کاربست الگوریتم‌های یادگیری ماشین برای تخمین تابش خورشیدی (مورد مطالعه: اقلیم خشک و نیمه‌خشک)

سمیه سلطانی گردفرامرزی [۲و۱*] و هاجر مومنی [۳]

[۱] دانشیار گروه علوم و مهندسی آب، دانشکده کشاورزی و منابع طبیعی، دانشگاه اردکان، اردکان، ایران
[۲] پژوهشکده آب، انرژی و محیط زیست، دانشگاه اردکان، اردکان، ایران
[۳] استادیار گروه مهندسی برق، دانشکده فنی و مهندسی، دانشگاه اردکان، اردکان، ایران

(دریافت: ۱۴۰۲/۰۱/۲۶، پذیرش: ۱۴۰۲/۰۲/۲۴)

چکیده

تابش خورشیدی، یکی از متغیرهای مهم در مدل‌های بیلان انرژی و شبیه‌سازی رشد گیاهان است. در پژوهش حاضر، عملکرد نه الگوریتم یادگیری ماشین نظارت شده شامل الگوریتم‌های رگرسیون خطی (LR)، رگرسیون خطی با اصلاح تابع زیان (LASSO)، رگرسیون خالص الاستیک (EN)، k نزدیک‌ترین همسایه (KNN)، درخت تصمیم‌گیری (DT)، ماشین بردار پشتیبان (SVR)، جنگل تصادفی (RF)، درختان اضافی (ET) و الگوریتم تقویت ماشین (GBM) برای برآورد تابش خورشیدی در ایستگاه همدید یزد در حد فاصل سال‌های ۲۰۰۵ تا ۲۰۲۱ با روش اعتبار سنجی متقابل (kfold) مورد بررسی قرار گرفت. پارامترهای میانگین دما، دمای کمینه، دمای بیشینه، ساعات آفتابی، رطوبت نسبی و تابش خورشیدی به‌صورت روزانه از سازمان هواشناسی کشور دریافت و متغیرهای تابش فرازمینی، فاصله نسبی زمین تا خورشید، زاویه میل خورشیدی و حداکثر ساعات آفتابی با روابط موجود محاسبه و برای ورودی مدل‌های پیش‌بینی انتخاب شدند. معیارهای ارزیابی برای تخمین تابش خورشیدی MSE (متوسط مربعات خطا)، MAPE (متوسط قدرمطلق خطا) و ضریب تعیین (R^2) در نظر گرفته شدند. نتایج نشان داد که مدل رگرسیون ماشین بردار پشتیبان (SVR) کمترین خطا را برای تخمین تابش روزانه خورشید دارد؛ به‌طوری‌که مدل ماشین بردار پشتیبان با میانگین مربعات خطای ۲/۸۵ مگا ژول بر مترمربع بر روز، قدر مطلق خطای ۰/۸۰۳ و ضریب تبیین ۰/۹۱۹ در مرحله آزمون و به‌ترتیب ۱/۵۴ مگا ژول بر مترمربع بر روز، ۴/۹۲ و ۰/۸۷۰ در مرحله آموزش مدل‌ها نسبت به سایر مدل‌ها عملکرد بهتری در تخمین تابش خورشیدی داشته است که نشان‌دهنده توانائی این مدل برای کاربردهای خورشیدی و گرمایی توسط مهندسان و سایر محققین است.

واژه‌های کلیدی: مشخصات هندسی، داده‌کاوی، زاویه میل خورشیدی، تابش، الگوریتم، یزد

* نگارنده رابط: ssoltani@ardakan.ac.ir

۱ مقدمه

پیچیدگی فرآیندهای هواشناسی و هیدرولوژی از نظر انتخاب تمام پارامترهای تأثیرگذار بر آن‌ها و نقص اطلاعات آماری، امکان مدل‌سازی این فرآیندها را غیرممکن می‌سازد. در چنین شرایطی استفاده از مدل‌سازی مبنی بر روابط ریاضی، مورد توجه قرار گرفته است (عبدالحفیظی و همکاران، ۲۰۲۱). از مطالعات انجام شده در این زمینه می‌توان به مدل‌سازی تبخیر و تعرق گیاه مرجع (شیخ‌الاسلامی و همکاران، ۱۳۹۳)، آبشستگی پایه پل (سلطانی گردفرامرزی و تقی زاده، ۱۳۹۵)، ضریب پخشیدگی آلودگی (سلطانی گردفرامرزی و همکاران، ۱۳۹۴)، فرسایش خاک (بروغنی و همکاران، ۲۰۲۲) و کیفیت آب زیرزمینی (امیری و همکاران، ۲۰۲۱) و .... با روش‌های مختلف هوش مصنوعی و مدل‌سازی ریاضی اشاره کرد. تابش خورشیدی یکی از متغیرهای مهم و مؤثر هواشناختی در برآورد تبخیر و تعرق و نیاز آبی گیاهان است و منشأ انرژی برای همه تحولات جو و سطح زمین است. اندازه‌گیری این متغیر اگرچه در ایران سابقه نسبتاً طولانی دارد ولی به‌دلیل هزینه‌های زیاد وسایل اندازه‌گیری در بسیاری از ایستگاه‌های موجود کشور دستگاه تابش سنج یا پیرانومتر وجود ندارد و یا مشکلاتی همچون واسنجی آن، تجمع آب و گردوغبار بر روی سنجنده آن وجود دارد (رحیمی خوب، ۲۰۱۰). 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 به‌دلیل نقض دستگاه و یا مشکلات دیگر مشاهده می‌شود (هانت و همکاران، ۱۹۹۸). هرچند در اغلب این ایستگاه‌ها ساعات آفتابی به‌طور روزانه اندازه گرفته می‌شود. تحقیقات مختلفی برای تخمین تابش خورشیدی با استفاده از داده‌های هواشناسی انجام شده و روش‌های زیادی توسعه یافته است. محققین به دنبال راهی برای تخمین بهتر و دقیق‌تر تابش خورشیدی می‌باشند. در نتیجه روابط تجربی و رگرسیونی، فنون سنجش از دور و میان‌یابی خطی توسعه یافتند (سبزی پرور و شتایی، ۲۰۰۷)؛ اما در تخمین تابش خورشیدی توسط معادلات تجربی و نیمه تجربی، تنها تعداد محدودی از متغیرهای هواشناسی کاربرد دارد. در سال‌های اخیر پژوهشگران زیادی مطالعات خود را بر مبنای استفاده از روش‌های داده کاوی و مدل‌سازی ریاضی برای تخمین تابش خورشیدی معطوف داشته‌اند (یاداو و چاندل، ۲۰۱۵؛ مهدی زاده و همکاران، ۲۰۱۶، اولالکان و همکاران، ۲۰۱۸؛ محمدی و امامقلی زاده، ۱۳۹۵۷، اکوندامیا و همکاران، ۲۰۱۶، مینال و سلواکومار، ۲۰۱۸، رادوسویچ و همکاران،۲۰۲۰، ژنگ و همکاران ۲۰۲۰، بالمهدی و همکاران، ۲۰۲۰، عبدالحفیظی و همکاران، ۲۰۲۱، تکی و همکاران ۲۰۲۱؛ نوکولو و همکاران، ۲۰۲۲، سلطانی گردفرامرزی، ۱۴۰۲ و جهان تیغ و پیری ۱۴۰۲). لازم به ذکر است که مطالعات زیادی در خصوص برآورد تابش خورشیدی در برخی ایستگاه‌های ایران و یا حتی در خارج از ایران انجام شده است که تنها از داده‌های هواشناسی استفاده شده و تأثیر داده‌های هندسی، نجومی، جغرافیایی بررسی نشده است. تحقیقات گذشته در ایران نشان می‌دهد بیشتر مطالعات انجام شده مبتنی بر استفاده از مدل‌های تجربی است و ضروری است که کارایی روش‌های هوش مصنوعی در برآورد تابش خورشیدی ارزیابی گردد. از یک طرف، 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 مورد استفاده قرار گیرد، لذا بررسی روابط غیرخطی موجود بین متغیرهای هواشناسی مؤثر بر تابش خورشیدی ضروری به نظر می‌رسد. با توجه به اهمیت تابش رسیده به سطح زمین و عوامل متعدد تأثیرگذار بر آن، هدف از این تحقیق تخمین تابش خورشیدی در ایستگاه یزد با یک اقلیم خشک و گرم

با مجموعه‌ای از داده‌های هواشناسی، هندسی و نجومی با استفاده از نه الگوریتم‌های یادگیری ماشین شامل الگوریتم‌های رگرسیون خطی (LR)، رگرسیون خطی با اصلاح تابع زیان (LASSO)، رگرسیون خالص الاستیک (EN)، k نزدیک‌ترین همسایه (KNN)، درخت تصمیم‌گیری (DT)، ماشین بردار پشتیبان (SVR)، جنگل تصادفی (RF)، درختان اضافی (ET) و الگوریتم تقویت ماشین (GBM) با روش اعتبار سنجی متقابل (kfold) و تعیین بهترین الگوریتم برای برآورد تابش خورشیدی است که تاکنون چنین مطالعه‌ای انجام نشده است. در این پژوهش ابتدا منطقه مورد مطالعه و نحوه آماده سازی نتایج معرفی می‌گردد. سپس روش انجام کار تشریح شده و در قسمت سوم نتایج تحقیق و نتیجه‌گیری کلی ارائه می‌گردد. معرفی الگوریتم‌های مورد استفاده نیز در قسمت پیوست انجام شده است.

۲ مواد و روش

۲-۱ معرفی منطقه مورد مطالعه

داده‌های مورد استفاده در این تحقیق متغیرهای هواشناختی اندازه‌گیری شده در ایستگاه یزد طی سال‌های ۲۰۰۵ تا ۲۰۲۱ در مقیاس روزانه است. دلیل استفاده از این بازه زمانی پیوستگی و کامل بودن داده‌های مربوط به ساعات آفتابی و همچنین طول دوره آماری مشترک آمار بلندمدت متغیرهای هواشناختی است. مشخصات متغیرهای هواشناختی مذکور در جدول (۱) نشان داده شده است.

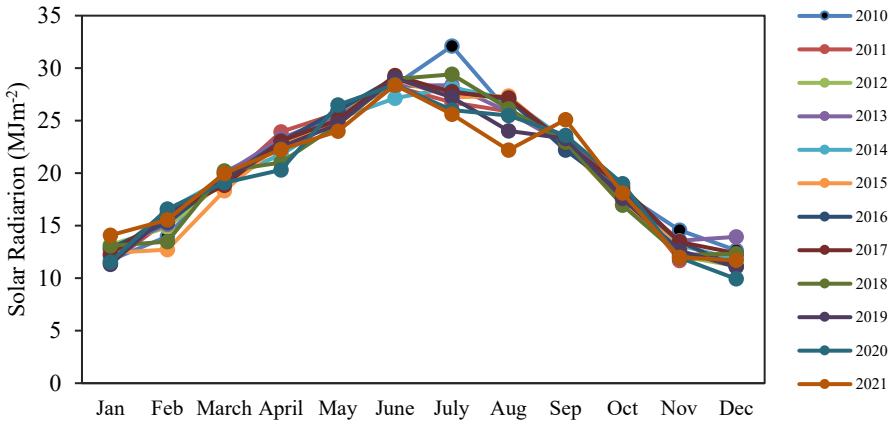
ایستگاه یزد در موقعیت طول و عرض جغرافیایی به‌ترتیب ۳۱/۸۹۷۴ درجه شمالی و ۵۴/۳۵۶۹ شرقی در ارتفاع ۱۲۱۶ متری از سطح دریا قرار گرفته است. روند تغییرات مقادیر اندازه‌گیری شده متوسط تابش خورشیدی در ایستگاه یزد طی سال‌های ۲۰۱۰ تا ۲۰۲۱ در شکل (۱) آورده شده است.

۲-۲ آماده سازی داده‌ها

به‌منظور تعیین مقادیر نادرست تابش خورشیدی، شاخص شفافیت روزانه (Kt) محاسبه گردید و مقادیر خارج از بازه ۰/۱۵ <Kt<۱ حذف شدند (محمدی و همکاران، ۲۰۱۵). لازم به ذکر است که این شاخص، نسبت تابش خورشیدی روزانه مشاهده شده به تابش خورشیدی فرازمینی روزانه می‌باشد (Kt=Rs/Ra). 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 ارائه شده توسط دافی و بکمن (۱۹۹۱) محاسبه شدند. به‌منظور پیاده‌سازی الگوریتم‌ها برای پیش‌بینی تابش خورشیدی از زبان برنامه‌نویسی پایتون (Python 3.9.12) استفاده شده و کدها در محیط توسعه یکپارچه اسپایدر (Spyder 5.1.5) که یک محیط متن باز است، نوشته و اجرا شد. همچنین زمانی که کاربر کدها را نوشته و اجرا می‌کند، خطاها را نمایش می‌دهد. این محیط، زمانی که کدها اجرا می‌شوند قابلیت نمایش متغیرها و مقادیر آن‌ها را دارد.

جدول ۱. شاخص‌های آماری متغیرهای مورد بررسی.

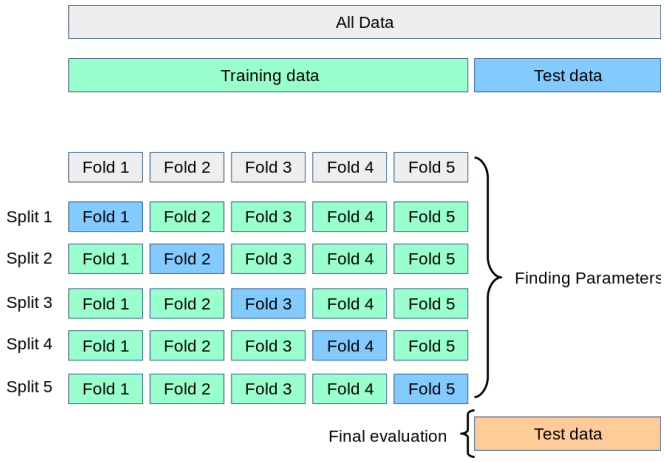
پارامترها	واحد	حداقل	حداکثر	میانگین	انحراف معیار	چولگی
تابش خورشیدی (Rs)	مگاژول بر مترمربع بر روز	۴/۰۵	۳۰/۶۶	۱۹/۳۵	۶/۱۷	−۰/۱۹۴
تابش فرازمینی (Ra)	مگاژول بر مترمربع بر روز	۳۱/۶۱	۳۳/۰۳	۳۲/۳۲	۰/۴۳۳	۰/۰۰۳
ساعات آفتابی نسبی (n/N)	–	۰/۰۱۵	۱	۰/۷۵۸	۰/۲۰۴	−۱/۵۱۱
رطوبت نسبی (RH)	–	۵	۹۵/۳۸	۲۷/۳۲	۱۷/۶۷	۱/۲۹۸
اختلاف دمای حداکثر و حداقل	سانتی‌گراد	۲/۸	۲۸	۱۳/۸۶	۳/۰۹	−۰/۳۶۵
میانگین دما (T_{mean})	سانتی‌گراد	−۳/۶	۴۵/۶	۲۸/۰۰	۹/۲۶	−۰/۰۴۳
زاویه میل خورشیدی (δ)	درجه	۰/۰۰۰۳۱	۰/۰۳۵۰	۰/۰۱۲۶	۰/۰۱۳	−۰/۰۱۲
فاصله نسبی زمین تا خورشید (dr)	درجه	۱/۰۳۲	۱/۰۳۳	۱/۰۳۲	۰/۰۰۰۰۶	−۰/۵۱۱



شکل ۱. تغییرات متوسط تابش خورشیدی در ایستگاه یزد از سال ۲۰۱۰ تا ۲۰۲۱ میلادی.

۳-۲ روش اعتبار سنجی متقابل یا kfold CV

اعتبارسنجی متقابل، یک روش ارزیابی مدل است که تعیین می‌نماید نتایج یک تحلیل آماری بر روی یک مجموعه داده تا چه اندازه قابل تعمیم و مستقل از داده‌های آموزشی است. این روش به‌طور ویژه در کاربردهای پیش‌بینی مورد استفاده قرار می‌گیرد تا مشخص شود مدل موردنظر تا چه اندازه در عمل مفید خواهد بود. در این روش، مجموعه داده به k مجموعه کوچک‌تر تقسیم می‌شود. هر بار به تصادف یکی از بخش‌ها به‌صورت مجموعه ارزیابی در نظر گرفته می‌شود و k-1 بخش دیگر، مجموعه آموزش برای آموزش مدل‌ها مورد استفاده قرار می‌گیرد. در مجموعه آموزش، پارامترهای مرتبط با هر مدل تنظیم می‌شود. نتیجه ارزیابی از روی مجموعه ارزیابی محاسبه می‌شود. این کار k بار تکرار می‌شود و میانگین نتایج ارزیابی گزارش می‌شود (تکی و همکاران، ۲۰۲۱). شکل (۲) شماتیکی از روش اعتبار سنجی متقابل را نشان می‌دهد که k برابر با ۵ است. نتایج الگوریتم‌های یادگیری ماشین به مجموعه داده‌های مورد استفاده در مرحله آموزش بستگی دارد (هان و همکاران، ۲۰۲۲)؛ بنابراین، نتایج در هر تکرار با انتخاب تصادفی داده متفاوت خواهد بود. در پژوهش حاضر از



شکل ۲. شماتیکی از روش اعتبار سنجی متقابل.

روش اعتبارسنجی متقاطع K-fold برای تخمین بهتر الگوریتم‌های یادگیری ماشین و ارزیابی پایداری و تعمیم‌پذیری استفاده شد. به این صورت که ۸۰ درصد داده‌ها، داده آموزش و به‌طور تصادفی انتخاب و مدل با آن آموزش داده شد و ۲۰ درصد باقی‌مانده داده‌ها، داده ارزیابی یا آزمون در نظر گرفته شدند و نتیجه به‌دست آمد. این کار چندین بار تکرار شده و هر بار به‌طور تصادفی ۸۰ درصد جداسازی انجام و در نهایت، میانگین و انحراف معیار نتایج برای هر الگوریتم به‌دست می‌آید. الگوریتم‌های مورد استفاده در این پژوهش در قسمت پیوست مقاله ارائه شده است.

۲-۴ 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 معیارهای ارزیابی برای تخمین تابش خورشیدی MSE (متوسط مربعات خطا)، MAPE (متوسط قدرمطلق خطا) و ضریب تعیین (R^2) می‌باشند که به‌ترتیب در روابط (۱) تا (۳) آورده شده است. دقیق‌ترین مدل با توجه به این معیارها، مدلی خواهد بود که مقدار سه معیار ارزیابی به‌ترتیب صفر، صفر و یک باشد.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(Rs_{ic} - Rs_{im})^2 \quad (۱)$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{Rs_{ic}-Rs_{im}}{Rs_{ic}}\right| \quad (۲)$$

$$R^2 = \left(\frac{\sum_{i=1}^{N}(Rs_{ic}-Rs_{cave})(Rs_{im}-Rs_{mave})}{\sqrt{(\sum_{i=1}^{N}(Rs_{ic}-Rs_{cave})^2)(\sum_{i=1}^{N}(Rs_{im}-Rs_{mave})^2)}}\right)^2 \quad (۳)$$

در اینجا، N تعداد مشاهدات، Rs_{im} و Rs_{ic} به‌ترتیب مقادیر مشاهده شده تابش خورشیدی و مقادیر برآورد شده تابش خورشیدی و Rs_{cave} و Rs_{mave} به‌ترتیب میانگین مقادیر برآورد شده تابش خورشیدی و مشاهده شده تابش خورشیدی است. همچنین به‌منظور ارزیابی میزان همبستگی بین متغیرها از ضریب همبستگی پیرسون استفاده شد که از رایج‌ترین روش‌های همبستگی می‌باشد.

۳ تحلیل نتایج

میزان همبستگی بین متغیرهای پژوهش حاضر و تابش خورشیدی در ایستگاه یزد در جدول (۲) ارائه شده است. برای شناخت مهم‌ترین متغیرهای هواشناسی و جغرافیایی مؤثر بر مقدار شدت تابش خورشیدی، مقدار ضریب همبستگی بین شدت تابش خورشیدی با هر یک از متغیرها محاسبه شد. میزان ارتباط بین این متغیرها با استفاده از نتایج به‌دست آمده از آزمون پیرسون نشانگر ارتباط معنی‌دار بین

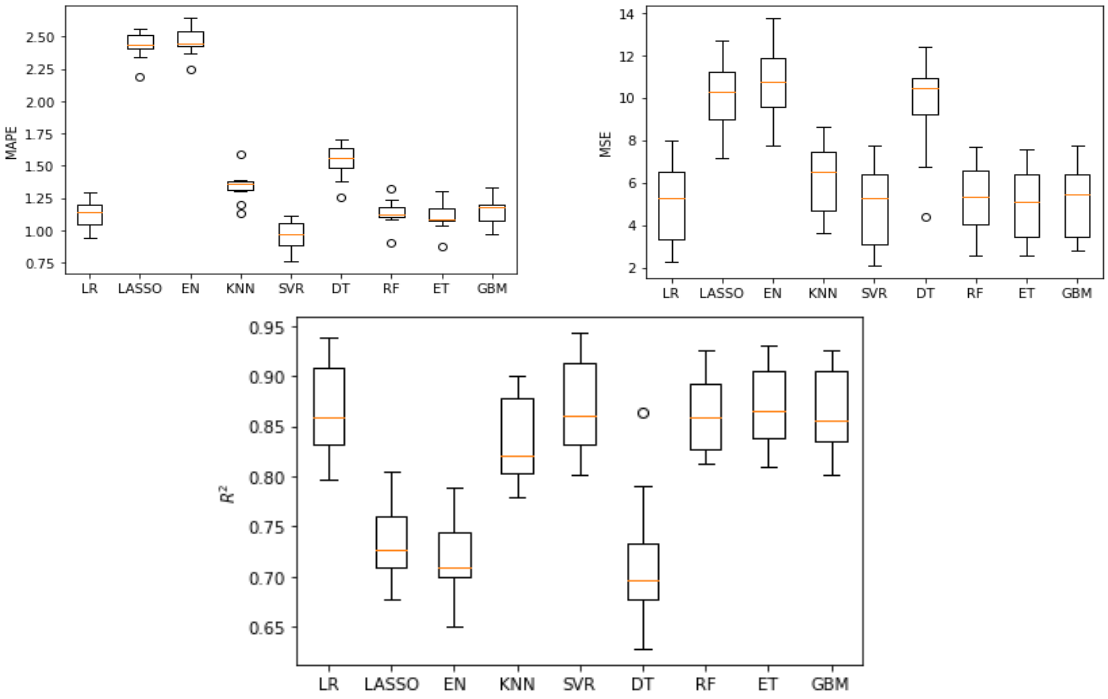
جدول ۲. مقادیر ضریب همبستگی پارامترهای هواشناسی، جغرافیایی و هندسی ایستگاه یزد با تابش خورشیدی.

	T mean	Tmax-Tmin	RH	Rs	dr	δ	Ra	n/N
Tmax-Tmin	**۰/۲۰۸	۱						
RH	**-۰/۶۷۲	**-۰/۵۰۸	۱					
Rs	**۰/۷۳۴	**۰/۲۷۰	**-۰/۶۳۴	۱				
dr	*۰/۰۳۷	**-۰/۰۴۵	**-۰/۰۸۱	**۰/۲۴۲	۱			
δ	**۰/۱۶۷	**۰/۰۶۱	**-۰/۰۴۶	**-۰/۰۵۹	**-۰/۹۷۳	۱		
Ra	**۰/۱۶۱	**۰/۰۶۰	*-۰/۰۴۳	**-۰/۰۶۵	**-۰/۹۷۵	**۱/۰۰	۱	
n/N	**۰/۴۰۷	**۰/۴۲۳	**-۰/۵۵۲	**۰/۷۱۶	^{ns}۰/۰۱۰	**۰/۰۷۶	**۰/۰۷۶	۱

متغیرها در سطح اطمینان ۹۹ درصد است. همان‌طور که نتایج نشان می‌دهد، همه پارامترهای مورد بررسی در سطح اطمینان ۹۹ درصد با تابش خورشیدی همبستگی دارند و در نتیجه برای ورودی الگوریتم‌های یادگیری ماشین استفاده شدند. در میان این پارامترها میانگین دما و ساعات آفتابی بالاترین همبستگی مثبت و رطوبت نسبی دارای همبستگی منفی معنی‌دار است. چون با افزایش بخار آب و رطوبت در هوا تابش خورشیدی کمتری به سطح زمین می‌رسد (سیدیان و همکاران، ۱۳۹۶؛ ژنگ و چیائو، ۲۰۱۳).

۳-۱ نتایج مرحله آموزش مدل

در شکل (۳) نتایج معیارهای ارزیابی نه الگوریتم آموزش به‌کار رفته در پژوهش حاضر در مرحله آموزش ارائه شده است. محور افقی شکل ۹ الگوریتم بکار رفته در پژوهش و محور عمودی مقادیر آماره‌های ارزیابی مدل‌ها می‌باشد. همان‌طور که شکل نشان می‌دهد، ضریب تعیین (R^2) بسته به الگوریتم به‌کار رفته بین ۰/۶۵ و ۰/۹۵ متغیر است؛ به عبارت دیگر، از نظر مقادیر ضریب تبیین، همه الگوریتم‌های مورد استفاده نتایج خوبی برای پیش‌بینی تابش خورشیدی نشان دادند. متوسط مربعات خطای مدل‌ها در بازه ۲ تا ۱۴ مگاژول بر مترمربع بر روز متغیر است. مطابق با شکل (۳) مدل‌های LASSO، EN و DT دارای بیشترین مقادیر متوسط مربعات خطا بوده و الگوریتم‌های LR، ET، RF، SVR و GBM کمترین مقدار متوسط مربعات خطا را دارا هستند. نتایج متوسط قدر مطلق خطا نیز نشان داد که الگوریتم‌های LASSO، EN بالاترین مقدار خطا (۲/۵) و الگوریتم SVR کمترین میزان خطا (۱) را در مرحله آموزش مدل کسب کردند. مقادیر میانگین ضریب تبیین همچنین نشان داد که مدل SVR و DT به ترتیب دارای بیشترین (۰/۸۷) و کمترین (۰/۷) ضریب تبیین در بین مدل‌های مورد استفاده هستند. به طور کلی با توجه به نتایج هر سه معیار مشاهده می‌شود که الگوریتم رگرسیون بردار پشتیبان (SVR) بهتر از سایر الگوریتم‌ها در مرحله آموزش مدل‌ها عمل کرده است.



شکل ۳. مقادیر معیارهای ارزیابی الگوریتم‌های مورد استفاده در مرحله آموزش مدل‌ها.

جدول ۳. مقادیر میانگین و انحراف معیارهای ارزیابی مدل‌های بکار رفته در مطالعه در مرحله آموزش.

		LR	LASSO	EN	KNN	DT	SVR	RF	ET	GBM
MAPE	میانگین	۵/۰۴	۱۰/۱۰	۱۰/۷۳	۶/۲۱	۱۰/۶۲	۴/۹۲	۵/۳۱	۵/۱۱	۵/۱۶
	انحراف معیار	۱/۸۸	۱/۵۷	۱/۶۸	۱/۷۴	۲/۷۰	۱/۸۷	۱/۵۶	۱/۶۸	۱/۶۵
MSE	میانگین	۱/۱۳	۲/۴۳	۲/۴۶	۱/۳۴	۱/۵۴	۰/۹۵	۱/۱۲	۱/۰۹	۱/۱۵
	انحراف معیار	۰/۱۱	۰/۱۰	۰/۱۰	۰/۱۱	۰/۱۶	۰/۱۲	۰/۱۰	۰/۱۱	۰/۱۱
R^2	میانگین	۰/۸۶۷	۰/۷۳۲	۰/۷۱۶	۰/۸۳۶	۰/۷۱۷	۰/۸۷۰	۰/۸۶۰	۰/۸۶۷	۰/۸۶۴
	انحراف معیار	۰/۰۴	۰/۰۴	۰/۰۴	۰/۰۴	۰/۰۶	۰/۰۴	۰/۰۴	۰/۰۴	۰/۰۴

جدول (۳) نیز مقادیر میانگین و انحراف معیارهای ارزیابی مدل‌های به‌کار رفته در پژوهش حاضر در مرحله آموزش ارائه شده است. مطابق با نتایج، متوسط قدر مطلق خطا در بازه ۴/۹۲ تا ۱۰/۷۳ و مقادیر متوسط مربعات خطا در مرحله آموزش از ۰/۹۵ تا ۲/۴۶ مگاژول بر مترمربع بر روز متغیر بود. مطابق با نتایج کمترین میانگین متوسط مربعات خطا ۰/۹۵ و انحراف معیار ۰/۱۲ برای الگوریتم SVR حاصل شد و نسبت به سایر روش‌ها نتایج قابل قبولی در پیش‌بینی تابش خورشیدی ارائه داد، چرا که بیشترین دقت و کمترین خطا را نشان داد. یکی از دلایل این موضوع را می‌توان به‌دلیل وجود رابطه غیرخطی تراز تابش خورشیدی با پارامترهای ورودی مانند دما، رطوبت، ساعات آفتابی و ... نسبت داد. همچنین این روش قابلیت کار کردن در فضای مشاهدات زیاد را دارا است و بهترین مدل را در کل فضای مشاهدات پیدا می‌کند و چون به‌صورت ناحیه‌ای عمل نمی‌کند، نتایج بهتری نسبت به سایر الگوریتم‌ها نشان داد. متوسط قدر مطلق خطا و انحراف معیار این الگوریتم نیز به‌ترتیب ۴/۹۲ و ۱/۸ به‌دست آمد. هم‌چنین میانگین و انحراف معیار ضریب تبیین الگوریتم SVR به‌ترتیب برابر با ۰/۸۷ و ۰/۰۴ محاسبه گردید. بعد از الگوریتم رگرسیون بردار پشتیبان (SVR)، الگوریتم رگرسیون خطی (LR) با میانگین متوسط قدر مطلق خطای ۵/۰۴ و میانگین متوسط مربعات خطای ۱/۱۳ و میانگین ضریب تبیین ۰/۸۶۷ در رتبه بعدی قرار گرفت. هم‌چنین الگوریتم رگرسیون خالص الاستیک (EN) با بیشترین میانگین قدر مطلق خطا (MAPE)، بیشترین میانگین متوسط مربعات خطا (MSE) و کمترین ضریب تبیین (R^2) در بین نه الگوریتم مورد استفاده در رتبه آخر قرار گرفت.

با توجه به اینکه الگوریتم رگرسیون بردار پشتیبان نتایج بهتری نسبت به سایر الگوریتم‌ها نشان داد، به ازای دو پارامتر مهم این الگوریتم، یعنی پارامتر حاشیه اطمینان (C) و کرنل آن، میزان‌سازی انجام شد. مقادیر حاشیه اطمینان ۰/۱، ۰/۳، ۰/۵، ۰/۷، ۰/۹، ۱/۰، ۱/۳، ۱/۵، ۱/۷ و ۲/۰ در نظر گرفته شد. برای مقادیر کرنل نیز توابع خطی (linear)، چندجمله‌ای (poly)، تابع شعاعی (rbf) و سیگموئید (sigmoid) به‌کار رفت. نتایج میزان‌سازی الگوریتم SVR با در نظر گرفتن مقادیر مختلف برای پارامتر C و انواع کرنل‌ها و بهترین نتیجه در جدول (۴) نشان داده شده است. برای مدل نهایی، این روش با پارامترهای C=2 و کرنل RBF استفاده گردید؛ چراکه میانگین متوسط مربعات خطای کمتری نسبت به سایر پارامترها حاصل شد (۴/۸۲ مگاژول بر مترمربع بر روز). مطابق نتایج میزان‌سازی دامنه تغییرات میانگین متوسط مربعات خطا برای تابع خطی از ۵/۷۹ تا ۶/۴۱ تغییر می‌کند. تغییرات میانگین متوسط مربعات خطا تابع سهمی از ۹/۳۳ تا ۱۲/۷۲ متغیر است و میانگین متوسط مربعات خطا تابع پایه شعاعی نیز از ۴/۸۲ تا ۶/۵۱ و هم‌چنین این تغییرات برای تابع سیگموئید از ۳۹/۳۳ تا ۱۶۳۰۳/۱۹ است که نشان دهنده ناکارآمدی این تابع

جدول ۴. نتایج میزان‌سازی الگوریتم رگرسیون بردار پشتیبان برای مقادیر مختلف C و انواع کرنل.

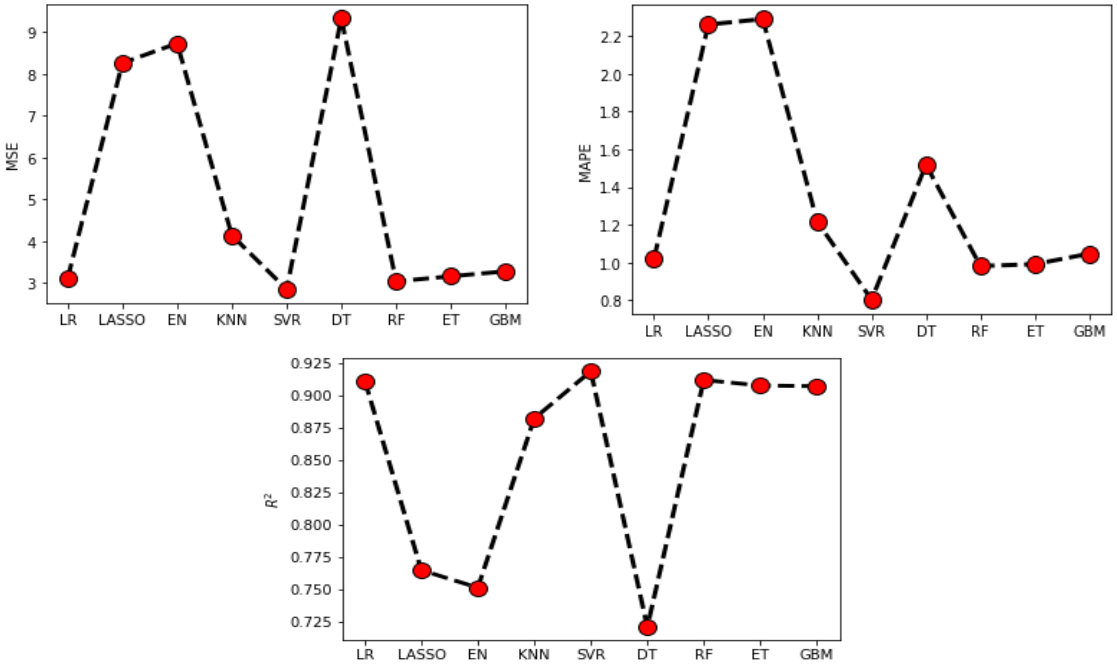
C	تابع	انحراف معیار± میانگینMSE	C	تابع	انحراف معیار± میانگینMSE
۰/۱	Linear	۶/۴۱± ۱/۷۹	۱	Linear	۵/۷۹± ۱/۹۷
	Poly	۱۲/۷۲± ۱/۸۳		Poly	۱۰/۱۹± ۱/۷۷
	Rbf	۶/۵۱± ۱/۷۴		Rbf	۴/۹۲± ۱/۸۷
	Sigmoid	۳۹/۳۳± ۳/۹۱		Sigmoid	۴۰۸۷/۴۵± ۵۳۸/۰۱
۰/۳	Linear	۵/۸۶± ۱/۸۹	۱/۳	Linear	۵/۷۹± ۱/۹۹
	Poly	۱۱/۵۱± ۱/۷۵		Poly	۹/۸۶± ۱/۸۰
	Rbf	۵/۳۵۱± ۱/۸۱		Rbf	۴/۸۸± ۱/۸۸
	Sigmoid	۳۷۶/۶۶± ۴۳/۲۴		Sigmoid	۶۸۹۰/۴۲± ۸۹۵/۰۲
۰/۵	Linear	۵/۸۰± ۱/۹۳	۱/۵	Linear	۵/۷۹± ۱/۹۹
	Poly	۱۰/۹۷± ۱/۷۵		Poly	۹/۶۱± ۱/۸۱
	Rbf	۵/۰۹± ۱/۸۴		Rbf	۴/۸۶± ۱/۸۸
	Sigmoid	۱۰۳۰/۵۹± ۱۳۰/۸۴		Sigmoid	۹۱۷۴/۴۶± ۱۱۸۳/۸۲
۰/۷	Linear	۵/۷۹± ۱/۹۶	۱/۷	Linear	۵/۷۹± ۱/۹۹
	Poly	۱۰/۶۰± ۱/۷۵		Poly	۹/۵۳± ۱/۸۳
	Rbf	۴/۹۹± ۱/۸۶		Rbf	۴/۸۴± ۱/۸۸
	Sigmoid	۲۰۱۰/۲۴± ۲۶۳/۷۲		Sigmoid	۱۱۷۷۹/۸۳± ۱۵۱۸/۵۶
۰/۹	Linear	۵/۷۹± ۱/۹۷	۲	Linear	۵/۷۹± ۱/۹۹
	Poly	۱۰/۳۲± ۱/۷۶		Poly	۹/۳۳± ۱/۸۵
	Rbf	۴/۹۴± ۱/۸۷		**Rbf**	**۴/۸۲± ۱/۸۹**
	Sigmoid	۳۳۱۲/۲۹± ۴۳۶/۵۷		Sigmoid	۱۶۳۰۳/۶۹± ۲۰۹۹/۳۲

برای پیش‌بینی تابش خورشیدی در الگوریتم ماشین بردار پشتیبان است.

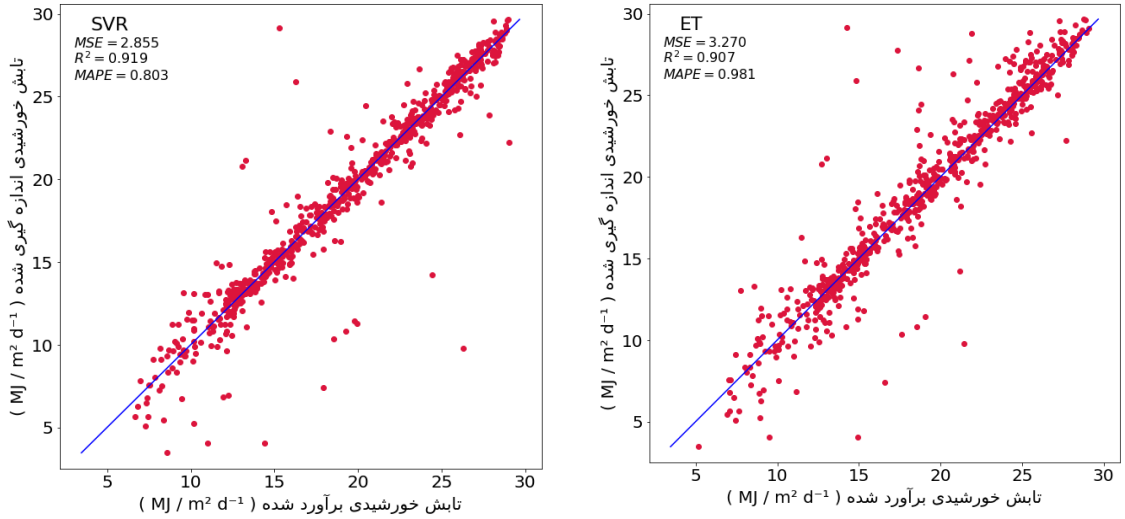
۳-۳ نتایج مرحله آزمون

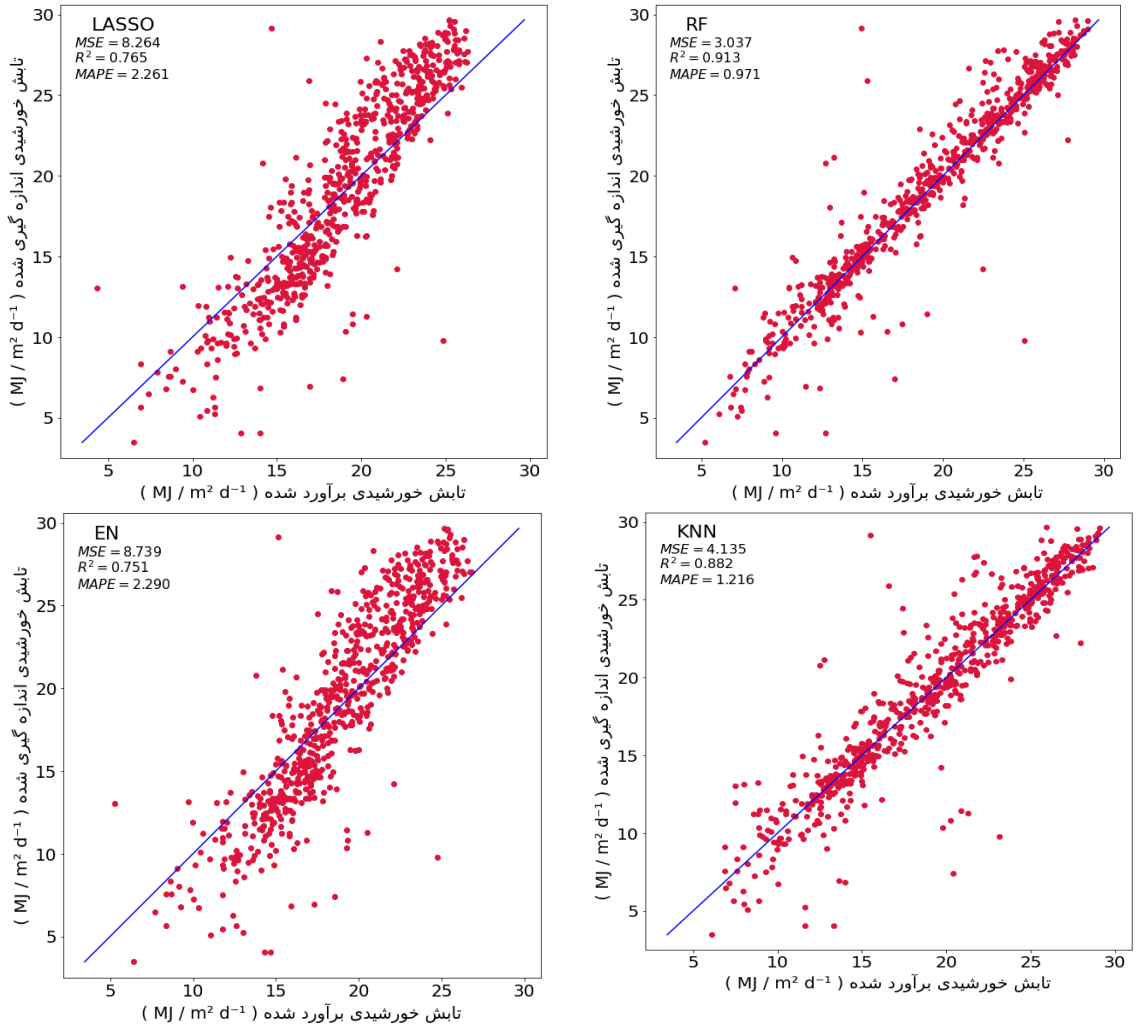
بعد از مرحله آموزش داده‌ها، با استفاده از روش اعتبار سنجی متقاطع K-fold، ۲۰ درصد باقی‌مانده داده‌ها مورد آزمون قرار گرفتند. نتایج مرحله آزمون مدل‌سازی الگوریتم‌های مذکور با معیارهای ارزیابی در شکل (۴) نشان داده شده است. نتایج نشان می‌دهد، تغییرات متوسط مربعات خطا برای نه الگوریتم یادگیری ماشین زیاد است و از حدود ۳ تا ۹ مگا ژول بر مترمربع بر روز متفاوت است. بیشترین مقدار آن در الگوریتم DT (درخت تصمیم‌گیری) و کمترین مقدار آن در الگوریتم ماشین بردار پشتیبان مشاهده گردید. مقدار میانگین قدر مطلق خطا نیز در بازه ۰/۸ تا ۲/۲ مگا ژول بر مترمربع بر روز و هم‌چنین مقادیر ضریب تبیین در دامنه ۰/۷ تا ۰/۹ متفاوت بود. بیشترین و کمترین مقدار میانگین قدر مطلق خطا به ترتیب در الگوریتم‌های رگرسیون خالص الاستیک (EN) و الگوریتم ماشین بردار پشتیبان (SVR) مشاهده گردید، در حالی‌که بیشترین و کمترین مقدار ضریب تبیین در الگوریتم‌های SVR و DT (الگوریتم درخت تصمیم‌گیری) حاصل گردید. به‌طور کلی و با توجه به نتایج هر سه معیار ارزیابی الگوریتم ماشین بردار پشتیبان بهترین نتیجه را در مرحله آزمون داده‌ها همانند مرحله آموزش نشان داد. مقایسه مقادیر اندازه‌گیری شده و برآورد شده تابش خورشیدی برای الگوریتم‌های به‌کار رفته در مطالعه در مرحله آزمون

در ایستگاه همدید یزد در شکل (۵) ارائه شده است. نتایج نشان داد، الگوریتم ماشین بردار پشتیبان با بالاترین ضریب تبیین (۰/۹۱۹) و کمترین مقدار میانگین قدر مطلق خطا (۰/۸۰۳) و میانگین متوسط مربعات خطا (۲/۸۵ مگا ژول بر مترمربع بر روز) بهترین نتیجه را نشان داد. الگوریتم‌های رگرسیون خطی (۰/۹۱۱)، جنگل تصادفی (۰/۹۱۳)، درختان اضافی (۰/۹۰۷) و ماشین تقویت گرادیان (۰/۹۰۷) نیز با مقادیر ضریب تبیین بالاتر از ۰/۹ نتایج مشابهی نشان دادند و الگوریتم رگرسیون خالص الاستیک با کمترین ضریب تبیین (۰/۷۵۱) در بین همه الگوریتم‌های مورد استفاده در پژوهش حاضر روش نامناسب برای تخمین تابش خورشیدی در ایستگاه همدید یزد معرفی گردید.

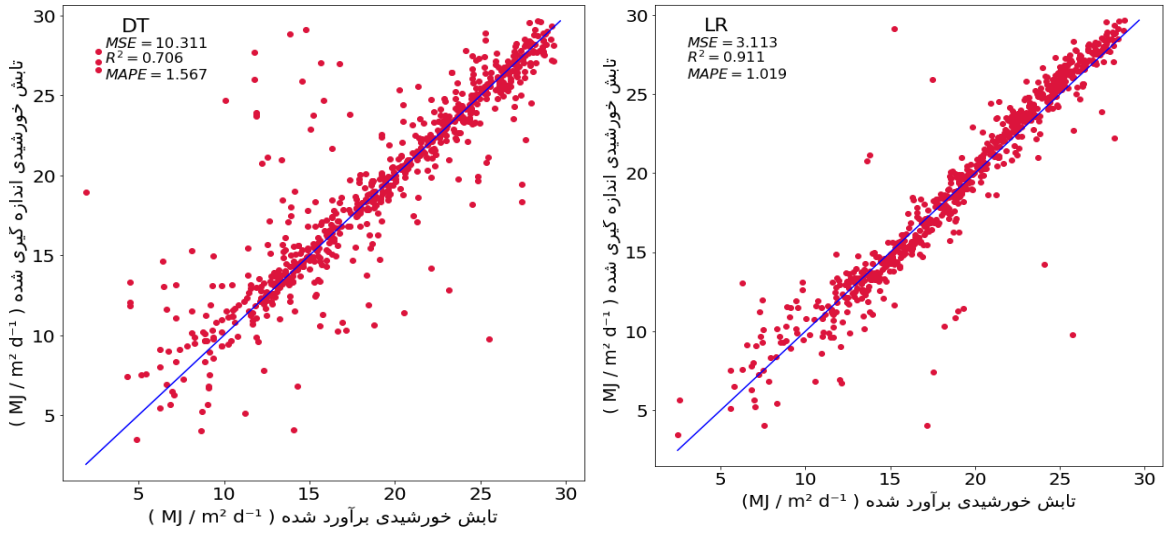


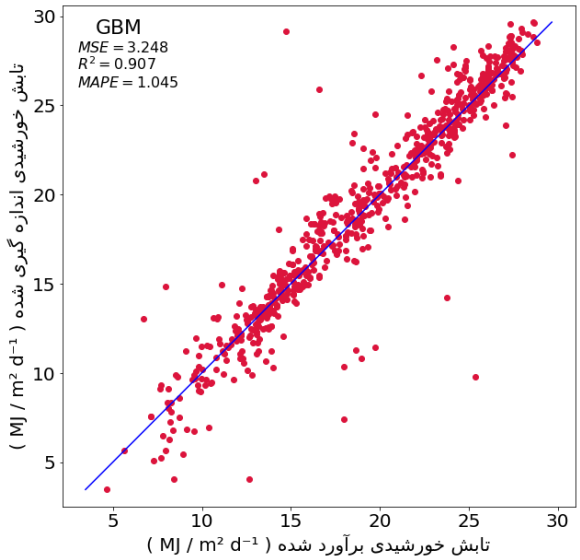
شکل ۴. مقایسه نتایج مرحله آزمون مدل‌سازی الگوریتم‌های مذکور با معیارهای ارزیابی.





شکل ۵. مقایسه مقادیر اندازه‌گیری شده و برآورد شده تابش خورشیدی برای الگوریتم‌های بکار رفته در مطالعه در مرحله آزمون.





ادامه شکل ۵. مقایسه مقادیر اندازه‌گیری شده و برآورد شده تابش خورشیدی برای الگوریتم‌های بکار رفته در مطالعه در مرحله آزمون.

۴ نتیجه‌گیری

با توجه به اهمیت تخمین درست تابش خورشیدی در پدیده‌های آب‌شناختی و لزوم استفاده از روش‌های نوین در برآورد آن، در پژوهش حاضر از نه الگوریتم یادگیری ماشین استفاده گردید. به‌طور کلی، از نتایج مدل‌سازی می‌توان نتیجه گرفت که مدل ماشین بردار پشتیبان با میانگین مربعات خطای ۲/۸۵ مگا ژول بر مترمربع بر روز، قدر مطلق خطای ۰/۸۰۳ و ضریب تبیین ۰/۹۱۹ در مرحله آزمون و به‌ترتیب ۱/۵۴ مگا ژول بر مترمربع بر روز، ۴/۹۲ و ۸۷۰/ در مرحله آموزش مدل‌ها نسبت به سایر مدل‌ها عملکرد بهتری در تخمین تابش خورشیدی داشته است. مرور تحقیقات گذشته و مقایسه آن با پژوهش حاضر در خصوص پیش‌بینی تابش خورشیدی با الگوریتم‌های مختلف نشان می‌دهد که معیارهای ارزیابی روش‌های مختلف متفاوت بوده و با توجه به مناطق مختلف از نظر میزان دریافت تابش خورشیدی نتایج متفاوتی به‌دست آمده و بهترین الگوریتم پیش‌بینی معرفی شده است. برای مثال چن و همکاران (۲۰۱۳a) روش ماشین بردار پشتیبان را نسبت به چندین روش تجربی، بهترین روش در تخمین تابش خورشیدی در منطقه‌ای در چین معرفی کردند. مورنو و همکاران (۲۰۱۱) در اسپانیا روش شبکه عصبی مصنوعی را در مقایسه با یک روش تجربی و رگرسیون کرنل روش برتر دانستند. محمدی و همکاران (۲۰۱۵) از بین روش‌های ماشین بردار پشتیبان، شبکه عصبی مصنوعی و الگوریتم ژنتیک در تخمین تابش خورشیدی روزانه و ماهانه در یک شهر ساحلی ایران، روش ماشین بردار پشتیبان را روش مناسب‌تری پیشنهاد دادند. کوئج و همکاران (۲۰۱۷) روش ماشین بردار پشتیبان را نسبت به روش‌های انفیس و شبکه عصبی مصنوعی در تخمین تابش خورشیدی در یک منطقه گرم نیمه مرطوب ارجح دانستند. اگبولوت و همکاران (۲۰۲۱) در چهار منطقه آب و هوایی مختلف در ترکیه تابش خورشیدی را با روش‌های شبکه عصبی مصنوعی، ماشین بردار پشتیبان، نزدیک‌ترین همسایگی کرنل و یادگیری عمیق برآورد کردند. نتایج آن‌ها حاکی از مناسب بودن روش شبکه عصبی مصنوعی نسبت به سایر روش‌ها بود. همان‌طور که مقایسه نتایج تحقیقات گذشته نشان می‌دهد، با توجه به شرایط آب و هوایی، طول دوره آماری، داده‌های گمشده و انتخاب داده‌های ورودی در الگوریتم‌های مورد استفاده

در پیش‌بینی تابش خورشیدی نتایج متفاوتی به‌دست می‌آید و همه روش‌های داده‌کاوی به‌کار رفته در مطالعات گذشته نتیجه یکسانی را نشان نمی‌دهد. هرچند همه این تحقیقات روش‌های داده‌کاوی را روش‌های مناسب برای تخمین تابش خورشیدی در همه مناطق نسبت به روش‌های تجربی معرفی کرده‌اند؛ همان‌گونه که نتایج این تحقیق نیز روش‌های داده‌کاوی را روش مؤثری در تخمین تابش خورشیدی خصوصاً در مناطقی که داده‌های تابش خورشیدی موجود نیست، پیشنهاد می‌دهد.

پیوست

الگوریتم‌های یادگیری ماشین به‌کار رفته در پژوهش

۱- روش رگرسیون خطی

این روش برای برازش یک مدل خطی با به حداقل رساندن مجموع مربعات خطا میان مقدار مشاهده شده و مقدار پیش‌بینی شده، به‌کار گرفته می‌شود. در رگرسیون خطی سعی می‌شود، به‌کمک معادله خطی که با استفاده از روش رگرسیون معرفی می‌شود، برآورد مقدار متغیر وابسته به ازای مقدارهای مختلف متغیر مستقل به‌دست آید. به‌منظور برآورد پارامترهای مناسب برای مدل، کوشش می‌شود بر اساس داده‌های موجود، مدلی انتخاب شود که کمترین خطا را داشته باشد (ضیافتی و ملکی، ۲۰۲۰).

۲- روش رگرسیون حداقل انقباض مطلق و عملگر انتخاب

این یک روش خطی است که مجموع مربعات خطا میان مقدار مشاهده شده و مقدار پیش‌بینی شده را به حداقل می‌رساند به شرط اینکه مجموع قدر مطلق ضرایب کمتر از یک ثابت باشد (کوکرجا و همکاران، ۲۰۰۶).

۳- روش رگرسیون خالص الاستیک

اگرچه برآوردگر حداقل مربعات دارای ویژگی‌های مطلوب، به‌خصوص نااریبی است، اما می‌تواند در برخی شرایط از مسئله بزرگ بودن واریانس رنج ببرد. برای مثال، اگر تعداد (متغیرهای توضیحی) در مقایسه با حجم نمونه زیاد باشد یا اگر تعداد زیادی متغیر همبسته وجود داشته باشد، در این صورت، برآوردهای حداقل مربعات نسبت به خطاهای تصادفی بسیار حساس هستند و ممکن است واریانس بالا و کارایی پایینی داشته باشند. یکی از راه‌حل‌هایی که برای این مسئله وجود دارد، استفاده از الگوهای رگرسیون خالص الاستیک است. این یک مدل رگرسیون خطی منظم است که هر دو تابع زیان نورم ۱ (L1) و نورم ۲ (L2) را ترکیب می‌کند. این الگوریتم انتخاب متغیر و منظم‌سازی را به‌طور هم‌زمان انجام می‌دهد (هان و همکاران، ۲۰۲۲).

۴-رگرسیون k نزدیک‌ترین همسایگان

این الگوریتم از شباهت ویژگی برای پیش‌بینی مقادیر هر نقطه داده جدید استفاده می‌کند، به این معنی که نقطه جدید بر اساس شباهت آن به نقاط مجموعه آموزشی، مقداری به آن اختصاص می‌دهد. این روش از فاصله اقلیدسی برای یافتن نزدیک‌ترین همسایگان به یک شی استفاده می‌کند. نزدیک‌ترین نقاط داده یا k بر اساس فاصله انتخاب می‌شوند. مقدار متوسط این نقاط داده، پیش‌بینی نهایی برای نقطه جدید است. این روش یکی از ساده‌ترین و قدیمی‌ترین روش‌های ناپارامتریک با رویکردهای طبقه‌بندی نظارت شده در میان الگوریتم‌های یادگیری ماشین شناخته می‌شود. (چن و همکاران، ۲۰۱۳b؛ هو و همکاران، ۲۰۱۶).

۵- رگرسیون درخت تصمیم

این یک الگوریتم تصمیم‌گیری است که از ساختار درختی روندنما مانند استفاده می‌کند. این الگوریتم ویژگی‌های یک شی را مشاهده می‌کند تا بتواند مدلی را در ساختار یک درخت برای پیش‌بینی داده‌ها در آینده آموزش دهد. با شروع از یک گره ریشه، یک درخت تصمیم با گره‌های تصمیم و گره‌ها برگ می‌سازد که از بالا به پایین، جستجوی

حریصانه در فضای شاخه‌های احتمالی و بدون پس‌گرد استفاده می‌کند. درخت تصمیم از یک گره ریشه از بالا به پایین ساخته می‌شود و شامل تقسیم داده‌ها به زیر مجموعه‌هایی است که حاوی نمونه‌هایی با مقادیر مشابه هستند (هان و همکاران، ۲۰۲۲).

۶- روش رگرسیون بردار پشتیبان

این روش با استفاده از کرنل‌ها، پراکندگی، کنترل حاشیه اطمینان و تعداد بردارهای پشتیبانی مشخص می‌شود. SVR از رگرسیون خطی و غیرخطی پشتیبانی می‌کند. یک کرنل باعث می‌شود بدون افزایش هزینه محاسبات، یک ابر صفحه را در فضای ابعاد بالاتر پیدا کنیم. این الگوریتم یک ابر صفحه یا مجموعه‌ای از ابرصفحه‌ها را در فضایی با ابعاد بالا یا حتی بی‌نهایت می‌سازد. دو خط در اطراف ابر صفحه با فاصله C کشیده شده است که برای ایجاد حاشیه بین نقاط داده استفاده می‌شود. این یک منطقه حساس به C متقارن را شناسایی می‌کند. انواع کرنل‌ها عبارت‌اند از: خطی، چند جمله، تابع پایه شعاعی و سیگموئید. (اونل و همکاران، ۲۰۱۸؛ کیم و همکاران، ۲۰۱۸).

۷- رگرسیون جنگل تصادفی

این روش یک الگوریتم یادگیری نظارت شده است که از روش یادگیری گروهی برای رگرسیون استفاده می‌کند و از پرکاربردترین الگوریتم‌های یادگیری ماشین محسوب می‌شود که با ساخت درخت‌های تصمیم‌گیری متعدد در طول زمان آموزش و تعیین خروجی نهایی به جای تکیه بر درخت‌های تصمیم‌گیری فردی عمل می‌کند. هر درخت با فن خود بهینه‌سازی ساخته می‌شود که نمونه‌برداری ردیفی را انجام می‌دهد و ویژگی‌هایی از نمونه‌ای از مجموعه داده را می‌سازد. خروجی نهایی میانگین تمام خروجی‌ها (تجمیع) است. جنگل‌های تصادفی برای درختان تصمیم که در مجموعه آموزشی دچار بیش‌برازش می‌شوند، مناسب هستند. عملکرد جنگل تصادفی معمولاً بهتر از درخت تصمیم است، اما این بهبود عملکرد تا حدی به نوع داده همبستگی دارد (شی و هروات، ۲۰۰۶).

۸- رگرسیون درختان اضافی

این یک الگوریتم یادگیری ماشینی است که پیش‌بینی‌های بسیاری از درخت‌های تصمیم را ترکیب می‌کند. این روش مشابه روش‌های دیگر، مانند درخت‌های تصمیم‌گیری و جنگل‌های تصادفی است، اما از داده‌های اضافی در مورد داده‌ها برای بهبود دقت پیش‌بینی استفاده می‌کند. این روش نتایج چندین درخت تصمیم‌گیری هم‌بسته جمع‌آوری شده در یک جنگل را در خروجی تجمیع می‌کند. از نظر انتخاب نقاط برش برای تقسیم گره‌ها، جنگل‌های تصادفی تقسیم بهینه را انتخاب می‌کنند، در حالی‌که درختان اضافی آن را به‌طور تصادفی انتخاب می‌کنند (گورتز و همکاران، ۲۰۰۶).

۹- ماشین تقویت گرادیان

این یک الگوریتم یادگیری گروهی است که با به حداقل رساندن گرادیان خطا، درخت‌های تصمیم تقویت‌شده را برازش می‌کند. مدل‌ها متناسب با هر تابع زیان متمایز دلخواه و الگوریتم بهینه‌سازی کاهش گرادیان ساخته می‌شوند. در تقویت گرادیان، هر یادگیرنده ضعیف جدید به جای اینکه بر روی نسخه وزن‌دار مجموعه آموزشی اولیه قرار بگیرد، مستقیماً بر روی خطاهای فعلی مدل قرار می‌گیرد تا اینکه درختان را بر اساس اشتباهات مدل تقویت کند (هان و همکاران، ۲۰۲۲).

منابع

جهان تیغ، ن.، پیری، ج. ۱۴۰۲. تخمین میزان تابش خورشیدی در اقلیم‌های مختلف ایران با استفاده از روش‌های هیبریدی یادگیری ماشین. نشریه علوم کاربردی و محاسباتی در مکانیک. ۳۴(۴): در حال چاپ.

سلطانی گردفرامرزی، س.، ر.، تقی زاده، م. قاسمی. ۱۳۹۴. برآورد ضریب پخشیدگی طولی رودخانه با استفاده از

انواع روش‌های داده کاوی. تحقیقات آب‌وخاک ایران. ۴۶(۳): ۳۸۵-۳۹۴.

سلطانی گردفرامرزی، س.، ر.، تقی زاده. ۱۳۹۵. کاربرد روش‌های داده کاوی در تخمین عمق آبشستگی گروه پایه‌ها. هیدرولیک، ۱۱(۱): ۶۷-۷۵.

سلطانی گردفرامرزی، س.، ۱۴۰۲. پیش‌بینی شدت تابش خورشیدی در ایستگاه یزد با به‌کارگیری مدل رگرسیونی مبتنی بر مولفه‌های اصلی (PCR). هواشناسی کشاورزی. در حال چاپ.

سیدیان، س. م.، فراستی، م.، روحانی، ح.، حشمت پور، ع. ۱۳۹۶. تخمین تابش خورشیدی با استفاده از پارامترهای هواشناسی. تحقیقات منابع آب ایران، ۱۳(۱): ۸۸-۱۰۰.

شیخ الاسلامی، ن.، قهرمان، ب.، مساعدی، ا.، داوری، ک.، مهاجرپور، م. 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.

محمدی، ب.، امامقلی زاده، ص. 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.

Abdelhafidi, N., Bachari, N.E.I. and Abdelhafidi, Z., 2021. Estimation of solar radiation using stepwise multiple linear regression with principal component analysis in Algeria. Meteorology and Atmospheric Physics, 133(2), pp.205-216.

Ağbulut, Ü., Gürel, A.E. and Biçen, Y., 2021. Prediction of daily global solar radiation using different machine learning algorithms: Evaluation and comparison. Renewable and Sustainable Energy Reviews, 135, p.110114.

Amiri, V., Kamrani, S., Ahmad, A., Bhattacharya, P. and Mansoori, J., 2021. Groundwater quality evaluation using Shannon information theory and human health risk assessment in Yazd province, central plateau of Iran. Environmental Science and Pollution Research, 28(1), pp.1108-1130.

Belmahdi, B., Louzazni, M. and El Bouardi, A., 2020. One month-ahead forecasting of mean daily global solar radiation using time series models. Optik, 219, p.165207.

Boroughani, M., Soltani, S., Ghezelseflu, N. and 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), pp.23-34.

Chen, J.L., Li, G.S. and Wu, S.J., 2013a. Assessing the potential of support vector machine for estimating daily solar radiation using sunshine duration. Energy conversion and management, 75, pp.311-318.

Chen, H.L., Huang, C.C., Yu, X.G., Xu, X., Sun, X., Wang, G. and Wang, S.J., 2013b. An efficient diagnosis system for detection of Parkinson's disease using fuzzy k-nearest neighbor approach. Expert systems with applications, 40(1), pp.263-271.

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

Geurts, P., Ernst, D. and Wehenkel, L., 2006. Extremely randomized trees. Machine learning, 63, pp.3-42.

Han, J., Kim, S.Y., Lee, J. and Lee, W.H., 2022. Brain Age Prediction: A Comparison between Machine Learning Models Using Brain Morphometric Data. Sensors, 22(20), p.8077.

Hunt, L.A., Kuchar, L., Swanton, C.J., 1998. Estimation of solar radiation for use in crop modeling. Agric. Meteorol. 91, 293–300.

Hu, L.Y., Huang, M.W., Ke, S.W. and Tsai, C.F., 2016. The distance function effect on k-nearest neighbor classification for medical datasets. SpringerPlus, 5(1), pp.1-9.

Kim, S., Mun, B.M. and Bae, S.J., 2018. Data depth based support vector machines for predicting corporate bankruptcy. Applied Intelligence, 48, pp.791-804.

Kukreja, S.L., Löfberg, J. and Brenner, M.J., 2006. A least absolute shrinkage and selection operator (LASSO) for nonlinear system identification. IFAC proceedings volumes, 39(1), pp.814-819.

Meenal, R. and Selvakumar, A.I., 2018. Assessment of SVM, empirical and ANN based solar radiation prediction models with most influencing input parameters. *Renewable Energy*, *121*, pp.324-343.

Mehdizadeh, S., Behmanesh, J. and Khalili, K.,

2016. Comparison of artificial intelligence methods and empirical equations to estimate daily solar radiation. Journal of Atmospheric and Solar-Terrestrial Physics, 146, pp.215-227.

Mohammadi, K., Shamshirband, S., Tong, C.W., Alam, K.A., Petkovic, D., 2015. Potential of adaptive neuro-fuzzy system for prediction of daily global solar radiation by day of the year. Energy Convers. Manag. 93, 406–413.

Moreno, A., Gilabert, M.A. and Martínez, B., 2011. Mapping daily global solar irradiation over Spain: A comparative study of selected approaches. Solar Energy, 85(9), pp.2072-2084.

Nwokolo, S.C., Obiwulu, A.U., Ogbulezie, J.C. and Amadi, S.O., 2022. Hybridization of statistical machine learning and numerical models for improving beam, diffuse and global solar radiation prediction. Cleaner Engineering and Technology, 9, p.100529.

Okundamiya MS, Emagbetere JO, Ogujor EA (2016) Evaluation of various global solar radiation models for Nigeria. Int J Green Energy 13(5):505–512.

Olalekan, S., Abdullahi, M. I. and Olabisi, A. (2018). Modeling of Solar Radiation Using Artificial Neural Network for Renewable Energy Application. Journal of Applied Physics, 10(2), 6-12.

Onel, M., Kieslich, C.A., Guzman, Y.A., Floudas, C.A. and Pistikopoulos, E.N., 2018. Big data approach to batch process monitoring: Simultaneous fault detection and diagnosis using nonlinear support vector machine-based feature selection. Computers & chemical engineering, 115, pp.46-63.

Quej, V.H., Almorox, J., Arnaldo, J.A. and Saito, L., 2017. ANFIS, SVM and ANN soft-computing techniques to estimate daily global solar radiation in a warm sub-humid environment. Journal of Atmospheric and Solar-Terrestrial Physics, 155, pp.62-70.

Radosevic, N., Duckham, M., Liu, G.J. and Sun, Q., 2020. Solar radiation modeling with KNIME and Solar Analyst: Increasing environmental model reproducibility using scientific workflows. Environmental Modelling & Software, 132, p.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.

Sabziparvar A.A., and Shetaee H. 2007. Estimation of global solar radiation in arid and semi-arid climates of East and West Iran, Energy 32: 649–655.

Seidian, S. M., Ferasati, M., Rouhani, H., Heshmatpour, A. 2016. Estimation of solar radiation using meteorological parameters. Iran Water Resources Research, 13(1): 88-100. (In Persian)

Sheikhul-Islami, N., Qahraman, B., Mosaedi, A., Davari, K., Mohajerpour, M. 2013. Forecasting reference plant evapotranspiration (ETO) using principal component analysis (PCA) method and development of multiple linear regression model (MLR-PCA) (case study: Mashhad station). Journal of Water and Soil, 28(2): 420-429. (In Persian)

Shi, T. and Horvath, S., 2006. Unsupervised learning with random forest predictors. Journal of Computational and Graphical Statistics, 15(1), pp.118-138.

Taki, M., Rohani, A. and Yildizhan, H., 2021. Application of machine learning for solar radiation modeling. Theoretical and Applied Climatology, 143(3-4), pp.1599-1613.

Yadav, A. K. and 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.

Zang, H., Cheng, L., Ding, T., Cheung, K.W., Wang, M., Wei, Z. and Sun, G., 2020. Application of functional deep belief network for estimating daily global solar radiation: A case study in China. Energy, 191, p.116502.

Zeng J, Qiao W (2013) Short-term solar power prediction using a support vector machine. Renewable Energy 52:118-127.

Ziafati, A. and Maleki, A., 2020. Fuzzy ensemble system for SSVEP stimulation frequency detection using the MLR and MsetCCA. Journal of Neuroscience Methods, 338, p.108686.

Iranian Journal of Geophysics, Vol 17 No 4, 2023, Page 2 (Research Article)
DOI: 10.30499/IJG.2023.393259.1512

Application of machine learning algorithms to estimate solar radiation (case study: arid and semi-arid climate)

Somayeh Soltani-Gerdefaramarzi [1,2*] and Hajar Momeni[3]

[1] *Associate Professor, Department of Water Engineering and Sciences, Faculty of Agriculture and Natural Resources, Ardakan University, Ardakan, Iran*
[2] *Water, Energy and Environment Research Institute, Ardakan University, P.O. Box 184, Ardakan, Iran*
[3] *Assistant Professor, Department of Electrical Engineering, Faculty of Engineering, Ardakan University, Ardakan, Iran*

(Received: 15 April 2023, Accepted: 14 May 2023)

Summary
Solar radiation is very important as one of the important variables in energy balance models and plant growth simulation. Although the measurement of this variable has a relatively long history in Iran, due to the high costs of measuring devices, there is no pyranometer in many existing stations in the country, and there are problems such as its recalibration, water, and dust accumulation that exists on the sensor. Even in meteorological 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 violations or other problems.

This research investigated the performance of nine machine learning algorithms including linear regression (LR), Least Absolute Shrinkage and Selection Operator (Lasso), Elastic Net (EN), K-Nearest Neighbors (kNN), Decision Tree (DT), Support Vector Regression (SVR), Random Forest (RF), Extra Trees (ET) and Gradient Boosting Machine (GBM) to estimate solar radiation in Yazd synoptic station between 2005 and 2021 with cross-validation method (kfold). The parameters of average temperature, minimum temperature, maximum temperature, sunny hours, relative humidity, and solar radiation are obtained from the National Meteorological Organization on a daily basis and extraterrestrial radiation variables, relative distance from the earth to the sun, solar inclination angle, and maximum sunny hours are calculated with existing relationships and were selected as input for the prediction models. The evaluation criteria for solar radiation estimation were MSE (Mean Square Error), MAPE (Mean Absolute Error), and determination Coefficient (R^2).

The results showed the coefficient of determination (R2) varies between 0.716 and 0.870 depending on the algorithm used in the training phase. In other words, in terms of the determination coefficient, all the used algorithms showed good results for predicting solar radiation. According to the results of all three criteria, it can be seen that the Support Vector Regression (SVR) algorithm has performed better than other algorithms. After the support vector regression (SVR) algorithm, the linear regression (LR) algorithm was ranked next with the MAPE of 5.04, the MSE of 1.13, and the R2 of 0.867. Also, the elastic pure regression algorithm (EN) with the highest mean absolute value of error (MAPE), the highest mean squared error (MSE), and the lowest coefficient of explanation (R^2) ranked last among the nine used algorithms. After the data training phase, using the K-fold cross-validation method, the remaining 20% of the data were tested. As the results show, the MSE changes for nine machine learning algorithms are high and vary from about 3 to 9 $Mj/m^2/day$. Its highest value was observed in the DT algorithm and its lowest value was observed in the support vector algorithm. The average value of the absolute value of the error was also in the range of 0.8 to 2.2 $Mj/m^2/day$, and also the values of the R^2 were different in the range of 0.7 to 0.9. In general, and according to the results of all three evaluation criteria, the support vector machine algorithm showed the best results in the data test stage as well as in the training stage.

Keywords: Geometric characteristics, data mining, solar inclination angle, radiation, algorithm, Yazd

*Corresponding author: ssoltani@ardakan.ac.ir