌های یادگیری ماشین بر روی‌داده‌های تابش خورشیدی در ایستگاه یزد استفاده شد. دلیل انتخاب ایستگاه سینوپتیک یزد این بود که دارای آسمان صاف و روزهای آفتابی زیادی در طول سال است. این شرایط جوی موجب می‌شود که داده‌های خام برای تحلیل و مدل‌سازی تابش خورشیدی بسیار معتبر و بهینه باشند. یزد تجربهٔ تغییرات جوی در طول سال را دارد که به تحلیل و ارزیابی تأثیرات فصلی بر تابش خورشیدی کمک می‌کند. این تغییرات می‌تواند اطلاعات ارزشمندی درباره پیش‌بینی تابش در شرایط مختلف فراهم کند. همچنین استان یزد در حال حاضر توجه‌های زیادی را برای توسعه تکنولوژی‌های انرژی خورشیدی جلب کرده و به همین دلیل، مطالعه در این منطقه ممکن است به نوآوری‌ها و تحقیق‌های جدید در این زمینه منجر شود. علاوه بر این ایستگاه یزد دارای داده‌های طولانی‌مدت در زمینه تابش خورشیدی است که این داده‌ها برای الگوسازی و بررسی الگوهای تابش و پیش‌بینی‌های بهتر، بسیار مفید هستند. نتایج حاکی از برتری روش تحلیل مؤلفه اصلی در کاهش بعد داده‌های ورودی نسبت به روش تئوری شانون بود و تعداد داده‌های ورودی را به دو مؤلفه کاهش داد. این روش یکی از متداول‌ترین تکنیک‌ها برای کاهش بعد است. این روش داده‌ها را به یک سیستم مختصات جدید تبدیل می‌کند که در آن بزرگ‌ترین تنوع در مؤلفه اصلی اول قرار دارد، بزرگ‌ترین تنوع بعدی در محور دوم و به همین ترتیب در مؤلفه‌های بعدی. همچنین نتایج نشان داد که در بعضی شرایط، تکیه‌بر روش آنتروپی شانون به‌عنوان معیار برای انتخاب ویژگی‌ها یا کاهش بعد ممکن است نادرست باشد، زیرا این روش نسبت به تغییرات کوچک در داده‌ها بسیار حساس است. همچنین تابش خورشیدی تحت تأثیر عوامل متعددی مانند داده‌های هواشناسی، زمان روز و موقعیت جغرافیایی قرار دارد. این پیچیدگی‌ها ممکن است باعث شود که روش‌های

مبتنی بر آنتروپی نتوانند به‌درستی الگوهای غیرخطی را شناسایی کنند. همچنین روش شانون عمدتاً در مورد اطلاعات کیفی و توزیع‌های تصادفی کاربرد دارد و کمتر به داده‌های کمی و ناپیوسته توجه دارد که در تخمین تابش خورشیدی معمولاً اهمیت بیشتری دارند. نتایج مدل‌سازی نشان داد که مدل SVR نسبت به سایر مدل‌ها عملکرد بهتری در تخمین تابش خورشیدی دارد. این مدل می‌تواند از هسته‌های مختلف مانند هسته‌های چندجمله‌ای و گوسی استفاده کند که این قابلیت به SVR اجازه می‌دهد تا به‌خوبی الگوهای غیرخطی در داده‌ها را مدل‌سازی کند. همچنین با حداکثر کردن حاشیه بین داده‌های آموزشی و پیش‌بینی‌ها می‌پردازد و به مدل کمک می‌کند تا تعمیم بهتری برای داده‌های جدید داشته باشد. علاوه بر این، این روش معمولاً در برابر نویز موجود در داده‌ها مقاوم‌تر است، به‌ویژه زمانی که تنظیمات پارامترها به‌درستی انجام‌شده باشد. این ویژگی سبب می‌شود که SVR در داده‌های واقعی و با نویز بالا عملکرد بهتری داشته باشد. همچنین SVR به‌ویژه برای داده‌های با ابعاد بالا ویژگی‌های ساختاری پیچیده مناسب است و به‌خوبی با مجموعه‌های داده کوچک سازگار است و این باعث می‌شود که در شرایطی که داده‌ها محدود هستند، مؤثر واقع شود.

نتیجه‌گیری

با توجه به اهمیت تخمین درست تابش خورشیدی در پدیده‌های هیدرولوژیکی و لزوم استفاده از روش‌های نوین در برآورد آن، در این پژوهش از روش‌های تحلیل مؤلفه اصلی و تئوری آنتروپی برای پیش‌پردازش داده‌ها استفاده گردید ورودهای مدل‌های تخمین‌گر توسط دو روش مذکور شناسایی شدند. مدل‌سازی با مدل‌های جنگل تصادفی (RF)، رگرسیون خطی (LR)، ماشین بردار پشتیبان (SVR)، نزدیک‌ترین همسایه (KNN)، درخت تصمیم (DT) و XGB انجام گرفت. نتایج تئوری آنتروپی نشان داد که در ایستگاه یزد متغیرهای زاویه میل خورشیدی، دمای حداقل، رطوبت نسبی حداقل و میانگین رطوبت نسبی متغیرهای مؤثر در برآورد تابش خورشیدی بودند. هم‌چنین با تحلیل مؤلفه اصلی تعداد متغیرهای ورودی به دو مؤلفه کاهش یافت و مدل‌سازی با دو متغیر ورودی ایجادشده از این روش انجام گرفت. به‌طورکلی از نتایج مدل‌سازی می‌توان نتیجه گرفت که مدل PCA-SVR نسبت به سایر مدل‌ها عملکرد بهتری در تخمین تابش خورشیدی داشته است. در کل پیش‌پردازش PCA نشان داد که این روش ورودی‌های بهتری را برای مدل‌های تخمین‌گر تعیین می‌کند. این نتیجه‌گیری با نتایج سایر پژوهشگران همخوانی دارد. محمدی و امامقلی زاده (۱۳۹۵)، محمدی و همکاران (۱۳۹۸)، Djeldjeli et al., (2024)، Demir et al., (2023)، Saraswat et al., (2024) نشان دادند که تکنیک‌های کاهش ابعاد و حذف ویژگی‌های زائد تأثیر مثبتی در بهبود دقت و کاهش پیچیدگی مدل‌ها داشته‌اند. با استفاده از PCA و حذف ویژگی‌های دارای همبستگی بالا، می‌توان کارایی مدل‌ها را با دقتی نزدیک به حالت اولیه حفظ کرد، درحالی‌که ابعاد داده‌ها کاهش‌یافته است. لازم به ذکر است که روش تئوری شانون نتوانست نتایج مدل‌سازی را نسبت به روش بدون پیش‌پردازش اولیه بهبود بخشد. این تحلیل نشان می‌دهد که استفاده از تکنیک‌های کاهش ابعاد و انتخاب مدل‌های مناسب می‌تواند منجر به‌دقت بیشتر و پیچیدگی محاسباتی کمتر در مسائل پیش‌بینی شود، هرچند در انتخاب مدل پیش‌پردازش داده‌های اولیه باید دقت کافی داشت. انجام تحقیقات مشابه با استفاده از داده‌های جدید یا در شرایط مختلف جغرافیایی نیز می‌تواند به اعتبارسنجی بیشتر نتایج کمک کند.

حامی مالی

بر اساس اظهار نویسندگان این پژوهش حامی مالی نداشته است.

سهم نویسندگان

در این پژوهش، سهم نویسندگان به‌صورت زیر مشخص می‌شود: سمیه سلطانی گردفرامرزی طراحی مطالعه، جمع‌آوری داده‌ها، تجزیه‌وتحلیل، نگارش پیش‌نویس اولیه ویرایش نهایی مقاله و مژگان عسکری زاده مدل‌سازی و نتایج را بر عهده داشته‌اند.

تضاد منافع

نویسندگان اعلام می‌دارند که هیچ تضاد منافعی در ارتباط با نویسندگی یا انتشار مقاله ندارند.

تقدیر و تشکر

نویسندگان از تمامی کسانی که در انجام پژوهش حاضر یاری‌رسان بوده‌اند، به‌ویژه کسانی که کار ارزیابی کیفیت مقاله را انجام دادند، تشکر و قدردانی می‌نمایند.

منابع

سلطانی گردفرامرزی، سمیه؛ تقی زاده، روح‌الله؛ قاسمی، محسن. (۱۳۹۴). برآورد ضریب پخشیدگی طولی رودخانه با استفاده از انواع روش‌های داده‌کاوی. *تحقیقات آب‌وخاک ایران*، ۴۶(۳)، ۳۸۵-۳۹۴. doi: 10.22059/ijswr.2015.56728

سلطانی گردفرامرزی، سمیه. (۱۴۰۲). پیش‌بینی تابش خورشیدی در ایستگاه یزد با به‌کارگیری مدل رگرسیونی مبتنی بر مؤلفه‌های اصلی (PCR). *هواشناسی کشاورزی*، ۱۱(۱)، ۶-۱۶. doi: 10.22125/agmj.2023.352446.1140

سلطانی گردفرامرزی، سمیه و مؤمنی، هاجر. (۱۴۰۲). کاربست الگوریتم‌های یادگیری ماشین برای تخمین تابش خورشیدی (موردمطالعه: اقلیم خشک و نیمه‌خشک). *ژئوفیزیک ایران*، ۱۷(۴)، ۲۹-۳۵. doi: 10.30499/ijg.2023.393259.1512

شیخ‌الاسلامی، نونا؛ قهرمان، بیژن؛ مساعدی، ابوالفضل؛ داوری، کامران و مهاجرپور، مهدی. (۱۳۹۳). پیش‌بینی تبخیر و تعرق گیاه مرجع (ET_O) با استفاده از روش آنالیز مؤلفه‌های اصلی (PCA) و توسعه مدل رگرسیونی خطی چندگانه (MLR-PCA) (مطالعه موردی: ایستگاه مشهد). *نشریه آب‌وخاک*، ۲۸(۲)، ۴۲۰-۴۲۹. doi: 10.22067/jsw.v0i0.25711

صفاری پور، محمدحسن و مهرابیان، مظفرعلی. (۱۳۸۸). پیش‌بینی مقدار کل تابش خورشیدی در کرمان با استفاده از مشخصات هندسی، نجومی، جغرافیایی و هواشناسی. *شریف*، ۵۱، ۳-۱۳.

محمدی، بابک و امامقلی زاده، صمد. (۱۳۹۵). استفاده از تحلیل مؤلفه اصلی برای تعیین ورودی‌های مؤثر بر تخمین بارش به کمک شبکه عصبی مصنوعی و ماشین بردار پشتیبان، *سامانه‌های سطوح آبگیر باران*، ۴(۱۳)، ۶۷-۷۵. doi:20.1001.1.24235970.1395.4.4.6.9

عوض پور، صدیقه؛ بختیاری، بابک و قادری، کوروش. (۱۳۹۸). بررسی کارایی روش‌های شبکه عصبی و رگرسیون چند متغیره در برآورد تابش کل خورشیدی در چند ایستگاه معرف اقلیم‌های خشک و نیمه‌خشک. *تحقیقات آب‌وخاک ایران*، ۵۰(۱۳۹)، ۱۸۵۵-۱۸۶۹. doi:20.1001.1.24234931.1399.7.2.3.2

محمدی، بابک؛ آقاشریعتمداری، زهرا و مؤذن‌زاده، روزبه. (۱۳۹۸). تعیین متغیرهای ورودی برای تخمین تابش خورشیدی با استفاده از تئوری آنتروپی و تحلیل مؤلفه اصلی. *تحقیقات آب‌وخاک ایران*، ۵۰(۳)، ۶۲۶-۶۳۹. doi: 10.22059/ijswr.2018.257150.667906

References

Abdelhafidi, N., Bachari, N.E.I., & Abdelhafidi, Z. (2021). Estimation of solar radiation using stepwise multiple linear regression with principal component analysis in Algeria. *Meteorology and Atmospheric Physics*, 133(2), 205-216. http://doi: 10.1007/s00703-020-00739-0.

Avazpour, S., Bakhtiari, B., & Qaderi, K. (2019). Performance evaluation of Neural Network and Multivariate Regression Methods for Estimation of Total Solar Radiation at several stations in Arid and Semi-Arid Climates. *Iranian Journal of Soil and Water Research*, 50(8), 1855-1869. http://doi: 20.1001.1.24234931.1399.7.2.3.2. [In Persian]

Boroughani, M., Soltani, S., Ghezelseflu, N., & Pazhouhan, I. (2022). A comparative assessment between artificial neural network, neuro-fuzzy, and support vector machine models in splash erosion modelling under simulation circumstances. *Folia Oecologica*, 49(1), 23-34. http://doi:10.2478/foecol-2022-0003.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32. DOI: 10.1023/A:1010933404324.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13(1), 21-27. http://doi: 10.1109/TIT.1967.1053964.

Demir, V., & Citakoglu, H. (2023). Forecasting of solar radiation using different machine learning approaches. *Neural Computing and Applications*, 35(1), 887-906. http://doi: 10.1007/s00521-022-07831-z.

Djeldjeli, Y., Taouaf, L., Alqahtani, S., Mokaddem, A., Alshammari, B.M., Menni, Y., & Kolsi, L. (2024). Enhancing solar power forecasting with machine learning using principal component analysis and diverse statistical indicators. *Case Studies in Thermal Engineering*, 61, 104924. http://doi: 10.1016/j.csite.2024.104924.

Draper, N. R., & Smith, H. (1998). *Applied regression analysis*. Wiley-Interscience.

Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A., & Vapnik, V. (1997). Support vector regression machines. *Advances in Neural Information Processing Systems*, 9, 155-161. https://www.researchgate.net/publication/309185766 Support vector regression machines.

Duffie, J.A., & Beckman, W.A. (1991). Solar Engineering of Thermal Processes. Wiley, New York.

Hunt, L.A., Kuchar, L., & Swanton, C.J. (1998). Estimation of solar radiation for use in crop modeling. *Agric. Meteorol*. 91, 293–300. http://doi: 10.1016/S0168-1923(98)00085-4.

Liu, C.W., Lin K.H., & Kuo Y.M. (2003). Application of factor analysis in the assessment of groundwater quality in a blackfoot disease area in Taiwan. *Science of the Total Environment*, 313, 77-89. http://doi: 10.1016/S0048-9697(02)00683-6.

Meenal, R., & Selvakumar, A.I. (2018). Assessment of SVM, empirical and ANN based solar radiation prediction models with most influencing input parameters. *Renewable Energy*, *121*, 324-343. http://doi: 10.1016/j.renene.2017.12.005.

Mohammadi, B., Aghashariatmadari, Z., & moazenzadeh, R. (2019). Determination of Input Variables to Estimate Solar Radiation Using Entropy Theory and Principal Component Analysis. *Iranian Journal of Soil and Water Research*, 50(3), 625-639. http://doi: 10.22059/ijswr.2018.257150.667906. [In Persian]

Mohammadi, B., & Emamgholizadeh, S. (2017). Using principal component analysis to inputs the effective rainfall estimates based on entries to help support vector machine and artificial neural network. *Journal of Rainwater Catchment Systems*; 4 (4), 67-75. http://doi:20.1001.1.24235970.1395.4.4.6.9. [In Persian]

Olalekan, S., Abdullahi, M. I., & Olabisi, A. (2018). Modeling of Solar Radiation Using Artificial Neural Network for Renewable Energy Application. *Journal of Applied Physics*, 10(2), 6-12. http://dx.doi.org/10.9790/4861-1002030612.

Quinlan, J. R. (1986). *Induction of decision trees. Machine Learning*, 1(1), 81-106. http://doi: 10.1007/BF00116251.

Radosevic, N., Duckham, M., Liu, G.J., & Sun, Q. (2020). Solar radiation modeling with KNIME and Solar Analyst: Increasing environmental model reproducibility using scientific workflows. *Environmental Modelling & Software*, 132, 104780. http://doi: 10.1016/j.envsoft.2020.104780.

Rahimikhoob, A. (2010). Estimating global solar radiation using artificial neural network and air temperature data in a semi-arid environment. *Renew Energy*, 35, 2131-2135. http://doi: 10.1016/j.renene.2010.01.013.

Sabziparvar, A.A., & Shetaee, H. (2007). Estimation of global solar radiation in arid and semi-arid climates of East and West Iran, *Energy*, 32, 649–655. http://doi: 10.1016/j.energy.2006.06.006.

Saffaripour, M., & Mehrabian, M. (2009). Predicting the total amount of solar radiation in Kerman using geometric, astronomical, geographical and meteorological characteristics. *Sharif*, 51 (1), 3-13 [In Persian]

Saraswat, R., Jhanwar, D., & Gupta, M. (2024). Enhanced Solar Power Forecasting Using XG Boost and PCA-Based Sky Image Analysis. *Traitement du Signal*, *41*(1). http://doi: 10.18280/ts.410104.

Shannon, C.E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(379–423), 623–656. http://doi: 10.1002/j.1538-7305.1948.tb01338.x.

Sheikholeslami, N., Ghahraman, B., Mosaedi, A., Davari, K., & Mohajeri, M. (2014). Estimating Reference Evapotranspiration by Using Principal Component Analysis (PCA) and The Development of a Regression Model (MLR-PCA) (Case Study: Mashhad Station). *Water and Soil*, 28(2), 420-429. http://doi: 10.22067/jsw.v0i0.25711. [In Persian]

Soltani-Gerdefaramarzi, S., Taghizadeh-Mehrjerdi, R., & Ghasemi, M. (2015). Prediction of Longitudinal Dispersion Coefficient in Natural Streams using Soft Computing Techniques. *Iranian Journal of Soil and Water Research*, 46(3), 385-394. http://doi: 10.22059/ijswr.2015.56728. [In Persian]

Soltani-Gerdefaramarzi, S., & Momeni, H. (2023). Application of machine learning algorithms to estimate solar radiation (case study: arid and semi-arid climate). *Iranian Journal of Geophysics*, 17(4), 25-39. http://doi:10.30499/ijg.2023.393259.1512. [In Persian]

Soltani-Gerdefaramarzi, S. (2023). Prediction of solar radiation intensity in Yazd station by using regression model based on principal components (PCR). *Journal of Agricultural Meteorology*, 11(1), 6-16. http://doi:10.22125/agmj.2023.352446.1140. [In Persian]

Xu, H. Xu, C. Y, Sælthun, N. R. Xu, Y. Zhou, B., & Chen, H. (2015). Entropy theory based multicriteria resampling of rain gauge networks for hydrological modelling – A case study of humid area in southern China. *Journal of Hydrology*, 525, 138-151. http://doi: 10.1016/j.jhydrol.2015.03.034.

Yadav, A. K., & Chandel, S. S. (2015). Solar energy potential assessment of western Himalayan Indian state of Himachal Pradesh using J48 algorithm of WEKA in ANN based prediction model. *Renewable Energy*, 75, 675-693. http://doi:10.1016/j.renene.2014.10.045.