دوفصلنامهٔ هیدروفیزیک

دورهٔ هشتم، شمارهٔ دوم (پاییز و زمستان ۱۴۰۱)؛ صفحات ۱۴۳-۱۳۱

مقالهٔ پژوهشی

DOR: 20.1001.1.24767131.1401.8.2.3.3

درصد همانندی: ۱٪

مقایسه مدل‌های رگرسیون خطی وغیرخطی مبتنی بر یادگیری ماشین برای برآورد میزان کلروفیل آ در سواحل قشم و هرمز

میترا نعیمی[۱]، زهرا عزیزی[۲*]، محمد صدیق مرتضوی[۳]، سیده لیلی محبی نوذر[۴]، مجتبی عظام[۵]

[۱] دانشجوی دکتری رشته سنجش از دور وGIS، دانشکده منابع طبیعی و محیط زیست، واحد علوم و تحقیقات، دانشگاه آزاد اسلامی، تهران، ایران naimi.mitra@gmail.com

[۲*] نویسنده مسئول، دانشیار، گروه سنجش از دور وGIS، دانشکده منابع طبیعی و محیط زیست، واحد علوم و تحقیقات، دانشگاه آزاد اسلامی، تهران، ایران zazizi@srbiau.ac.ir

[۳] پژوهشکده اکولوژیکی خلیج‌فارس و دریای عمان، پژوهشکده علوم شیلات ایران، سازمان تحقیقات،آموزش و ترویج کشاورزی، بندرعباس، ایران mseddiq1@yahoo.com

[۴] پژوهشکده اکولوژی خلیج‌فارس و دریای عمان، پژوهشکده علوم شیلات ایران، سازمان تحقیقات،آموزش و ترویج کشاورزی، بندرعباس، ایران lmohebbi@yahoo.com

[۵] استادیار،گروه علوم دریایی، دانشکده منابع طبیعی و محیط زیست، دانشگاه آزاد اسلامی، واحد علوم و تحقیقات تهران، ایرانezam@srbiau.ir

تاریخ دریافت: ۱۴۰۲/۰۳/۲۰ **تاریخ پذیرش:**۱۴۰۲/۰۴/۳۰

چکیده

کلروفیل آ، به عنوان یک شاخص مهم برای اندازه‌گیری شکوفایی جلبکی و کیفیت آب، در مطالعات دریایی بسیار اهمیت دارد. این پژوهش با هدف مقایسه مدل‌های رگرسیون خطی و غیرخطی بر اساس الگوریتم‌های یادگیری ماشین برای بررسی میزان کلروفیل آ در آب‌های ساحلی بندرعباس، جزیره قشم و هرمز انجام شد. برای این منظور از داده‌های ماهواره ترا سنجنده مودیس و برداشت‌های میدانی از نقاط مختلف محدوده مطالعه استفاده شده است. مدل‌های مورد بررسی شامل رگرسیون خطی، مدل خطی تعمیم یافته با توزیع پواسون، جنگل تصادفی و ماشین بردار پشتیبان است. عملکرد این مدل‌ها با استفاده از معیارهای ریشه میانگین مربعات خطا (RMSE)، میانگین درصد خطا (MPE)، میانگین خطای مطلق (MAE) و ضریب تعیین (rsq)ارزیابی شد. نتایج نشان می‌دهد که رگرسیون خطی و مدل خطی تعمیم یافته ضعیف عمل می‌کنند، در حالی که جنگل تصادفی و ماشین بردار پشتیبان عملکرد بهتری را نشان می‌دهند. به طور خاص، جنگل تصادفی بالاترین عملکرد را با RMSE= ۰/۵۷۲۵ و rsq =۰/۶۶۳۷ نشان می‌دهد. این مدل قابلیت تشخیص الگوهای غیرخطی و پیچیده‌تر را دارد و با استفاده از تعداد زیادی درخت تصمیم‌گیری می‌تواند به صورت مؤثر با داده‌های حجیم کار کند. به طور کلی، این پژوهش اثربخشی مدل‌های یادگیری ماشین، به ویژه جنگل‌های تصادفی را در پیش‌بینی دقیق میزان کلروفیل آ به عنوان یک عامل مهم در مدیریت اکوسیستم‌های دریایی در منطقه مورد مطالعه برجسته می‌کند.

واژه‌های کلیدی: کلروفیل آ، یادگیری ماشین، مدل رگرسیون،جنگل تصادفی، ماشین بردار پشتیبان ، سواحل قشم وهرمز

۱. مقدمه

در مناطق ساحلی، به دلیل تاثیرتغییرات اقلیمی و فعالیت‌های انسانی شدید مانند بارندگی، دفع فاضلاب و صید بی رویه، آب‌های آلوده که از طریق روان آب‌های سطحی وارد آب‌های ساحلی می‌شوند و به آسیب رساندن به کیفیت آب ساحلی که پیش از این هم در حال تباهی بود، منجر می‌شود [۱] .کلروفیل آ رنگدانه اصلی در فیتوپلانکتون برای فتوسنتز است و به عنوان نماینده‌ای برای زیست توده در آب در نظر گرفته می‌شود [۲و۳] . بنابراین کلروفیل آ به عنوان یک شاخص کلیدی برای ارزیابی کیفیت آب و پدیده شکوفایی جلبکی استفاده می‌شود. و نظارت بر میزان آن در مدیریت آب‌های ساحلی بسیار مهم است.

در سال‌های اخیر با استفاده از فن آوری ماهواره‌ای و علم سنجش‌ازدور، دریافت و ارزیابی پارامترهای مختلفی از جمله دمای سطحی آب و میزان کلروفیل و ... گسترش روزافزون یافته است. سنجنده CZCS اولین سنسور رنگ اقیانوسی بود که در سال ۱۹۷۸ توسط NASA به همراه ماهواره نیمبوس ۷ به فضا پرتاب شد. این سنسور برای بررسی غلظت کلروفیل آ در آبهای اقیانوسی به کار رفت. CZCS تا سال ۱۹۸۶ در حال عمل بوده و اطلاعات ارزشمندی درباره رنگ اقیانوس و غلظت کلروفیل آ ارائه داده است [۶]. در دهه‌های گذشته، الگوریتم‌های بازیابی کلروفیل آ بر اساس داده‌های مختلف از سنسورهای از دور مختلف توسعه یافته‌اند. الگوریتم‌های اندازه‌گیری از راه دور فلورسانس، از زمان پرتاب ماهواره‌های نسل سوم (MODIS، MERIS، و غیره)، به عنوان روشی متداول برای بازیابی غلظت کلروفیل آ استفاده می‌شوند [۷]. در مطالعه‌ی انجام شده توسط موری[۱] و همکاران در سال ۲۰۰۷، از داده‌های طیف‌سنجی تصویربرداری با وضوح متوسط (مودیس) برای ارتباط غلظت کلروفیل آ با پارامترهای موثر در شمال آدریاتیک استفاده شد [۸]. در این مطالعه، خلیج تریست مورد بررسی قرار گرفت، اما از داده‌های درجا استفاده نشد. در مطالعه‌ی دیگری که توسط موزتی[۲] و همکاران در سال ۲۰۱۰ انجام شد، از داده‌های حسگر میدان دید گسترده (SeaWiFS) به‌طور میانگین ماهیانه استفاده شد و به‌طور میانگین در چندین منطقه بزرگ‌تر نیز بررسی صورت گرفت [۹]. در این مطالعه، همچنین مقایسه‌ای با داده‌های درجا انجام شد. بارازا مورگا[۳] همکارانش در سال ۲۰۲۲ مطالعه‌ای را بر روی تخمین غلظت کلروفیل آ در دریاچه لندال هو[۴] با استفاده از تصاویر ماهواره‌ای سنتینل[۵] انجام دادند. نویسندگان به منظور نظارت بر کیفیت آب‌های داخلی داده‌های ماهواره‌ای را برای تخمین غلظت کلروفیل آ بررسی کردند. آن‌ها با تجزیه و تحلیل داده‌های جمع‌آوری‌شده در طول فصول مختلف و محدود کردن محدوده کلروفیل آ به اندازه‌گیری‌های میدانی، همبستگی قوی بین باندهای طیفی و غلظت کلروفیل آ، با R2 بیشتر از ۰/۸۷ و خطاهای کم نشان دادند. این مطالعه پتانسیل سنجش از دور ماهواره‌ای را برای نظارت بر کیفیت آب در محیط‌های آبی داخلی برجسته می‌کند [۱۶].

محصولات مشتق شده از سنجنده‌های MODIS و SeaWiFS که عمدتاً برای آب‌های ساحلی و آزاد استفاده می‌شوند، می‌توانند به‌عنوان یکی از منابع مهم داده در کنار داده‌های میدانی در تشخیص و پیش‌بینی این پدیده استفاده شوند.

از طرفی پیشرفت‌های اخیر در زمینه هوش مصنوعی، یادگیری ماشین که از ابزارهای اصلی توسعهٔ هوشمندی در ماشین‌ها محسوب می‌شوند، می‌تواند با استفاده از مدل‌های ریاضی و با بررسی روابط پیچیده بین پدیده‌ها امکان پیش‌بینی و پایش میزان کلروفیل در آب‌های ساحلی وآزاد را فراهم کند.

در مطالعه آب‌های ساحلی هنگ کنگ، رگرسیون بردار پشتیبان (SVM)، جنگل تصادفی[۶](RF)، رگرسیون مکعبی و شبکه‌های عصبی مصنوعی (ANN) برای تخمین پارامترهای کیفیت آب استفاده شد [۵].

الگوریتم شبکه عصبی دیگری به نام شبکه چگالی مخلوط (MDN) نیز برای بازیابی غلظت کلروفیل آ با استفاده از ابزار چندطیفی و تصاویر ابزار رنگ اقیانوس و زمین (OLCI) در انواع مختلف آب استفاده شد. نتایج نشان داد که روش

MDN عملکرد بهتری نسبت به مدل‌های تجربی نشان می‌دهد [۱۰].

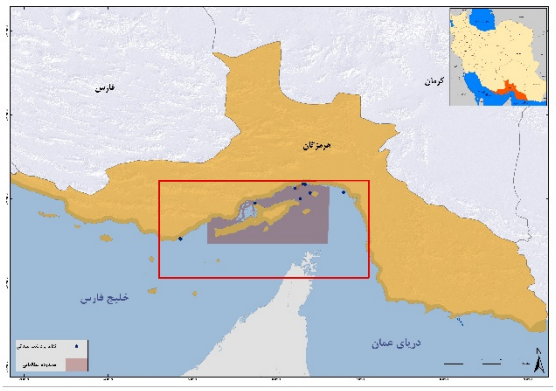
سو[۷] و همکارانش در سال ۲۰۲۱ یک مدل یادگیری ماشین جدید به نام لایت جی بی ام[۸] را برای پیش‌بینی غلظت کلروفیل آ (chl-a) در آب‌های ساحلی فوجیان با استفاده از داده‌های OLCI و داده‌های موقعیتی معرفی می‌کند. این مدل با استفاده از شاخص‌های طیفی مبتنی بر باندهای OLCI و داده‌های معرفی شده، دقت پیش‌بینی را بهبود می‌بخشد. نتایج نشان می‌دهد که با افزودن شاخص‌های طیفی، این مدل عملکرد بهتری نسبت به روش‌های سنتی و محصولات OLCI دیگر دارد [۱۸].

در این پژوهش، مسئله‌ی اصلی بررسی میزان کلروفیل آ در آب‌های ساحلی بندرعباس، جزیره هرمز و جزیره قشم است و برای این منظور از داده‌های ماهواره‌ای از سنجنده مودیس[۹] و ماهواره ترا[۱۰] و برداشت‌های زمینی از نقاط مختلف محدوده مطالعه استفاده شده است. تکنولوژی هوش مصنوعی به دنبال بهبود دقت و صحت پیش‌بینی‌ها و نظارت بر میزان کلروفیل آ در آب‌های ساحلی به عنوان ابزاری کارآمد در مدیریت آب در مناطق ساحلی مورد استفاده قرار گرفت.

هدف اصلی در این پژوهش ارزیابی دقت و عملکرد مدل‌های رگرسیونی مختلف و انتخاب بهترین مدل برای پیش‌بینی میزان کلروفیل آ در این منطقه می‌باشد که می‌تواند به بهبود دقت پیش‌بینی‌های مربوط به کلروفیل آ در مناطق ساحلی کمک می‌کند

۲. مواد و روش‌ها

محدوده مورد مطالعه شامل آب‌های ساحلی بندرعباس، جزیره هرمز و جزیره قشم با مساحتی حدود ۶۶۰۰ کیلومترمربع و در محدوده طول جغرافیایی ۵۵ درجه و ۱۰ دقیقه تا ۵۶ درجه و ۳۵ دقیقه شرقی و عرض جغرافیایی ۲۶ درجه و ۳۰ دقیقه تا ۲۷ درجه و ۱۰ دقیقه شمالی واقع شده است. این منطقه دارای اهمیت بالایی از نظر بررسی کلروفیل آ است، زیرا در این منطقه در طی سال‌های گذشته پدیده شکوفایی جلبکی دیده شده و همچنین اکوسیستم‌های دریایی بسیار متنوعی وجود دارد که باعث می‌شود که اندازه‌گیری کلروفیل آ در این منطقه از اهمیت ویژه برخوردار باشد.



نقشه ۱. موقعیت محدوده مورد مطالعه

داده‌های اندازه‌گیری شده کلروفیل آ که توسط پژوهشکده اکولوژی خلیج فارس و دریای عمان موسسه علوم تحقیقاتی کشور در ۲۵ روز مجزا با تکرار برداشت شد. نمونه‌برداری‌ها در بازه زمانی آبان ماه ۱۳۸۸ تا بهمن ماه ۱۳۸۸ در ۲۸ ایستگاه دریایی در محدوده مطالعاتی انجام شده است.

داده‌های میدانی این پژوهش با استفاده از ابزارهای نمونه‌برداری میدانی با استفاده دستگاه CTD مدل هیدروبیوز در عمق کمتر از ۵ متر برداشت شد. پس از برداشت، صحت سنجی شده و داده‌های غیر مرتبط حذف یا تصحیح شده و سپس داده های CTD برای تجزیه و تحلیل آماری با دقت مناسبی آماده شد [۱۷].

در این پژوهش مقادیر کلروفیل آ اندازه‌گیری شده توسط سنسور مودیس بر روی ماهواره ترا در بازه زمانی مطابق داده‌های برداشت میدانی استفاده شده است. این داده‌های ماهواره‌ای نیز در محدوده مطالعه قرار دارند. استفاده از داده‌های ماهواره‌ای به ما امکان می‌دهد تا به طور جامع‌تری از الگوها و تحولات این متغیر زیستی در این منطقه بررسی شود. در نمودار ۱ رنگ بنفش مقادیر داده‌های ماهواره‌ای و رنگ سبز داده‌های میدانی برداشت شده را نشان می‌دهد.

نمودار ۱. تغییرات میزان کلرفیل آ (mg/m3)در محدوده مورد مطالعه

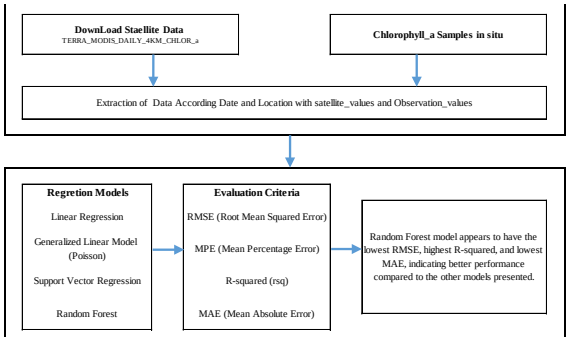
پس از تطابق زمانی و مکانی داده‌های برداشت شده و داده‌های ماهواره‌ای، مجموعه‌ای از داده‌ها (۹۲۸ نمونه برداشت شده) که هر دو مقدار داده ماهواره‌ای و درجا از متغییر کلروفیل آ (میلی‌گرم بر متر مکعب) در یک زمان و مکان در محدوده مطالعاتی وجود داشت، استخراج شد. جدول ۱ خلاصه‌ای از داده‌ها را نشان می‌دهد.

جدول ۱. خلاصه داده‌ها به تفکیک زمان ومکان در محدوده مورد مطالعه

ایستگاه نمونه‌برداری	۸۸/۲/۲۰	۸۸/۸/۲۴	۸۸/۸/۲۵	۸۸/۹/۲۴	۸۸/۱۰/۲۰	۸۸/۱۱/۳	مجموع
۱	۵۲	۶۹	۰	۲۶	۳۵	۰	۱۸۲
۲	۶۸	۵۶	۰	۳۲	۶۷	۰	۲۲۳
۳	۳۹	۶۲	۰	۳۴	۴۲	۰	۱۷۷
۴	۶۷	۳۱	۰	۲۵	۰	۰	۱۲۳
۵	۰	۴۱	۰	۲۷	۰	۰	۶۸
۶	۰	۰	۰	۰	۰	۰	۰
۷	۰	۰	۳۰	۰	۰	۶۰	۹۰
۸	۰	۰	۰	۰	۰	۲۸	۲۸
۹	۰	۰	۰	۰	۰	۳۴	۳۴
۱۰	۰	۰	۰	۰	۰	۳	۳
مجموع	۲۲۶	۲۵۹	۳۰	۱۴۴	۱۴۴	۱۲۵	۹۲۸

۳. تئوری و محاسبات

در این پژو هش با مقایسه مدل‌های رگرسیون خطی[۱۱] و غیر خطی می‌توان بهترین مدل را برای پیش‌بینی شاخص تاثیر بر شکوفایی جلبکی (کلروفیل آ) تعیین کرد. یک نمای کلی از فرایند انجام این پژوهش در شکل۱ نشان داده شده است . در ابتدا داده‌های درجا و به موازات آن داده‌های ماهواره‌ای جمع آوری و آماده سازی شد. مجموعه داده‌ها به دودسته آموزشی[۱۲] و آزمایشی[۱۳] تقسیم شد. برای چهار مدل رگرسیون خطی و غیر خطی در محیط R، بر اساس همبستگی‌های موجود بین غلظت کلروفیل آ اندازه‌گیری شده در محل و مقادیر تصاویر ماهواره‌ای به عنوان متغیرمستقل که در زمان (تاریخ) و مکان (پیکسل نمونه‌برداری) همپوشانی داشتند، برنامه نویسی شد. در هر یک از الگوریتم‌هایی که توسعه داده شده اند، شاخص‌های تطابق محاسبه شد و بر اساس آن‌ها، بهترین مدل عملکرد تعیین شد.



شکل ۱. نمای کلی از فرایند انجام

رگرسیون خطی (lm): رگرسیون خطی یک مدل آماری ساده و پرکاربرد برای پیش‌بینی متغیر وابسته پیوسته بر اساس یک یا چند متغیر مستقل است اما محدودیت‌هایی در مدیریت روابط غیر خطی بین متغیرهای پاسخ و پیش‌بینی در عوارض وجود دارد. در این پژوهش، تابع lm در R برای برازش مدل رگرسیون خطی استفاده شده است.

رگرسیون خطی تعمیم یافته: این مدل یک روش آماری است که برای مدل‌سازی روابط بین متغیرهای وابسته و مستقل استفاده می‌شود. این مدل، توسط نیلز ویدرول در سال ۱۹۷۲ ارائه شد [۱۱]. مدل رگرسیون خطی تعمیم یافته از توزیع‌های احتمال غیرنرمال برای متغیر وابسته استفاده می‌کند، مانند توزیع پواسون[۱۴]، بینومیال، نرمال، و گاما. این روش، به مدل‌سازی متغیرهای پاسخ غیرپیوسته و یا متغیرهای پیوسته با تابعی غیرخطی از متغیرهای وابسته کمک می‌کند.

مدل رگرسیون خطی تعمیم یافته به صورت زیر تعریف می‌شود:

$$g(\mu) = \beta 0 + \beta 1x1 + \beta 2x2 + \cdots + \beta pxpg(\mu) = \beta 0 + \beta 1x1 + \beta 2x2 + \cdots + \beta pxp \quad (۱)$$

در اینجا، g(.) تابع لینک است که با توجه به توزیع احتمال مشخص می‌شود و μ میانگین مشروط[۱۵] متغیر وابسته را نشان می‌دهد [۱۱].

مدل رگرسیون خطی تعمیم یافته یک روش موثر برای تحلیل داده‌ها و استخراج اطلاعات از روابط میان متغیرها است. از آن برای مدل‌سازی در حوزه‌های مختلفی از جمله علوم پزشکی، علوم اجتماعی، علوم زمین، و بسیاری از زمینه‌های دیگر استفاده می‌شود. مدل خطی تعمیم یافته با استفاده از توزیع پواسون رگرسیون خطی را برای مدیریت توزیع‌های غیر نرمال و واریانس غیر ثابت گسترش می‌دهند. در این پژوهش، تابع glm در R پایه برای تطبیق GLM با توزیع پواسون استفاده شده است.

مدل بردار پشتیبان (SVM) : ماشین بردار پشتیبان[۱۶] برای اولین بار در سال ۱۹۹۵ توسط واپنک[۱۷] برای طبقه‌بندی ارائه شد [۱۲]. بعدها این مدل برای مسائل رگرسیون با نام رگرسیون بردار پشتیبان[۱۸] (SVR) تعمیم داده شد. رگرسیون بردار پشتیبان یک روش یادگیری ماشینی است که از ماشین‌های بردار پشتیبان برای انجام وظایف رگرسیونی استفاده می‌کند که پس از برازش مدل‌ها، عملکرد آن‌ها را در مجموعه آزمایشی ارزیابی می‌شود. هدف اصلی SVM، تعیین یک صفحه (در مسئله دسته‌بندی) یا یک سطح رگرسیون (در مسئله رگرسیون) در فضای ویژگی است که بهترین تعمیم‌پذیری را به داده‌های جدید داشته باشد و همچنین مارژین (حاشیه) بین دسته‌ها (در مسئله دسته‌بندی) یا بین داده‌های آموزش و صفحه رگرسیون (در مسئله رگرسیون) را حداکثر کند. در این روش، برای مسائل غیرخطی، از هسته[۱۹] استفاده می‌شود. تابع هسته، این امکان را می‌دهد که داده‌ها به یک فضای ویژگی با ابعاد بالا نگاشت شود، جایی که دسته‌بندی خطی در این فضا ممکن باشد [۱۲] .یکی از توابع هسته معروف، تابع هسته شعاعی[۲۰] است که در این پژوهش استفاده شده است.

مدل جنگل تصادفی (RF) : در سال ۲۰۰۱، لئو بریمن درختان طبقه‌بندی را در یک جنگل تصادفی ترکیب کرد، یعنی استفاده از متغیرها و داده‌ها برای به دست آوردن تعداد معینی از درختان طبقه‌بندی تصادفی شد. سپس نتایج درختان طبقه‌بندی خلاصه شد و الگوریتم جنگل تصادفی پیشنهاد شد [۱۳]. RF نوعی از یادگیری گروهی است. یادگیری گروهی یک استراتژی یادگیری ماشینی بسیار محبوب است و تقریباً همه مشکلات را می‌توان با استفاده از ایده‌های آن بهبود بخشید و هنگام تجزیه و تحلیل داده‌های نمونه کوچک پایدارتر است [۱۴].

یک روش یادگیری ماشینی مبتنی بر درخت تصمیم است که دارای دقت پیش‌بینی بالا، تحمل بالا نسبت به نقاط پرت و اثر برازش خوب است [۱۵]. اصول اساسی جنگل‌های تصادفی به شرح زیر است: [۱۳]

۱. تعیین مجموعه داده نمونه اصلی D و تعداد متغیرهای M
۲. نمونه‌گیری مجدد برای استخراج N واحد نمونه از آن با همان تعداد نمونه در D (Ntree)

۳. تعیین تعداد متغیرها در گره‌های درختی که Mtry در آن نشان دهنده تعداد متغیرهای مورد استفاده در تصمیم‌گیری در گره‌های درخت تصمیم است.
۴. هر درخت تا حداکثر رشد کند، تمام درخت‌های تصمیم را به طور کامل تولید کنید، و تکرارهای متعدد را برای به دست آوردن یک جنگل تصادفی از n درخت تصمیم انجام دهید.
۵. نتیجه نهایی یک جنگل تصادفی، میانگین نتایج هر درخت تصمیم است.
۶. دقت مدل جنگل تصادفی به Ntree و Mtry بستگی دارد، Ntree اندازه کلی جنگل تصادفی را تعیین می‌کند و Mtry رشد تک تک درختان را تعیین می‌کند و هر دو دقت مدل جنگل تصادفی را در سطوح کلان و خرد تعیین می‌کند.

ریشه میانگین مربعات خطا[۲۱]: RMSE میانگین خطای پیش‌بینی را اندازه می‌گیرد و نشان می‌دهد که مقادیر پیش‌بینی شده چقدر به مقادیر واقعی نزدیک هستند . RMSE کمتر نشان دهنده عملکرد بهتر مدل است.

$$MAE = n\sum i = 1|yi - xi|n \quad (۲)$$

$$RMSE = \sqrt{\sum ni = 1(yi - xi)\,2n} \quad (۳)$$

میانگین درصد خطا[۲۲] (MPE): میانگین درصد اختلاف بین مقادیر پیش‌بینی شده و واقعی را محاسبه می‌کند. نشان دهنده میانگین خطای نسبی پیش‌بینی‌های مدل است.

ضریب تعیین[۲۳](rsq): همچنین به عنوان R-squared شناخته می‌شود، نسبت واریانس در متغیر وابسته را که توسط متغیر(های) مستقل توضیح داده شده است، اندازه‌گیری می‌کند. مقدار R-squared بالاتر نشان دهنده تناسب بهتر است.

میانگین خطای مطلق[۲۴] (MAE): میانگین اختلاف مطلق بین مقادیر پیش‌بینی‌شده و واقعی را محاسبه می‌کند و معیاری از دقت پیش‌بینی مدل ارائه می‌کند.

٤. یافته‌ها

در این پژوهش، مدل رگرسیون خطی ساده با استفاده از تابع lm در R بر روی داده‌های آموزش اعمال شده است. در این مدل، متغیر وابسته توسط متغیر مستقل "satellite" توضیح داده می‌شود. در این مدل توزیع نرمال فرض شده است. نتایج حاکی از وجود ارتباط مثبت قابل توجه بین متغیر وابسته و متغیر مستقل است. مقدار p-value برای متغیر مستقل "satellite" بسیار کمتر از سطح معناداری ۰/۰۵ است. به عبارت ساده‌تر، مقدار p کوچک نشان می‌دهد که متغیر مستقل "satellite" تاثیر قابل توجهی بر متغیر وابسته دارد.

علاوه بر این، ضریب تعدیل شده[۲۵] بسیار کم است و تنها حدود ۳/۱٪ از تغییرپذیری متغیر وابسته توسط متغیر مستقل توضیح داده می‌شود.

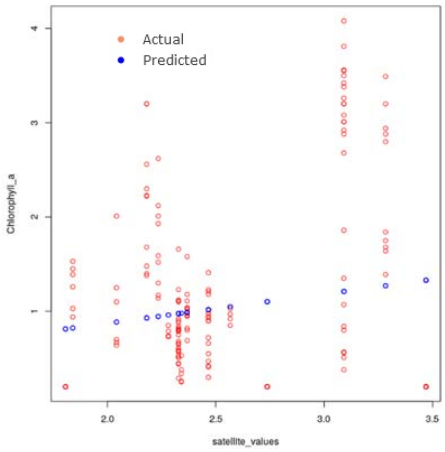
میزان خطای باقی‌مانده[۲۶] نیز حدود ۰/۸۹۷۹ است که نشان دهنده وجود تنوع قابل توجه در داده‌ها و نامطلوب بودن تطابق کامل مدل با داده‌ها است.

F-statistic نیز بدان معناست که حداقل یکی از ضرایب در مدل از صفر متمایز است.

```
Call:
lm(formula = formula, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-1.1289 -0.6133 -0.1531  0.4072  2.7241

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.25299    0.16816   1.505    0.133
satellite_   0.31010    0.06361   4.875 1.33e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8979 on 742 degrees of freedom
Multiple R-squared:  0.03104,   Adjusted R-squared:  0.02973
F-statistic: 23.77 on 1 and 742 DF,  p-value: 1.329e-06
```

شکل ۲. خلاصه مدل رگرسیون خطی در R

برای تحلیل رابطه بین متغیرهای مستقل و وابسته، مدل رگرسیون خطی بر روی داده‌های آزمایشی نیز پیاده شد.با اجرای آن، پیش‌بینی‌هایی برای مقادیر وابسته ارائه شد که نمودار ۲ نشان دهنده مقادیر واقعی و پیش‌بینی شده توسط مدل رگرسیون خطی است. مقادیر واقعی با رنگ قرمز و مقادیر پیش‌بینی شده با رنگ آبی نشان داده شده‌اند.



نمودار ۲. مقادیر واقعی و پیش‌بینی شده برای میزان کلروفیل آ (میلی‌گرم بر مترمکعب) توسط مدل رگرسیون خطی برای داده‌های آزمایشی

glm_model یک مدل رگرسیون خطی تعمیم یافته را با توزیع پواسون بر روی داده‌های آموزش اجرا می‌کند. این مدل براساس متغیرهای وابسته و مستقل داده‌های آموزش، ضرایب را تخمین می‌زند.

خروجی این مدل عبارت است از:

مقادیر انحراف باقیمانده‌ها[۲۷] که نشان‌دهنده تفاوت بین پیش‌بینی‌های مدل و مقادیر واقعی است. این مقادیر شامل حداقل، چارک اول، میانه، چارک سوم و بیشینه است.

ضرایب: این قسمت شامل تخمین‌های ضرایب برای هر یک از متغیرها است. هر ضریب همراه با خطاهای استاندارد[۲۸] خود، آمار t-value و مقدار p-value آن ضریب نیز درج شده است. این مقادیر برای ارزیابی اهمیت آماری هر متغیر در مدل استفاده می‌شوند. مقادیر ستاره (*) نشان‌دهنده میزان اهمیت آماری هر ضریب است.

پارامتر پراکندگی[۲۹] برای خانواده توزیع پواسون در نظر گرفته شده است.

تفاوت بین انحراف نول (Null Deviance) و انحراف باقی‌مانده (Residual Deviance) نشان‌دهنده میزان تطابق مدل با داده‌های آموزش هستند. از طریق مقایسه این دو می‌توان به خوبی مدل را ارزیابی کرد.

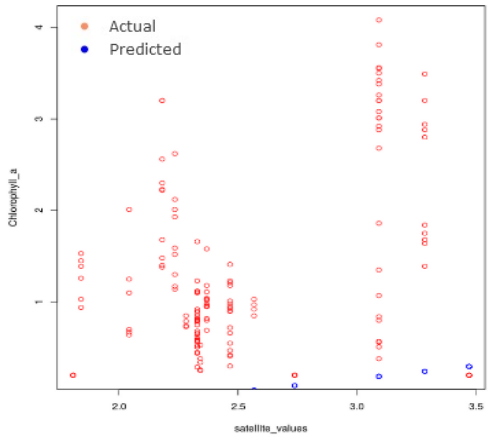
مقدار AIC (Akaike Information Criterion) یک معیار است که براساس آن می‌توان مدل‌ها را مقایسه کرد. هدف کمینه کردن این مقدار است. اگر مقدار AIC برابر Infباشد، به معنی این است که مدل بسیار ناسازگار است.

تعداد تکرارهای فیشر (Fisher Scoring iterations) که نشان‌دهنده تعداد تکرارهای الگوریتم برای یافتن مقادیر بهینه ضرایب است.

```
Call:
glm(formula = formula, family = poisson(), data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3515  -0.8353  -0.1456   0.3857   2.1257

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.69846    0.18364  -3.803 0.000143 ***
satellite_   0.28646    0.06744   4.248 2.16e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 535.69  on 743  degrees of freedom
Residual deviance: 517.87  on 742  degrees of freedom
AIC: Inf

Number of Fisher Scoring iterations: 5
```

شکل۳. خلاصه مدل Generalized Linear Model در محیط R

با استفاده از مدل رگرسیون خطی تعمیم یافته بر روی داده‌های تست، پیش‌بینی‌هایی برای مقادیر وابسته ارائه شد. نمودار ۳ نتایج را نشان می‌دهد که مقادیر واقعی با رنگ قرمز و مقادیر پیش‌بینی شده توسط مدل رگرسیون خطی با رنگ آبی نشان داده شده‌اند.



نمودار ۳. مقادیر واقعی و پیش‌بینی شده برای میزان کلروفیل آ (میلی‌گرم بر مترمکعب) توسط مدل خطی تعمیم یافته برای داده‌های آزمایشی

جنگل تصادفی (RF) : جنگل تصادفی یک روش یادگیری است که چندین درخت تصمیم را برای پیش‌بینی ترکیب

می‌کند.در این پژوهش،با تقسیم داده‌های به دو مجموعه آموزشی و آزمایشی از کتابخانه caret در R برای برازش مدل جنگل تصادفی (RF) استفاده شده است.

مدل جنگل تصادفی (RF) که در این پژوهش به کار گرفته شده، با استفاده از اطلاعات زیر آموزش داده شده و ارزیابی شده است:

۱. داده‌ها: مدل تصادفی جنگلی بر روی مجموعه‌داده‌ای با ۷۴۴ نمونه (مشاهدات) و ۱ متغیر پیش‌بین استفاده می‌شود.

۲. نمونه‌برداری: روش نمونه‌برداری مورد استفاده بوت‌استراپینگ[۳۰] است. در این روش، با جایگذاری تصادفی از مجموعه‌داده، چندین نمونه بوت‌استراپ ساخته می‌شود. در این حالت، بوت‌استراپینگ ۲۵ بار انجام می‌شود.

۳. اندازه نمونه‌ها: خلاصه‌ای از اندازه نمونه‌ها نشان می‌دهد که هر نمونه بوت‌استراپ شامل ۷۴۴ مشاهده است که با اندازه مجموعه‌داده اصلی برابر است.

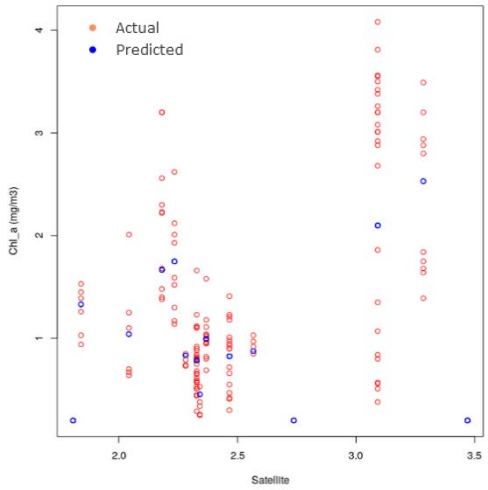
۴. نتایج نمونه‌برداری: عملکرد مدل جنگل تصادفی با استفاده از سه معیار، یعنی خطا میانگین مربعاتی مطلق (RMSE)، ضریب تعیین (Rsq) و میانگین خطا مطلق (MAE) ارزیابی می‌شود. مقدار RMSE برابر با ۰/۵۳۹۳۱۷۷ است که خطای میانگین پیش‌بینی مدل را نشان می‌دهد. مقدار Rsq برابر با ۰/۶۵۵۳۶۶۹ است که نشان‌دهنده نسبت واریانس متغیر وابسته توسط مدل توضیح داده شده است. مقدار بالاتر برای Rsq نشان‌دهنده تطبیق بهتر مدل با داده‌ها است. مقدار MAE برابر با ۰/۳۴۵۱۰۷۵ است که نشان‌دهنده میانگین اختلاف مطلق بین مقادیر پیش‌بینی شده و واقعی است. مقدار کمتر برای MAE بهترین دقت پیش‌بینی را نشان می‌دهد.

۵. پارامتر تنظیم "mtry" که تعیین‌کننده تعداد متغیرهای پیش‌بین در هر تقسیم است، ثابت با مقدار ۲ در نظر گرفته شده است. این بدان معناست که در هر تقسیم، مدل تصادفی جنگلی ۲ متغیر پیش‌بین را به‌صورت تصادفی انتخاب می‌کند.

```
Random Forest

744 samples
  1 predictor

No pre-processing
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 744, 744, 744, 744, 744, 744, ...
Resampling results:

  RMSE       Rsquared   MAE
  0.5393177  0.6553669  0.3451075

Tuning parameter 'mtry' was held constant at a value of 2
```

شکل ۴. خلاصه مدل جنگل تصادفی در محیط R

برای ارزیابی دقت مدل RF ، نمودار مقادیر واقعی و پیش‌بینی شده برای متغیر وابسته رسم شده است. در نمودار ۴، نقاط قرمز نمایانگر مقادیر واقعی است و نقاط آبی نمایانگر مقادیر پیش‌بینی شده توسط مدل هستند.



نمودار ۴. مقادیر واقعی و پیش‌بینی شده برای میزان کلروفیل آ (میلی‌گرم بر مترمکعب) توسط مدل جنگل تصادفی برای داده‌های آزمایشی

مدل دیگری که در این پژوهش استفاده شده است، مدل ماشین بردار پشتیبان (SVM) با استفاده از هسته RBF می‌باشد. عملکرد این مدل و مقادیر پارامترها و نتایج به دست آمده به شرح زیر است:

۱. مجموعه داده: این مدل با استفاده از ۷۴۴ نمونه آموزشی آموزش داده شده است. هر نمونه شامل یک متغیر پیش‌بینی‌کننده است.

۲. پیش‌پردازش: هیچ فرایند پیش‌پردازشی روی داده‌ها قبل از آموزش مدل انجام نشد.

۳. روش نمونه‌برداری: از روش بوت‌استراپینگ با ۲۵ تکرار برای نمونه‌برداری استفاده شده است. در این روش، برای هر تکرار ۷۴۴ نمونه به صورت تصادفی و با جایگزینی از مجموعه داده اصلی ساخته شد.

۴. پارامترهای تنظیم: مدل SVM با استفاده از دو پارامتر C و sigma تنظیم شد: پارامترC تعادل بین کاهش خطا در آموزش و حفظ مارجین مدل را کنترل می‌کند. مقادیر مختلفی از C را بررسی شد، به‌طور خاص ۰/۲۵، ۰/۵ و ۱/۰ و پارامتر sigma عرض هسته RBF را تعیین می‌کند. در این مقاله، این پارامتر ثابت با مقدار ۷/۱۴۸۹۷۱ در نظر گرفته شده است.

۵. سه معیار عملکرد RMSE، Rsq و MAE برای ارزیابی مدل استفاده شد.

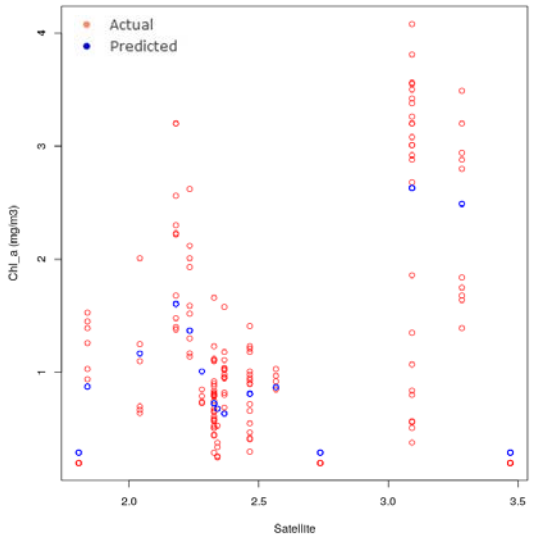
۶. مدل بهینه: مدل بهینه با استفاده از کمترین مقدار RMSE انتخاب شد. مقادیر پارامترهای این مدل برابر با C= ۱/۰ و sigma= ۷/۱۴۸۹۷۱ بود.

به طور خلاصه، ما از مدل SVM با هسته RBF بر روی ۷۴۴ نمونه و یک متغیر پیش‌بینی‌کننده استفاده کردیم. مدل با استفاده از پارامترهای C و sigma تنظیم شد و با استفاده از معیارهایRMSE، Rsq و MAE ارزیابی شد. مدل بهینه با مقادیر C = ۱/۰ و sigma= ۷/۱۴۸۹۷۱ بر اساس کمترین مقدار RMSE انتخاب شد.

```
Support Vector Machines with Radial Basis Function Kernel

744 samples
  1 predictor

No pre-processing
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 744, 744, 744, 744, 744, 744, ...
Resampling results across tuning parameters:

  C     RMSE       Rsquared   MAE
  0.25  0.6056787  0.5758694  0.4027237
  0.50  0.6031758  0.5807204  0.3977620
  1.00  0.5973204  0.5913207  0.3906627

Tuning parameter 'sigma' was held constant at a value of 7.148971
RMSE was used to select the optimal model using the smallest value.
The final values used for the model were sigma = 7.148971 and C = 1.
```

شکل ۵. خلاصه مدل ماشین بردار پشتیبان در محیط R

برای ارزیابی دقت مدل SVM ، نمودار مقادیر واقعی و مقادیر پیش‌بینی شده توسط مدل رسم شده است. در نمودار ۵، مقادیر واقعی با نقاط قرمز و مقادیر پیش‌بینی شده توسط مدل SVM با نقاط آبی نمایش داده شده‌اند.



نمودار ۵ . مقادیر واقعی و پیش‌بینی شده برای میزان کلروفیل آ (میلی‌گرم بر مترمکعب) توسط مدلSVM برای داده‌های آزمایشی

در این پژوهش، چهار مدل رگرسیون مختلف اعمال شدند و عملکرد آن‌ها با استفاده از معیارهای مختلفی ارزیابی شد. نتایج ارزیابی هر مدل شامل مقادیر مختلفی از جمله خطای میانگین مربعات ریشه (RMSE)، درصد خطای میانگین مطلق (MPE)، ضریب توضیح مدل (rsq)، خطای میانگین مطلق (MAE) است.

جدول ۵ مقادیر کمی معیارهای ارزیابی مدل‌ها را نشان می‌دهد.

جدول ۵. معیارهای ارزیابی مدل‌ها

	مدل	خطای میانگین مربعات ریشه (RMSE)	درصد خطای میانگین مطلق (MPE)	ضریب توضیح مدل (rsq)	خطای میانگین مطلق (MAE)
۱	مدل رگرسیون خطی	۰/۹۸۲۲	۱۳۴/۶۵۶۱	۰/۰۰۱۴۸	۰/۷۵۶۵۹
۲	مدل رگرسیون خطی تعمیم‌یافته	۱/۴۲۷۶	-۷۷/۶۰۳	۰.۰۰۱۴۸	۰/۷۵۶۵۹
۳	جنگل تصادفی	۰/۵۷۲۴	۱۳.۴۶۴۱۸	۰/۶۶۳۱۶	۰/۳۴۹۲۶
۴	ماشین بردار پشتیبانی	۰/۵۹۱۱	۲۶/۳۳۵۵۵	۰/۶۳۳۹۶	۰/۳۸۹۰۴

برای ارزیابی عملکرد هر مدل، ضریب همبستگی و نسبت انحراف استاندارد محاسبه شده و در جدول ۶ نمایش داده شده است. نتایج نشان داد که مدل‌هایRF و SVM با ضریب همبستگی بالا و نسبت انحراف استاندارد کمتر، عملکرد بهتری نسبت به مدل‌های LM و GLM دارند و بنابراین مدل‌های RF و SVM قادر به پیش‌بینی دقیق‌تری از مقادیر کلروفیل آ هستند و پیش‌بینی‌های آن‌ها با داده‌های واقعی تطابق بالاتری دارند. از اینرو، استفاده از این مدل‌ها به عنوان روش‌های پیش‌بینی بهتر در این حوزه توصیه می‌شود.

جدول ۶. معیارهای ارزیابی عملکرد مدل‌ها

	مدل	ضریب همبستگی[۳۱]	نسبت انحراف استاندارد[۳۲]
۱	رگرسیون خطی	۰/۰۳۹	۰/۱۶۶
۲	مدل رگرسیون خطی تعمیم‌یافته	۰/۰۳۹	۰/۱۵۴
۳	جنگل تصادفی	۰/۸۱۴	۰/۷۴۵
۴	ماشین بردار پشتیبانی	۰/۷۹۶	۰/۸۳۳

۵. بحث

داده‌های ماهواره‌ای و برداشت شده کلروفیل-آ (mg/m^3) برای تاریخ‌های مختلف در سال ۱۳۸۸ مورد بررسی قرار گرفته است. بر اساس داده‌های مشاهده شده، غلظت کلروفیل-آ در این زمان‌ها دارای تغییراتی بوده است. برای مثال، در تاریخ ۲۰ آبان ماه ۱۳۸۸، غلظت کلروفیل ـ آ بین ۱/۸۰۷ تا ۳/۴۷ mg/m^3 متغیر بوده و میانگین آن حدوداً ۲/۸۵ mg/ m^3 بوده است. همچنین، در تاریخ ۲۴ آذر ماه ۱۳۸۸ غلظت کلروفیل-آ بین ۰/۵۴ تا ۳ /۹۱ mg/m^3 متغیر بوده و میانگین آن حدوداً ۱/۶۴ mg/m^3 بوده است. بر اساس همبستگی‌های محاسبه شده، برخی از داده‌ها همبستگی مثبت با یکدیگر دارند، در حالی که برخی دیگر دارای همبستگی منفی هستند. با بررسی داده‌ها می‌توان الگوها و تغییرات میزان غلظت کلروفیل-آ را در طول زمان شناسایی کرد و درک بهتری از وضعیت محدوده مورد مطالعه و فعالیت‌های زیستی مرتبط با آن به دست آورد. جدول ۳ و ۴ به ترتیب خلاصه آماری داده‌های ماهواره‌ای و برداشت شده (نمونه‌برداری شده) را نشان می‌دهد.

جدول ۳. خلاصه آماری داده‌های ماهواره‌ای در محدوده مورد مطالعه

خلاصه آماری داده‌های برداشت شده کلروفیل آ(mg/ m^3)					
تاریخ	حداقل	میانگین	حداکثر	انحراف معیار	ضریب همبستگی
۱۳۸۸/۲/۲۰	۰/۲	۰/۲	۰/۲	۰	
۱۳۸۸/۸/۲۴	۰/۱۳	۱/۶۵۶	۴/۰۸۵	۱/۰۷۰۹۳۷	۰/۴۱۴۷۷۵۲
۱۳۸۸/۸/۲۵	۰/۶۴	۰/۸۲۸	۱/۱۶۱	۰/۱۱۴۴۸۲	
۱۳۸۸/۹/۲۴	۰/۵۴	۱/۶۴	۳/۹۱۰	۰/۸۵۷۵۱۷	۰/۶۹۳۷۰۷۲
۱۳۸۸/۱۰/۲۰	۰/۲۶	۱/۰۹۵	۳/۶۷	۰/۵۹۱۷۲۸	-۰/۷۰۵۰۷۳۷
۱۳۸۸/۱۱/۳	۰/۰۱	۰/۷۶۴۶	۲/۲۵	۰/۴۰۷۳۷۷	-۰/۲۵۲۹۷۵۴

جدول ۴. خلاصه آماری داده‌های برداشت شده در محدوده مورد مطالعه

خلاصه آماری داده‌های ماهواره‌ای کلروفیل آ(mg/m^3)					
تاریخ	حداقل	میانگین	حداکثر	انحراف معیار	ضریب همبستگی
۱۳۸۸/۲/۲۰	۱/۸۰۷	۲/۸۵	۳/۴۷	۰/۷۲۸۶۱۹۲	
۱۳۸۸/۸/۲۴	۲/۰۴۲	۲/۶۲۹	۳/۰۹۱	۰/۴۵۳۹۴۳۷	۰/۴۱۴۷۷۵۲
۱۳۸۸/۸/۲۵	۲/۲۸۱	۲/۲۸۱	۲/۲۸۱	۰	
۱۳۸۸/۹/۲۴	۱/۸۴	۲/۶۸۲	۳/۲۸۳	۰/۵۴۴۴۱۴۴	۰/۶۹۳۷۰۷۲
۱۳۸۸/۱۰/۲۰	۲/۲۳۴	۲/۳۹۸	۲/۴۶۶	۰/۱۰۵۶۶۳۹	-۰/۷۰۵۰۷۳۷
۱۳۸۸/۱۱/۳	۲/۳۲۸	۲/۳۳	۲/۳۹۲	۰/۰۰۹۷۹۲۲۲۲	-۰/۲۵۲۹۷۵۴

با توجه به نتایج، مدل رگرسیون خطی با RMSE برابر با ۰/۹۸۲۲ و rsq برابر با ۰/۰۰۱۵، دارای بهترین عملکرد نیست. مدل خطی تعمیم یافته با توزیع پواسون نیز با RMSE برابر با ۱/۴۲۷۶ و rsq برابر با ۰/۰۰۱۵عملکرد ضعیفی دارد.

دو مدل دیگر، یعنی جنگل تصادفی و ماشین بردار پشتیبان، با عملکرد بهتر مواجه شده‌اند. جنگل تصادفی با RMSE برابر با ۰/۵۷۲۵ و rsq برابر با ۰/۶۶۳۲ بهترین عملکرد را دارد. همچنین، ماشین بردار پشتیبان نیز با RMSE برابر با ۰/۵۹۱۱و rsq برابر با ۰/۶۳۴به عنوان یک روش دیگر با عملکرد قابل قبول می‌تواند در نظرگرفته شود.

انتخاب بین مدل جنگل تصادفی و SVM بستگی به ماهیت داده‌ها، خصوصیات مسئله، و توجه به عملکرد مطلوب دارد. در این پژوهش، مدل جنگل تصادفی عملکرد بهتری نسبت به ماشین بردار پشتیبان (SVM) نشان داده است و این تفاوت در عملکرد به عوامل زیر وابسته است:

انعطاف‌پذیری و قابلیت اطمینان :جنگل تصادفی یک مدل غیرخطی و غیرپارامتریک است که مبتنی بر ترکیب چندین درخت تصمیم‌گیری است. این مدل توانایی کشف الگوهای غیرخطی و پیچیده‌تر را دارد و می‌تواند با توجه به تنوع و تعدد درخت‌ها، بهتر به داده‌های پیچیده تطبیق پیدا کند. در مقابل، SVM یک مدل خطی است که بر پایه تابع هسته عمل می‌کند. این مدل معمولاً در داده‌های دارای رابطه خطی قوی عملکرد بهتری دارد.

قدرت تعمیم‌پذیری :جنگل تصادفی به دلیل ترکیب چندین درخت تصمیم‌گیری، قابلیت تعمیم‌پذیری بیشتری نسبت به SVMدارد. این به این معنی است که جنگل تصادفی می‌تواند به صورت مؤثری با داده‌های جدید و ناشناخته کار کند و قابلیت تعمیم یافتن به مناطق دیده نشده را داشته باشد. در حالی که SVM به دلیل خطی بودن وابسته به تابع هسته، ممکن است در تعمیم به داده‌های جدید مشکلاتی مواجه شود.

پردازش داده‌های بزرگ :جنگل تصادفی به دلیل استفاده از تعداد زیادی درخت تصمیم‌گیری، قابلیت پردازش داده‌های بزرگتر را دارد. این مدل می‌تواند با جمع‌آوری نتایج از چند درخت و تصمیم‌گیری بر اساس اکثریت آراء، به صورت موازی و مؤثر با داده‌های حجیم کار کند. در مقابل، SVM معمولاً درپردازش داده‌های بزرگ با مشکلاتی همچون زمان و محدودیت‌های حافظه روبرو است.

٦. نتیجه‌گیری

روش‌های نمونه‌برداری سنتی محدودیت‌هایی دارد که می‌تواند دقت و صحت نتایج تحلیل و درک الگوها و تغییرات در داده‌ها را تحت تاثیر قرار دهد [۱۹ و ۲۰]. برای حل این مشکلات، استفاده از روش‌های جدیدتر و پیشرفته‌تر می‌تواند منجر به بهبود قابلیت پوشش مکانی و زمانی داده‌ها و کاهش تاخیر و هزینه‌های مرتبط با روش نمونه‌برداری شود.

این پژوهش، بررسی و تحلیل مقادیر کلروفیل آ در منطقه مورد مطالعه را با استفاده از داده‌های برداشت میدانی و داده‌های ماهواره‌ای انجام داده است. نتایج حاصل از تحلیل نشان می‌دهد که مدل‌های جنگل تصادفی و ماشین بردار پشتیبان به عنوان روش‌های پیش‌بینی با عملکرد بهتر نسبت به مدل‌های خطی مانند رگرسیون خطی و مدل خطی تعمیم‌یافته با توزیع پواسون عمل می‌کنند.

با توجه به اهمیت کلروفیل آ در منطقه مطالعه و تنوع اکوسیستم‌های دریایی، مطالعه حاضر ارزش بالایی دارد. استفاده از داده‌های ماهواره‌ای نیز امکان می‌دهد تا الگوها و تحولات این متغیر زیستی را به طور جامع‌تر و دقیق‌تر بررسی کرد. در نتیجه، این پژوهش می‌تواند به درک بهتر از پدیده شکوفایی جلبکی و اکوسیستم‌های دریایی موجود در منطقه و همچنین پیش‌بینی و مدیریت مناسب آن‌ها کمک کند.

به طور کلی، این پژوهش نشان می‌دهد که برای بررسی و پیش‌بینی مقادیر کلروفیل آ در منطقه مورد مطالعه، استفاده از ترکیب داده‌های برداشت میدانی و داده‌های ماهواره‌ای به وسیله مدل‌های پیش‌بینی مانند جنگل تصادفی و ماشین بردار پشتیبان می‌تواند روشی مؤثر و دقیق باشد. نتایج حاصل از این پژوهش می‌توانند به سازمان‌ها و تصمیم‌گیران مرتبط در زمینه حفاظت و مدیریت منابع دریایی کمک کنند.

از آنجا که کیفیت مجموعه داده‌ها به طور قابل توجهی عملکرد مدل‌های مبتنی بر یادگیری ماشین را محدود می‌کند، بنابراین برای بهبود داده‌های جمع‌آوری شده از ایستگاه‌های مختلف، یکنواختی مناسب برای توزیع مکانی داده‌های نمونه‌برداری ضروریست. همچنین، به جز پارامترهای نوری داده‌های ماهواره‌ای که مستقیماً با غلظت کلروفیل آ مرتبط هستند، عوامل محیطی دیگری (شامل دما، نور، مواد مغذی و ...) ممکن است بر غلظت کلروفیل آ تأثیر بگذارند، که پیشنهاد می‌شود در مطالعات آتی مورد بررسی قرار گیرد.

سپاسگزاری

از پژوهشکده اکولوژی خلیج فارس و دریای عمان که داده‌های برداشت و جمع‌آوری شده میدانی برای انجام این پژوهش را در اختیار قرار داده اند، سپاسگزاری می‌شود.

مراجع

1. Harvey ET, Kratzer S, Philipson P. Satellite-based water quality monitoring for improved spatial and temporal retrieval of chlorophyll-a in coastal waters. Remote Sensing of Environment. 2015;158:417-30.
2. Blix K, Li J, Massicotte P, Matsuoka A. Developing a new machine-learning algorithm for estimating chlorophyll-a concentration in optically complex waters: A case study for high northern latitude waters by using Sentinel 3 OLCI. Remote Sensing. 2019;11(18):2076.
3. Harding Jr L, Mallonee M, Perry E. Toward a predictive understanding of primary productivity in a temperate, partially stratified estuary. Estuarine, Coastal and Shelf Science. 2002;55(3):437-63.
4. Moses WJ, Gitelson AA, Berdnikov S, Povazhnyy V. Estimation of chlorophyll-a concentration in case II waters using MODIS and MERIS data—successes and challenges. Environmental research letters. 2009;4(4):045005.
5. Hafeez S, Wong MS, Ho HC, Nazeer M, Nichol J, Abbas S, et al. Comparison of machine learning algorithms for retrieval of water quality indicators in case-II waters: A case study of Hong Kong. Remote sensing. 2019;11(6):617.
6. Gordon HR, Clark DK, Brown JW, Brown OB, Evans RH, Broenkow WW. Phytoplankton pigment concentrations in the Middle Atlantic Bight: comparison of ship determinations and CZCS estimates. Applied optics. 1983;22(1):20-36.
7. Gower J, Doerffer R, Borstad G. Interpretation of the 685nm peak in water-leaving radiance spectra in terms of fluorescence, absorption and scattering, and its observation by MERIS. International Journal of Remote Sensing. 1999;20(9):1771-86.
8. Mauri E, Poulain PM, Južnič-Zonta Ž. MODIS chlorophyll variability in the northern Adriatic Sea and relationship with forcing parameters. Journal of Geophysical Research: Oceans. 2007;112(C3).
9. Mozetič P, Solidoro C, Cossarini G, Socal G, Precali R, Francé J, et al. Recent trends towards oligotrophication of the northern Adriatic: evidence from chlorophyll a time series. Estuaries and coasts. 2010;33:362-75.
10. Pahlevan N, Smith B, Schalles J, Binding C, Cao Z, Ma R, et al. Seamless retrievals of chlorophyll-a from sentinel-2 (MSI) and sentinel -3 (OLCI) in inland and coastal waters: A machine-learning approach. Remote Sensing of Environment. 2020;240:111604.
11. Nelder JA, Wedderburn RW. Generalized linear models. Journal of the Royal Statistical Society: Series A (General). 1972;135(3):370-84.
12. Vapnik V. The Nature of Statistical Learning Theory, Springer-Verlag; New York, Inc. 1995.
13. Breiman L. Random forests. Machine learning. 2001;45:5-32.
14. Zhao X, Li Y, Chen Y, Qiao X, Qian W. Water Chlorophyll a Estimation Using UAV-Based Multispectral Data and Machine Learning. Drones. 2023;7(1):2.
15. Lv H, Feng Q. A review of random forests algorithm. Journal of the Hebei Academy of Sciences. 2019;36:37-41.
16. Barraza-Moraga F, Alcayaga H, Pizarro A, Félez-Bernal J, Urrutia R. Estimation of Chlorophyll-a Concentrations in Lanalhue Lake Using sentiel2 MSI Satellite Images. Remote Sensing. 2022;14(22):5647.
17. Hamzehei S. Field study and numerical simulation of developing red tide in the northern Strait of Hormuz. 2012, Islamic Azad University ,Science and Research Branch, Tehran, Marine Physics.

18.Su H, Lu X, Chen Z, Zhang H, Lu W, Wu W. Estimating coastal chlorophyll-a concentration from time-series OLCI data based on machine learning. Remote Sensing. 2021 Feb 6;13(4):576.

19. Azizi Z, Montazeri Z. Effects of microtopography on the spatial pattern of woody species in West Iran. Arabian Journal of Geosciences. 2018 May;11(10):244.

20. Azizi Z, Najafi A. Fuzzy classification in forest area for road design, (Case study: Lirehsar forest, Tonekabon). Iranian Journal of Forest and Poplar Research. 2011;19(1): 43-54. (In Persian)

پی‌نوشت‌ها

1. Muri
2. Mozetič
3. Barraza-Moraga
4. Lanalhue
5. Sentinel
6. Random Forest
7. Su
8. LightGBM
9. MODIS
10. TERRA
11. Linear Model
12. Train
13. Test
14. Generalized Linear Model
15. conditional mean
16. Support Vector machine
17. Vapnik
18. Support Vector Regression
19. Kernel
20. Radial Basis Function Kernel
21. Root Mean Square Error
22. Mean Percentage Error
23. R-squared
24. Mean Absolute Error
25. Adjusted R squared
26. Residual standard error
27. Deviance Residuals
28. Standard Error
29. Dispersion parameter
30. Bootstrapping
31. correlation
32. SD Ratio