

ارزیابی دقت الگوریتم‌های مختلف یادگیری ماشین در پیش‌بینی تبخیر - تعرق محصول خیار گلخانه‌ای

مرتضی خوش‌سیمای چنار^۱ | حمیده نوری^۲ | عبدالمجید لیاقت^{۳*} | فروزنده سلطانی صالح‌آبادی^۴ | بابک متشع‌زاده^۵

۱. گروه مهندسی آبیاری و آبادانی، دانشکده کشاورزی، پردیس کشاورزی و منابع طبیعی، دانشگاه تهران، کرج، ایران. رایانامه: Khoshsimamortaza@ut.ac.ir

۲. گروه مهندسی آبیاری و آبادانی، دانشکده کشاورزی، پردیس کشاورزی و منابع طبیعی، دانشگاه تهران، کرج، ایران. رایانامه: hnoory@ut.ac.ir

۳. نویسنده مسئول، گروه مهندسی آبیاری و آبادانی، دانشکده کشاورزی، پردیس کشاورزی و منابع طبیعی، دانشگاه تهران، کرج، ایران. رایانامه: aliaghat@ut.ac.ir

۴. گروه علوم باغبانی و فضای سبز، دانشکده کشاورزی، پردیس کشاورزی و منابع طبیعی، دانشگاه تهران، کرج، ایران. رایانامه: soltanyf@ut.ac.ir

۵. گروه علوم و مهندسی خاک، دانشکده کشاورزی، پردیس کشاورزی و منابع طبیعی، دانشگاه تهران، کرج، ایران. رایانامه: moteshare@ut.ac.ir

چکیده

در این پژوهش، تبخیر - تعرق گیاه (ET_c) خیار گلخانه‌ای (رقم تالسیا) در شرایط کنترل‌شده گلخانه‌ای در دو دوره رشد پاییز - زمستان ۱۴۰۱ و بهار - تابستان ۱۴۰۲ واقع در پردیس کشاورزی و منابع طبیعی دانشگاه تهران، با هدف مدل‌سازی و پیش‌بینی ET_c مورد بررسی قرار گرفت. ET_c با پایش روزانه رطوبت خاک با استفاده از سنسورهای TDR نصب‌شده در عمق ۰ تا ۳۰ سانتی‌متر و به‌کارگیری معادله بیلان آب خاک و ET_c از طریق میکرولاسیمتر چمن با دقت دو گرم اندازه‌گیری شد. داده‌های ورودی مدل شامل ۱۰ ویژگی فصل، روز پس از انتقال نشاء، دمای حداقل، حداکثر و میانگین، رطوبت نسبی حداقل، حداکثر و میانگین، تابش خورشیدی و تبخیر - تعرق مرجع بود. همبستگی پی‌رسون نشان داد که متغیرهای روز پس از انتقال نشاء، میانگین دما و تابش بیشترین ارتباط مثبت با ET_c داشتند. برای پیش‌بینی ET_c ، شش الگوریتم یادگیری ماشین شامل رگرسیون مؤلفه‌های اصلی (PCR)، حداقل مربعات جزئی (PLS)، ماشین بردار پشتیبان (SVM)، جنگل تصادفی (RF)، گرادیان بوستینگ (GB) و اکستریم گرادیان بوستینگ (XGB) در زبان برنامه‌نویسی پایتون پیاده‌سازی شد. بهینه‌سازی هایپرپارامترها با بهره‌گیری از الگوریتم TPE کتابخانه Optuna با ۲۰۰ تکرار صورت گرفت. ارزیابی عملکرد مدل‌ها بر اساس شاخص‌های آماری R^2 ، RMSE، MAE و NSE با اعتبارسنجی متقابل پنج بخشی و سه تکرار انجام شد. نتایج مدل‌سازی نشان داد الگوریتم GB با میانگین R^2 ، RMSE، MAE و NSE به‌ترتیب برابر با ۰/۹۰، ۰/۵۹ میلی‌متر در روز، ۰/۴۱ میلی‌متر در روز و ۰/۸۹ بالاترین دقت و عملکرد را نسبت به سایر مدل‌ها دارد. به‌دنبال آن، الگوریتم‌های XGB، RF و SVM نیز با عملکرد نزدیک به GB قرار گرفتند و با آن اختلاف معنی‌دار نداشتند. تحلیل SHAP به‌عنوان ابزار تفسیر مدل نشان داد که ویژگی‌های روز پس از انتقال نشاء، تبخیر - تعرق مرجع و تابش بیشترین سهم را در مدل‌سازی ET_c دارند. در مجموع، نتایج این پژوهش نشان داد که الگوریتم‌های یادگیری ماشین مبتنی بر درخت تصمیم می‌توانند ابزار دقیقی برای پیش‌بینی نیاز آبی خیار گلخانه‌ای باشند و در مدیریت بهینه آبیاری و بهره‌وری منابع آب نقش بسزایی ایفا نمایند.

کلیدواژه‌ها: بیلان آب خاک، تبخیر - تعرق مرجع، تحلیل SHAP، الگوریتم XGBoost

۱. مقدمه

کشاورزی گلخانه‌ای نقش فزاینده‌ای در تأمین امنیت غذایی و کیفیت محصولات کشاورزی در کشاورزی مدرن ایفا می‌کند. گلخانه‌ها با فراهم آوردن محیط‌های کنترل‌شده، امکان تولید در تمام طول سال و مدیریت بهینه منابع را میسر می‌سازند. در کشور، تولیدات گلخانه‌ای به طور چشمگیری به اقتصاد محلی کمک کرده و کارایی کشت را بهبود بخشیده‌اند. بر اساس آمارنامه‌های وزارت جهاد کشاورزی، محصولات گلخانه‌ای در سال‌های اخیر ۹/۸-۱۱/۹ درصد از کل تولیدات باغی کشور را تشکیل داده است، همچنین محصول خیار (*Cucumis sativus* L.) به عنوان یکی از مهم‌ترین محصولات گلخانه‌ای در کشور بوده و حدود ۵۷ درصد از سطح زیر کشت گلخانه‌ای را به خود اختصاص داده است (Khoshsimae Chenar et al., 2025). یکی از جنبه‌های حیاتی در مدیریت کشت خیار گلخانه‌ای، برآورد دقیق تبخیر - تعرق گیاه (ET_c) است. ET_c که شامل مجموع تبخیر آب از سطح خاک (E) و تعرق از گیاه (T) است (Allen et al., 1998)، نقش مهمی در برنامه‌ریزی آبیاری و مدیریت کلی منابع آب دارد. پیش‌بینی دقیق ET_c امکان بهینه‌سازی آبیاری را فراهم می‌آورد که منجر به صرفه‌جویی در مصرف آب و کاهش اثرات زیست‌محیطی می‌شود. تولیدات گلخانه‌ای نیازمند برنامه‌ریزی دقیق آبیاری برای حفظ عملکرد بالا و کیفیت مطلوب محصول هستند (Incrocci et al., 2020). آبیاری بیش از حد می‌تواند منجر به کاهش عملکرد به دلیل شرایط بی‌هوای و آب‌گرفتگی خاک و همچنین از دست رفتن نیتروژن ناشی از دنتریفیکاسیون و آبشویی مواد مغذی شود. در مقابل، آبیاری ناکافی باعث تنش آبی شده و هر دو حالت بر بهره‌وری و کارایی منابع تاثیر منفی می‌گذارد (Zhang et al., 2025). این امر نشان‌دهنده نیاز به تعادل دقیق بوده و برآورد دقیق ET_c برای دستیابی به این تعادل بسیار مهم است. کشاورزان اغلب به عنوان یک اقدام احتیاطی بیش از حد آبیاری می‌کنند که این امر نشان‌دهنده فقدان روش‌های قابل اعتماد برای برآورد دقیق ET_c است. لایسیمترها ابزاری برای اندازه‌گیری مستقیم ET_c به شمار می‌روند، اما به کارگیری آن‌ها با چالش‌های متعددی همراه است. هزینه بالای نصب، فرآیند پیچیده راه‌اندازی و نگهداری، و همچنین محدودیت‌هایی نظیر ناهمگونی خاک، اثرات مرزی و تغییر رژیم‌های آبی، موجب می‌شود این دستگاه‌ها نتوانند شرایط واقعی مزرعه یا گلخانه را به طور کامل بازآفرینی کنند. در این میان، لایسیمترهای وزنی اگرچه دقت بالایی دارند، اما به دلیل هزینه‌های زیاد نصب و بهره‌برداری و نیز غیرقابل حمل بودن، امکان استفاده گسترده از آن‌ها وجود ندارد. بنابراین، محدودیت‌های عملی لایسیمترها مانعی جدی در به کارگیری آن‌ها برای پایش معمول ET_c در گلخانه‌های تجاری به شمار می‌آید (Shahrajabian & Sun, 2024). روش بیلان آب خاک، ET_c را بر اساس ورودی‌ها (آبیاری، بارندگی) و خروجی‌های (زهکشی، رواناب) آب در خاک تخمین می‌زند. ضعف این روش این است که اجزا بیلان اغلب تخمین زده می‌شوند و به طور مستقیم در منطقه ریشه اندازه‌گیری نمی‌شوند که منجر به عدم دقت می‌شود. خطاها در بیلان آب می‌توانند با گذشت زمان جمع شوند و قابلیت اطمینان تخمین‌های ET_c را کاهش دهند (Nikolaou et al., 2019). روش تشت تبخیر از میزان تبخیر از یک تشت آب برای تخمین ET_c با استفاده از ضرایب گیاهی و تشت استفاده می‌کند. ضریب تشت (K_p) می‌تواند بسته به شرایط محیطی و ویژگی‌های گلخانه متفاوت باشد و انتخاب مقدار مناسب را دشوار می‌سازد. تبخیر داخل گلخانه می‌تواند به دلیل پوشش پلاستیکی که بر تعادل تشعشع تاثیر می‌گذارد، به طور قابل توجهی با بیرون گلخانه متفاوت باشد. پدیده "پارادوکس تبخیر" که در آن تبخیر تشت با وجود افزایش دما کاهش می‌یابد، پیچیدگی استفاده از این روش برای تخمین ET_c در شرایط آب و هوایی متغیر را نشان می‌دهد. این نشان می‌دهد که تکیه صرف بر تشت تبخیر ممکن است نیاز واقعی آب محصول را به طور دقیق منعکس نکند. همچنین، مساحت زیاد مورد نیاز برای یک تشت تبخیر کلاس A می‌تواند در گلخانه‌های با فضای محدود یک عیب محسوب شود (Nikolaou et al., 2019). روش نسبت بوون^۱ بر اساس نسبت شار حرارت محسوس به شار حرارت نهان، ET_c را تخمین می‌زند. این روش زمانی که نسبت بوون به ۱- نزدیک می‌شود یا زمانی که انرژی در دسترس کم است، می‌تواند از نظر محاسباتی ناپایدار شود. اندازه‌گیری دقیق تشعشع خالص و شار حرارت زیرسطحی بسیار مهم است و خطا در این اندازه‌گیری‌ها می‌تواند به تخمین‌های ET_c منتقل شود. محیط گلخانه که در مقایسه با فضای باز یکنواخت‌تر و پایدارتر است، ممکن است به دلیل گرادین‌های بالقوه کوچک دما و فشار بخار، چالش‌هایی را برای روش نسبت بوون ایجاد کند. این روش به اختلافات قابل اندازه‌گیری در این گرادین‌ها متکی است که ممکن است در یک محیط کنترل‌شده کمتر مشهود باشد (Alfieri et al., 2009; Gallegos et al., 2013). روش‌های سنتی دیگر مانند اندازه‌گیری جریان شیره و روش‌های

آیروپنایمیکی نیز در محیط‌های گلخانه‌ای محدودیت‌هایی دارند و در عمل استفاده نمی‌شوند. اندازه‌گیری جریان شیره می‌تواند پیچیده باشد و ممکن است کل ET_c را به طور دقیق منعکس نکند. روش‌های آیروپنایمیکی می‌توانند تحت تأثیر سرعت کم و نزدیک به صفر باد که معمولاً در گلخانه‌ها وجود دارد، قرار گیرند (Jones, 2004).

مطالعاتی که از شبکه‌های عصبی مصنوعی (ANNs) برای پیش‌بینی ET_c در گلخانه‌ها استفاده کرده باشند، به طور خاص برای محصول خیار در منابع موجود یافت نشد. با این حال، مطالعات مرتبط بر روی سایر محصولات گلخانه‌ای مانند گوجه فرنگی و فلفل دلمه‌ای نشان‌دهنده پتانسیل این روش برای خیار نیز است. در این مطالعات، متغیرهای ورودی مورد استفاده شامل دما، رطوبت، تشعشع و سرعت باد بوده و معماری‌های مختلف ANN (مانند پرسپترون چندلایه، LSTM) به کار گرفته شده‌اند. برای مثال، مطالعه Assari et al. (2012) نشان داد که ANN می‌تواند با استفاده از پارامترهای هواشناسی مانند دما، تشعشع و فشار بخار، تبخیر - تعرق مرجع (ET_o) را در گلخانه‌ها با خطای کم تخمین بزند. مطالعه دیگری نشان داد که ANN در پیش‌بینی ET_c فلفل دلمه‌ای گلخانه‌ای با استفاده از دما، رطوبت، سرعت باد و تشعشع خورشیدی به عنوان ورودی، دقت قابل قبولی داشته است. تحلیل حساسیت نشان داد که حداکثر دما، میانگین رطوبت نسبی و تشعشع خالص، مهم‌ترین متغیرهای آب و هوایی برای تخمین ET_c با استفاده از ANN هستند (Pandorfi et al., 2016). همچنین، مدل ANN برای گوجه‌فرنگی گلخانه‌ای در مقایسه با مدل‌های پنمن - مانتیث، هارگریوز - سامانی و تشت تبخیر، دقت محاسباتی به مراتب بالاتری را در تخمین ET_c نشان داد (Shi et al., 2023). اگرچه مطالعه خاصی برای محصول خیار با استفاده از ANN در منابع موجود یافت نشد، کاربرد موفقیت‌آمیز ANN برای پیش‌بینی ET_c در سایر محصولات گلخانه‌ای نشان‌دهنده پتانسیل آن برای محصول خیار گلخانه‌ای نیز می‌باشد. نکته کلیدی، آموزش مدل با داده‌های خاص محصول خیار و متغیرهای محیطی مرتبط خواهد بود. فرآیندهای فیزیولوژیکی حاکم بر تعرق در بسیاری از گونه‌های گیاهی مشابه است، اگرچه میزان آن ممکن است متفاوت باشد. بنابراین، مدل‌های آموزش‌دیده بر روی محصولات مرتبط می‌توانند نقطه شروع خوب یا نشان‌دهنده امکان‌پذیر بودن این رویکرد باشند.

الگوریتم ماشین بردار پشتیبان (SVM) نیز برای تخمین ET_c در محیط‌های کنترل‌شده مورد استفاده قرار گرفته‌اند. مطالعات مرتبط بر روی سایر محصولات گلخانه‌ای مانند گوجه‌فرنگی نشان داده‌اند که SVM با استفاده از متغیرهای ورودی مشابه ANN، عملکرد خوبی در پیش‌بینی ET_c دارند. Eslamian et al. (2009) در مطالعه خود به این نتیجه رسیدند که مدل SVM در مدل‌سازی ET_c در یک محیط گلخانه‌ای کنترل‌شده با استفاده از متغیرهای هواشناسی، برتر از ANN عمل کرده است. SVM به دلیل توانایی در مدیریت داده‌های غیرخطی و جلوگیری از بیش‌برازش شناخته می‌شود که می‌تواند در مدل‌سازی تعاملات پیچیده بیولوژیکی و محیطی مفید باشد. این امر این الگوریتم را به یک نامزد قوی برای پیش‌بینی دقیق ET_c در گلخانه‌های خیار نیز تبدیل می‌کند.

الگوریتم‌های جنگل تصادفی (RF) نیز برای پیش‌بینی ET_c در گلخانه‌ها به کار گرفته شده‌اند. تحقیقات مرتبط بر روی سایر محصولات گلخانه‌ای مانند گوجه فرنگی نشان می‌دهد که RF دقت محاسباتی بالایی در تخمین ET_c در گلخانه‌ها دارند. مدل RF در صورت وجود پارامترهای زمانی (استفاده از شاخص‌های نشان‌دهنده زمان مانند روز پس از انتقال نشاء یا روز ژولایوسی) در مجموعه داده، دقت چشمگیری در پیش‌بینی ET_c نشان داده‌اند. الگوریتم XGBoost (که به عنوان یک روش مبتنی بر درخت‌های گروهی مرتبط با RF است) دقت بالایی در مدل‌سازی ET_c روزانه برای گوجه فرنگی گلخانه‌ای نشان داد و تشعشع خورشیدی به عنوان مهم‌ترین عامل شناخته شد (Ge et al., 2022). RF که در برابر داده‌های پرت مقاوم بوده و قادر به مدیریت داده‌های با ابعاد بالا هستند، به نظر می‌رسد ابزار قدرتمندی برای پیش‌بینی ET_c در گلخانه‌ها، از جمله به طور بالقوه برای خیار، باشند. اهمیت تشعشع خورشیدی که در مطالعات گوجه فرنگی مشخص شده، احتمالاً برای خیار نیز مرتبط است. روش‌های گروهی مانند RF اغلب در مقایسه با مدل‌های منفرد، تعمیم و دقت پیش‌بینی بهتری ارائه می‌دهند و آن‌ها را برای سیستم‌های پیچیده کشاورزی مناسب می‌سازند.

مدل‌های رگرسیونی مختلف (خطی، چندگانه خطی و غیرخطی) نیز برای تخمین ET_c گلخانه‌ای به کار گرفته شده‌اند. مطالعات نشان داده‌اند که این مدل‌ها می‌توانند با استفاده از متغیرهای محیطی و پارامترهای محصول، تخمین‌های قابل قبولی از ET_c ارائه دهند. برای مثال، مدل‌های رگرسیونی چندگانه برای تخمین میزان مصرف آب روزانه گیاهان گوجه‌فرنگی گلخانه‌ای بر اساس متغیرهای

آب و هوایی و وزن تر گیاه استفاده شده و مقادیر R^2 تعدیل شده بالایی به دست آمده است (JuarezMaldonado et al., 2014). تجزیه و تحلیل رگرسیونی ارتباط بالایی بین ضرایب گیاهی و شاخص سطح برگ نشان داد ($K_e = 0.447 \times LAI$) که نشان می‌دهد ET_c را می‌توان بر اساس این رابطه مدل‌سازی کرد (Nikolaou et al., 2023). در حالی که مدل‌های رگرسیونی ساده‌تر از سایر تکنیک‌های یادگیری ماشین هستند، همچنان می‌توانند بینش‌های ارزشمندی در مورد روابط بین ET_c و عوامل مختلف مؤثر در محیط‌های گلخانه‌ای ارائه دهند. دقت آن‌ها ممکن است کمتر از مدل‌های پیشرفته‌تر یادگیری ماشین باشد، اما تفسیر و پیاده‌سازی آن‌ها می‌تواند آسان‌تر باشد. مدل‌های رگرسیونی می‌توانند به شناسایی عوامل کلیدی مؤثر بر ET_c در محصول خیار گلخانه‌ای کمک کرده و یک مبنا برای مقایسه با مدل‌های پیچیده‌تر فراهم کنند.

روش‌های سنتی اندازه‌گیری ET_c در محیط‌های گلخانه‌ای با چالش‌هایی روبرو هستند. این روش‌ها می‌توانند پرهزینه، زمان‌بر و پیچیده باشند و احتمال بروز خطا در آن‌ها وجود دارد. همچنین، ریزاقلیم منحصر به فرد داخل گلخانه‌ها (مانند سرعت کم باد، دما و رطوبت کنترل شده) می‌تواند بر دقت مدل‌های استاندارد ET_c تاثیر بگذارد. یادگیری ماشین به عنوان یک راه‌حل بالقوه برای غلبه بر محدودیت‌های روش‌های سنتی مطرح شده و رویکردهای مبتنی بر داده را برای برآورد دقیق‌تر و قابل اعتمادتر ET_c در کشت گلخانه‌ای ارائه می‌دهد. الگوریتم‌های یادگیری ماشین قادر به درک روابط پیچیده و غیرخطی بین متغیرهای محیطی و ET_c هستند. مطالعات متعددی نشان داده‌اند که استفاده از مدل‌های یادگیری ماشین در مقایسه با روش‌های سنتی، دقت پیش‌بینی ET_c را بهبود بخشیده است. هدف از این مطالعه، بررسی و تحلیل کاربرد و دقت مدل‌های مختلف یادگیری ماشین در برآورد و پیش‌بینی ET_c برای محصول خیار گلخانه‌ای به عنوان یک محصول اصلی گلخانه‌های کشور است.

۲. مواد و روش‌ها

۱.۲. اندازه‌گیری تبخیر - ترق مرجع و گیاه در گلخانه

این پژوهش با هدف تعیین و مدل‌سازی ET_c خیار گلخانه‌ای (رقم تالسیا) در دو دوره رشد پاییز - زمستان ۱۴۰۱ (A-W) و بهار - تابستان ۱۴۰۲ (S-S)، در گلخانه تحقیقاتی پردیس کشاورزی و منابع طبیعی دانشگاه تهران، واقع در شهرستان کرج، اجرا گردید. گلخانه مورد استفاده دارای مساحتی معادل ۸۰ مترمربع با جهت‌گیری شمالی - جنوبی بود. پوشش سقف آن از جنس پلی‌کربنات و دیواره‌ها از شیشه بودند. طرح آزمایشی شامل سه کرت به ابعاد ۸ × ۱ متر بود که در هر کرت دو ردیف کاشت با فاصله میان‌ردیفی ۵۰ سانتی‌متر در نظر گرفته شد. فاصله بین بوته‌های خیار در هر ردیف ۴۰ سانتی‌متر تنظیم گردید. دوره رشد A-W از تاریخ ۲۷ مهر ماه ۱۴۰۱ تا ۳۰ دی ماه همان سال به مدت ۹۶ روز، و دوره رشد S-S از ۹ فروردین ماه تا ۱۰ مرداد ماه ۱۴۰۲ به مدت ۱۲۶ روز، از زمان انتقال نشاء به گلخانه، به طول انجامید.

به‌منظور تعیین ET_c گیاه خیار گلخانه‌ای رطوبت خاک به‌صورت روزانه با استفاده از سنسورهای [TDR](#) نصب‌شده در لایه ۰ تا ۳۰ سانتی‌متری خاک اندازه‌گیری شد. میانگین داده‌های رطوبتی حاصل از سه کرت آزمایشی، در کنار عمق ریشه در طول دوره رشد، مبنای محاسبه ET_c بر اساس بیلان آب خاک قرار گرفت (تغییرات رطوبت × عمق ریشه). این محاسبات مطابق رابطه ۱ شامل اجزای آب آبیاری (I)، بارندگی (P)، صعود موئینه (Cr)، نفوذ عمقی (Dp)، رواناب (Rf) و تغییرات رطوبت خاک (ΔS) در یک بازه زمانی مشخص در ناحیه ریشه گیاه است (James, 1988).

$$ET_c = I + P + Cr - Dp - Rf \pm \Delta S \quad \text{رابطه ۱}$$

با توجه به شرایط گلخانه‌ای که بارندگی در آن وجود ندارد و همچنین بهره‌گیری از سیستم آبیاری قطره‌ای، مؤلفه‌های Cr، Rf و Dp صفر در نظر گرفته شده و معادله به شکل ساده‌شده ۲ تبدیل می‌شود:

$$ET_c = I \pm \Delta S \quad \text{رابطه ۲}$$

برای تعیین عمق ریشه گیاه در طول دوره رشد، از مدل سینوسی طبق رابطه ۳ استفاده شد (Borg & Grimes, 1986; Merrill et al., 2002).

$$Z_r = Z_{r,m} \left(0.5 + 0.5 \times \sin \left[3.03 \left(\frac{DAP}{DTM} \right) - 1.47 \right] \right) \quad \text{رابطه ۳}$$

که در آن Z_r عمق ریشه، $Z_{r,m}$ حداکثر عمق ریشه، DAP روزهای پس از کاشت و DTM تعداد روزهایی است که گیاه به عمق ریشه نهایی می‌رسد. علی‌رغم اینکه برخی منابع حداکثر عمق ریشه خیار گلخانه‌ای را تا یک متر نیز گزارش کرده‌اند، با توجه به خصوصیات خاک بستر در این مطالعه، مقدار $Z_{r,m}$ برابر ۳۰ سانتی‌متر و DTM برابر ۳۵ روز در نظر گرفته شد (Rezvani et al., 2022). همچنین، در طول هر دو دوره رشد، برای اندازه‌گیری روزانه تبخیر - تعرق مرجع (ET_0)، از میکرو لایسیمتر چمن (ترازویی با دقت دو گرم و ظرفیت ۴۰ کیلوگرم) استفاده شد. فرآیند اندازه‌گیری ET_0 پس از آن آغاز گردید که ارتفاع چمن در لایسیمتر به ۱۲ سانتی‌متر رسید. در طول دوره مطالعه، ارتفاع چمن به‌صورت مستمر در بازه ۱۱ تا ۱۳ سانتی‌متر حفظ شد تا شرایط مرجع استاندارد تبخیر - تعرق مطابق دستورالعمل‌های FAO فراهم شود (Allen et al., 1998). به‌منظور جلوگیری از بروز تنش آبی و اطمینان از تداوم شرایط ET_0 ، رطوبت خاک داخل لایسیمتر همواره در محدوده ظرفیت زراعی و شرایط مطلوب نگهداری شد.

۲.۲. الگوریتم‌های یادگیری ماشین

۳.۲. رگرسیون مؤلفه‌های اصلی

رگرسیون مؤلفه‌های اصلی^۲ (PCR) یک روش غیرنظارتی برای کاهش ابعاد داده‌هاست که در گام نخست با استفاده از تحلیل مؤلفه‌های اصلی (PCA) ساختار داده را به مجموعه‌ای از مؤلفه‌های جدید و مستقل تبدیل کرده و در گام دوم، مدل رگرسیون خطی بر اساس این مؤلفه‌های اصلی ساخته می‌شود (Jolliffe, 1986). در این مطالعه، تحلیل مؤلفه‌های اصلی با استفاده از تابع PCA از ماژول sklearn.decomposition و مدل‌سازی رگرسیون خطی با استفاده از تابع LinearRegression از ماژول sklearn.linear_model انجام شد. پیش از انجام PCA، داده‌های با بهره‌گیری از تابع StandardScaler از ماژول sklearn.preprocessing نرمال‌سازی شدند (Kramer, 2016).

۱.۳.۲. رگرسیون حداقل مربعات جزئی

رگرسیون حداقل مربعات جزئی^۳ (PLS) یک تکنیک نظارت‌شده برای کاهش ابعاد است که متغیرهای پیش‌بین را به مجموعه‌ای کوچکتر از مؤلفه‌های بدون همبستگی تقلیل داده و سپس رگرسیون حداقل مربعات را بر روی این مؤلفه‌ها اجرا می‌کند. برخلاف PCA که مؤلفه‌هایی را انتخاب می‌کند که بیشترین واریانس موجود در متغیرهای پیش‌بین را توصیف می‌کنند، روش PLS مؤلفه‌هایی را می‌یابد که ضمن حداکثرسازی واریانس در متغیرهای پیش‌بین، بیشترین همبستگی ممکن با متغیر پاسخ را نیز حفظ می‌کنند (Kuhn & Johnson, 2013). در این مطالعه، PLS با استفاده از تابع PLSRegression از ماژول sklearn.cross_decomposition پیاده‌سازی شد. پیش از اعمال مدل، داده‌ها با استفاده از تابع StandardScaler از ماژول sklearn.preprocessing نرمال‌سازی شدند تا مقیاس متغیرها یکسان گردد (Kramer, 2016). این روش با هدف کاهش ابعاد و حفظ بیشترین همبستگی بین مؤلفه‌های استخراج‌شده و متغیر پاسخ به کار گرفته شد.

۲.۳.۲. ماشین بردار پشتیبان

الگوریتم ماشین بردار پشتیبان (SVM) که توسط Vapnik (1999) توسعه‌یافته، یک مدل یادگیری ماشین نظارت‌شده برای شناسایی الگوها و تحلیل داده‌ها است که به‌طور گسترده‌ای در حوزه‌های مختلفی همچون کشاورزی، هیدرولوژی، هواشناسی و مطالعات محیط‌زیست برای مقاصد رگرسیون و پیش‌بینی به‌کار رفته است (Fan et al., 2018; Ghorbani et al., 2017; Shrestha & Shukla, 2015). مدل SVM با استفاده از مجموعه‌ای از توابع هسته‌ای^۴ رگرسیون را برآورد می‌کند. این توابع قابلیت تبدیل داده‌های ورودی با ابعاد پایین به فضای ویژگی با ابعاد بالاتر را به‌صورت ضمنی دارند. همین ویژگی، امکان مدل‌سازی روابط غیرخطی بین متغیرها را فراهم می‌سازد. مدل SVM در پیش‌بینی ET_0 نیز مورد توجه قرار گرفته و عملکرد موفقی را در مطالعات متعدد از خود نشان داده است (Mehdizadeh et al., 2017; Wen et al., 2015; Yin et al., 2017). جزئیات بیشتر در مورد مبانی نظری و فرآیند محاسباتی مدل SVM در Vapnik (1999) ارائه شده است. در این مطالعه از تابع SVR از ماژول sklearn.svm استفاده شد (Kramer, 2016).

۳.۳.۲. جنگل تصادفی

الگوریتم جنگل تصادفی^۵ (RF)، یک روش ensemble یا تجمیعی (گروهی) مبتنی بر یادگیری درختی است که مجموعه‌ای از درختان تصمیم‌گیری را ایجاد کرده و با رأی‌گیری (برای مسائل طبقه‌بندی) یا میانگین‌گیری (برای مسائل رگرسیون) خروجی نهایی را تولید می‌کند (Breiman, 2001). در RF، هر درخت با استفاده از یک نمونه تصادفی^۶ از داده‌ها آموزش داده می‌شود و در هر گره، تنها بخشی از ویژگی‌ها برای تقسیم داده‌ها در نظر گرفته می‌شود. این ساختار منجر به افزایش تعمیم‌پذیری و جلوگیری از بیش‌برازش می‌شود. از مزایای RF می‌توان به پایداری بالا، مقاومت در برابر داده‌های پرت، و قابلیت محاسبه اهمیت ویژگی‌ها^۷ اشاره کرد (Dhaliwal & Williams, 2024). در این مطالعه از تابع RandomForestRegressor از ماژول sklearn.ensemble استفاده شد (Kramer, 2016).

۴.۳.۲. گرادیان بوستینگ

الگوریتم گرادیان بوستینگ^۸ (GB) یک روش تقویتی^۹ است که در آن مدل‌های ضعیف (معمولاً درخت‌های تصمیم‌گیری کم‌عمق) به‌صورت تدریجی آموزش داده می‌شوند و هر مدل جدید تلاش می‌کند تا خطای مدل‌های قبلی را کاهش دهد (Friedman, 2001). در واقع، مدل نهایی به‌صورت ترکیبی وزن‌دار از مدل‌های ضعیف است که به‌مرور با استفاده از گرادیان تابع هزینه^{۱۰} بهبود یافته‌اند. علی‌رغم دقت بالا، یکی از محدودیت‌های GB حساسیت به نویز و نیاز به تنظیم دقیق پارامترها است که ممکن است منجر به بیش‌برازش شود (Natekin & Knoll, 2013). در این مطالعه از تابع GradientBoostingRegressor از ماژول sklearn.ensemble استفاده شد (Kramer, 2016).

۴.۳.۲. اکستریم گرادیان بوستینگ

اکستریم گرادیان بوستینگ^{۱۱} یا XGBoost (XGB) نسخه‌ای بهینه‌شده و توسعه‌یافته از GB است که به‌واسطه سرعت بالا، پایداری بهتر و قابلیت تنظیم پیشرفته‌تر پارامترها، در بسیاری از رقابت‌های علمی و صنعتی مورد استفاده قرار گرفته است. این الگوریتم از تکنیک‌هایی مانند منظم‌سازی^{۱۲}، هرس^{۱۳} پیشرفته درخت‌ها و پردازش موازی بهره می‌برد که باعث افزایش دقت و کاهش احتمال بیش‌برازش می‌شود. XGBoost همچنین امکان مدیریت مقادیر گم‌شده و استفاده از توقف زود هنگام^{۱۴} را برای جلوگیری از یادگیری بیش از حد فراهم می‌کند (Chen & Guestrin, 2016). در این مطالعه از کلاس XGBRegressor از کتابخانه XGBoost استفاده شد.

۴.۲. مراحل مدل‌سازی

برای مدل‌سازی ET_c محصول خیار در شرایط گلخانه‌ای، شش الگوریتم یادگیری ماشین شامل RF، GB، XGB، SVM، PLS و PCR مورد ارزیابی قرار گرفتند. مراحل مدل‌سازی در ادامه ارائه شده است.

۱.۴.۲. آماده‌سازی داده‌ها

ابتدا داده‌های مربوط به پارامترهای هواشناسی اندازه‌گیری شده داخل گلخانه، ET_o و ET_c محصول خیار از فایل Excel بارگذاری شد. داده‌های ورودی مدل شامل ویژگی‌هایی نظیر فصل (season)، روز پس از انتقال نشاء (DAT)، دمای حداقل، حداکثر و میانگین (minT، maxT، meanT) به درجه سلسیوس، درصد رطوبت نسبی حداقل، حداکثر و حداقل (minRH، maxRH، meanRH)، تابش خورشیدی (Radiation) به مگاژول بر مترمربع در روز و تبخیر - تعرق مرجع (ET_o) به میلی‌متر در روز بودند. حداقل و حداکثر دما و رطوبت نسبی داخل گلخانه به‌طور روزانه و در ساعت هفت و نیم صبح با استفاده از سنسور دیجیتال مدل HTC-1 اندازه‌گیری و ثبت شد (Rahimikhoob et al., 2020). مقدار تابش خورشیدی داخل گلخانه نیز بر اساس ضریب انتقال پوشش گلخانه (با توجه به پوشش گلخانه ۰/۸۵ در نظر گرفته شد) برآورد گردید (Fernández et al., 2010; Incrocci et al., 2020). بدین‌منظور، داده‌های تابش خورشیدی خارج از گلخانه از ایستگاه هواشناسی مستقر در فاصله ۳/۶ کیلومتری و در مزرعه تحقیقاتی پردیس کشاورزی و منابع طبیعی دانشگاه تهران اخذ شد و سپس با اعمال ضریب انتقال به‌عنوان تابش داخلی گلخانه در نظر گرفته شد. متغیر هدف نیز ET_c خیار گلخانه‌ای به میلی‌متر در روز تعریف شد. نمودار جعبه‌ای پراکندگی داده‌های ورودی مدل و متغیر هدف در شکل ۱ ارائه شده است.

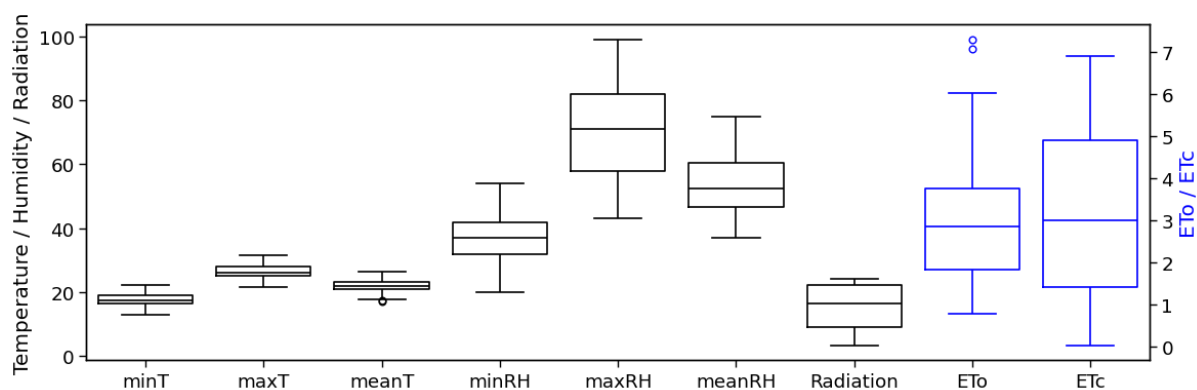


Figure 1. Distribution of climatic data including temperature, relative humidity, solar radiation, ET_0 and ET_c inside the greenhouse

شکل ۲ ماتریس همبستگی پیرسون بین متغیرهای ورودی و خروجی در فرآیند مدل‌سازی ET_c خیار گلخانه‌ای را نشان می‌دهد. همبستگی پیرسون ضریبی عددی بین ۱- تا ۱ است که نشان‌دهنده شدت و جهت رابطه خطی بین دو متغیر می‌باشد. مقدار ۱ نشان‌دهنده همبستگی کاملاً مثبت، مقدار ۱- همبستگی کاملاً منفی و مقدار ۰ بیانگر نبود رابطه خطی بین دو متغیر است. با توجه به نتایج ارائه شده در شکل ۲ مشاهده می‌شود، ET_c بیشترین همبستگی مثبت را به ترتیب با $meanT$ ، DAT و $Radiation$ به ترتیب به مقدار ۰/۸۶، ۰/۵۶ و ۰/۴۵ داشته است. این یافته‌ها نشان می‌دهند که دما، تابش و روز پس از انتقال نشاء از مهم‌ترین عوامل تأثیرگذار بر ET_c در شرایط گلخانه‌ای مورد بررسی هستند. از سوی دیگر، تابش خورشیدی همبستگی منفی نسبتاً قوی با فصل (۰/۸۹-) دارد که بیانگر تفاوت قابل توجه شدت تابش بین فصول (پاییز - زمستان و بهار - تابستان) مختلف است. همچنین برخی متغیرها مانند $minRH$ همبستگی ضعیفی با ET_c نشان داده‌اند که احتمالاً نقش کمتری در مدل‌سازی این متغیر هدف دارند. استفاده از این تحلیل همبستگی به شناسایی متغیرهای کلیدی و کاهش ابعاد در مراحل پیش‌پردازش و انتخاب ویژگی کمک می‌کند.

season	1.00	0.22	0.28	0.46	0.45	-0.05	0.57	0.43	0.89	0.77	0.38
DAT	0.22	1.00	0.48	0.32	0.49	0.30	0.02	0.14	0.28	0.11	0.86
minT	0.28	0.48	1.00	0.35	0.83	0.01	-0.03	-0.02	0.33	0.22	0.54
maxT	0.46	0.32	0.35	1.00	0.81	-0.13	0.37	0.24	0.58	0.47	0.37
meanT	0.45	0.49	0.83	0.81	1.00	-0.07	0.20	0.13	0.55	0.42	0.56
minRH	-0.05	0.30	0.01	-0.13	-0.07	1.00	0.31	0.67	-0.07	-0.06	0.31
maxRH	0.57	0.02	-0.03	0.37	0.20	0.31	1.00	0.92	0.60	0.51	0.20
meanRH	0.43	0.14	-0.02	0.24	0.13	0.67	0.92	1.00	0.44	0.38	0.29
Radiation	0.89	0.28	0.33	0.58	0.55	-0.07	0.60	0.44	1.00	0.81	0.45
ET ₀	0.77	0.11	0.22	0.47	0.42	-0.06	0.51	0.38	0.81	1.00	0.32
ET _c	0.38	0.86	0.54	0.37	0.56	0.31	0.20	0.29	0.45	0.32	1.00
season	DAT	minT	maxT	meanT	minRH	maxRH	meanRH	Radiation	ET ₀	ET _c	

Figure 2. Pearson correlation matrix between input and output variables in the modeling process of greenhouse cucumber ET_c

۲.۴.۲. بهینه‌سازی هایپر پارامترها

به منظور افزایش دقت مدل‌ها و جلوگیری از بیش‌برازش، بهینه‌سازی هایپر پارامترها^{۱۵} برای تمامی الگوریتم‌ها با استفاده از کتابخانه Optuna ورژن ۴.۲.۱ در زبان برنامه‌نویسی پایتون انجام شد (Akiba et al., 2019). برای هر مدل، تابع هدف جداگانه‌ای تعریف شده که در آن مقادیر مختلف هایپر پارامترها توسط الگوریتم TPE^{۱۶} به صورت خودکار پیشنهاد شده و مدل با این تنظیمات بر روی مجموعه داده آموزش دیده و ارزیابی شد. هایپر پارامترهای انتخابی هر مدل برای بهینه‌سازی در جدول ۱ ارائه شده است. معیار اصلی برای ارزیابی

عملکرد مدل‌ها در فرآیند بهینه‌سازی، مجذور میانگین مربعات خطا (RMSE) با استفاده از اعتبارسنجی متقابل^{۱۷} با پنج بخش^{۱۸} و سه تکرار بود. ۲۰۰ تکرار برای فرآیند بهینه‌سازی در نظر گرفته شد.

Table 1. Selected hyperparameters for optimization using the Optuna library

Hyperparameters	Algorithm
n_estimators, max_depth, min_samples_split, min_samples_leaf, max_features	RF
n_estimators, max_depth, min_samples_split, min_samples_leaf, learning_rate, subsample	GB
n_estimators, max_depth, colsample_bytree, gamma, reg_alpha, reg_lambda, learning_rate, subsample	XGB
kernel, C, epsilon, gamma	SVM
n_components	PLS
n_components	PCR

۳.۴.۲. ارزیابی عملکرد مدل‌ها

پس از انتخاب بهترین هایپرپارامترها برای هر مدل، مدل نهایی آموزش داده شد. برای ارزیابی مدل دو روش وجود دارد. روش K-Fold Cross-Validation، مجموعه داده به K زیرمجموعه هم‌اندازه تقسیم می‌شود. در هر تکرار، یکی از این زیرمجموعه‌ها به عنوان داده‌های اعتبارسنجی (تست) در نظر گرفته شده و سایر زیرمجموعه‌ها برای آموزش مدل استفاده می‌شوند. این فرآیند به تعداد K بار تکرار شده و در پایان، عملکرد نهایی مدل به صورت میانگین نتایج حاصل از تمامی تکرارها گزارش می‌گردد. مزیت اصلی این روش، استفاده بهینه از تمامی داده‌ها برای هر دو نقش آموزش و ارزیابی است. با این حال، از آن‌جا که تقسیم اولیه داده‌ها تنها یک بار انجام می‌شود، نتایج ممکن است به نوع تقسیم‌بندی خاص داده‌ها وابسته باشد و این امر می‌تواند سبب ناپایداری در ارزیابی گردد. برای رفع این محدودیت، روش Repeated K-Fold Cross-Validation معرفی شده است. در این روش، فرآیند K-Fold چندین بار با تقسیم‌بندی‌های تصادفی و متفاوت بر روی داده‌ها تکرار می‌شود. به عبارت دیگر، داده‌ها در هر بار تکرار مجدداً به K بخش تقسیم می‌گردند و کل فرآیند آموزش و اعتبارسنجی دوباره انجام می‌شود. در نهایت، میانگین نتایج حاصل از تمامی تکرارها به عنوان معیار عملکرد مدل گزارش می‌شود. این رویکرد موجب کاهش واریانس ناشی از نحوه تقسیم داده‌ها شده و تخمین دقیق‌تر و پایدارتری از عملکرد واقعی مدل فراهم می‌کند. در این مطالعه از روش Repeated K-Fold Cross-Validation با پنج بخش و سه تکرار (۸۰ درصد داده‌ها برای آموزش و ۲۰ درصد مابقی برای آزمایش مدل) استفاده شد. معیارهای آماری ضریب تبیین (R^2)، مجذور میانگین مربعات خطا (RMSE)، میانگین قدر مطلق خطا (MAE) و ضریب کارایی نش - ساتکلیف (NSE) طبق روابط ۴ تا ۷ برای مقایسه عملکرد مدل‌ها مورد استفاده قرار گرفت. این شاخص‌ها برای هر بار تکرار اعتبارسنجی محاسبه و به صورت میانگین $\pm$ انحراف معیار گزارش شدند.

$$R^2 = \frac{[\sum_{i=1}^n ((S_i - \bar{S}) \times (O_i - \bar{O}))]^2}{\sum_{i=1}^n ((S_i - \bar{S})^2) \times \sum_{i=1}^n ((O_i - \bar{O})^2)} \quad \text{رابطه ۴}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (O_i - S_i)^2}{n}} \quad \text{رابطه ۵}$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |S_i - O_i| \quad \text{رابطه ۶}$$

$$NSE = 1 - \frac{\sum_{i=1}^n (O_i - S_i)^2}{\sum_{i=1}^n (O_i - \bar{O})^2} \quad \text{رابطه ۷}$$

که در رابطه‌های بالا S_i و O_i به ترتیب مقادیر پیش‌بینی مدل و اندازه‌گیری شده، n تعداد مقادیر مشاهداتی و $\bar{O}$ و $\bar{S}$ به ترتیب میانگین مقادیر اندازه‌گیری و شبیه‌سازی شده است. شاخص RMSE و MAE میزان تطابق منحنی مدل با داده‌ها یا فاصله پیش‌بینی‌ها با مقادیر اندازه‌گیری را اندازه می‌گیرد و هرچه قدر نزدیک صفر باشد؛ مدل عملکرد خوبی داشته است. شاخص NSE نشان‌دهنده کارایی کلی مدل در مدل‌سازی بوده و می‌تواند از منفی بی‌نهایت تا ۱ باشد که عدد منفی و ۱ این شاخص به ترتیب نشان‌دهنده کارایی ضعیف مدل و عملکرد خوب آن است (Jamieson et al., 1991).

۵.۲. اهمیت ویژگی‌ها و تفسیر مدل با استفاده از SHAP

به منظور بررسی و تحلیل تأثیر متغیرهای ورودی در پیش‌بینی ET_c گیاه خیار، دو رویکرد تکمیلی تحلیل اهمیت ویژگی‌ها در مدل‌های درخت‌مبنا و تفسیر مدل‌ها بر اساس مقادیر SHAP^{۱۹} مورد استفاده قرار گرفت (Lundberg & Lee, 2017). برای الگوریتم‌های GB، RF

و XGB اهمیت هر ویژگی با استفاده از روش کلاسیک مبتنی بر درخت محاسبه شد. این روش اهمیت هر متغیر را بر اساس میزان کاهش میانگین خطا در هنگام تقسیم گره‌ها اندازه‌گیری می‌کند. برای افزایش قابلیت تفسیر مدل‌ها و ارائه شفافیت در تصمیم‌گیری مدل پیش‌بینی، از روش SHAP استفاده شد. این روش مبتنی بر تئوری بازی‌های شاپلی بوده و سهم هر ویژگی را در خروجی نهایی مدل به صورت کمی و بصری نمایش می‌دهد (Lundberg & Lee, 2017). بر اساس نتایج به‌دست‌آمده از تحلیل SHAP، یافته‌های حاصل از تحلیل اهمیت ویژگی‌ها مورد تأیید قرار گرفت و درک دقیق‌تری از نحوه تأثیرگذاری پارامترهای مختلف بر پیش‌بینی ET_c به دست آمد.

۳. نتایج و بحث

۱.۳. تبخیر - تعرق گیاه

مقادیر روزانه و تجمعی ET_c اندازه‌گیری‌شده در گلخانه برای دو دوره رشد A-W و S-S در شکل ۳ نمایش داده شده است. بر اساس نتایج حاصل، در دوره رشد A-W مقدار حداقل، حداکثر، میانگین و مجموع ET_c به‌ترتیب برابر با ۰/۱۴ (در آغاز دوره)، ۳/۹۲، ۲/۷۲ میلی‌متر در روز و ۲۰۰ میلی‌متر در کل دوره رشد بوده است. در مقابل، در دوره رشد S-S مقادیر متناظر به‌ترتیب معادل ۶/۹، ۳/۷ میلی‌متر در روز و ۴۴۰/۱۱ میلی‌متر ثبت شده‌اند. با توجه به افزایش تابش خورشیدی و دمای محیط در دوره S-S و نیز طولانی‌تر بودن مدت دوره رشد نسبت به A-W، افزایش چشم‌گیری در ET_c مشاهده شده که به طور قابل توجهی نیاز آبی کل دوره رشد را افزایش داده است؛ به‌گونه‌ای که مقدار تجمعی ET_c در این دوره تقریباً دو برابر دوره A-W بوده است. مطالعه انجام‌شده توسط Rezvani et al. (2022) نشان داد که در طول دوره رشد A-W به مدت ۹۶ روز، مجموع ET_c خیار گلخانه‌ای برابر با ۲۱۴/۲ میلی‌متر بوده که این مقدار با یافته‌های تحقیق حاضر هم‌راستا است. در حالی که برای دوره رشد S-S، مقدار تجمعی ET_c توسط ایشان معادل ۲۲۲/۳ میلی‌متر گزارش شده است. این اختلاف را می‌توان ناشی از تفاوت در شرایط اقلیمی محیطی، نوع سازه گلخانه و همچنین استفاده یا عدم استفاده از تجهیزات تنظیم‌کننده شرایط آب‌وهوایی دانست. علاوه بر این، Rezvani et al. (2022) در تحقیق خود، برای برآورد ET_c از مدل Stanghellini (1987) بهره گرفته‌اند که یکی از پارامترهای کلیدی آن شاخص سطح برگ LAI (۲۰) است. این پژوهشگران از معادله رگرسیونی ارائه‌شده توسط Medrano et al. (2005) برای برآورد LAI استفاده کرده‌اند که می‌تواند دلیل تفاوت در نتایج حاصل باشد. در مطالعه دیگری، Nikolaou et al. (2021) مقدار ET_c خیار در دوره‌ای به طول هفت ماه کشت گلخانه‌ای را برابر با ۵۶۶ میلی‌متر گزارش کرده‌اند. شایان ذکر است که مقدار ET_c یا ET_o در گلخانه می‌تواند به عوامل متعددی نظیر رقم گیاه، نوع و سن پوشش گلخانه، ساختار سازه، شدت تابش ورودی، مدیریت زراعی، فصل کاشت، اقلیم منطقه و میزان بهره‌گیری از تجهیزات کنترل شرایط محیطی وابسته باشد.

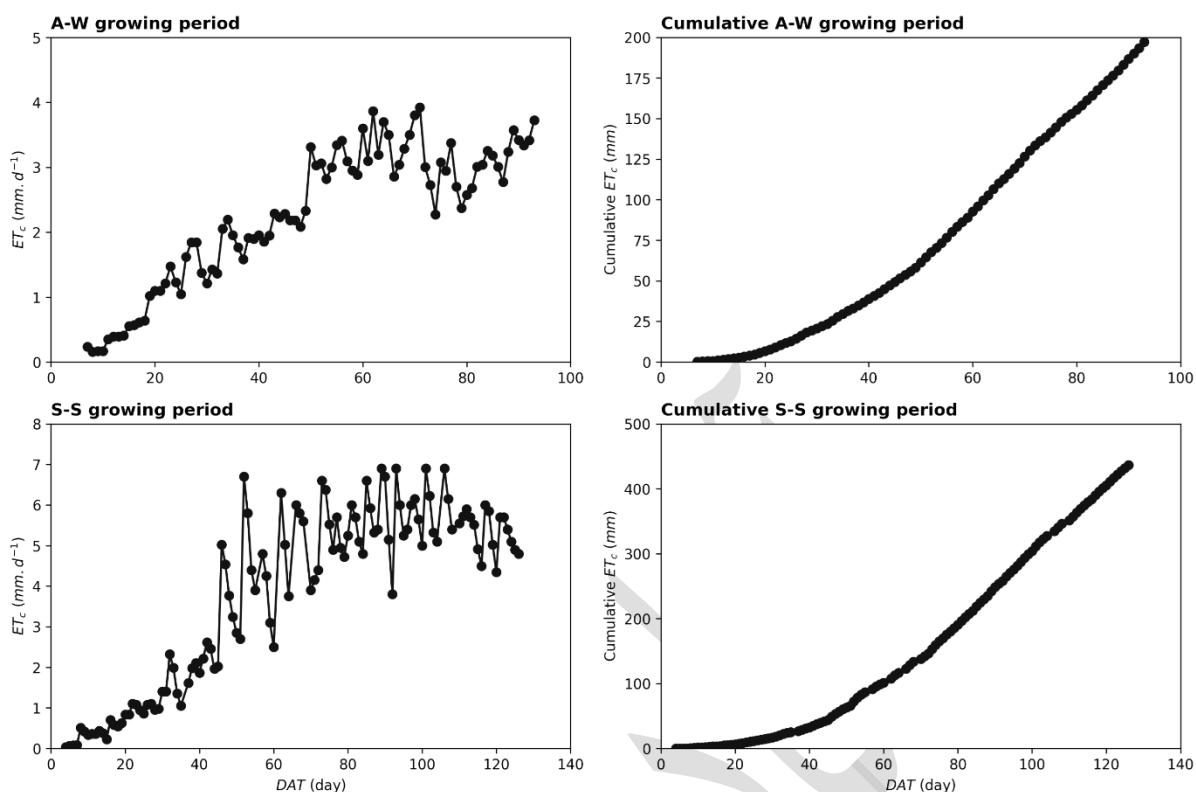


Figure 3. Daily and cumulative ET_c of greenhouse cucumber during the two growth periods A-W and S-S

۲.۳. عملکرد مدل‌ها

نتایج حاصل از ارزیابی شش مدل یادگیری ماشین برای پیش‌بینی ET_c خیار گلخانه‌ای در شکل ۴ ارائه شده است. معیارهای آماری مورد استفاده شامل R^2 , RMSE, MAE و NSE هستند که به کمک آن‌ها دقت و کارایی مدل‌ها بررسی شده است. در بین مدل‌های مورد بررسی، مدل GB با میانگین R^2 , RMSE, MAE و NSE به ترتیب برابر با ۰/۹۰، ۰/۵۹، ۰/۴۱ میلی‌متر در روز و ۰/۸۹ بالاترین دقت و عملکرد را نسبت به سایر مدل‌ها از خود نشان داده است. پس از آن، مدل XGB و RF با NSE به ترتیب برابر با ۰/۸۹ و ۰/۸۸ و RMSE برابر با ۰/۶۱ و ۰/۶۵ میلی‌متر در روز، به ترتیب در جایگاه دوم و سوم قرار گرفته‌اند. مدل SVM نیز با NSE و RMSE به ترتیب برابر با ۰/۸۶ و ۰/۷ میلی‌متر در روز عملکرد نسبتاً خوبی داشته است. از طرف دیگر، مدل‌های PLS و PCR که مدل‌های مبتنی بر رگرسیون خطی هستند با RMSE برابر ۰/۸۶ میلی‌متر در روز و NSE برابر با ۰/۷۹ عملکرد ضعیف‌تری نسبت به مدل‌های درخت تصمیم و غیرخطی نشان داده‌اند.

نکته قابل توجه در این بررسی، نتایج حاصل از تحلیل آماری تفاوت شاخص‌های مورد بررسی (R^2 , RMSE, MAE و NSE) میان مدل‌ها است. مطابق انحراف معیار ارائه شده در شکل ۴ مشخص است که تفاوت معنی‌داری بین عملکرد مدل‌های RF, GB, XGB و SVM وجود ندارد؛ اما عملکرد این چهار مدل به‌طور معنی‌داری بهتر از مدل‌های PLS و PCR در همه شاخص‌های مورد بررسی بوده است. همچنین بین مدل‌های PLS و PCR تفاوت آماری معنی‌داری وجود نداشته و عملکرد هر دو مدل مشابه یکدیگر است. این یافته نشان می‌دهد که مدل‌های پیشرفته‌تر و غیرخطی یادگیری ماشین توانایی بالاتری در مدل‌سازی روابط پیچیده بین پارامترهای اقلیمی و ET_c دارند. در مجموع، مدل‌های مبتنی بر تقویت گرادیان و درخت تصمیم (GB, XGB و RF) و همچنین SVM به‌عنوان مدل‌های برتر برای پیش‌بینی دقیق ET_c در شرایط گلخانه‌ای معرفی می‌شوند.

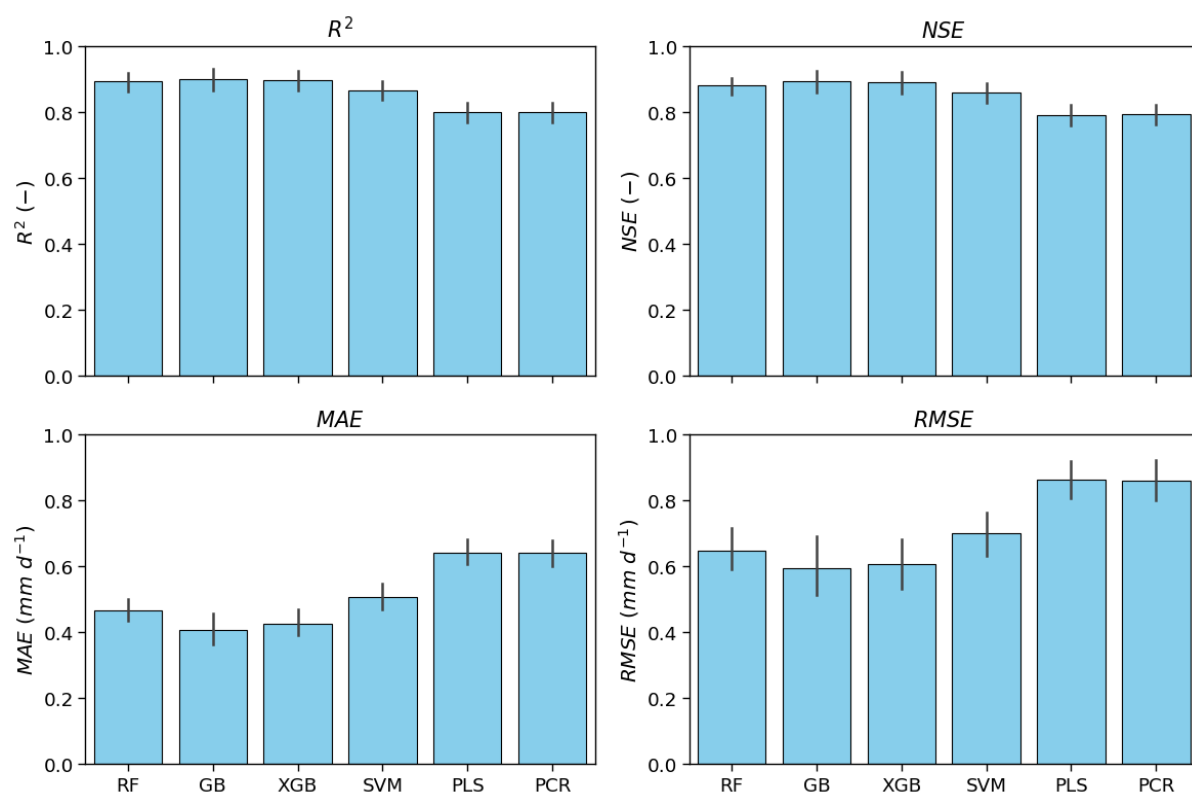


Figure 4. Performance comparison of different machine learning models in estimating greenhouse cucumber ET_c based on statistical indicators: R^2 , RMSE, MAE, and NSE

بر اساس بررسی منابع موجود، به نظر می‌رسد که روش‌های گروهي مانند جنگل‌های تصادفی و XGBoost و همچنین رویکردهای یادگیری عمیق مانند LSTM، اغلب در مقایسه با مدل‌های رگرسیون سنتی یا حتی ANN و SVM منفرد، دقت بالاتری در پیش‌بینی ET_c و ET_o دارند. با این حال، مدل بهینه می‌تواند به مجموعه داده خاص، متغیرهای ورودی و شرایط گلخانه بستگی داشته باشد (Jung et al., 2022). عوامل مختلفی می‌توانند بر دقت مدل‌های یادگیری ماشین برای پیش‌بینی ET تأثیر بگذارند، از جمله کیفیت و کمیت داده‌های آموزشی، انتخاب متغیرهای ورودی مرتبط، تنظیم ابرپارامترهای مدل و محیط خاص گلخانه. دقت هر روش عمدتاً به شرایط آب و هوایی خاص منطقه بستگی دارد. برای دستیابی به بهترین دقت مدل‌های تجربی و مدرن، کالیبراسیون محلی مورد نیاز است (Averbuch & Moshelion, 2024; Lakhiar et al., 2025). وضوح زمانی داده‌ها (به عنوان مثال، ساعتی در مقابل روزانه) می‌تواند به طور قابل توجهی بر دقت مدل‌های پیش‌بینی ET ، به‌ویژه در محیط‌های پویای گلخانه‌ای، تأثیر بگذارد. ریزاقلیم گلخانه می‌تواند به سرعت تغییر کند و مدل‌های آموزش‌دیده بر روی داده‌های با وضوح پایین ممکن است نوسانات مهم مؤثر بر ET را از دست بدهند (Rho et al., 2023). استفاده از معیارهای ارزیابی استاندارد و ایجاد مجموعه‌های داده معیار برای تسهیل مقایسه سازگارتر و عینی‌تر مدل‌های مختلف یادگیری ماشین برای پیش‌بینی ET گلخانه‌ای از اهمیت ویژه‌ای برخوردار است.

۳.۳. اهمیت ویژگی‌ها

شکل ۵ اهمیت ویژگی‌ها در مدل‌های مختلف مبتنی بر درخت (RF، GB و XGB) در برآورد ET_c خیار گلخانه‌ای را نشان می‌دهد. در مدل RF مهمترین ویژگی DAT است که بیش از ۶۰ درصد از اهمیت کل ویژگی‌ها را به خود اختصاص داده است. این موضوع نشان‌دهنده نقش بسیار برجسته روز پس از انتقال نشاء در تغییرات ET_c است، که احتمالاً ترکیبی از تأثیرات فصلی، دما و تابش خورشیدی در طول زمان را در بر دارد. سایر ویژگی‌های مهم شامل Radiation، meanT و minT هستند که نقش مهمی در توجیه رفتار مدل دارند. ویژگی‌هایی مانند maxRH و season از کمترین اهمیت برخوردار هستند.

مدل GB نیز مشابه RF بیشترین اهمیت را برای DAT (بیش از ۷۰ درصد) قائل شده است. ویژگی Radiation در رتبه دوم و ET_o در رتبه سوم قرار دارند که نشان از اهمیت زیاد انرژی ورودی به سیستم در تخمین ET_c دارد. برخلاف RF، در این مدل دما و رطوبت

نقش کم‌رنگ‌تری داشته‌اند. اهمیت بسیار پایین ویژگی‌هایی مانند minRH، season و meanRH می‌تواند بیانگر این باشد که مدل GB بیشتر به ویژگی‌های قوی‌تر و غیرخطی واکنش نشان داده است.

در مدل XGB نیز ویژگی روز پس از انتقال نشاء مهمترین ویژگی است، ولی با اهمیتی کمتر از دو مدل دیگر (حدود ۳۰ درصد). نکته قابل توجه اینجاست که در این مدل season (دوره کاشت؛ پاییز - زمستان یا بهار - تابستان) نقش دوم را بازی می‌کند که نشان می‌دهد الگوریتم XGB توانسته است از اطلاعات دوره کاشت بهره‌برداری بهتری داشته باشد. سایر ویژگی‌های مهم شامل meanT، Radiation و minRH هستند. ویژگی maxT کمترین اهمیت را در بین متغیرهای ورودی در این مدل دارد.

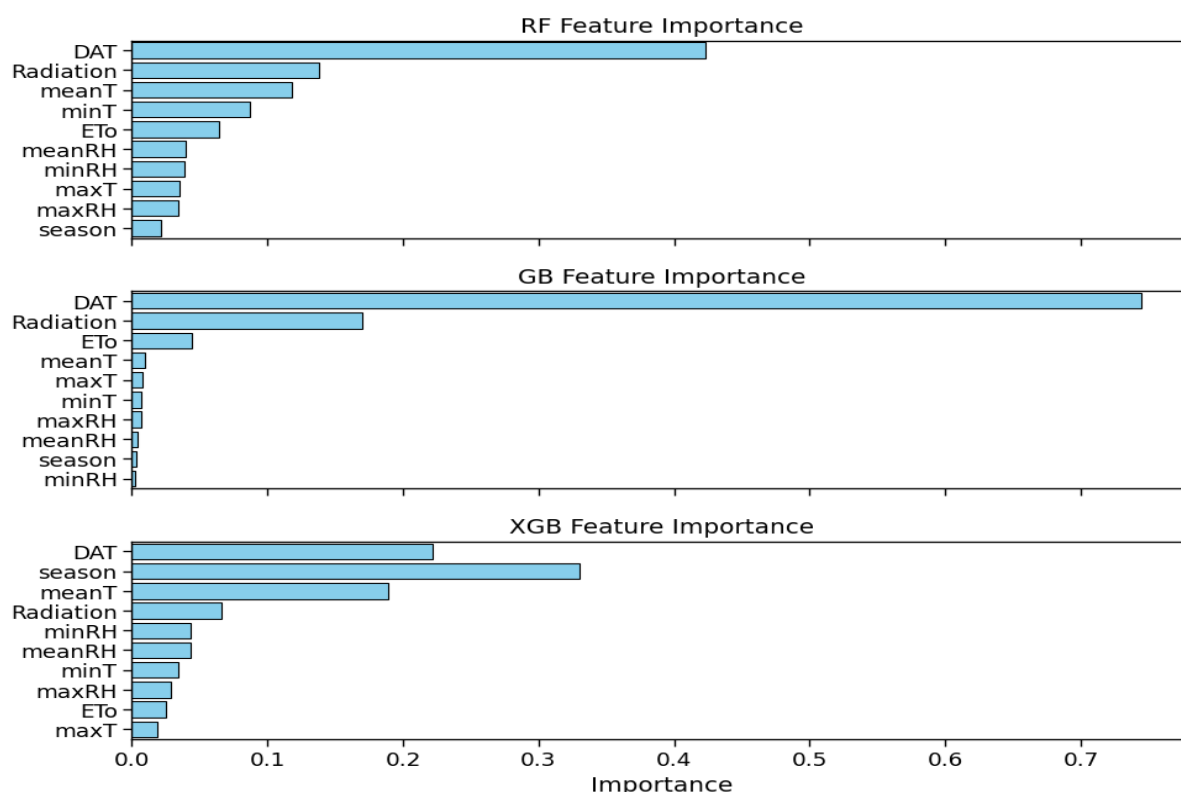


Figure 5. Comparison of feature importance in tree-based models (RF, GB, and XGB) for estimating greenhouse cucumber ET_c .

به‌عنوان جمع‌بندی، در هر سه مدل، DAT به‌عنوان مؤثرترین ویژگی شناسایی شده است، که منعکس‌کننده عوامل فصلی، رشد گیاه و تغییرات اقلیمی روزانه است. Radiation و meanT در هر سه مدل جایگاه بالایی داشته‌اند و اهمیت آن‌ها تأکیدی بر نقش دما و انرژی تابشی در فرآیند ET_c دارد. تفاوت بین مدل‌ها در وزن‌دهی به ویژگی‌هایی مانند season، minRH و ETo می‌تواند به دلیل تفاوت در نحوه یادگیری و حساسیت به وابستگی‌های خطی یا غیرخطی باشد. همچنین لازم به ذکر است که اهمیت نسبتاً پایین دما و رطوبت نسبی در برخی مدل‌ها به معنای بی‌تأثیری این پارامترها در گلخانه نیست، بلکه به دلیل همبستگی بالای آن‌ها با متغیرهایی مانند تابش و روز پس از انتقال نشاء است که سبب می‌شود مدل‌های درختی اثرات این متغیرها را به‌طور غیرمستقیم بازتاب دهند.

۴.۳. تحلیل SHAP

شکل ۸، تحلیل مقادیر SHAP را برای سه مدل یادگیری ماشین مبتنی بر درخت تصمیم شامل RF، GB و XGB نمایش می‌دهد. مقادیر SHAP ابزاری برای تفسیر مدل‌های پیچیده یادگیری ماشین هستند که نشان می‌دهد هر ویژگی (متغیر ورودی) چگونه و با چه شدتی بر خروجی مدل (ET_c) تأثیر می‌گذارد (Lundberg & Lee, 2017). در نمودار پراکندگی SHAP مربوط به الگوریتم RF، مشاهده می‌شود که ویژگی DAT (روز پس از انتقال نشاء) بیش‌ترین تأثیر را بر پیش‌بینی مدل دارد و با افزایش مقدار آن (نقاط قرمز)، مقدار برآورد ET_c نیز افزایش یافته است. ویژگی‌های Radiation، meanT و ETo نیز تأثیر قابل توجهی دارند. نمودار میله‌ای در سمت راست نیز تأیید می‌کند که این چهار ویژگی دارای بالاترین میانگین مقادیر مطلق SHAP هستند که نشان از اهمیت آن‌ها در یادگیری مدل RF دارد.

مدل GB، همچنان DAT مقدار SHAP برابر ۱/۴ مهم‌ترین ویژگی است. پس از آن Radiation، ET_c و meanT نقش‌های مهمی ایفا می‌کنند. اهمیت کم‌تر ویژگی‌هایی مانند season و minRH نشان می‌دهد که مدل تأثیر کمتری برای آن‌ها در نظر گرفته است. پراکندگی رنگ نقاط نیز وابستگی مقدار ویژگی به تأثیر آن را نشان می‌دهد (ویژگی‌های با مقدار بالا معمولاً اثر مثبت و با رنگ قرمز نمایش داده شده‌اند).

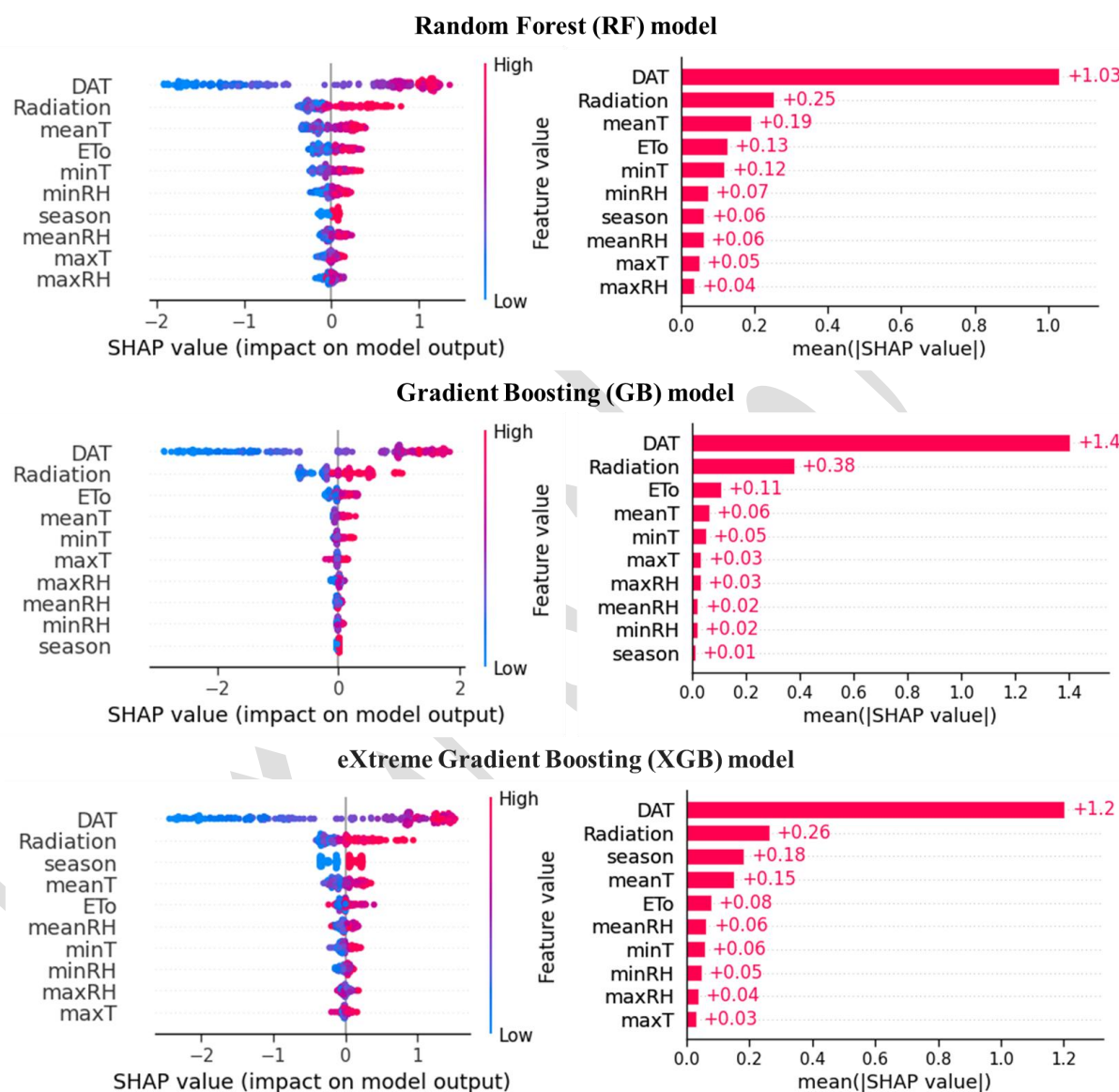


Figure 6. Feature importance analysis using SHAP values in tree-based machine learning models for estimating greenhouse cucumber ET_c

در مدل XGB نیز، DAT مهم‌ترین عامل مؤثر در پیش‌بینی ET_c است. برخلاف مدل GB، در این مدل season اهمیت بیشتری نسبت به سایر مدل‌ها پیدا کرده است که می‌تواند ناشی از توانایی بالاتر XGBoost در شناسایی تعاملات غیرخطی بین فصل و سایر متغیرها باشد. نمودار میله‌ای بار دیگر اهمیت نسبی متغیرها را از نظر میانگین قدر مطلق مقادیر SHAP نشان می‌دهد، که در آن Radiation، meanT و season پس از DAT در رتبه‌های بالاتر قرار دارند. به طور خلاصه، در هر سه مدل، DAT مهم‌ترین متغیر در پیش‌بینی ET_c خیار گلخانه‌ای است، که با یافته‌های مطالعات پیشین هم‌راستا است و نشان می‌دهد ET_c رابطه مستقیمی با مرحله رشد گیاه دارد (Allen et al., 1998). همچنین ویژگی‌هایی مانند تشعشع خورشیدی، دما و تبخیر - تعرق مرجع در همه مدل‌ها نقش مؤثری ایفا کرده‌اند، که اهمیت پارامترهای اقلیمی در فرآیند ET_c را نشان می‌دهد. تفاوت در اهمیت برخی متغیرها بین مدل‌ها می‌تواند به تفاوت در معماری الگوریتم‌ها و نحوه یادگیری تعاملات پیچیده بین ویژگی‌ها بازگردد.

شکل ۹ تحلیل اهمیت ویژگی‌ها با استفاده از مقادیر SHAP در مدل‌های SVM، PLS و PCR برای برآورد ET_c خیار گلخانه‌ای را نشان می‌دهد. در مدل SVM، طبق نمودار میله‌ای، DAT دارای بیشترین مقدار میانگین $(+0.27)$ SHAP است و مهم‌ترین عامل در پیش‌بینی ET_c به شمار می‌رود. ویژگی‌های اقلیمی شامل $maxRH$ ، $meanRH$ و $minRH$ و Radiation نیز تأثیر بالایی در عملکرد مدل دارند که اهمیت شرایط رطوبتی و تابش خورشیدی در گلخانه را تأیید می‌کند. نمودار پراکندگی SHAP نشان می‌دهد که مقدار بالای $minRH$ و $meanRH$ (به رنگ قرمز) تأثیر منفی بر خروجی دارد، در حالی که متغیرهای روز پس از انتقال نشاء، تابش و دما بیشتر اثر مثبت دارند. این تفسیر با روابط فیزیولوژیک گیاه و ET_c مطابقت دارد، چرا که در شرایط رطوبت بالا معمولاً نرخ تبخیر کاهش می‌یابد.

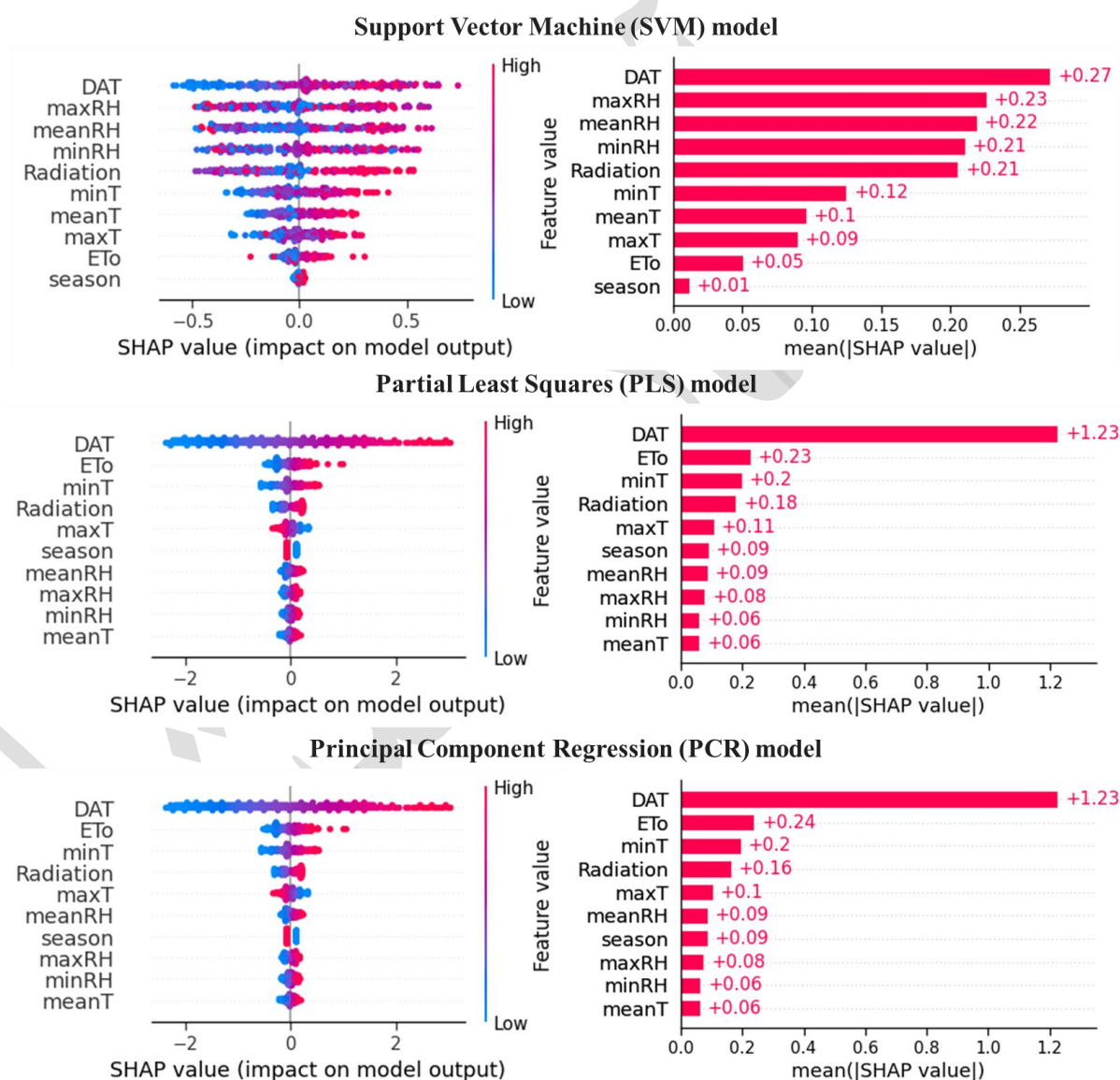


Figure 7. Feature importance analysis using SHAP values in SVM, PLS and PCR models for estimating greenhouse cucumber ET_c

در مدل PLS، DAT به وضوح با فاصله زیاد (میانگین SHAP برابر با $+1.23$) مؤثرترین متغیر است. پس از آن، ET_c و $minT$ و Radiation در رتبه‌های بعدی قرار دارند. نمودار پراکندگی SHAP به وضوح نقش مثبت مقدار بالای DAT در افزایش برآورد ET_c را نشان می‌دهد، در حالی که اثر متغیرهای دیگر ترکیبی از اثر مثبت و منفی بسته به مقدارشان است. اهمیت بالای $minT$ نسبت به سایر پارامترهای دمایی جالب توجه است و می‌تواند به حساسیت بالای خیار به دماهای پایین در شب و کاهش نرخ تعرق در این شرایط مربوط باشد. مدل PCR نیز نتایج مشابهی با مدل PLS دارد. بار دیگر، DAT مهم‌ترین متغیر با مقدار SHAP $+1.23$ است. متغیرهای

ET₀, minT, Radiation, maxT در رتبه‌های بعدی قرار دارند. برخلاف برخی مدل‌ها مانند SVM، اثر متغیرهای رطوبتی در این مدل کمتر است. PCR به‌عنوان مدلی بر پایه تحلیل مؤلفه‌های اصلی، ممکن است روابط غیرخطی را به‌خوبی مدل نکند، اما با کاهش ابعاد داده‌ها و حذف هم‌خطی بین ویژگی‌ها، توانسته اثر ویژگی‌های اصلی از جمله تابش و دما را برجسته سازد. به‌عنوان جمع‌بندی کلی از نتایج تحلیل SHAP در مدل‌های مورد بررسی می‌توان گفت که متغیر DAT در تمامی مدل‌ها مؤثرترین عامل برآورد ET₀ در خیار گلخانه‌ای بوده که تأییدی بر نقش حیاتی مراحل فنولوژیکی در نیاز آبی گیاه است. متغیرهای اقلیمی از جمله تابش، دما و ET₀ نیز نقش‌های ثانویه اما قابل توجهی دارند. تفاوت در میزان اهمیت سایر متغیرها بین مدل‌ها نشان‌دهنده نوع حساسیت و توانایی مدل‌ها در شناسایی روابط پیچیده و تعاملات غیرخطی بین ویژگی‌ها است.

در تحلیل خروجی SHAP شش مدل مختلف مورد بررسی، یک الگوی نسبتاً ثابت در اهمیت متغیرها قابل مشاهده است که نشان از پایداری عوامل کلیدی مؤثر بر تبخیر - تعرق در شرایط گلخانه‌ای دارد. متغیر DAT در تمامی مدل‌ها به‌عنوان مؤثرترین ویژگی ظاهر شده است، که اهمیت مراحل فنولوژیکی گیاه را در نیاز آبی آن به وضوح نشان می‌دهد. این موضوع با مبانی فیزیولوژیکی خیار گلخانه‌ای همخوانی دارد، چرا که با افزایش سن گیاه، سطح برگ، تعرق و در نتیجه نیاز آبی افزایش می‌یابد. در میان متغیرهای اقلیمی Radiation و ET₀ به‌طور مداوم در رتبه‌های بالای اهمیت قرار دارند. این متغیرها نشان می‌دهند که تأمین انرژی برای فرآیند تبخیر - تعرق از طریق تابش و شاخص‌های مشتق شده مانند ET₀ نقش کلیدی دارند. در مدل‌های مبتنی بر Boosting مانند XGB و GB، Radiation و ET₀ سهم قابل توجهی داشته‌اند و این الگوریتم‌ها توانسته‌اند تعاملات غیرخطی بین این متغیرها و متغیر هدف را به‌خوبی شناسایی کنند. مدل RF نیز رفتاری مشابه مدل‌های تقویتی نشان داده و علاوه بر اهمیت بالای Radiation و ET₀، تأثیر minT و meanT را نیز شناسایی کرده است. این نشان‌دهنده توانمندی RF در مدل‌سازی ساختارهای داده‌ای پیچیده با حفظ توازن بین ویژگی‌ها است. از سوی دیگر، مدل SVM که ساختاری مبتنی بر مرز تصمیم‌گیری دارد، نسبت به دیگر مدل‌ها به ویژگی‌های رطوبتی (minRH, maxRH و meanRH) حساسیت بیشتری نشان داده است. در این مدل، اثر مثبت و منفی ویژگی‌های رطوبتی به شکل متقارن‌تری توزیع شده که ممکن است نشان‌دهنده وجود روابط غیرخطی میان این متغیرها و ET₀ باشد. این موضوع بیانگر دقت بالای SVM در مدل‌سازی محدوده‌های خاص و حساس داده، خصوصاً در محیط‌های بسته‌ای مانند گلخانه است که رطوبت نقش تعیین‌کننده دارد. در دو مدل خطی PLS و PCR، اهمیت ویژگی‌ها بسیار متمرکزتر بوده و مجدداً DAT، ET₀ و minT در صدر قرار گرفته‌اند. مدل PLS به‌واسطه ارتباط قوی بین متغیرهای مستقل و پاسخ و توان کاهش هم‌خطی، عملکرد قابل‌قبولی داشته و ویژگی‌هایی چون Radiation و minT را نیز به‌خوبی شناسایی کرده است. مدل PCR نیز با استفاده از مؤلفه‌های اصلی توانسته اثرات ترکیبی متغیرهای هم‌بسته را نمایان سازد، اما نسبت به مدل‌های غیرخطی، حساسیت کمتری نسبت به ویژگی‌های رطوبتی از خود نشان داده است. در مجموع، مدل‌های تقویتی و مبتنی بر درخت (GB، RF و XGB) به دلیل توانایی در مدل‌سازی روابط پیچیده و تعاملات بین ویژگی‌ها، توزیع SHAP گسترده‌تر و معنادارتری را ارائه داده‌اند. این در حالی است که مدل‌های خطی (SVM، PLS و PCR) با دقت بالا بر تمرکز بر متغیرهای خاص، دید مناسبی از ساختار کلی داده‌ها و تأثیر ویژگی‌ها فراهم کرده‌اند.

بر اساس بررسی منابع موجود، متغیرهای محیطی رایج مورد استفاده به‌عنوان ورودی در مدل‌های یادگیری ماشین برای پیش‌بینی تبخیر - تعرق گلخانه‌ای عبارتند از تشعشع خورشیدی، دمای هوا (حداکثر، حداقل، میانگین)، رطوبت نسبی (حداکثر، حداقل، میانگین)، کمبود فشار بخار (VPD)، سرعت باد (اگرچه اغلب در گلخانه‌ها کم است) و تشعشع فرازمینی (Ge et al., 2022; Rho et al., 2023; Yu et al., 2020). برخی از مدل‌ها متغیرهای خاص محصول مانند LAI، مرحله رشد محصول و وزن تر یا زیست‌توده گیاه را نیز شامل می‌شوند (JuarezMaldonado et al., 2014; Jung et al., 2022; Nikolaou et al., 2023). در کشت بدون خاک، متغیرهای مربوط به محیط کشت مانند دمای بستر کشت، هدایت الکتریکی بستر کشت و رطوبت بستر کشت نیز ممکن است استفاده شوند (Nikolaou et al., 2017). متغیرهای زمانی مانند روز سال یا چرخه رشد و ساعت روز نیز در مدل‌ها اهمیت دارند. تشعشع خورشیدی به‌طور مداوم به‌عنوان یک عامل بسیار مؤثر در پیش‌بینی تبخیر - تعرق مرجع و گیاه در مدل‌های مختلف یادگیری ماشین و محصولات مختلف ظاهر می‌شود که بر اهمیت نور و تعادل انرژی در فرآیند تعرق تأکید دارد (Averbuch & Moshelion, 2024).

۴. نتیجه گیری

این پژوهش با هدف بررسی کارایی مدل‌های مختلف یادگیری ماشین در پیش‌بینی تبخیر – تعرق گیاه در شرایط گلخانه‌ای برای محصول خیار انجام شد. با توجه به پیچیدگی‌های اقلیمی محیط‌های کنترل‌شده گلخانه‌ای و محدودیت‌های ذاتی روش‌های سنتی، استفاده از مدل‌های داده‌محور مبتنی بر یادگیری ماشین به عنوان راهکاری مؤثر برای بهبود دقت و پایداری برآورد ET_c مطرح می‌شود. نتایج تجربی نشان داد که مقدار ET_c در دوره رشد S-S به طور قابل‌توجهی بالاتر از دوره A-W است که این امر ناشی از شرایط اقلیمی گرم‌تر و تابش بیشتر در این دوره رشد است. مقایسه عملکرد شش الگوریتم یادگیری ماشین نشان داد که مدل‌های مبتنی بر درخت تصمیم به‌ویژه RF، GB و XGB بالاترین دقت را در پیش‌بینی ET_c نشان دادند (NSE بالای ۰/۸۸ و RMSE کمتر از ۰/۶۵ میلی‌متر در روز). مدل‌های خطی مانند PLS و PCR عملکرد ضعیف‌تری داشتند که نشان از ناتوانی آن‌ها در مدل‌سازی روابط غیرخطی میان متغیرها دارد. تحلیل اهمیت ویژگی‌ها و شاخص‌های SHAP در تمامی مدل‌ها نشان داد که متغیر DAT مؤثرترین عامل در برآورد ET_c است. همچنین متغیرهای اقلیمی مانند meanT، Radiation، ET_o نقش مهمی در پیش‌بینی دقیق مدل‌ها ایفا کرده‌اند. این یافته‌ها حاکی از آن است که مدل‌های یادگیری ماشین نه تنها قادر به شناسایی الگوهای غیرخطی میان متغیرهای ورودی و خروجی هستند، بلکه می‌توانند پویایی‌های پیچیده محیط گلخانه را نیز با دقت بالا مدل‌سازی کنند. در مجموع، نتایج این مطالعه تأکید می‌کند که استفاده از الگوریتم‌های یادگیری ماشین به‌ویژه مدل‌های تقویت‌گر گرادینان و درخت تصمیم، می‌تواند روشی قابل اعتماد و دقیق برای پیش‌بینی نیاز آبی خیار گلخانه‌ای باشد. با این حال، برای دستیابی به عملکرد بهینه، ضرورت کالیبراسیون محلی مدل‌ها، انتخاب مناسب متغیرهای ورودی و استفاده از داده‌های با وضوح زمانی بالا پیشنهاد می‌شود. همچنین، ایجاد پایگاه‌های داده مرجع و بهره‌گیری از روش‌های تفسیرپذیر مانند SHAP، می‌تواند زمینه را برای توسعه مدل‌های قابل اتکا و قابل تعمیم به سایر محصولات گلخانه‌ای فراهم آورد.

۵. پی‌نوشت‌ها

- ¹ Bowen Ratio
- ² Principal Components Regression
- ³ Partial Least Squares Regression
- ⁴ Kernel Functions
- ⁵ Random Forest
- ⁶ Bootstrapped
- ⁷ Feature Importance
- ⁸ Gradient Boosting
- ⁹ Boosting
- ¹⁰ Loss Function
- ¹¹ Extreme Gradient Boosting, XGBoost
- ¹² Regularization
- ¹³ Pruning
- ¹⁴ Early Stopping
- ¹⁵ Hyperparameters
- ¹⁶ Tree-structured Parzen Estimator
- ¹⁷ Cross-Validation
- ¹⁸ Fold
- ¹⁹ SHapley Additive exPlanations
- ²⁰ Leaf Area Index

۶. تشکر و قدردانی

این اثر تحت حمایت مادی صندوق حمایت از پژوهشگران و فناوران کشور (INSF) برگرفته شده از طرح شماره «۴۰۱۳۶۷۴» انجام شده است.

۷. تعارض منافع

هیچ‌گونه تعارض منافع توسط نویسندگان وجود ندارد.

۸. منابع

- Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A Next-generation Hyperparameter Optimization Framework. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2623–2631. <https://doi.org/10.1145/3292500.3330701>
- Alfieri, J. G., Blanken, P. D., Smith, D., & Morgan, J. (2009). Concerning the Measurement and Magnitude of Heat, Water Vapor, and Carbon Dioxide Exchange from a Semiarid Grassland. *Journal of Applied Meteorology and Climatology*, 48(5), 982–996. <https://doi.org/10.1175/2008JAMC1873.1>
- Allen, R. G., Pereira, L. S., Raes, D., Smith, M., & others. (1998). Crop evapotranspiration-Guidelines for computing crop water requirements-FAO Irrigation and drainage paper 56. *Fao, Rome*, 300(9), D05109.
- Assari, M., Kouchakzadeh, M., Shahabifar, M., & Bayat, K. (2012). Estimation of Reference Evapotranspiration in Greenhouse by Artificial Neural Network. *Journal of Water and Soil Conservation*, 16(1), 107–121. https://jwsc.gau.ac.ir/article_582.html (In Persian)
- Averbuch, N., & Moshelion, M. (2024). *Evaluating Evapotranspiration in a Commercial Greenhouse: A Comparative Study of Microclimatic Factors and Machine-Learning Algorithms*. <https://doi.org/10.1101/2024.01.11.575151>
- Borg, H., & Grimes, D. W. (1986). Depth Development of Roots with Time: An Empirical Description. *Transactions of the ASAE*, 29(1), 194–197. <https://doi.org/10.13031/2013.30125>
- Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- 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*, 785–794. <https://doi.org/10.1145/2939672.2939785>
- Dhaliwal, D. S., & Williams, M. M. (2024). Sweet corn yield prediction using machine learning models and field-level data. *Precision Agriculture*, 25(1), 51–64. <https://doi.org/10.1007/s11119-023-10057-1>
- Eslamian, S. S., Abedi-Koup, J., Amiri, M. J., & Gohari, S. A. (2009). Estimation of Daily Reference Evapotranspiration Using Support Vector Machines and Artificial Neural Networks in Greenhouse. *Research Journal of Environmental Sciences*, 3(4), 439–447. <https://doi.org/10.3923/rjes.2009.439.447>
- Fan, J., Wu, L., Zhang, F., Cai, H., Wang, X., Lu, X., & Xiang, Y. (2018). Evaluating the effect of air pollution on global and diffuse solar radiation prediction using support vector machine modeling based on sunshine duration and air temperature. *Renewable and Sustainable Energy Reviews*, 94, 732–747. <https://doi.org/10.1016/j.rser.2018.06.029>
- Fernández, M. D., Bonachela, S., Orgaz, F., Thompson, R., López, J. C., Granados, M. R., Gallardo, M., & Fereres, E. (2010). Measurement and estimation of plastic greenhouse reference evapotranspiration in a Mediterranean climate. *Irrigation Science*, 28(6), 497–509. <https://doi.org/10.1007/S00271-010-0210-Z>
- Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, 29(5), 1189–1232. <http://www.jstor.org/stable/2699986>
- Gallegos, R. K. B., Saludes, R. B., & Paras Jr, F. O. (2013). Bowen ratio estimates in green house and open field conditions of the University of the Philippines Los Banos. *ASIA LIFE SCIENCES*, 22(2), 699–712.
- Ge, J., Zhao, L., Yu, Z., Liu, H., Zhang, L., Gong, X., & Sun, H. (2022). Prediction of Greenhouse Tomato Crop Evapotranspiration Using XGBoost Machine Learning Model. *Plants*, 11(15), 1923. <https://doi.org/10.3390/plants11151923>
- Ghorbani, M. A., Shamshirband, S., Zare Haghi, D., Azani, A., Bonakdari, H., & Ebtehaj, I. (2017). Application of firefly algorithm-based support vector machines for prediction of field capacity and permanent wilting point. *Soil and Tillage Research*, 172, 32–38. <https://doi.org/10.1016/j.still.2017.04.009>
- Incrocci, L., Thompson, R. B., Fernandez-Fernandez, M. D., De Pascale, S., Pardossi, A., Stanghellini, C., Roupheal, Y., & Gallardo, M. (2020). Irrigation management of European greenhouse vegetable crops. *Agricultural Water Management*, 242, 106393. <https://doi.org/10.1016/j.agwat.2020.106393>
- James, L. (1988). *Principles of farm irrigation systems design*. <https://www.cabidigitallibrary.org/doi/full/10.5555/19891934086>
- Jamieson, P. D., Porter, J. R., & Wilson, D. R. (1991). A test of the computer simulation model ARCWHEAT1 on wheat crops grown in New Zealand. *Field Crops Research*, 27(4), 337–350. [https://doi.org/10.1016/0378-4290\(91\)90040-3](https://doi.org/10.1016/0378-4290(91)90040-3)
- Jolliffe, I. T. (1986). *Principal Components in Regression Analysis* (pp. 129–155). https://doi.org/10.1007/978-1-4757-1904-8_8
- Jones, H. G. (2004). Irrigation scheduling: advantages and pitfalls of plant-based methods. *Journal of Experimental Botany*, 55(407), 2427–2436. <https://doi.org/10.1093/jxb/erh213>
- JuarezMaldonado, A., BenavidesMendoza, A., deAlbaRomenus, K., & MoralesDiaz, A. (2014). Estimation of the water requirements of greenhouse tomato crop using multiple regression models. *Emirates Journal of Food and Agriculture*, 26(10), 885. <https://doi.org/10.9755/ejfa.v26i10.18270>
- Jung, D.-H., Lee, T. S., Kim, K., & Park, S. H. (2022). A Deep Learning Model to Predict Evapotranspiration and Relative Humidity for Moisture Control in Tomato Greenhouses. *Agronomy*, 12(9), 2169. <https://doi.org/10.3390/agronomy12092169>
- Khoshsimaie Chenar, M., Liaghat, A., Noory, H., Soltani Salehabadi, F., & Motesharezadeh, B. (2025). Impact of Different Salinity Levels and Irrigation Water Amounts on Yield and Water Productivity of Greenhouse Cucumber

- in Two Autumn-Winter and Spring-Summer Growing Periods. *Water Management in Agriculture*. https://wmaj.iaid.ir/article_217895.html (In Persian)
- Kramer, O. (2016). Scikit-Learn. *Studies in Big Data*, 20, 45–53. https://doi.org/10.1007/978-3-319-33383-0_5
- Kuhn, M., & Johnson, K. (2013). Linear Regression and Its Cousins. In *Applied Predictive Modeling* (pp. 101–139). Springer New York. https://doi.org/10.1007/978-1-4614-6849-3_6
- Lakhiar, I. A., Yan, H., Zhang, C., Zhang, J., Wang, G., Deng, S., Syed, T. N., Wang, B., & Zhou, R. (2025). A review of evapotranspiration estimation methods for climate-smart agriculture tools under a changing climate: vulnerabilities, consequences, and implications. *Journal of Water and Climate Change*, 16(2), 249–288. <https://doi.org/10.2166/wcc.2024.048>
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.
- Medrano, E., Lorenzo, P., Sánchez-Guerrero, M. C., & Montero, J. I. (2005). Evaluation and modelling of greenhouse cucumber-crop transpiration under high and low radiation conditions. *Scientia Horticulturae*, 105(2), 163–175. <https://doi.org/10.1016/J.SCIENTA.2005.01.024>
- Mehdizadeh, S., Behmanesh, J., & Khalili, K. (2017). Using MARS, SVM, GEP and empirical equations for estimation of monthly mean reference evapotranspiration. *Computers and Electronics in Agriculture*, 139, 103–114. <https://doi.org/10.1016/j.compag.2017.05.002>
- Merrill, S. D., Tanaka, D. L., & Hanson, J. D. (2002). Root Length Growth of Eight Crop Species in Haplustoll Soils. *Soil Science Society of America Journal*, 66(3), 913–923. <https://doi.org/10.2136/SSSAJ2002.9130>
- Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in Neuroinformatics*, 7. <https://doi.org/10.3389/fnbot.2013.00021>
- Nikolaou, G., Neocleous, D., Christou, A., Polycarpou, P., Kitta, E., & Katsoulas, N. (2021). Energy and Water Related Parameters in Tomato and Cucumber Greenhouse Crops in Semiarid Mediterranean Regions. A Review, Part I: Increasing Energy Efficiency. *Horticulturae*, 7(12), 521. <https://doi.org/10.3390/horticulturae7120521>
- Nikolaou, G., Neocleous, D., Katsoulas, N., & Kittas, C. (2017). Modelling transpiration of soilless greenhouse cucumber and its relationship with leaf temperature in a mediterranean climate. *Emirates Journal of Food and Agriculture*, 29(12), 911–920. <https://doi.org/10.9755/EJFA.2017.V29.I12.1561>
- Nikolaou, G., Neocleous, D., Katsoulas, N., & Kittas, C. (2019). Irrigation of Greenhouse Crops. *Horticulturae*, 5(1), 7. <https://doi.org/10.3390/horticulturae5010007>
- Nikolaou, G., Neocleous, D., Kitta, E., & Katsoulas, N. (2023). Estimating cucumber crop coefficients under different greenhouse microclimatic conditions. *International Journal of Biometeorology*, 67(11), 1745–1756. <https://doi.org/10.1007/s00484-023-02535-y>
- Pandorfi, H., Bezerra, A. C., Atarassi, R. T., Vieira, F. M. C., Barbosa Filho, J. A. D., & Guiselini, C. (2016). Artificial neural networks employment in the prediction of evapotranspiration of greenhouse-grown sweet pepper. *Revista Brasileira de Engenharia Agrícola e Ambiental*, 20(6), 507–512. <https://doi.org/10.1590/1807-1929/agriambi.v20n6p507-512>
- Rahimikhoob, H., Sohrabi, T., & Delshad, M. (2020). Assessment of reference evapotranspiration estimation methods in controlled greenhouse conditions. *Irrigation Science*, 38(4), 389–400. <https://doi.org/10.1007/s00271-020-00680-5>
- Rezvani, S., Zarei, G., & Salemi, H. (2022). Evapotranspiration and crop coefficient of greenhouse cucumber in the Hamedan region. *Iranian Journal of Irrigation & Drainage*, 16(5), 904–916. https://idj.iaid.ir/article_159921.html (In Persian)
- Rho, H., Su, J., Sim, H. S., Moon, Y. H., Woo, U. J., & Kim, S. K. (2023). Development of a Cucumber Transpiration Model Based on a Simplified Penman-Monteith Model in a Semi-closed Greenhouse. *HortScience*, 58(10), 1208–1216. <https://doi.org/10.21273/HORTSCI17218-23>
- Shahrajabian, M. H., & Sun, W. (2024). A Review of Lysimeter Studies and Experiments by Considering Agricultural Production. *Journal of Stress Physiology & Biochemistry*, 20(2), 114–132.
- Shi, W., Zhang, X., Xue, X., Feng, F., Zheng, W., & Chen, L. (2023). Analyzing Evapotranspiration in Greenhouses: A Lysimeter-Based Calculation and Evaluation Approach. *Agronomy*, 13(12), 3059. <https://doi.org/10.3390/agronomy13123059>
- Shrestha, N. K., & Shukla, S. (2015). Support vector machine-based modeling of evapotranspiration using hydro-climatic variables in a sub-tropical environment. *Agricultural and Forest Meteorology*, 200, 172–184. <https://doi.org/10.1016/j.agrformet.2014.09.025>
- Stanghellini, C. (1987). *Transpiration of greenhouse crops: an aid to climate management*. Wageningen University and Research.
- Vapnik, V. (1999). *The nature of statistical learning theory*. Springer science & business media.
- Wen, X., Si, J., He, Z., Wu, J., Shao, H., & Yu, H. (2015). Support-Vector-Machine-Based Models for Modeling Daily Reference Evapotranspiration with Limited Climatic Data in Extreme Arid Regions. *Water Resources Management*, 29(9), 3195–3209. <https://doi.org/10.1007/s11269-015-0990-2>

- Yin, Z., Wen, X., Feng, Q., He, Z., Zou, S., & Yang, L. (2017). Integrating genetic algorithm and support vector machine for modeling daily reference evapotranspiration in a semi-arid mountain area. *Hydrology Research*, 48(5), 1177–1191. <https://doi.org/10.2166/nh.2016.205>
- Yu, J., Zheng, W., Xu, L., Zhangzhong, L., Zhang, G., & Shan, F. (2020). A PSO-XGBoost Model for Estimating Daily Reference Evapotranspiration in the Solar Greenhouse. *Intelligent Automation & Soft Computing*, 26(5), 989–1003. <https://doi.org/10.32604/iasc.2020.010130>
- Zhang, Y., Chen, X., Geng, S., & Zhang, X. (2025). A review of soil waterlogging impacts, mechanisms, and adaptive strategies. *Frontiers in Plant Science*, 16. <https://doi.org/10.3389/fpls.2025.1545912>

Evaluation of the Accuracy of Different Machine Learning Algorithms in Predicting Greenhouse Cucumber Crop Evapotranspiration

Morteza Khoshsimaie Chenar¹ | Hamideh Noory² | Abdolmajid Liaghat^{3*} | Forouzandeh Soltani Salehabadi⁴ | Babak Motesharezadeh⁵

1. Department of Irrigation and Reclamation Engineering, Faculty of Agriculture, College of Agriculture and Natural Resources, University of Tehran, Karaj, Iran, Email: Khoshsima.mortaza@ut.ac.ir
2. Department of Irrigation and Reclamation Engineering, Faculty of Agriculture, College of Agriculture and Natural Resources, University of Tehran, Karaj, Iran, Email: hnoory@ut.ac.ir
3. Corresponding Author*, Department of Irrigation and Reclamation Engineering, Faculty of Agriculture, College of Agriculture and Natural Resources, University of Tehran, Karaj, Iran, Email: aliaghat@ut.ac.ir
4. Department of Horticultural Science and Landscape Architecture, Faculty of Agriculture, College of Agriculture and Natural Resources, University of Tehran, Karaj, Iran, Email: soltanyf@ut.ac.ir
5. Department of Soil Science, Faculty of Agriculture, College of Agriculture and Natural Resources, University of Tehran, Karaj, Iran, Email: moteshare@ut.ac.ir

Abstract

In this study, the crop evapotranspiration (ET_c) of greenhouse cucumber was investigated under controlled greenhouse conditions during two growing periods: autumn – winter 2022 and spring – summer 2023, at the College of Agriculture and Natural Resources, University of Tehran. The objective was to model and accurately predict ET_c . Daily soil moisture was measured using TDR sensors installed at a depth of 0 – 30 cm and the soil water balance equation was applied. Reference evapotranspiration (ET_o) was measured using a grass micro-lysimeter with an accuracy of 2 grams. The input data for modeling included 10 meteorological variables recorded during the growing period: days after transplanting, minimum, maximum, and average temperature, minimum, maximum, and average relative humidity, solar radiation, and reference evapotranspiration. Pearson correlation analysis revealed that days after transplanting, average temperature, and solar radiation had the strongest positive correlations with ET_c . To predict ET_c , six machine learning algorithms were implemented in Python: Principal Component Regression (PCR), Partial Least Squares Regression (PLS), Support Vector Machine (SVM), Random Forest (RF), Gradient Boosting (GB) and Extreme Gradient Boosting (XGB). Hyperparameter optimization was conducted using the Tree-structured Parzen Estimator (TPE) algorithm from the Optuna library with 200 iterations. Model performance was evaluated based on R^2 , RMSE, MAE, and NSE metrics using five-fold cross-validation with three repetitions. The modeling results indicated that the GB algorithm achieved the highest accuracy and performance, with average R^2 , RMSE, MAE, and NSE values of 0.90, 0.59 mm/day, 0.41 mm/day, and 0.89, respectively. Following GB, the XGB, RF, and SVM models also performed well, with no statistically significant differences compared to GB. SHAP analysis, used as a model interpretability tool, revealed that days after transplanting, reference evapotranspiration, and solar radiation were the most influential features in predicting ET_c . Overall, the results demonstrated that tree-based machine learning algorithms can serve as accurate tools for forecasting the water requirements of greenhouse cucumber and can play a key role in optimizing irrigation management and improving water use efficiency.

Keywords: Reference evapotranspiration, SHAP analysis, Soil water balance, XGBoost algorithm