Gorgan University of Agricultural Sciences and Natural Resources

Journal of Water and Soil Conservation

Print ISSN: 2322-2069
Online ISSN: 2322-2794

Simulating chlorophyll a in dam reservoirs using remote sensing and data-driven approaches

Javad Zahiri[*1], Mitra Cheraghi[*2], Meysam Salarijazi[3]

1. Corresponding Author, Associate Prof., Dept. of Water Engineering, Agricultural Sciences and Natural Resources University of Khuzestan, Khuzestan, Iran. E-mail: j.zahiri@asnrukh.ac.ir
2. Corresponding Author, Assistant Prof., Dept. of Nature Engineering, Agricultural Sciences and Natural Resources University of Khuzestan, Khuzestan, Iran. E-mail: mitra.cheraghi@asnrukh.ac.ir
3. Associate Prof., Dept. of Water Engineering, Gorgan University of Agricultural Sciences and Natural Resources, Gorgan, Iran. E-mail: meysam.salarijazi@gau.ac.ir

Article Info

Article type:
Research Full Paper

Article history:
Received: 05.25.2024
Revised: 09.22.2024
Accepted: 11.06.2024

Keywords:
Algae,
Sentinel-2 Images,
Water quality,
XGBoost, M5 and MARS Algorithms

ABSTRACT

Background and Objectives: Chlorophyll a is used as a Indicator to measure the amount of algae growing in water, which can be applied to classify the nutritional status of water bodies. Although algae are a natural part of freshwater ecosystems, excessive amounts of algae can cause numerous problems such as skin problems and bad odors, and reduce dissolved oxygen levels. In addition, harmful algal blooms create a severe environmental problem and have significant economic and environmental consequences on coastal areas. Accordingly, predicting the occurrence of these blooms has become increasingly vital for coastal communities. Based on the investigations, the concentration of very high levels of chlorophyll a indicates low water quality, and the long-term existence of high concentrations of chlorophyll a is a major problem for the primary production of biomass. Accordingly, the concentration of chlorophyll a is considered a key indicator for water quality. In this research, an attempt has been made to estimate the amount of chlorophyll a in Sardasht dam reservoir using remote sensing techniques and data driven models.

Materials and Methods: The data categories was used in this study: the first part is the measured values of chlorophyll a in Sardasht dam reservoir, and the second part includes 7 stages of measurement at 10 points with different coordinates in the reservoir of the dam and is related to March 2016 to June 2018. The second part of the data which used in data driven models, was extracted from the Sentinel-2 satellite images. The information of different bands was extracted from Sentinel-2 images and based on the band values, the criteria for measuring the amount of chlorophyll a were calculated and provided to the data driven models for training. In this research, three data driven models XGBoost, M5 and MARS were used to estimate the amount of chlorophyll a. In this study, 9 chlorophyll a estimation equations were considered as input of data driven models and the measured logarithm value of chlorophyll a was considered as output. 80% of the available data was used to train and the remaining 20% was used to verify the effectiveness of the used models.

Results: Based on the input data, the M5 model tree has divided the problem space into 5 parts and presented a linear equation for each part. Based on the structure provided by M5 and MARS algorithms, blue, red and green band combinations as well as infrared and red have a high

impact on the models provided by these two algorithms. The results obtained from XGBoost algorithm show the importance of blue, red and green band combinations on the presented results. Based on this, the combination of blue, red and green bands has been used in all three algorithms as the most important or one of the most important input variables to calculate chlorophyll a. The coefficients of determination for three models XGBoost, M5 and MARS was calculated as 0.61, 0.49 and 0.31, respectively. The value of Nash-Sutcliffe coefficient for XGBoost, M5 and MARS models was calculated as 0.54, 0.47 and 0.27, respectively, which shows that the results of XGBoost and M5 models are favorable.

Conclusion: The results show that the XGBoost and M5 models provided more accurate results than the MARS model. The use of Taylor's diagram also shows the close efficiency of the XGBoost and M5 models in calculating the amount of chlorophyll a. The spatial distribution of chlorophyll a in the Sardasht dam reservoir shows that the lack of information used has caused differences between the measured and calculated values in limited areas. The spatial distribution of chlorophyll a in the Sardasht dam reservoir shows that the limited data used and the lack of complete temporal compatibility of the Sentinel-2 images with the measurement data in the dam reservoir has caused differences between the measured and calculated values in limited areas. The use of a large number of data in the reservoirs of different dams, the use of various data driven models and applying images from other sensors can provide a suitable tool for the managers of the reservoirs so that they can evaluate the water quality more accurately.

Cite this article: Zahiri, Javad, Cheraghi, Mitra, Salarijazi, Meysam. 2024. Simulating chlorophyll a in dam reservoirs using remote sensing and data-driven approaches. *Journal of Water and Soil Conservation*, 31 (3), 85-108.

DOI: 10.22069/jwsc.2024.22570.3739
Publisher: Gorgan University of Agricultural Sciences and Natural Resources

نشریه پژوهش‌های حفاظت آب و خاک

شاپا چاپی: ۲۳۲۲-۲۰۶۹
شاپا آنلاین: ۲۳۲۲-۲۷۹۴

برآورد میزان کلروفیل آ در مخزن سد با استفاده از تکنیک‌های سنجش از دور و مدل‌های داده‌کاوی

جواد ظهیری*[۱]، میترا چراغی*[۲]، میثم سالاری‌جزی[۳]

۱. نویسنده مسئول، دانشیار گروه علوم و مهندسی آب، دانشگاه علوم کشاورزی و منابع طبیعی خوزستان، خوزستان، ایران. رایانامه: j.zahiri@asnrukh.ac.ir

۲. نویسنده مسئول، استادیار گروه مهندسی طبیعت، دانشگاه علوم کشاورزی و منابع طبیعی خوزستان، خوزستان، ایران. رایانامه: mitra.cheraghi@asnrukh.ac.ir

۳. دانشیار گروه مهندسی آب، دانشگاه علوم کشاورزی و منابع طبیعی گرگان، گرگان، ایران. رایانامه: meysam.salarijazi@gau.ac.ir

اطلاعات مقاله

نوع مقاله:
مقاله کامل علمی- پژوهشی

تاریخ دریافت: ۰۳/۰۳/۰۵
تاریخ ویرایش: ۰۳/۰۷/۰۱
تاریخ پذیرش: ۰۳/۰۸/۱۶

واژه‌های کلیدی:
الگوریتم‌های XGBoost، M5 و MARS
تصاویر سنتینل-۲،
جلبک،
کیفیت آب

چکیده

سابقه و هدف: کلروفیل آ به‌عنوان معیاری جهت اندازه‌گیری میزان جلبک در حال رشد در آب مورداستفاده قرار می‌گیرد که می‌توان از آن برای طبقه‌بندی وضعیت تغذیه‌ای پهنه‌های آبی استفاده کرد. اگرچه جلبک‌ها بخشی طبیعی از اکوسیستم‌های آب شیرین هستند، اما مقادیر بیش‌ازحد جلبک می‌تواند باعث مشکلات متعددی مانند عوارض پوستی و بوی بد شده و سطح اکسیژن محلول را کاهش دهد. علاوه بر این شکوفه‌های جلبکی مضر یک مسأله زیست‌محیطی شدید ایجاد می‌کنند و پیامدهای اقتصادی و زیست‌محیطی قابل‌توجهی بر پهنه‌های ساحلی دارند. بر همین اساس پیش‌بینی وقوع این شکوفه‌ها به‌طور فزاینده‌ای برای جوامع ساحلی حیاتی شده است. براساس بررسی‌های صورت گرفته غلظت سطوح بسیار بالای کلروفیل آ نشان‌دهنده کیفیت پایین آب است و وجود طولانی‌مدت غلظت‌های بالای کلروفیل آ مشکلات اساسی برای تولید اولیه زیست‌توده است. بر همین اساس غلظت کلروفیل آ یک شاخص کلیدی برای کیفیت آب به‌حساب می‌آید. در این پژوهش سعی شده است تا با استفاده از تکنیک‌های سنجش‌ازدور و مدل‌های داده‌کاوی اقدام به برآورد میزان کلروفیل آ در مخزن سد سردشت گردد.

مواد و روش‌ها: اطلاعات مورداستفاده در این پژوهش به دو قسمت تقسیم می‌شوند. قسمت اول مقادیر اندازه‌گیری شده کلروفیل آ در مخزن سد سردشت است. این اطلاعات شامل ۷ مرحله اندازه‌گیری در ۱۰ نقطه با مختصات مختلف در مخزن سد بوده و مربوط به اسفندماه ۱۳۹۶ تا خردادماه ۱۳۹۸ می‌باشد. قسمت دوم اطلاعات مورداستفاده، تصاویر ماهواره سنتینل-۲ بوده که از اطلاعات آن در مدل‌های داده‌کاوی استفاده شده است. اطلاعات باندهای مختلف از

تصاویر سنتینل-۲ استخراج گردید و بر اساس مقادیر باندی، معیارهای سنجش میزان کلروفیل آ محاسبه شده و در اختیار مدل‌های داده‌کاوی جهت آموزش قرار گرفت. در این پژوهش از سه مدل داده‌کاوی XGBoost، M5 و MARS جهت برآورد میزان کلروفیل آ استفاده شد. در این مطالعه، ۹ معادله برآورد کلروفیل آ به‌عنوان ورودی مدل‌های داده‌کاوی در نظر گرفته شد و مقدار لگاریتم کلروفیل آ اندازه‌گیری شده به‌عنوان خروجی لحاظ گردید. ۸۰ درصد داده‌های موجود جهت آموزش مدل‌های داده‌کاوی و ۲۰ درصد باقی‌مانده جهت صحت‌سنجی کارایی مدل‌های مورداستفاده به‌کار رفت.

یافته‌ها: براساس اطلاعات ورودی، مدل درختی M5 فضای مسأله را به ۵ قسمت تقسیم کرده و به ازای هر بخش معادله خطی ارائه داده است. براساس ساختار ارائه‌شده توسط الگوریتم‌های M5 و MARS، ترکیب‌های باندی آبی، قرمز و سبز و نیز مادون‌قرمز و قرمز تأثیر بالایی بر روی مدل‌های ارائه‌شده توسط این دو الگوریتم داشته‌اند. نتایج به‌دست‌آمده از الگوریتم XGBoost نشان‌دهنده اهمیت ترکیب‌های باندی آبی، قرمز و سبز بر روی نتایج ارائه‌شده می‌باشد. بر این اساس ترکیب باندی آبی، قرمز و سبز در هر سه الگوریتم به‌عنوان مهم‌ترین و یا یکی از مهم‌ترین متغیرهای ورودی جهت محاسبه کلروفیل آ مورداستفاده قرار گرفته است. ضریب تبیین برای سه مدل XGBoost، M5 و MARS به ترتیب برابر با ۰/۶۱، ۰/۴۹ و ۰/۳۱ محاسبه شد. مقدار ضریب ناش-ساتکلیف برای سه مدل XGBoost، M5 و MARS به‌ترتیب برابر با ۰/۵۴، ۰/۴۷ و ۰/۲۷ محاسبه شد که نشان می‌دهد نتایج دو مدل XGBoost و M5 دارای وضعیت مطلوبی می‌باشند.

نتیجه‌گیری: نتایج مدل‌های مورداستفاده نشان می‌دهد که دو مدل XGBoost و M5 نتایج دقیق‌تری را نسبت به مدل MARS ارائه نمودند. استفاده از دیاگرام تیلور نیز نشان‌دهنده نزدیک بودن کارایی دو مدل XGBoost و M5 در محاسبه میزان کلروفیل آ می‌باشد. توزیع کلروفیل آ در محدوده مخزن سد سردشت توسط مدل‌های مورداستفاده نشان می‌دهد که در نواحی محدودی از سد، مقادیر ارائه‌شده با مقادیر اندازه‌گیری شده همخوانی ندارد که محدود بودن داده‌های اندازه‌گیری مورداستفاده و عدم انطباق کامل زمانی تصاویر سنتینل-۲ با داده‌های اندازه‌گیری در مخزن سد می‌تواند تأثیر مهمی در این زمینه داشته باشد. استفاده از تعداد داده‌های متعدد در مخازن سدهای مختلف، به‌کارگیری مدل‌های داده‌کاوی متنوع و استفاده از تصاویر سایر سنجنده‌ها می‌تواند ابزار مناسبی را در اختیار مدیران مخازن قرار داده تا بتوانند با دقت بیش‌تری اقدام به ارزیابی کیفی آب مخازن نمایند.

استناد: ظهیری، جواد، چراغی، میترا، سالاری‌جزی، میثم (۱۴۰۳). برآورد میزان کلروفیل آ در مخزن سد با استفاده از تکنیک‌های سنجش از دور و مدل‌های داده‌کاوی. *پژوهش‌های حفاظت آب و خاک*، ۳۱ (۳)، ۱۰۸-۸۵.
DOI: 10.22069/jwsc.2024.22570.3739

ناشر: دانشگاه علوم کشاورزی و منابع طبیعی گرگان

مقدمه

کلروفیل به گیاهان (ازجمله جلبک‌ها) اجازه می‌دهد تا فتوسنتز کنند، یعنی از نور خورشید برای تبدیل مولکول‌های ساده به ترکیبات آلی استفاده کنند. کلروفیل آ نوع غالب کلروفیل است که در گیاهان سبز و جلبک‌ها یافت می‌شود. کلروفیل آ به‌عنوان معیاری جهت اندازه‌گیری میزان جلبک در حال رشد در آب مورداستفاده قرار می‌گیرد که می‌توان از آن برای طبقه‌بندی وضعیت تغذیه‌ای پهنه‌های آبی استفاده کرد. اگرچه جلبک‌ها بخشی طبیعی از اکوسیستم‌های آب شیرین هستند، اما مقادیر بیش‌ازحد جلبک می‌توانند باعث مشکلات متعددی مانند عوارض پوستی و بوی بد شده و سطح اکسیژن محلول را کاهش دهند (۱). برخی از جلبک‌ها نیز سمومی تولید می‌کنند که وقتی در غلظت‌های بالا یافت می‌شوند می‌توانند برای سلامت عمومی نگران‌کننده باشند. یکی از علائم کیفی آب آلوده، افزایش زیست‌توده جلبک است که با غلظت کلروفیل آ اندازه‌گیری می‌شود. آب‌هایی با سطوح بالای مواد مغذی از کودها، سیستم‌های سپتیک، تصفیه‌خانه‌های فاضلاب و رواناب شهری ممکن است دارای غلظت بالایی از کلروفیل آ و مقادیر اضافی جلبک باشند. اندازه‌گیری غلظت کلروفیل آ در آب جایگزینی برای اندازه‌گیری واقعی زیست‌توده جلبک است و برای تخمین وضعیت تغذیه‌ای استفاده می‌شود (۲). طبق گفته سازمان مهندسین ارتش ایالات‌متحده[1] و سازمان زمین‌شناسی ایالات‌متحده[2] مشکلات مربوط به شکوفه‌های مضر جلبکی[3] در سال‌های اخیر افزایش قابل‌توجهی داشته است (۳). شکوفه‌های مضر جلبکی در حال حاضر یک مشکل جهانی در ۴۵ کشور در سراسر جهان هستند و در حداقل ۲۷ ایالت از ایالات‌متحده در مرگ حیوانات نقش دارند (۴). این جلبک‌ها بر اساس آژانس حفاظت از محیط‌زیست ایالات‌متحده[4] می‌توانند درماتوکسین، هپاتوکسین و نوروتوکسین تولید می‌کنند (۱). تماس با این جلبک‌ها می‌تواند باعث مسموم شدن کبد و کلیه و مسمومیت عصبی شود که منجر به سردرد، بی‌حسی، سرگیجه، مشکل در تنفس و در موارد نادر مرگ شود (۳).

دستورالعمل‌های سازمان بهداشت جهانی[5] برای آب آشامیدنی استانداردی را برای میکروسیستین LR، یک سم سیانوباکتری رایج، برابر با یک ppb[6] یا کم‌تر از آن تعیین کرده است. وزارت بهداشت اوهایو هنگامی‌که سطح میکروسیستین از ۶ ppb بیش‌تر شود، توصیه‌های بهداشت عمومی صادر می‌کند و زمانی که سطح میکروسیستین از ۲۰ ppb بیش‌تر شود، توصیه‌های در جهت عدم تماس مستقیم با آب صادر می‌کند (۵).

اهمیت شکوفه‌های جلبکی، توسعه الگوریتم‌های بازتاب ماهواره‌ای را برای تخمین کلروفیل آ و زیست‌توده فیتوپلانکتون مرتبط با آن به یک اولویت تحقیقاتی بالا تبدیل کرده‌اند (شکل ۱). اگرچه الگوریتم‌های کلروفیل آ بین شکوفه‌های جلبکی مضر و کم‌مضر تفاوتی قائل نمی‌شوند، اما به‌راحتی با سیستم‌های تصویربرداری ماهواره‌ای موجود سازگار می‌شوند و ممکن است به مدیران منابع آب کمک کنند تا بر این اساس بر کاهش خطرهای بالقوه این جلبک‌ها تمرکز کنند.

در زمینه نظارت بر غلظت کلروفیل آ، در نظر گرفتن خواص نوری انواع مختلف آب ضروری است. آب‌های اقیانوس آزاد معمولاً تحت سلطه فیتوپلانکتون‌ها هستند (۶)، درحالی‌که ویژگی‌های نوری آب‌های ساحلی و دریاچه‌ها علاوه بر فیتوپلانکتون‌ها تحت‌تأثیر رسوبات معلق و مواد زردرنگ قرار

1- U.S. Army Corps of Engineers (USACE)
2- U.S. Geological Survey (USGS)
3- Harmful algal blooms (HABs)
4- U.S. Environmental Protection Agency (USEPA)
5- World Health Organization
6- One part per billion

می‌گیرند که تخمین کلروفیل آ را به یک کار چالش‌برانگیز تبدیل می‌کند (۷ و ۸). برای غلبه بر این چالش‌ها و تسهیل نظارت بر محیط‌زیست، سنجش‌ازدور به‌عنوان ابزاری مؤثر برای مشاهده و تخمین غلظت کلروفیل آ در محیط‌های آبی مختلف، ازجمله اقیانوس‌ها، مناطق ساحلی، دریاچه‌های داخلی و رودخانه‌ها مورداستفاده قرار گرفته شده است (۹).

طاهری و همکاران (۱۳۹۷) از تصاویر لندست ۷ جهت شبیه‌سازی مقادیر غلظت کلروفیل a در مخزن سد اکباتان استفاده کردند (۱۰). بر همین اساس تبدیلات مختلفی ازجمله مجذور، مربع، لگاریتم، تفاضل، تابع‌نمایی، NDI[1] و نسبت باندها روی بازتابش باندها صورت گرفت و رابطه بین غلظت کلروفیل با بازتابش موردبررسی قرار گرفت. رابطه NDI با الهام از NDVI[2] و به‌صورت ترکیب دوبه‌دوی باندها استخراج گردید. نتایج مطالعه نشان داد که نسبت باندها از دقت بالاتری در تخمین مقادیر کلروفیل a برخوردار بوده است.

مبارک حسن و همکاران (۱۴۰۰) به بررسی اثر گردوخاک بر روی میزان غلظت کلروفیل در دریای عمان و خلیج‌فارس در جنوب و دریای خزر در شمال ایران پرداختند (۱۱). اطلاعات مورداستفاده در پژوهش شامل غلظت گردوخاک سطحی با استفاده از مدل MERRA-2، عمق نوری هواویزها و کلروفیل از تصاویر ماهواره مودیس در بازه ۲۰۰۷ تا ۲۰۱۷ بوده است. نتایج نشان‌دهنده تأثیر دو کانون اصلی گردوخاک بر روی منابع آبی ایران بوده است. یکی از کانون‌ها مناطق مرکزی و جنوب شرق عراق بوده و دیگری بیابان‌های ترکمنستان است که جنوب دریای خزر را تحت‌تأثیر قرار می‌دهد. ماتسوشیتا و همکاران (۲۰۱۵) از یک الگوریتم ترکیبی برای بازیابی مقادیر کلروفیل آ با استفاده از داده‌های سنجش از راه دور استفاده کردند (۱۲). این الگوریتم ترکیبی شامل سه الگوریتم تخمین کلروفیل آ است که قبلاً برای آب‌های شفاف (الگوریتم باندهای سبز و آبی)، آب‌های گل‌آلود (الگوریتم مبتنی بر باندهای قرمز و مادون‌قرمز)، و آب‌های بسیار کدر (الگوریتم مبتنی بر باندهای سه‌گانه) توسعه داده شده‌اند. برای ارزیابی عملکرد الگوریتم‌های ترکیبی پیشنهادی، از داده‌های سنجش‌ازدور و مقادیر کلروفیل آ جمع‌آوری‌شده از پنج دریاچه آسیایی استفاده گردید. نتایج نشان داد که میانگین مطلق خطای الگوریتم ترکیبی برای طیف گسترده‌ای از نمونه‌های مشاهداتی کم‌تر از ۱۳/۳ درصد محاسبه شد که نشان‌دهنده عملکرد مناسب الگوریتم پیشنهادی است. بک و همکاران (۲۰۱۶) از ۱۰ الگوریتم بازتاب ماهواره‌ای متداول و دو الگوریتم جدید برای تخمین کلروفیل a در یک مخزن آب در جنوب‌غربی اوهایو با استفاده از تصاویر هواپیمای ابرطیفی هم‌زمان با اندازه‌گیری سطحی که در عرض ۱ ساعت پس از دریافت تصویر صورت گرفت، استفاده کردند (۵). الگوریتم‌های توسعه داده‌شده جهت برآورد شکوفه‌های جلبکی (به‌ویژه شکوفه‌های جلبکی سمی یا مضر) مورداستفاده قرار گرفت. نتایج این پژوهش نشان داد که بازتاب‌های مربوط به تصاویر CASI، WorldView 2، Sentinel-2 و MERIS نسبت به تصاویر MODIS دقت بالاتری در تخمین کلروفیل a داشته‌اند. ریو و همکاران (۲۰۲۳) از تصاویر ماهواره مودیس جهت بررسی کارایی الگوریتم‌های برآورد کلروفیل آ در دریای ژاپن استفاده نمودند (۱۳). در این مطالعه الگوریتم محاسبه کلروفیل اقیانوس ناسا[3] که جهت برآورد کلروفیل آ در اقیانوس‌ها براساس نسبت باندهای آبی تا سبز مورداستفاده قرار می‌گیرد، با استفاده از داده‌های مشاهداتی دریای ژاپن مورد ارزیابی قرار گرفت. نتایج این مطالعه نشان داد که الگوریتم کلروفیل آ ناسا در

1- Normalized Index Difference
2- Normalized Difference Vegetation Index
3- NASA ocean chlorophyll-type (OCx)

غلظت‌های پایین کلروفیل آ بین ۰/۱ تا ۱/۰ میلی‌گرم بر مترمکعب دارای بیش‌برآورد است. الگوریتم توسعه داده‌شده براساس رگرسیون درجه سوم با توجه به داده‌های اندازه‌گیری شده و حداکثر نسبت‌های باندی استخراج گردید که نسبت به الگوریتم پیشنهادی ناسا از دقت بهتری برخوردار بوده است. لی و همکاران (۲۰۲۴) برای پیش‌بینی توزیع روزانه کلروفیل آ در سطح آب یک مدل شبکه عصبی توسعه دادند (۱۴). در این مطالعه اطلاعات جمع‌آوری‌شده در مورد دما، شوری، نیتروژن معدنی محلول، فسفر آلی محلول، و زئوپلانکتون با داده‌های سنجش‌ازدور کلروفیل آ برای آموزش مدل عصبی استفاده شد. نتایج این مطالعه نشان داد که مدل عصبی توسعه داده‌شده می‌تواند به‌طور مؤثر تغییرات کلروفیل آ روزانه و فصلی را برآورد نماید. از این مدل جهت شبیه‌سازی اطلاعات مکانی-زمانی مربوط به شکوفه‌های مضر جلبکی ناشی از بارش شدید طوفان Lekima در سال ۲۰۱۹ استفاده گردید که نتایج آن نشان‌دهنده دقت مناسب مدل بوده است. علاوه بر این، از روش توسعه داده‌شده می‌توان جهت بازسازی داده‌های گمشده جهت شبیه‌سازی‌های طولانی‌مدت به‌ویژه در مناطق نزدیک ساحل استفاده نمود.

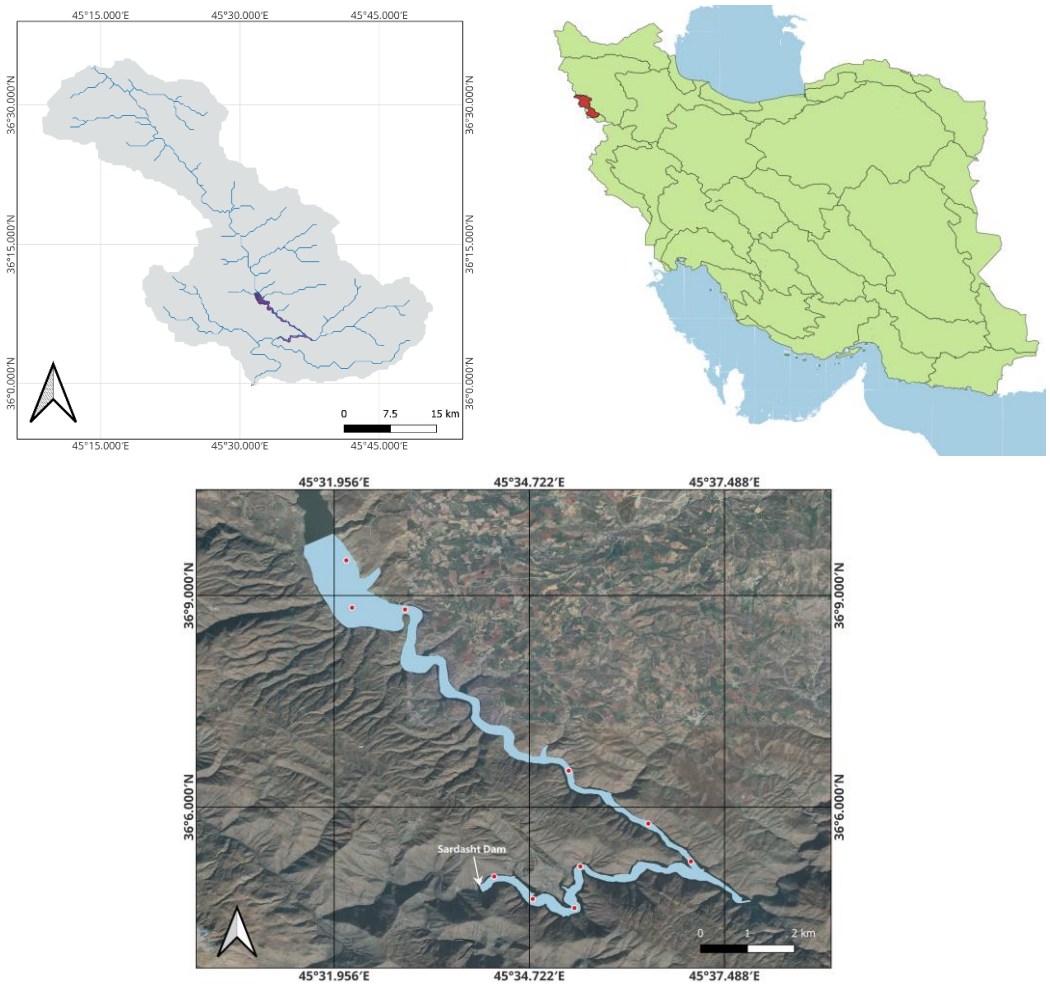
مدل‌های مبتنی بر فرآیند هیدروبیوژئوشیمیایی[1] دقت معقولی را در پیش‌بینی متغیرهای هیدرودینامیکی و مواد مغذی نشان می‌دهند، اما در پیش‌بینی کلروفیل آ مؤثر نیستند. این مدل‌ها از ترکیب معادلات دیفرانسیل جزئی حاکم بر هیدرودینامیک جریان و معادله انتشار پخش استفاده می‌کنند. تکنیک‌های یادگیری ماشینی صرفاً مبتنی بر داده نیز محدودیت‌هایی در پیش‌بینی دقیق کلروفیل آ دارند (۱۴). بر همین اساس در این پژوهش سعی شده است تا از مدل‌های داده‌کاوی در کنار تکنیک‌های سنجش‌ازدور جهت شبیه‌سازی کلروفیل آ استفاده شود. مدل‌های داده‌کاوی مورداستفاده شامل مدل درختی M5، مدل MARS و مدل XGBOOST بوده که ازجمله قوی‌ترین مدل‌های رگرسیونی به‌حساب می‌آیند. جهت بررسی کارایی روش‌های پیشنهادی در این مطالعه از اطلاعات مخزن سردشت استفاده شده است.

مواد و روش‌ها

محدوده مورد مطالعه در پژوهش حاضر مخزن سد سردشت در آذربایجان غربی می‌باشد. شهرستان سردشت در جنوب غربی استان آذربایجان غربی با مساحتی حدود ۱۳۷۶ کیلومترمربع و ارتفاعی حدود ۱۵۱۵ متر از سطح آزاد دریا قرار گرفته است. سد سردشت یکی از مهم‌ترین زیرساخت‌های عمرانی این شهرستان بوده که سدی خاکی با هسته رسی است. این سد بر روی رودخانه زاب در فاصله ۱۳ کیلومتری جنوب شرقی شهرستان سردشت احداث شده است. ارتفاع سد ۱۱۶ متر بوده که ۱۱۴ متر از بستر ارتفاع داشته و حجم مخزن آن در شرایط نرمال ۳۳۸ میلیون مترمکعب و حجم بدنه سد ۳/۰۸ میلیون مترمکعب است (۱۵). منطقه مورد مطالعه شامل مخزن سد سردشت و موقعیت جغرافیایی آن در شکل ۱ ارائه شده است. دریاچه سد سردشت منبع اصلی تأمین آب شرب شهرهای سردشت و ربط با جمعیتی در حدود ۸۰ هزار نفر است. آب دریاچه پس از انتقال به تصفیه خانه آب سردشت به دو شهر مذکور منتقل می‌شود. بر همین اساس بررسی کیفی آب مخزن از اهمیت بالایی برخوردار است. علاوه بر این از آب مخزن سردشت جهت تامین نیاز کشاورزی و نیز جهت تولید برق استفاده می‌گردد. بر اساس مطالعات شرکت مهاب قدس، نیاز سالانه شرب، کشاورزی، زیست‌محیطی و برقابی به ترتیب ۰/۴۵، ۲/۳، ۰/۶ و ۹۰ مترمکعب بر ثانیه می‌باشد. فاضلاب‌های متعددی به مخزن سد وارد می‌شوند که از جمله مهم‌ترین آن‌ها

1- Hydro-biogeochemical

می‌توان به فاضلاب کشتارگاه و شیرابه محل دفن (دره سردشت)، فاضلاب تصفیه شده و خام سردشت (شواشان) و فاضلاب ربط و فاضلاب نلاس اشاره کرد (۱۶).



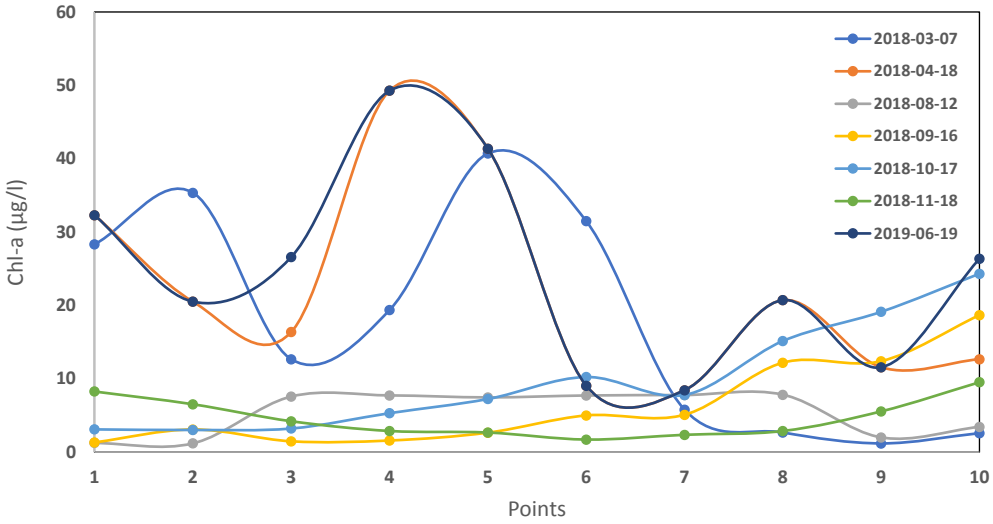
شکل ۱- منطقه مورد مطالعه شامل مخزن سد سردشت و موقعیت قرارگیری سد (موقعیت نقاط اندازه‌گیری کلروفیل آ با نقاط قرمز مشخص شده است).

Figure 1. The studied area including the Sardasht dam reservoir and the location of the dam (the location of the chlorophyll a measurement points are marked with red dots).

اطلاعات مورد استفاده در این پژوهش به دو قسمت تقسیم می‌شوند. قسمت اول مقادیر اندازه‌گیری شده کلروفیل آ در مخزن سد سردشت می‌باشد. این اطلاعات شامل ۷ مرحله اندازه‌گیری در نقاط مختلف مخزن سد بوده و مربوط به اسفندماه ۱۳۹۶ تا خردادماه ۱۳۹۸ می‌باشد (۱۵). در هر مرحله، اندازه‌گیری در ۱۰ نقطه با مختصات مختلف در مخزن سد صورت گرفته که موقعیت نقاط اندازه‌گیری در شکل ۱ با دایره‌های قرمزرنگ مشخص شده است. اطلاعات مورد استفاده در این مطالعه در شکل ۲ ارائه شده است. قسمت دوم اطلاعات مورد استفاده تصاویر ماهواره سنتینل-۲[1] بوده که از اطلاعات آن در مدل‌های داده‌کاوی استفاده شده است. ماهواره سنتینل-۲ تحت برنامه کوپرنیک[2] توسط سازمان

1- Sentinel-2
2- Copernicus

فضای اروپا جهت تهیه تصاویر اپتیک با قدرت تفکیک مکانی بالا جهت پایش زمین، تهیه نقشه‌های پوششی، مدیریت بحران و سیستم‌های هشداردهنده و نیز کمک‌های بشردوستانه در سال ۲۰۱۵ به فضا ارسال شد (۱۷). سنتینل-۲ یک سیستم حسگر چندطیفی است که از دامنه طول‌موج مرئی[1] تا دامنه طول‌موج کوتاه مادون‌قرمز[2] کار می‌کند. تصاویر این ماهواره دارای قدرت تفکیک مکانی ۱۰ تا ۲۰ متر (۶۰ متر برای سه باند اتمسفر) با زمان بازدید نظری پنج روز هنگام ترکیب سنتینل-۲ A و B ارائه می‌دهد (۱۸).

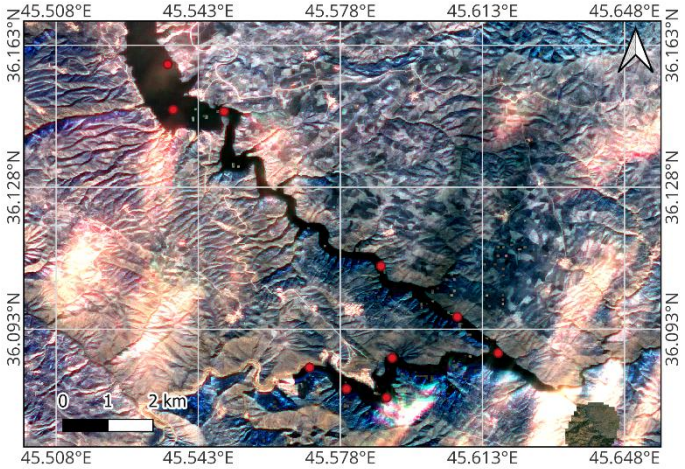


شکل ۲- اطلاعات مورد استفاده در مطالعه حاضر (۱۵).

Figure 2. Information used in the present study (15).

در این مطالعه جهت استخراج مقادیر بازتابش باندها از تصاویر سنتینل-۲ در همان تاریخ‌های اندازه‌گیری مقادیر کلروفیل آ در مخزن سد استفاده گردید. در بعضی از موارد با توجه به عدم وجود تصویر در تاریخ موردنظر، تصاویر با تأخیر زمانی ۱ و یا ۲ روز استفاده گردید. سامانه مورد استفاده در این پژوهش، سامانه گوگل ارث انجین[3] بوده که کدهای برنامه در زبان جاوا اسکریپت توسعه پیدا کردند. تصاویر مورد استفاده همراه با تصحیحات اتمسفریک بوده و تصاویر بالای ۲۰٪ ابر از مجموعه تصاویر حذف گردید. نمونه‌ای از تصویر تصحیح‌شده سنتینل-۲ به همراه نقاط اندازه‌گیری کلروفیل آ در محدوده مورد مطالعه در شکل ۳ ارائه شده است.

1- Visible (VIS)
2- Shortwave-infrared (SWIR)
3- Google Earth Engine



شکل ۳- تصویر ماهواره سنتینل-۲ از محدوده مورد مطالعه به همراه نقاط اندازه‌گیری مقادیر کلروفیل آ.

Figure 3. Sentinel-2 satellite image of the studied area along with the measuring points of chlorophyll a.

در این مطالعه سعی شده است تا از اغلب معادلات متداول موجود جهت محاسبه کلروفیل آ بر اساس بازتابش باندی استفاده شود. معادلات مورد استفاده در این مطالعه در جدول ۱ ارائه شده است. در جدول ۱، باندهای آبی (B)، سبز (G)، قرمز (R) و مادون قرمز نزدیک (NIR) مشخص شده است. پارامتر A از معادله A=(NIR-G)/(2NIR-R-G) محاسبه شده است.

جدول ۱- معادلات مورد استفاده جهت توسعه مدل‌های داده‌کاوی.

Table 1. Equations used to develop data driven models.

عنوان Features	فرم الگوریتم Algorithm form	منبع Source
BG	B/G	اریلی و وردل، ۲۰۱۹ (۱۹) O'Reilly and Werdell (2019)
BRG (CI)	G-[B+(G-B)/(R-B)*((R-B))]	هو و همکاران، ۲۰۱۹ (۲۰) Hu et al. (2019)
NIRGR	NIR-G+(G-R)*A	ژینگ و هو، ۲۰۱۶ (۲۱) Xing and Hu (2016)
RG	R/G	واتانابه و همکاران، ۲۰۱۷ (۲۲) Watanabe et al. (2017)
NIRR	NIR/R	دوان و همکاران، ۲۰۰۷ (۲۳) Duan et al. (2007)
RB	R/B	تان و همکاران، ۲۰۱۷ (۲۴) Tan et al. (2017)
NIRG	NIR/G	انگوین و همکاران، ۲۰۲۰ (۲۵) Nguyen et al. (2020)
BRGII	(B-R)/G	بوچاروف و همکاران، ۲۰۱۷ (۲۶) Bocharov et al. (2017)
NIRRII	(NIR-R)/(NIR+R)	مثیو و اودرمت، ۲۰۱۵ (۲۷) Matthews and Odermatt (2015)

در این پژوهش جهت برآورد مقادیر کلروفیل آ در مخزن سد از مدل‌های داده‌کاوی XGBoost، M5 و MARS استفاده شده است. اطلاعات مورد استفاده جهت آموزش و صحت‌سنجی مدل‌های داده‌کاوی، الگوریتم‌های ارائه‌شده در جدول ۱ بوده که مقادیر الگوریتم‌ها بر اساس مقادیر بازتابش سطحی از تصاویر سنتینل-۲ استخراج گردید. علاوه بر این مقادیر کلروفیل آ اندازه‌گیری شده در مخزن سد سردشت به‌عنوان خروجی به مدل‌های داده‌کاوی معرفی گردید. ۸۰ درصد اطلاعات جهت آموزش مدل‌ها و ۲۰ درصد باقی‌مانده جهت صحت‌سنجی روش‌ها داده‌کاوی مورد استفاده قرار گرفت. در این مطالعه جهت اجرای مدل XGBoost کد برنامه‌نویسی به زبان پایتون توسعه یافت. هم‌چنین برای مدل‌های M5 و MARS به ترتیب از نرم‌افزارهای Weka و STATISTICA استفاده گردید. نرم‌افزار Weka به‌صورت منبع باز بوده که توسط دانشگاه Waikato توسعه یافته و از آن می‌توان جهت ساخت و آموزش مدل‌های داده‌کاوی استفاده نمود (۲۸).

الگوریتم افزایش شدید گرادیان[1] (XGBoost) که اولین بار توسط چن و گسترین (۲۰۱۶) ارائه شد روشی جدید برای پیاده‌سازی ماشین تقویت گرادیان[2] به‌ویژه طبقه‌بندی K[3] و درختان رگرسیون به‌حساب می‌آید (۲۹). این الگوریتم مبتنی بر ایده "تقویت" است، که از طریق استراتژی‌های آموزشی، تمام پیش‌بینی‌های یادگیرندگان "ضعیف" را برای ایجاد یک یادگیرنده "قوی" ترکیب می‌کند. هدف XGBoost جلوگیری از تطبیق بیش‌ازحد و هم‌چنین بهینه‌سازی منابع محاسباتی است. این کار با ساده‌سازی توابع هدف به‌دست می‌آید که با حفظ سرعت محاسباتی بهینه، امکان ترکیب بخش‌های پیش‌بینی و تنظیم را فراهم می‌کند. هم‌چنین، محاسبات موازی برای توابع در XGBoost در مرحله آموزش به‌طور خودکار اجرا می‌شود (۳۰).

در فرآیند یادگیری الگوریتم XGBoost یادگیرنده اول ابتدا به کل فضای داده‌های ورودی برازش داده می‌شود و سپس مدل دوم برای رفع اشکالات یادگیرنده ضعیف بر روی باقی‌مانده‌ها برازش داده می‌شود. این فرآیند برازش برای چند بار تکرار می‌شود تا زمانی که معیار توقف برآورده شود. پیش‌بینی نهایی مدل از مجموع پیش‌بینی هر یادگیرنده به‌دست می‌آید. تابع کلی برای پیش‌بینی در مرحله t به‌صورت زیر ارائه می‌شود:

$$f_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = f_i^{(t-1)} + f_t(x_i) \qquad (۱)$$

که، $f_t(x_i)$ یادگیرنده در گام t، $f_i^{(t)}$ و $f_i^{(t-1)}$ پیش‌بینی‌های مدل در گام‌های t و t-1، و x_i متغیر ورودی می‌باشد. جهت جلوگیری از اتفاق افتادن بیش‌برازش در مدل بدون تأثیر گذاشتن بر روی سرعت محاسباتی، مدل XGBoost عبارت تحلیلی زیر را جهت محاسبه نکویی مدل از تابع اصلی استخراج می‌کند:

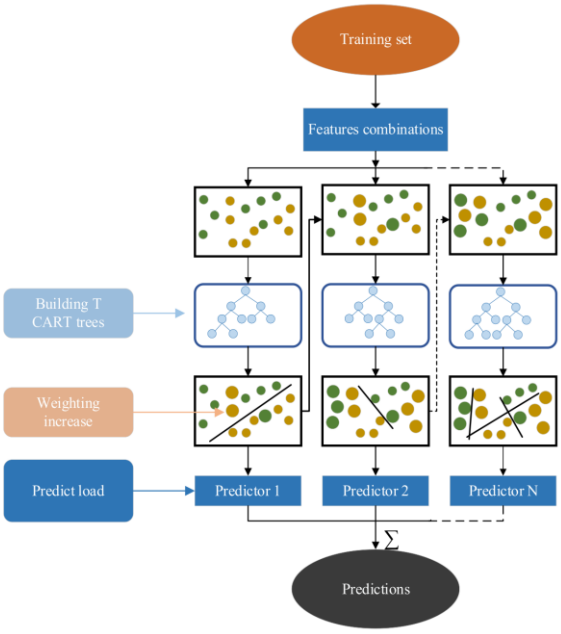
$$\mathrm{OBj}^{(t)} = \sum_{k=1}^{n} l(\bar{y}, y_i) + \sum_{k=1}^{t} \Omega(f_i) \qquad (۲)$$

که، l تابع کاهنده، n تعداد مشاهدات مورد استفاده و Ω ترم تنظیم بوده که از عبارت زیر محاسبه می‌شود:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda\|\omega\|^2 \qquad (۳)$$

که ω بردار امتیازها در برگ‌ها، γ پارامتر تنظیم و γ حداقل تلفات موردنیاز برای تقسیم‌بندی بیش‌تر گره برگ است (۳۰). نمایی از نحوه عملکرد مدل XGBoost در شکل ۴ ارائه شده است.

1- Extreme Gradient Boosting
2- Gradient Boosting Machine
3- K Classification



شکل ۴- نحوه عملکرد مدل XGBoost (۳۱).

Figure 4. Schematic illustration of the XGboost mode (31).

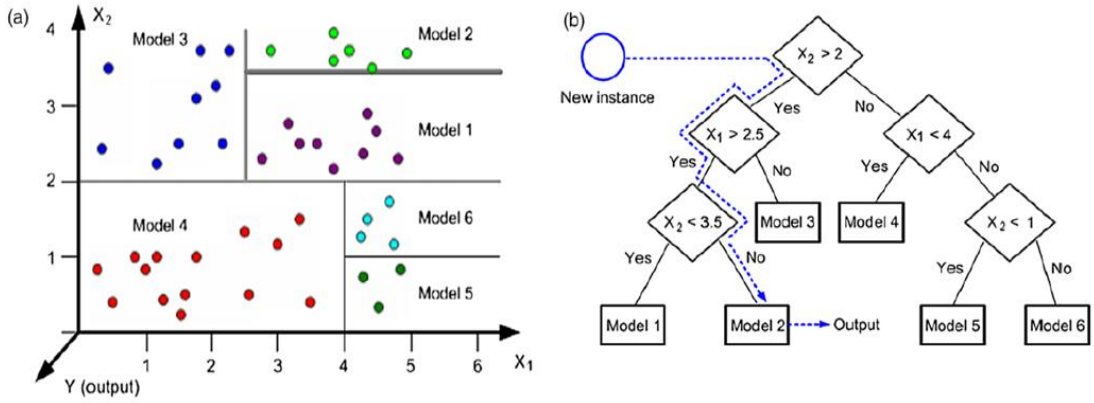
الگوریتم درختی مورد استفاده در این پژوهش، الگوریتم M5 بوده که اولین بار توسط کوئینلن (۱۹۹۲) ارائه شد (۳۲) و پس‌ازآن توسط ونگ و ویتن (۱۹۹۶) توسعه یافت (۳۳). روش M5 شاخه‌های خود را به‌صورت دوتایی و تنها بر اساس یک متغیر ایجاد می‌کند، بدین گونه که بر اساس شرطی که در هر گره تعریف می‌شود، اطلاعات در آن گره به دو قسمت تقسیم می‌شود. در روش M5 فضای مسأله به زیر دامنه‌هایی تقسیم شده و برای هر زیر دامنه یک مدل رگرسیون خطی چندمتغیره برازش داده می‌شود. فرآیند جداسازی در گره‌ها ممکن است بارها تکرار شده و در نتیجه درخت با شاخه‌های متعدد ایجاد شود. در این حالت مدل دچار بیش برازش[1] شده که از طریق هرس کردن[2] می‌توان این مشکل را رفع کرد. هرس کردن باعث کاهش خطای مورد انتظار جهت داده‌های غیرآموزشی می‌شود (۳۴). پس از هرس کردن، مدل‌های خطی مورد استفاده در برگ‌های مجاور درخت هرس شده به‌شدت دچار ناپیوستگی می‌گردند که این امر سبب از بین رفتن پیوستگی سیستم می‌گردد. بر همین اساس از مکانیسم هموارسازی[3] جهت رفع ناپیوستگی ایجادشده در مدل‌های خطی استفاده می‌شود. در این فرآیند مقدار تخمین زده‌شده در هر برگ تصحیح می‌شود. چنان‌چه نمونه موردنظر در شاخه s_i از زیر درخت s باشد، n_i تعداد نمونه‌های آموزشی در s_i، $PV(s_i)$ مقدار محاسبه‌شده در s_i و $M(s)$ مقدار محاسبه‌شده توسط مدل در s باشد، آنگاه مقدار اصلاح‌شده (PV) از رابطه زیر محاسبه خواهد شد.

$$PV = \frac{ni \times PV(si) + k \times M(s)}{ni + k} \qquad (۴)$$

در رابطه فوق، k ثابت هموارسازی بوده که به‌صورت پیش‌فرض برابر ۱۵ در نظر گرفته می‌شود. هموارسازی به‌ویژه در مواقعی که مدل‌های خطی در برگ‌های

1- Overfitting
2- Pruning
3- Smoothing process

مجاور، مقادیر کاملاً متفاوتی ارائه می‌دهند و یا مدل‌هایی که بر اساس داده‌های آموزشی محدود ساخته می‌شوند، می‌تواند به میزان زیادی مؤثر باشد (۳۲). نحوه تقسیم شدن فضای مسأله توسط الگوریتم درختی M5 در شکل ۵ ارائه شده است.



شکل ۵- تقسیم فضای مسأله و ارائه معادله خطی برای هر زیردامنه توسط مدل M5 (۳۵).

Figure 5. Splitting of the input domain by the model tree in this study, and presenting the linear equation for each subdomain by the M5 model (35).

الگوریتم رگرسیون تطبیقی چندمتغیره اسپلاین[1] (MARS) شکلی از الگوریتم‌های رگرسیونی بوده که توسط فریدمن (۱۹۹۱) جهت پیش‌بینی خروجی‌های عددی پیوسته معرفی گردید (۳۶). این الگوریتم به‌وسیله تقسیم فضای جواب به بازه‌هایی از متغیرهای پیش‌بینی‌کننده (ورودی) و برازش یک اسپلاین (تابع پایه) در هر بازه مدل‌های رگرسیونی، انعطاف‌پذیری را برای پیش‌بینی متغیر هدف ایجاد می‌نماید. تابع پایه نشان‌دهنده اطلاعاتی دربردارنده یک یا چند متغیر مستقل است. یک تابع پایه در یک بازه معین تعریف شده و نقاط ابتدایی و انتهایی آن گره نامیده می‌شود. گره مفهوم کلیدی در این روش است و بیانگر نقطه‌های است که رفتار تابع در آن نقطه تغییر می‌کند. توابع پایه ارتباط بین متغیرهای پیش‌بینی‌کننده و متغیر هدف را بیان می‌کنند، و توسط یکی از روابط زیر بیان می‌شود:

$$h_m = max(0, X - c) \quad (۵)$$

$$h_m = max(0, c - X) \quad (۶)$$

در روابط بالا، c یک مقدار آستانه[2] است.

توابع پایه به‌صورت مرحله‌ای به هر متغیر ورودی اعمال شده و مکان گره‌ها یعنی جایی که مقدار تابع تغییر می‌کند (یا شیب خطوط تغییر می‌کند)، تعیین می‌شوند. تعداد گره‌ها بر اساس یک فرآیند سعی و خطا حاصل می‌شوند. فرم عمومی مدل MARS به شکل زیر است:

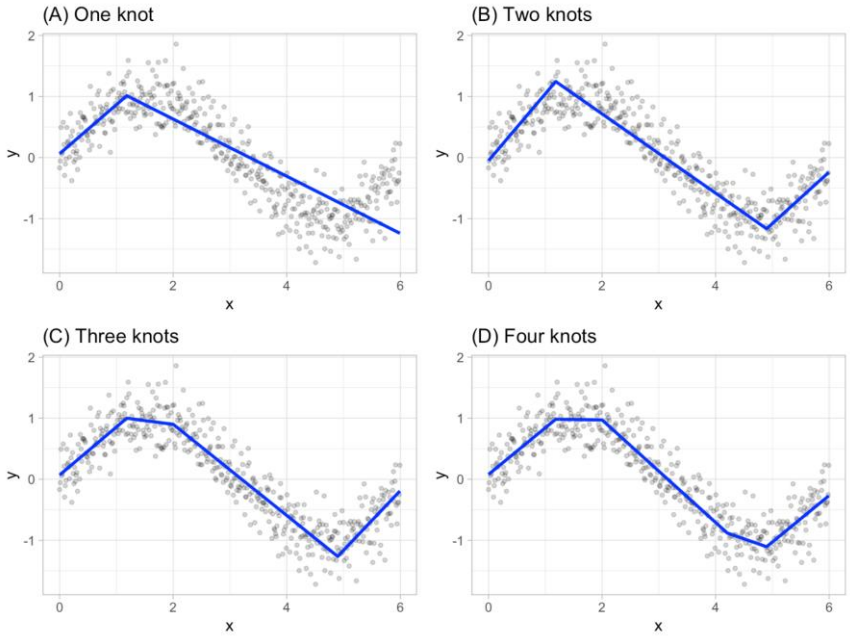
$$Y = f(x) = \beta_0 + \sum_{m=1}^{M} \beta_m h_m(x) \quad (۷)$$

در این رابطه، Y مقدار پیش‌بینی‌شده (متغیر هدف) توسط تابع f(x) است که به‌صورت ترکیبی از یک مقدار ثابت اولیه β_0 و مجموع M عبارت که هر کدام

1- Multivariate adaptive regression splines

2- Threshold

از یک ضریب β_m و یک تابع پایه $h_m(x)$ تشکیل شده‌اند، تعریف می‌شود. مثال‌هایی از رگرسیون تطبیقی اسپیلاین به ازای تعداد گره‌های مختلف در شکل ۶ ارائه شده است.



شکل ۶- مثال‌هایی از رگرسیون تطبیقی اسپلاین به ازای تعداد گره‌های مختلف (۳۷).

Figure 6. Multivariate adaptive regression splines for different nodes numbrs (37).

در این مطالعه، ۹ معادله برآورد کلروفیل آ به‌عنوان ورودی مدل‌های داده‌کاوی در نظر گرفته شد و مقدار لگاریتم کلروفیل آ اندازه‌گیری شده به عنوان خروجی لحاظ گردید. ۸۰ درصد داده‌های موجود جهت آموزش مدل‌های داده‌کاوی و ۲۰ درصد باقی‌مانده جهت صحت‌سنجی کارایی مدل‌های مورد استفاده به‌کار رفت.

جهت برآورد میزان کارایی مدل‌های مختلف، در این پژوهش از معیارهای آماری مانند ضریب تبیین[1] (R^2)، خطای جذر میانگین مربعات[2] (RMSE) و ضریب ناش-ساتکلیف[3] (NSE) استفاده شده است. معادلات مربوط به معیارهای آماری مورد استفاده به‌صورت زیر می‌باشد:

$$R^2 = 1 - \frac{\sum(Chl_{me} - \overline{Chl_{es}})^2}{\sum(Chl_{me} - \overline{Chl_{me}})^2} \quad (۸)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(Chl_{me} - Chl_{es})^2}{N}} \quad (۹)$$

$$NSE = 1 - \frac{\sum(Chl_{me} - Chl_{es})^2}{\sum(Chl_{me} - \overline{Chl_{me}})^2} \quad (۱۰)$$

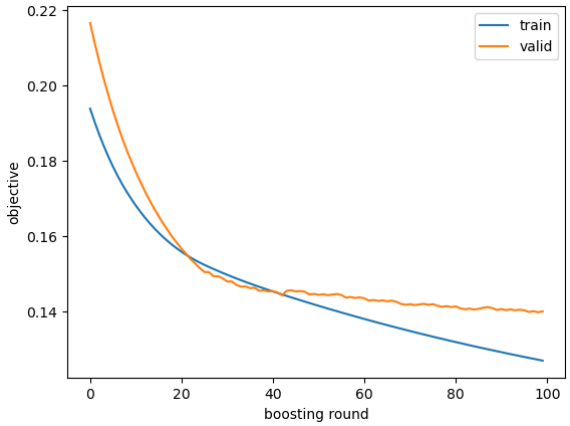
در معادلات فوق، Chl_{me} و Chl_{es} به ترتیب مقادیر کلروفیل آ اندازه‌گیری‌شده و محاسبه‌شده، $\overline{Chl_{me}}$ و $\overline{Chl_{es}}$ به‌ترتیب میانگین مقادیر کلروفیل آ اندازه‌گیری‌شده و محاسبه‌شده و N تعداد نمونه‌های اندازه‌گیری می‌باشد. RMSE هم بعد با مقادیر ورودی بوده و مقادیر کوچک‌تر آن نشان‌دهنده خطای

1- Coefficients of determination
2- Root mean squared error
3- Nash-Sutcliffe Efficiency

کم‌تر مدل می‌باشد. هرچه ضریب R^2 به یک نزدیک‌تر باشد، نشان‌دهنده تطابق بهتر بین نتایج مدل و مقادیر مشاهداتی است. NSE یک معیار نرمال شده است که نشان‌دهنده بزرگی نسبی واریانس باقی‌مانده در مقایسه با واریانس داده‌های مشاهده‌شده است. بر اساس مطالعات صورت گرفته NSE > ۰/۷ نشان‌دهنده خوب بودن نتایج بوده، ۰/۷≥ NSE >۰/۴ نشان‌دهنده رضایت‌بخش بودن نتایج و مقادیر ۰/۴ ≥ NSE بیانگر ضعیف بودن مدل‌سازی است (۳۸).

نتایج و بحث

نتایج مربوط به الگوریتم XGBoost به‌ازای داده‌های آموزش و اعتبارسنجی در شکل ۷ ارائه شده است. یکی از مهم‌ترین ایرادات مدل‌های داده‌کاوی، بیش برازش بودن آن‌هاست که باید هموار تمهیداتی را در نظر گرفت تا مدل دچار آن نشود. با توجه به شکل ۷، از دور تقویتی[1] ۴۲ به بعد، دقت مدل به ازای داده‌های آموزش به‌شدت افزایش یافته، این در حالی است که در مورد داده‌های اعتبارسنجی این مسئله دیده نمی‌شود. بر همین اساس دور ۴۲ به‌عنوان دور بهینه در نظر گرفته شد.



شکل ۷- نحوه تغییرات خطای مدل با افزایش تعداد دورهای تقویتی.

Figure 7. Model error changes with increasing number of boosting rounds.

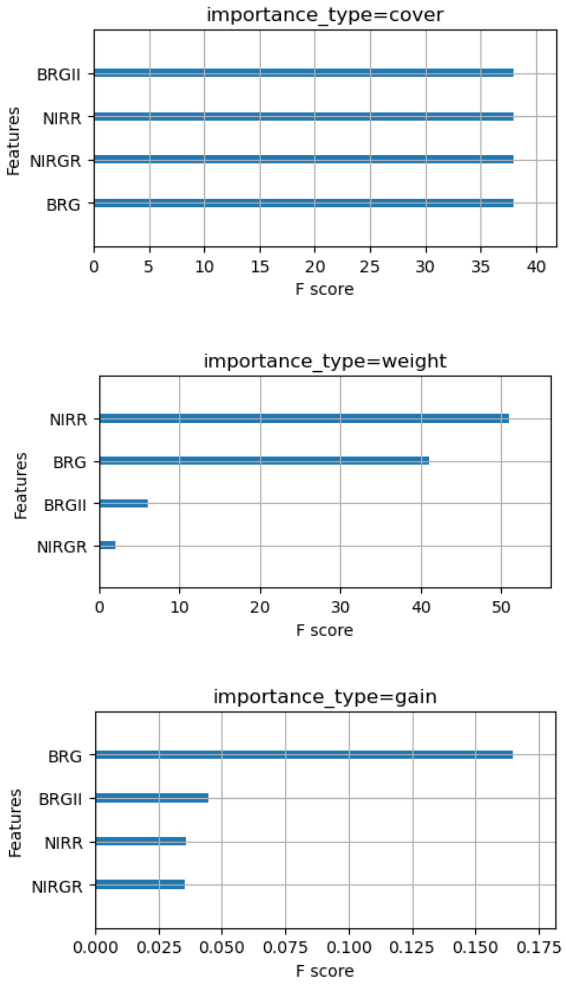
الگوریتم XGBoost به‌طور خودکار اهمیت ویژگی‌ها را با سه معیار مختلف در طول آموزش محاسبه می‌کند که این سه معیار به‌صورت زیر هستند:

وزن: تعداد تقسیم‌هایی که از یک ورودی استفاده می‌کنند، سود: میانگین سود در تابع هدف از تقسیم‌هایی که از آن ورودی استفاده می‌کنند، پوشش: میانگین تعداد نمونه‌های آموزشی تحت‌تأثیر تقسیماتی که از آن ورودی استفاده می‌کنند.

اهمیت پارامترهای ورودی در محاسبه مقادیر کلروفیل آ بر اساس سه معیار فوق در شکل ۸ ارائه شده است. بر اساس معیارهای مورد استفاده، سه ورودی BRG، NIRR و BRGII تأثیر بیش‌تری بر روی مدل ارائه‌شده دارند. استفاده از سه معیار فوق ممکن است همراه با خطا نیز باشد، به دلیل آن‌که تنها از داده‌های آموزش برای محاسبه این معیارها استفاده می‌شود و داده‌های صحت‌سنجی نقشی در محاسبه معیارها ندارند. معیار دیگری که برای تعیین اهمیت متغیرهای ورودی بر روی خروجی مورد استفاده قرار می‌گیرد و نسبت به سه معیار اشاره‌شده، دقت بیش‌تری دارد، اهمیت جایگشت ورودی[2] می‌باشد که در این مطالعه از آن استفاده شده است. این معیار

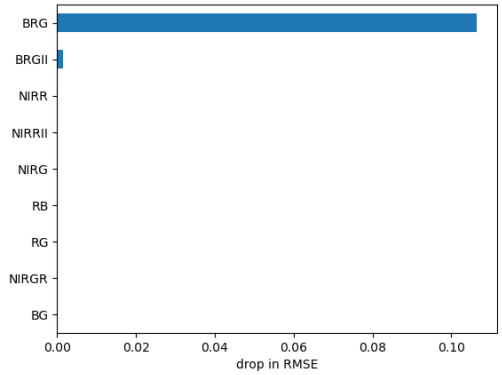
1- Boosting round
2- Permutation feature importance

بیانگر عملکرد مدل در حالتی است که مقادیر یک ورودی به‌طور تصادفی انتخاب شوند. انتخاب تصادفی مقادیر یک ورودی باعث از بین رفتن رابطه بین آن ورودی و خروجی می‌شود، بنابراین نشان می‌دهد که مدل چقدر برای پیش‌بینی به آن ورودی متکی است. همچنین از این معیار می‌توان برای داده‌های آموزشی و نیز داده‌های صحت‌سنجی استفاده نمود. نتایج معیار اهمیت جایگشت برای ورودی‌های مورد استفاده جهت محاسبه کلروفیل آ در شکل ۹ ارائه شده است. معیار مورد استفاده نشان می‌دهد که تنها دو متغیر ورودی BRG و BRGII بر روی خروجی مؤثر بوده که تأثیر متغیر BRG به‌مراتب بیش‌تر است.



شکل ۸- اهمیت پارامترهای ورودی در محاسبه مقادیر کلروفیل آ در الگوریتم XGBoost.

Figure 8. Importance of input parameters in estimation of chlorophyll a with XGBoost algorithm.



شکل ۹- نتایج معیار اهمیت جایگشت برای ورودی‌های مورد استفاده جهت محاسبه کلروفیل آ در الگوریتم XGBoost.

Figure 9. Results of permutation importance for inputs used to calculate chlorophyll a in XGBoost.

بر اساس اطلاعات ورودی، مدل درختی M5 فضای مسأله را به ۵ قسمت تقسیم کرده و به ازای هر بخش معادله خطی ارائه داده است. نحوه تقسیم فضای مسأله در مدل درختی در شکل ۱۰ ارائه شده است. بر اساس ساختار ارائه شده، ترکیب‌های باندی آبی، قرمز و سبز و نیز مادون‌قرمز و قرمز تأثیر بالایی بر روی مدل‌های ارائه شده توسط الگوریتم M5 داشته‌اند.

BRG<= 0.0034 : LM1	LM 1
BRG> 0.0034 :	LogChla = 26.60 × BRG + 0.11 × NIRR + 0.28
| BRGII <= 0.0062 : LM2	LM 2
| BRGII > 0.0062 :	LogChla = 16.41 × BRG + 0.07 × NIRR + 0.62
| | BRG <= 0.0075 : LM3	LM 3
| | BRG > 0.0075 :	LogChla = 16.41 × BRG + 0.07 × NIRR + 0.69
| | | NIRR <= 0.9155 : LM4	LM 4
| | | NIRR > 0.9155 : LM5	LogChla = 16.41 × BRG + 0.07 × NIRR + 0.65
	LM 5
	LogChla = 16.41 × BRG + 0.07 × NIRR + 0.67

شکل ۱۰- ساختار درختی ارائه‌شده توسط مدل درختی M5.

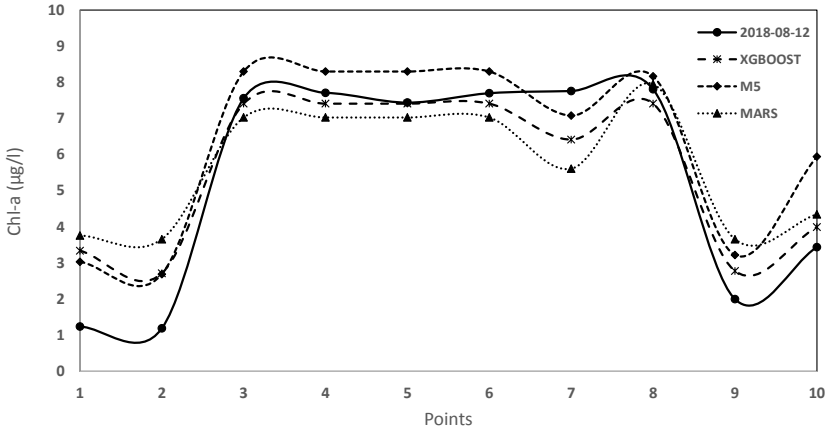
Figure 10. The tree structure provided by the M5 tree model.

ساختار ارائه‌شده توسط مدل MARS نیز براساس اطلاعات ورودی به‌صورت زیر می‌باشد:

$$\begin{aligned}\text{LogChla} = {} & 3.84\text{e-}1 + \\ & 54.2 \times \max(0, \text{BRG} - 3.12\text{e-}3) + \\ & 2.67 \times \max(0, \text{BRG} - 5.40\text{e-}3) + \\ & 0.58 \times \max(0, \text{NIRRII} + 2.65\text{e-}2) - \\ & 1.76 \times \max(0, \text{NIRRII} - 5.75\text{e-}2)\end{aligned} \quad (۱۱)$$

بر اساس ساختار ارائه‌شده توسط مدل MARS، دو متغیر ورودی BRG و NIRRII تنها بر روی مقدار کلروفیل آ مؤثر بوده و در معادله نهایی تنها این دو پارامتر مورد استفاده قرار گرفتند. مقایسه مدل‌های ارائه‌شده توسط سه الگوریتم XGBoost، M5 و MARS نشان می‌دهد که متغیر BRG در هر سه

الگوریتم به‌عنوان مهم‌ترین و یا یکی از مهم‌ترین متغیرهای ورودی جهت محاسبه کلروفیل آ ارائه شده است. نتایج مربوط به الگوریتم‌های مورد استفاده جهت محاسبه کلروفیل آ در تاریخ ۲۱ مرداد ۱۳۹۷ در شکل ۱۱ ارائه شده است.



شکل ۱۱- نتایج الگوریتم‌های مورد استفاده جهت محاسبه کلروفیل آ در تاریخ ۲۱ مرداد ۱۳۹۷.

Figure 11. The results of the algorithms used to estimate chlorophyll a on August 12, 2018.

نتایج این پژوهش با مطالعات صورت گرفته توسط هو و همکاران (۲۰) مطابقت دارد. نتایج این پژوهش‌گران نشان می‌دهد که پارامتر BRG را می‌توان جهت برآورد کلروفیل آ در مخازن مورد استفاده قرار داد. در این مطالعه نیز در هر سه الگوریتم مورد استفاده، پارامتر BRG بیش‌ترین تأثیر را داشته است. علاوه بر این، پارامتر BRGII (۲۶) نیز تأثیر به‌سزایی در ساخت مدل‌های XGBoost و M5 داشته است.

در این مطالعه از معیارهای مختلف آماری مانند ضریب تبیین، خطای جذر میانگین مربعات و ضریب ناش–ساتکلیف استفاده گردید که در بخش مواد و روش‌ها معرفی شدند. نتایج معیارهای آماری بر روی خروجی الگوریتم‌های مورد استفاده در جدول ۲ ارائه شده است. در جدول ۲، الگوریتم XGBoost برای دو حالت ارائه شده است که حالت اول مربوط به استفاده از همه ۹ معادله به عنوان ورودی الگوریتم و حالت دوم استفاده از موثرترین ورودی‌ها شامل سه ورودی BRG، NIRR و BRGII می‌باشد. مقایسه دو حالت نشان می‌دهد که تفاوت زیادی بین دو حالت وجود نداشت و ضریب ناش–ساتکلیف برای حالت اول تنها به میزان ۰/۰۷ بیش‌تر از حالت دوم محاسبه شده است. مقایسه نتایج الگوریتم‌های مختلف نشان می‌دهد الگوریتم XGBoost نسبت به دو الگوریتم دیگر از کارایی بالاتری برخوردار می‌باشد. علاوه بر این الگوریتم M5 نسبت به الگوریتم MARS عملکرد بهتری داشته است. بر اساس ضریب ناش–ساتکلیف، دو الگوریتم XGBoost و M5 در محدوده رضایت‌بخش قرار گرفته و الگوریتم MARS ازلحاظ این معیار، عملکرد ضعیفی داشته است.

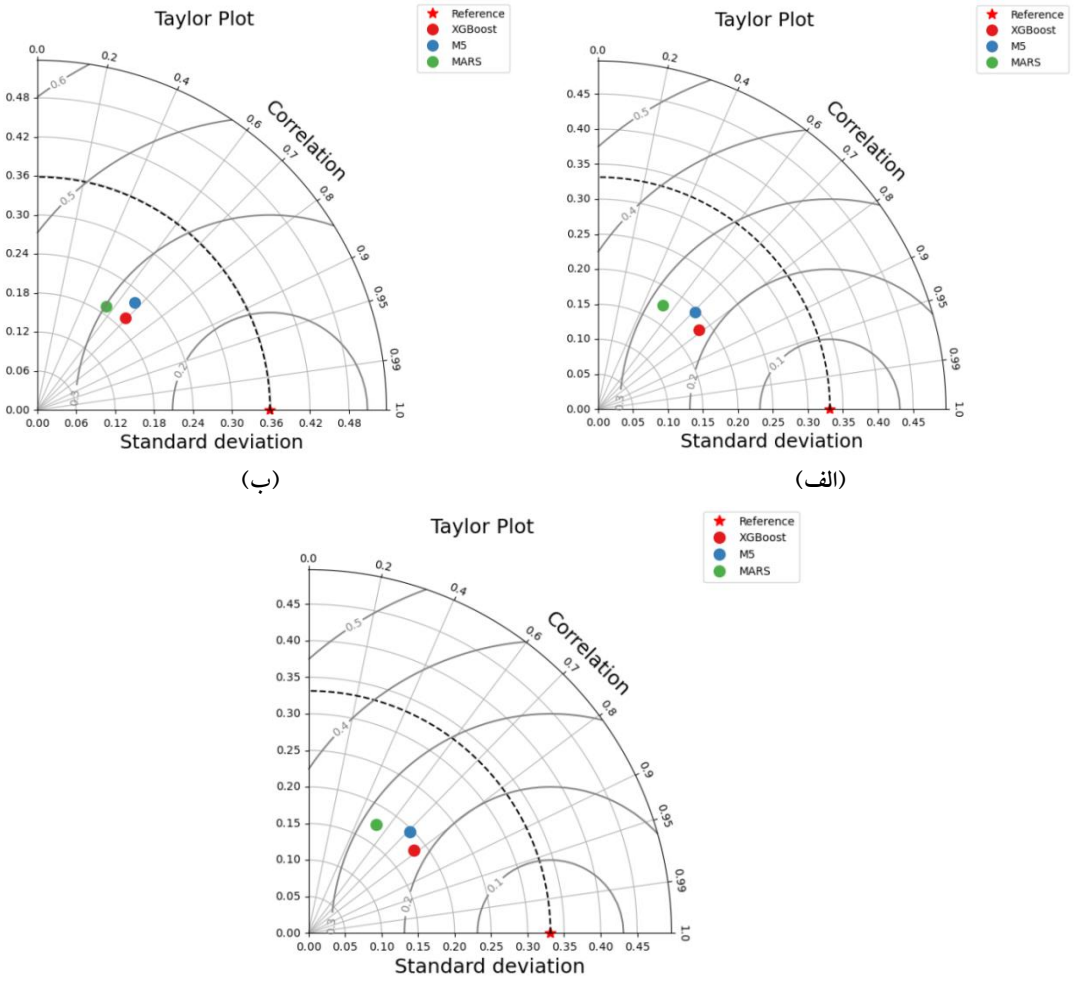
جدول ۲- نتایج معیارهای آماری مورد استفاده بر روی الگوریتم‌های مختلف.

Table 2. Results of statistical criteria used on different algorithms.

کل داده‌ها All data			صحت‌سنجی Validation			آموزش Training			الگوریتم Algorithm
RMSE	R^2	NSE	RMSE	R^2	NSE	RMSE	R^2	NSE	
0.22	0.61	0.54	0.26	0.48	0.44	0.21	0.62	0.56	XGBoost (All)
0.24	0.56	0.47	0.29	0.51	0.37	0.23	0.61	0.53	XGBoost (3 bands)
0.24	0.49	0.47	0.27	0.45	0.41	0.23	0.50	0.48	M5
0.28	0.31	0.27	0.31	0.30	0.22	0.28	0.28	0.28	MARS

جهت مقایسه دقیق‌تر کارایی مدل‌های مورد استفاده بر اساس معیارهای آماری، در این پژوهش از دیاگرام تیلور استفاده شده است. این دیاگرام از سه معیار انحراف معیار استاندارد مقادیر مشاهداتی و محاسباتی، ضریب همبستگی میان مقادیر مشاهداتی و محاسباتی و خطای مرکزی جذر میانگین مربعات[1] بهره می‌برد. استفاده از این سه معیار به‌صورت هم‌زمان و ارائه شماتیک گرافیکی باعث شده است تا دیاگرام تیلور به‌عنوان روشی قدرتمند جهت مقایسه کارایی مدل‌ها مورد استفاده قرار گیرد. شکل ۱۲ دیاگرام تیلور برای هر سه روش داده‌کاوی XGBoost، M5 و MARS در هر دو مرحله آموزش و صحت‌سنجی ارائه شده است. میزان کارایی مدل‌ها بر اساس فاصله برآیند معیارهای آماری نسبت به نقطه مرجع که نماینده داده‌های مشاهداتی است، سنجیده می‌شود. در مرحله آموزش، مدل XGBoost نسبت به سایر مدل‌ها کارایی بالاتری را نشان می‌دهد. دیاگرام تیلور در مرحله صحت‌سنجی نشان می‌دهد که کارایی دو مدل درختی M5 و XGBoost نزدیک به هم بوده و وضعیت بهتری نسبت به مدل MARS دارند. در مطالعه صورت گرفته توسط عالی‌ضمیر و همکاران (۲۰۲۱) که جهت برآورد غلظت کلروفیل آ در رودخانه از روش‌های یادگیری عمیق، شبکه‌های عصبی و مدل درختی M5 استفاده نمودند، نیز مشخص گردید که هرچند روش‌های یادگیری عمیق و شبکه‌های عصبی از دقت بهتری نسبت به مدل M5 برخوردار بوده ولی مدل درختی نیز با دقت بالایی توانسته است غلظت کلروفیل آ را برآورد نماید (۳۹). کوی و همکاران (۲۰۲۲) نیز که جهت شبیه‌سازی غلظت کلروفیل آ در اقیانوس از الگوریتم XGBoost در کنار الگوریتم‌های جنگل تصادفی، بردار پشتیبان و رگرسیون خطی استفاده کردند، نشان دادند که الگوریتم XGBoost نسبت به دیگر روش‌های مورد استفاده از دقت بهتری برخوردار است (۴۰).

1- Centered root mean square error (CRMSE)

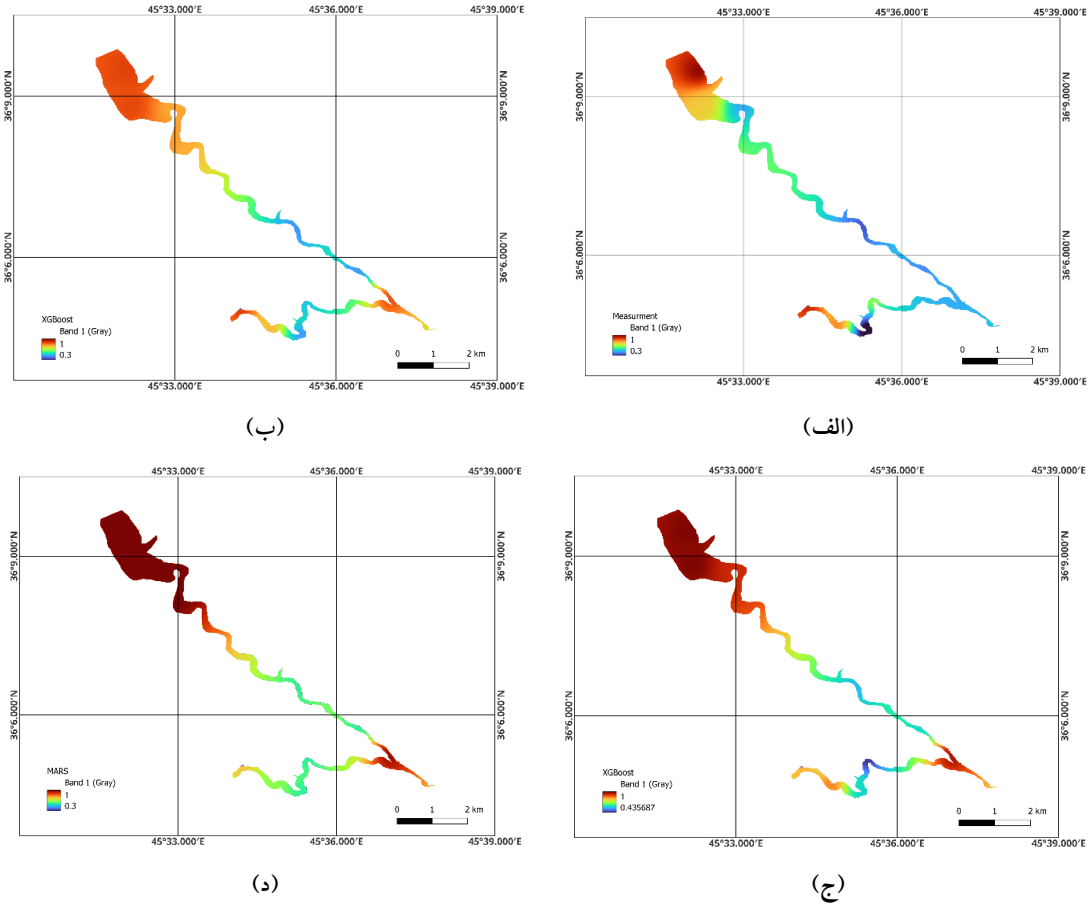


شکل ۱۲- دیاگرام تیلور برای هر سه روش داده‌کاوی XGBoost، M5 و MARS به ازای الف) مرحله آموزش، ب) مرحله صحت‌سنجی و ج) کل داده‌ها.

Figure 12. Taylor diagram for all three data driven methods: XGBoost, M5 and MARS for a) training stage, b) validation stage and c) total data.

نتایج درون‌یابی مقادیر کلروفیل آ برای مقادیر مشاهداتی و نیز نتایج مدل‌های مختلف به ازای کل محدوده مخزن سد سردشت در شکل ۱۳ ارائه شده است. در این مطالعه از روش درونیابی معکوس فاصله[1] در محیط QGIS استفاده شده است. نتایج ارائه‌شده در شکل ۱۳ مربوط به تاریخ ۲۷ آبان ۱۳۹۷ می‌باشد. مقایسه نتایج درون‌یابی مدل‌های مختلف با مقادیر مشاهداتی، نشان‌دهنده این است که در اکثر بخش‌های محدوده مخزن سد، توزیع مقادیر کلروفیل آ در مدل‌های ارائه شده شبیه به مقادیر مشاهداتی (۱۳ (الف)) بوده، هرچند در بخش‌های مانند گوشه سمت راست مخزن، نتایج مدل‌سازی با مقادیر مشاهداتی هم‌خوانی ندارد. عوامل متعددی در این زمینه می‌تواند نقش داشته باشد که مهم‌ترین آن، محدود بودن تعداد نمونه‌های مشاهداتی جهت آموزش مدل‌ها بوده است. علاوه بر این تصاویر ماهواره‌ای سنتینل-۲ در محدوده مورد مطالعه در بعضی از موارد با تأخیر یک تا دو روز مورد استفاده قرار گرفته‌اند.

1- Inverse distance weighting (IDW)



شکل ۱۳- نتایج درون‌یابی مقادیر کلروفیل آ به ازای کل محدوده مخزن سد سردشت، الف) مقادیر مشاهداتی، ب) الگوریتم XGBoost، ج) الگوریتم M5، د) الگوریتم MARS.

Figure 13. Results of interpolation of chlorophyll a values for the entire Sardasht dam reservoir area, a) observational values, b) XGBoost algorithm, c) M5 algorithm, d) MARS algorithm.

نتیجه‌گیری کلی

در این پژوهش سعی گردید با استفاده از ترکیب تکنیک‌های سنجش‌ازدور و مدل‌های داده‌کاوی اقدام به برآورد میزان کلروفیل آ که یکی از پارامترهای مهم کیفی به‌حساب می‌آید، گردد. بر همین اساس از اطلاعات اندازه‌گیری شده مربوط به مخزن سد سردشت در کنار اطلاعات باندهای مختلف تصاویر سنجنده سنتینل-۲ استفاده شد. از میان مدل‌های داده‌کاوی موجود از سه مدل XGBoost، M5 و MARS جهت برآورد میزان کلروفیل آ با استفاده از اطلاعات باندهای مختلف استفاده شد. نتایج به‌دست آمده نشان می‌دهد که استفاده از اطلاعات باندی به همراه مدل‌های داده‌کاوی می‌تواند جهت برآورد کلروفیل آ در مخازن سدها استفاده شود، هرچند جهت آموزش مدل‌ها نیاز به داده‌های اندازه‌گیری متعددی است. از میان مدل‌های مورداستفاده، دو مدل XGBoost و M5 نتایج دقیق‌تری را نسبت به مدل MARS ارائه نمودند. مقدار ضریب ناش-ساتکلیف برای سه مدل XGBoost، M5 و MARS به ترتیب برابر با ۰/۵۴، ۰/۴۷ و ۰/۲۷ محاسبه شد که نشان می‌دهد نتایج دو مدل XGBoost و M5 دارای وضعیت مطلوبی می‌باشند. استفاده از دیاگرام تیلور نیز نشان‌دهنده نزدیک بودن کارایی دو مدل XGBoost و M5 در محاسبه میزان کلروفیل آ است. توزیع

کلروفیل آ در محدوده مخزن سد سردشت توسط مدل‌های مورداستفاده نشان می‌دهد که در نواحی محدودی از سد، مقادیر ارائه‌شده با مقادیر اندازه‌گیری شده همخوانی ندارد که محدود بودن داده‌های اندازه‌گیری مورداستفاده و عدم انطباق کامل زمانی تصاویر سنتینل-۲ با داده‌های اندازه‌گیری در مخزن سد می‌تواند تأثیر مهمی در این زمینه داشته باشد. استفاده از تعداد داده‌های متعدد در مخازن سدهای مختلف، به‌کارگیری مدل‌های داده‌کاوی متنوع و استفاده از تصاویر سایر سنجنده‌ها می‌تواند ابزار مناسبی را در اختیار مدیران مخازن قرار داده تا بتوانند با دقت بیش‌تری اقدام به ارزیابی کیفی آب مخازن نمایند.

داده‌ها و اطلاعات

در این پژوهش از اطلاعات میدانی ارائه‌شده توسط نجف‌زاده قاچکانلو (۱۵) استفاده‌شده است.

تعارض منافع

در این مقاله تعارض منافع وجود ندارد.

مشارکت نویسندگان

نویسنده اول: توسعه مدل‌ها، انجام تحلیل‌ها و نوشتن مقاله. نویسنده دوم: مشارکت در تحلیل‌ها. نویسنده سوم: اصلاح و ویرایش نهایی مقاله.

اصول اخلاقی

نویسندگان اصول اخلاقی را در انجام و انتشار این اثر علمی رعایت نموده‌اند و این موضوع مورد تأیید همه آن‌ها می‌باشد.

حمایت مالی

این پژوهش با حمایت مالی دانشگاه علوم کشاورزی و منابع طبیعی خوزستان در قالب پروژه تحقیقاتی به شماره ۱۴۰۳/۱۱ صورت پذیرفته است.

منابع

1. USEPA. (2022). National Lakes Assessment 2022. Field Operations Manual. Version 1.2. EPA 841-B-16-011. U.S. Environmental Protection Agency, Washington, DC. 56-58.
2. Stumpf, R. P., Wynne, T. T., Baker, D. B., & Fahnenstiel, G. L. (2012). Interannual variability of cyanobacterial blooms in Lake Erie. *PloS One*, 7, 1-11.
3. Linkov, I., Satterstrom, F. K., Loney, D., & Steevans, J. A. (2009). The impact of harmful algal blooms on USACE operations. ANSRP technical notes collection. ERDC/TN ansrp-09-1. Vicksburg, MS: U.S. Army Engineer Research and Development Center, 16 p.
4. Graham, J. L. (2006). Harmful algal blooms. USGS Fact Sheet, 2006-3147, 2 p.
5. Beck, R., Zhan, S., Liu, H., Tong, S., Yang, B., Xu, M., ... Su, H. (2016). Comparison of satellite reflectance algorithms for estimating chlorophyll-a in a temperate reservoir using coincident hyperspectral aircraft imagery and dense coincident surface observations. *Remote Sensing of Environment*, 178, 15-30.
6. Werdell, P. J., & Bailey, S. W. (2005). An improved in-situ bio-optical data set for ocean color algorithm development and satellite data product validation. *Remote sensing of environment*, 98 (1), 122-140.
7. Attila, J., Koponen, S., Kallio, K., Lindfors, A., Kaitala, S., & Ylöstalo, P. (2013). MERIS Case II water processor comparison on coastal sites of the northern Baltic Sea. *Remote Sensing of Environment*, 128, 138-149.
8. Lei, S., Wu, D., Li, Y., Wang, Q., Huang, C., Liu, G., ... & Lv, H. (2019). Remote sensing monitoring of the suspended particle size in Hongze Lake based on GF-1 data. *International journal of remote sensing*, 40 (8), 3179-3203.

9.Shi, K., Zhang, Y., Li, Y., Li, L., Lv, H., & Liu, X. (2015). Remote estimation of cyanobacteria-dominance in inland waters. *Water research*, 68, 217-226.
10.Taheri, A., Serajian, M. R., Ghashghaie, M., & Weysi, K. (2018). Estimation of Chlorophyll-a Concentration Using Remote Sensing Images. *Iranian Journal of Soil and Water Research*, 49 (1), 39-50. [In Persian]
11.Mobarak Hassan, E. (2021). Impact of atmospheric factors with emphasis on dust concentration on chlorophyll in the southeast of the Caspian Sea (2007-2007). *Journal of Oceanography*, 12 (46), 74-85. [In Persian]
12.Matsushita, B., Yang, W., Yu, G., Oyama, Y., Yoshimura, K., & Fukushima, T. (2015). A hybrid algorithm for estimating the chlorophyll-a concentration across different trophic states in Asian inland waters. *ISPRS journal of photogrammetry and remote sensing*, 102, 28-37.
13.Ryu, J., Son, S., Jo, C.O., Kim, H., Kim, Y., Lee, S. H., & Joo, H. (2023). Revised chlorophyll-a algorithms for satellite ocean color sensors in the East/Japan Sea. *Regional Studies in Marine Science*, 60, 102876.
14.Li, H., Li, X., Song, D., Nie, J., & Liang, S. (2024). Prediction on daily spatial distribution of chlorophyll-a in coastal seas using a synthetic method of remote sensing, machine learning and numerical modeling. *Science of the Total Environment*, 910, 168642.
15.Najafzadeh Ghachkanloo, A. (2019). Estimation of turbidity and chlorophyll-a in lakes using remote sensing, Case study: Sardasht reservoir. Master's thesis, Kharazmi University, 109 p. [In Persian]
16.Hedayati Goudarzi, F. (2021). The effect of climate change on water quality in Sardasht dam using CE-Qual-W2 model. Master's thesis, Kharazmi University, 171 p. [In Persian]
17.Rangzan, K., Kabolizade, M., Rahshidian, M., & Delfan, H. (2019). Modeling and zoning water quality parameters using Sentinel-2 satellite images and computational intelligence (Case study: Karun River). *Journal of RS and GIS for Natural Resources*, 10 (4), 21-37. [In Persian]
18.Glasmann, F., Senf, C., Seidl, R., & Annighöfer, P. (2023). Mapping subcanopy light regimes in temperate mountain forests from Airborne Laser Scanning, Sentinel-1 and Sentinel-2. *Science of Remote Sensing*, 8, 100107.
19.O'Reilly, J. E., & Werdell, P. J. (2019). Chlorophyll algorithms for ocean color sensors-OC4, OC5 & OC6. *Remote sensing of environment*, 229, 32-47.
20.Hu, C., Feng, L., Lee, Z., Franz, B. A., Bailey, S. W., Werdell, P. J., & Proctor, C. W. (2019). Improving satellite global chlorophyll a data products through algorithm refinement and data recovery. *Journal of Geophysical Research: Oceans*, 124 (3), 1524-1543.
21.Xing, Q., & Hu, C. (2016). Mapping macroalgal blooms in the Yellow Sea and East China Sea using HJ-1 and Landsat data: Application of a virtual baseline reflectance height technique. *Remote sensing of Environment*, 178, 113-126.
22.Watanabe, F., Alcantara, E., Rodrigues, T., Rotta, L., Bernardo, N., & Imai, N. (2017). Remote sensing of the chlorophyll-a based on OLI/Landsat-8 and MSI/Sentinel-2A (Barra Bonita reservoir, Brazil). *Anais da Academia Brasileira de Ciências*, 90, 1987-2000.
23.Duan, H., Zhang, Y., Zhang, B., Song, K., & Wang, Z. (2007). Assessment of chlorophyll-a concentration and trophic state for Lake Chagan using Landsat TM and field spectral data. *Environmental monitoring and assessment*, 129, 295-308.
24.Tan, W., Liu, P., Liu, Y., Yang, S., & Feng, S. (2017). A 30-year assessment of phytoplankton blooms in Erhai Lake using Landsat imagery: 1987 to 2016. *Remote Sensing*, 9 (12), 1265.
25.Nguyen, H. Q., Ha, N. T., & Pham, T. L. (2020). Inland harmful cyanobacterial bloom prediction in the eutrophic Tri An Reservoir using satellite band ratio and machine learning approaches. *Environmental Science and Pollution Research*, 27, 9135-9151.

26. Bocharov, A. V., Tikhomirov, O. A., Khizhnyak, S. D., & Pakhomov, P. M. (2017). Monitoring of chlorophyll in water reservoirs using satellite data. *Journal of Applied Spectroscopy*, 84, 291-295.
27. Matthews, M. W., & Odermatt, D. (2015). Improved algorithm for routine monitoring of cyanobacteria and eutrophication in inland and near-coastal waters. *Remote Sensing of Environment*, 156, 374-382.
28. Eibe, F., Hall, M. A., & Witten, I. H. (2016). The WEKA workbench. Online appendix for data mining: practical machine learning tools and techniques. In Morgan Kaufmann. San Francisco, California: Morgan Kaufmann Publishers, 363-368.
29. Chen, T., & Guestrin, C. (2016). Xgboost: a scalable tree boosting system Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016: 785-794. ACM, New York, NY.
30. Fan, J., Wang, X., Wu, L., Zhou, H., Zhang, F., Yu, X., ... & Xiang, Y. (2018). Comparison of Support Vector Machine and Extreme Gradient Boosting for predicting daily global solar radiation using temperature and precipitation in humid subtropical climates: A case study in China. *Energy conversion and management*, 164, 102-111.
31. Yao, X., Fu, X., & Zong, C. (2022). Short-term load forecasting method based on feature preference strategy and LightGBM-XGboost. *IEEE Access*, 10, 75257-75268.
32. Quinlan, J. R. (1992). Learning with continuous classes. Singapore, 343-348.: Proceedings Australian Joint Conference on Artificial Intelligence, World Scientific.
33. Wang, Y., & Witten, I. H. (1997). Inducing model trees for continuous classes. In Proceedings of the ninth European conference on machine learning, 9 (1), 128-137.
34. Zahiri, J. (2015). Nonparametric CART and M5' Methods Application on Bridge Piers Scour Depth Computation. *Irrigation and Water Engineering*, 5 (4), 35-50.
35. Jung, N. C., Popescu, I., Kelderman, P., Solomatine, D. P., & Price, R. K. (2010) Application of model trees and other machine learning techniques for algal growth prediction in Yongdam reservoir, Republic of Korea. *Journal of Hydroinformatics*. 12 (3), 262-274.
36. Friedman, J. H. (1991). Multivariate adaptive regression splines. *The annals of statistics*, 19 (1), 1-67.
37. Boehmke, B., & Greenwell, B. M. (2019). Hands-on machine learning with R. Chapman and Hall/CRC, 1-392.
38. Zahiri, J., Mollaee, Z., & Ansari, M. R. (2020). Estimation of suspended sediment concentration by M5 model tree based on hydrological and moderate resolution imaging spectroradiometer (MODIS) data. Water Resources Management, 34 (12), 3725-3737.
39. Alizamir, M., Heddam, S., Kim, S., Gorgij, A. D., Li, P., Ahmed, K. O., & Singh, V. P. (2021). Prediction of daily chlorophyll-a concentration in rivers by water quality parameters using an efficient data-driven model: online sequential extreme learning machine. *Acta Geophysica*, 69, 2339-2361.
40. Cui, Z., Du, D., Zhang, X., & Yang, Q. (2022). Modeling and Prediction of Environmental Factors and Chlorophyll a Abundance by Machine Learning Based on Tara Oceans Data. *Journal of Marine Science and Engineering*, 10 (11), 1749.