Journal of Research in Mechanics of Agricultural Machinery/Vol. 14 /No.3 /Serial No. 36/ Autumn 2025

Journal of Research in Mechanics of Agricultural Machinery

Shahrekord University

Research Paper

Comparison of machine learning models in predicting the yield of Iranian greenhouse products

Khashayar Zare[1], Ahmad Hosseinnejad [2*]

[1]M.Sc, Department of Agricultural Machinery Engineering, Faculty of Agriculture, College of Agriculture and Natural Resources, University of Tehran, Karaj, Iran.

[2]Assistant Professor, Department of Environmental Engineering, Graduate Faculty of Environment, University of Tehran, Tehran, Iran

Article information

Received: 2025/03/16

Accepted: 2025/07/09

Keywords:
Machine learning, greenhouse crop performance, Random Forest, Decision Tree, Support Vector Machines, Multi-Layer Perceptron

***Corresponding author:**
ah.hosseinnejad@ut.ac.ir

ORCID:
0000-0003-2664-9433

Extended Abstract

Introduction

The increasing scarcity of natural resources and the growing demand for food security have driven agriculture toward innovative production strategies. Among them, greenhouse cultivation has emerged as a sustainable solution, offering up to ten times higher productivity while reducing water consumption to nearly 10% compared with open-field farming. However, its higher investment cost underscores the importance of resource optimization and accurate performance prediction. In recent years, the integration of big data and machine learning (ML) techniques has transformed agricultural management by enabling precise forecasting and decision-making. Unlike traditional statistical models, ML algorithms can handle nonlinear interactions and identify hidden patterns, thereby improving prediction accuracy. In greenhouse systems, models such as artificial neural networks, support vector machines, and random forests have been successfully applied to predict climatic factors and crop yields. This study aims to develop an ML-based model to predict the performance of greenhouse crops in Iran, thereby supporting sustainable agriculture and informed policy-making.

Method

This study developed regression machine learning models to predict greenhouse crop yields (tons per hectare) across various provinces and counties, considering crop types and cultivation area. Data were collected from official agricultural yearbooks (2002–2023) including 12,426 records covering 32 provinces, 474 counties, and 33 crop types. The dataset was preprocessed with label encoding for categorical variables and normalization for numerical features using MinMaxScaler and manual scaling.

https://dx.doi.org/10.22034/jrmam.2025.15147.746

Data were split into 90% training and 10% testing sets. Four regression algorithms were implemented: Support Vector Regression (SVR) with RBF kernel, Decision Tree (max_depth=128, min_samples_split=64), Multi-layer Perceptron (two hidden layers with 64 and 32 neurons), and Random Forest (200 trees, max_depth=64). Models were tuned manually for hyperparameters. Performance was evaluated using R^2, MSE, MAE, RMSE, and standard deviation, and statistical significance was assessed by one-way ANOVA.

Results

After training and evaluating the models, the Random Forest (RF) model demonstrated the best performance with an R^2 of 0.9598 and the lowest Mean Squared Error (MSE) of 15.6×10^6, indicating highly accurate predictions. The RF model's complexity, involving 200 trees and a max depth of 64, resulted in a large model size (184,488 KB) but ensured robust handling of complex, nonlinear relationships in the data. The prediction plot shows excellent alignment between actual and predicted values, confirming its superior accuracy.

The Multi-layer Perceptron (MLP) model achieved the second-best performance with an R^2 of 0.8878 and MSE of 27.3×10^6, outperforming SVR and Decision Tree (DT) in accuracy. Its architecture, with two hidden layers comprising 64 and 32 neurons, enabled the learning of nonlinear data patterns, although some prediction errors persisted due to the relatively shallow network depth. The model size was moderate (85 KB), reflecting the trade-off between complexity and computational efficiency. The DT model showed reasonable accuracy with an R^2 of 0.8864 but recorded a higher MSE of 44.3×10^6, indicating some inaccurate predictions. Due to its simple tree-based structure, it had the smallest model size (55 KB) and the fastest computation times, making it suitable for scenarios that prioritize speed and smaller storage over absolute accuracy. The Support Vector Regression (SVR) achieved the lowest accuracy, with R^2 of 0.8546 and the highest MSE of 33.6×10^6. While SVR can model nonlinear data effectively, its performance in this case was limited by its sensitivity to hyperparameter tuning. The model size was moderate (182 KB), but the prediction plots showed greater divergence from actual values compared to those of other models. Error metrics, including Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and standard deviation (STD), further confirmed the RF model's strong predictive stability with the lowest MAE and RMSE values. Statistical analysis using one-way ANOVA revealed no significant difference between RF, DT, and actual values at the 90% confidence level, while SVR showed statistically significant deviation, suggesting lower reliability. Two-way ANOVA confirmed that RF and DT had similarly high accuracy, statistically outperforming MLP and SVR.

In conclusion, RF is the most accurate and stable model for predicting greenhouse crop yield, though at the cost of higher computational demands and model size. DT offers a faster and smaller alternative with acceptable accuracy. MLP and SVR lag behind in performance, likely due to less optimal model complexity and parameter tuning. These findings suggest that RF is suitable for detailed predictive tasks and DT when computational resources are limited.

Conclusions

Greenhouses are recognized as a sustainable agricultural method, with yield performance being a key indicator of efficiency. This study evaluated four machine learning models MLP, SVR,

Decision Tree (DT), and Random Forest (RF) to predict greenhouse crop yields across Iran using geographic, crop type, and cultivation area data. Among them, RF showed the best performance (R^2 = 0.9598, MSE = 15.6×10^6) by effectively modeling complex relationships. DT also performed well, but it had some large prediction errors, resulting in a higher MSE. MLP and SVR exhibited weaker accuracy and significant deviations from actual values. These findings suggest that machine learning models, especially RF, can serve as powerful tools for policymakers and stakeholders in optimizing crop selection, resource management, and yield prediction. The developed RF model enables precise, location-based forecasting that can guide strategic agricultural planning and improve greenhouse productivity nationwide.

Acknowledgements
The authors sincerely extend their gratitude the Sabz Payesh Afra Watergy Systems Optimizers Company, for their kind cooperation and invaluable assistance in facilitating the collection of the necessary data for this study.

Author Contributions
Khashayar Zare: Methodology, Conceptualization, Writing, Data gathering, Data analysis
Ahmad Hosseinnejad: Methodology, Conceptualization, Supervision, Data analysis

Data Availability Statement
"Not applicable".

Ethical Considerations
This section states ethical approval details (e.g., Ethics Committee, ethical code) and confirms adherence to ethical standards, including avoidance of data fabrication, falsification, plagiarism, and misconduct.

Conflict of Interest
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper

Funding Statement
The authors received no specific funding for this research

How to cite this paper:
Zare, K., Hosseinnejad, A. (2025). Comparison of machine learning models in predicting the yield of Iranian greenhouses products. Journal of Research in Mechanics of Agricultural Machinery. 36: 21-37. https://dx.doi.org/ 10.22034/jrmam.2025.15147.746

نشریه پژوهش‌های مکانیک ماشین‌های کشاورزی جلد ۱۴/ شماره ۳/ پیاپی ۳۶/ فصل پاییز (۱۴۰۴)

نشریه پژوهش‌های مکانیک ماشین‌های کشاورزی

دانشگاه شهرکرد

مقاله پژوهشی

مقایسه مدل‌های یادگیری ماشین در پیش‌بینی میزان عملکرد محصولات گلخانه‌ای ایران

خشایار زارع[۱]، احمد حسین‌نژاد[۲*]

۱- کارشناسی ارشد، گروه مهندسی ماشین‌های کشاورزی، دانشکده کشاورزی، دانشکدگان کشاورزی و منابع طبیعی، دانشگاه تهران، کرج، ایران.

۲- استادیار، گروه مهندسی محیط زیست، دانشکده تحصیلات تکمیلی محیط زیست، دانشگاه تهران، تهران، ایران.

اطلاعات مقاله

تاریخ دریافت: ۱۴۰۳/۱۲/۲۶

تاریخ پذیرش: ۱۴۰۴/۰۴/۱۸

واژه‌های کلیدی:

یادگیری ماشین، عملکرد محصولات گلخانه‌ای

جنگل تصادفی

درخت تصمیم‌گیری

ماشین بردار پشتیبانی

پرسپتون چندلایه

***پست الکترونیکی نویسنده مسئول:**

ah.hosseinnejad@ut.ac.ir

ORCID: ۰۰۰۰-۰۰۰۳-۲۶۶۴-۹۴۳۳

چکیده

با افزایش روزافزون جمعیت، نیاز به تأمین پایدار مواد غذایی اهمیت بیش‌تری یافته است. یکی از راهکارهای افزایش بهره‌وری در تولیدات گلخانه‌ای، بهره‌گیری از روش‌های پیشرفته یادگیری ماشین در تحلیل داده‌های گلخانه‌ها است. در این پژوهش، به‌منظور پیش‌بینی عملکرد محصولات گلخانه‌ای در سطح کشور ایران، از داده‌های آماری وزارت جهاد کشاورزی شامل اطلاعات استان‌ها، شهرستان‌ها، نوع محصولات و سطح زیر کشت و میزان تولیدات آن‌ها استفاده شد. پس از پیش‌پردازش داده‌ها، چهار مدل یادگیری ماشین شامل شبکه عصبی پرسپترون چندلایه (MLP)، رگرسیون بردار پشتیبان (SVR)، درخت تصمیم (DT) و جنگل تصادفی (RF) به‌کار گرفته شدند. مدل‌ها با استفاده از نسبت ۹۰٪ داده آموزشی و ۱۰٪ داده آزمایشی ارزیابی شدند. دقت عملکرد مدل‌ها با استفاده از معیارهای آماری مختلف و آزمون آنالیزتحلیل واریانس یک طرفه بررسی شد. نتایج نشان داد مدل RF با R^2 معادل ۰/۹۵۹۸ و MSE برابر با $۱۵/۶ \times ۱۰^{۶}$ دقیق‌ترین پیش‌بینی را ارائه کرده و قادر به مدل‌سازی الگوهای پیچیده بین داده‌ها بوده است، اما این مدل دارای حجم زیادی بود. مدل DT با وجود برخی خطاها در نقاط خاص، با R^2 معادل ۰/۸۸۶۴ و MSE برابر با $۴۴/۳ \times ۱۰۶$، طبق آزمون آماری نتایج معناداری نداشت و نتایج پیش‌بینی نزدیک به مقادیر واقعی داشت. در مقابل، مدل‌های SVR و MLP با R^2های به‌ترتیب ۰/۸۵۴۶ و ۰/۸۸۷۸، اختلاف معناداری بین نتایج پیش‌بینی و واقعی داشتند. نتایج تأکید می‌کند انتخاب درست الگوریتم متناسب با ویژگی‌های داده، نقش مهمی در تصمیم‌گیری‌های کشاورزی دارد. مدل RF معرفی‌شده در این پژوهش، الگویی مناسب برای پیش‌بینی عملکرد گلخانه‌ها در سطح ملی است.

https://dx.doi.org/10.22034/jrmam.2025.15147.746

مقدمه

کمبود فزآینده منابع و نیاز روزافزون به تأمین امنیت غذایی، بشر را به سمت توسعه راهکارهای نوین در تولید محصولات کشاورزی سوق داده است. فصلی بودن تولیدات کشاورزی و ناهماهنگی زمانی بین عرضه و تقاضا، همراه با ضرورت افزایش بهره‌وری، امنیت غذایی، و توسعه صادرات، موجب شده تا کشت گلخانه‌ای به عنوان راه‌کاری پایدار در دستور کار سیاست‌گذاران و کشاورزان قرار گیرد (پیفه، ۱۳۸۷). گلخانه‌ها با افزایش بازده تولید تا ۱۰ برابر و کاهش مصرف آب به حدود ۱۰ درصد نسبت به کشت سنتی، نقش حیاتی در پیوند میان مدیریت منابع محدود (آب، انرژی، زمین) و تولید پایدار ایفا می‌کنند(Roshandel, 2020 & Esmaeli). با این حال، سرمایه‌گذاری بالاتر در واحد سطح گلخانه‌ها نسبت به کشت باز، لزوم بهینه‌سازی همزمان با مصرف منابع و تحلیل اقتصادی دقیق را پررنگ می‌سازد که در این بین افزایش عملکرد گلخانه ها در واحد سطح رونق اقتصادی این بخش را همراه دارد (Hosseinnejad *et al.,* 2023).

توسعه کلان‌داده‌ها و به کارگیری فناوری‌های نوین، به‌ویژه هوش مصنوعی (AI) و الگوریتم‌های یادگیری ماشین ، تحولی اساسی در روش‌های پیش‌بینی و برنامه‌ریزی برای مدیریت بهینه سامانه‌های کشاورزی ایجاد کرده است. مدل‌های ریاضی سنتی مانند رگرسیون چندمتغیره و مدل‌های مکانیکی، هرچند در شرایط استاندارد و با داده‌های دقیق محیطی دقت قابل قبولی ارائه می‌دهند (Neetu Rani *et al.,* 2022)، اما در مواجهه با تعاملات غیرخطی و پیچیدگی‌های سامانه‌های واقعی و با وجود تعدد عامل‌های مختلف که امروزه بر سامانه تأثیر گذار هستند، دیگر کارایی لازم را ندارند. در مقابل، الگوریتم‌های یادگیری ماشین (ML) با توانایی پردازش داده‌های حجیم و شناسایی الگوهای پنهان، دقت پیش‌بینی را به‌طور چشم‌گیری بهبود بخشیده‌اند (Maya Gopal & Bhargavi, 2019). به عنوان مثال، شبکه‌های عصبی مصنوعی (ANN) در پیش‌بینی عامل‌های کلیدی گلخانه مانند دما، رطوبت، و غلظت CO_2 با خطای پایینی موفق عمل کرده‌اند (Singh & Tiwari, 2017). این مدل‌ها با بهینه‌سازی مصرف انرژی و کاهش هزینه‌ها، پایه‌های کشاورزی هوشمند و پایدار را تقویت می‌کنند(Kim, Park & Ryu, 2018).

در سطح گلخانه‌ها، مدل‌های هوش مصنوعی مانند شبکه‌های عصبی پیچشی (CNN)، ماشین بردار پشتیبان (SVM)، و جنگل تصادفی (RF) با تحلیل داده‌های اقلیمی، خاک و مدیریتی را با دقت بالا امکان‌پذیر ساخته‌اند (Thomas, Ayalew & Cagatay, 2020). مطالعاتی نظیر پژوهش کایال و ساکینا در سال ۲۰۲۲ (Kulyal & Saxena, 2022) نشان داده‌اند که الگوریتم Random Forest با شناسایی روابط پیچیده بین متغیرهایی مانند نوع خاک، بارندگی، و دما، دقت پیش‌بینی را افزایش می‌دهد. هم‌چنین، ادغام الگوریتم‌های ژنتیک با درخت‌های رگرسیون (GenSRT) در بهینه‌سازی الگوی کشت و توزیع منابع، سود خالص را بهبود بخشیده است (Frausto-Solis, Gonzalez-Sanchez & Larre, 2009).

گلخانه‌ها به دلیل پایش شرایط محیطی مانند دما و رطوبت نوسانات عملکردی کم‌تری نسبت به کشت سنتی دارند و این پایداری نسبی شرایط محیطی گلخانه‌ها (مانند دما و رطوبت پایش‌شده)، امکان کاهش تنش‌های اقلیمی و دستیابی به پیش‌بینی‌های دقیق‌تر را فراهم می‌سازد.

از این رو از یادگیری ماشین برای پیش‌بینی و پایش دما و رطوبت در کارهای مختلف بهره برده شده است. در سال‌های اخیر، الگوریتم‌های یادگیری ماشین به‌طور گسترده برای پیش‌بینی و پایش شرایط اقلیمی در گلخانه‌ها به‌کار گرفته شده‌اند. برای مثال، مطالعه‌ای توسط Yang (*et al.,* 2023) از شبکه‌های عصبی FAM-LSTM چندبعدی برای پیش‌بینی دما گلخانه بهره برده و نتایج دقیقی را در بازه‌های زمانی مختلف ارائه داده است. هم‌چنین، پژوهش دیگری با استفاده از الگوریتم یادگیری ماشین CB-ELM توانسته است دقت بالایی در پیش‌بینی دما و رطوبت گلخانه‌های خورشیدی نشان دهد و آن را نسبت به سایر روش‌ها مؤثرتر معرفی کرده است (Zou *et al.,* 2017). در زمینه سامانه‌های پایش اقلیم هوشمند، پژوهش جدیدی در دانشگاه گیلان از ترکیب چندین مدل یادگیری ماشین مانند XGBoost، Random Forest و SVM برای پیش‌بینی رطوبت داخلی گلخانه استفاده کرده و از روش *Stacking* برای ترکیب آن‌ها بهره برده است. این مدل ترکیبی عملکرد بسیار دقیقی با ضریب تعیین (R^2) بیش از ۰/۹۶ به‌دست آورده است (Rezaei Melal, Aminian & Shekarian, 2024).

با وجود پیشرفت‌های چشم‌گیر، چالش‌هایی نظیر هزینه‌های بالای پیاده‌سازی، نیاز به تخصص فنی، و کمبود داده‌های جامع محلی همچنان پابرجاست (Maraveas,

2023)، با این حال آمار نامه سالانه بخش کشاورزی و گلخانه‌ها را که جهاد کشاورزی ارائه می‌دهد منبع قابل استفاده برای داده‌های مدل‌های یادگیری است. به ویژه در ایران، علی‌رغم تجربه گسترده گلخانه‌داران در مدیریت محصولات خاص، پژوهشی جامع با تمرکز بر پیش‌بینی عملکرد محصولات گلخانه‌ای مبتنی بر یادگیری ماشین انجام نشده است.

پژوهش‌های پیشین عمدتاً بر جنبه‌های محدودی مانند پیش‌بینی تبخیر و تعرق (Pandorfi *et al.,* 2016) یا مدیریت آفات (Palani *et al.,* 2023) متمرکز بوده‌اند. هر چند که مدیریت کشت و پایش دما و رطوبت بهبود فرایند کشت را همراه دارد، ولی در نهایت معیار ارزیابی برای کارایی گلخانه عملکرد تولید است. این در حالی است که تحلیل هم‌زمان موقعیت مکانی و سطح زیر کشت برای پیش‌بینی عملکرد محصولات متنوع گلخانه‌ای در مقیاس ملی مغفول مانده است و با این حال، علی‌رغم تجربه گسترده گلخانه‌داران، تاکنون پژوهشی جامع برای پیش‌بینی عملکرد محصولات گلخانه‌ای در ایران بر اساس داده‌های مکانی و اقلیمی، نوع محصولات و سطح زیر کشت آن‌ها انجام نشده است.

با توجه به اهمیت توسعه گلخانه‌ها در کشور، به عنوان کشاورزی پایدار، پژوهش حاضر با هدف توسعه یک مدل مبتنی بر یادگیری ماشین برای پیش‌بینی عملکرد محصولات گلخانه‌ای به عنوان یکی از معیارهای اصلی کارایی گلخانه‌ها، در ایران انجام شده است. در این پژوهش، با بهره‌گیری از الگوریتم‌های یادگیری ماشین، مدلی جامع طراحی شده است که با ادغام داده‌های مکانی، اقلیمی، محصول کشت شده و سطح زیر کشت، عملکرد محصولات را در یک گلخانه در یک موقعیت مکانی را پیش‌بینی می‌کند. نتایج این مطالعه می‌تواند به کاهش خطر سرمایه‌گذاری، افزایش سودآوری و تسهیل تصمیم‌گیری سیاست‌گذاران کمک و زمینه‌ساز توسعه کشاورزی پایدار در ایران شود.

مواد و روش‌ها

در این پژوهش با توجه به ضرورت پیش‌بینی عملکرد محصولات گلخانه‌ای در سطح استان‌ها و شهرستان‌های مختلف و نوع محصولات گلخانه‌ای و سطح زیر کشت مدل‌های یادگیری ماشین رگرسیونی توسعه داده شد که توانایی پیش‌بینی مقدار عملکرد (تن در هکتار) محصولات را دارد. نوع الگوریتم‌ها و مراحل انجام پژوهش حاضر به صورت زیر ارائه و در ادامه مراحل آن تشریح شده است.

۱. جمع‌آوری و آماده‌سازی داده‌ها

داده‌های این پژوهش شامل نام استان، شهرستان، نام محصولات گلخانه‌ای، سطح زیر کشت و عملکرد محصولات در نظر گرفته شده است. این اطلاعات از آمارنامه‌های رسمی وزارت جهاد کشاورزی (آمارنامه کشاورزی، وزارت جهاد کشاورزی، ایران، ۱۳۸۱-۱۴۰۲) استخراج و در قالب یک فایل Excel سازمان‌دهی شدند. مجموعه داده شامل ۱۲,۴۲۶ ردیف جمع‌آوری گردیده شد که هر ردیف بیان‌گر اطلاعات مرتبط با یک محصول در یک منطقه خاص است. داده‌ها به‌صورت خام وارد شده و برای انجام مدل‌سازی، مراحل پیش‌پردازش بر روی آن انجام گرفت. داده‌ها شامل ۳۲ استان، ۴۷۴ شهرستان و ۳۳ محصول مختلف به همراه سری داده‌های سالیانه آمارنامه‌ها بود. داده‌ها از سال ۱۳۸۱ که در آمارنامه‌های کشور موجود بودند در فایل Excel ذخیره شدند و در محیط Google Drive بارگذاری شد.

بعد از جمع‌آوری، داده‌ها برای ساخت یک مدل رگرسیونی آماده شدند. به منظور تسریع فرآیند آموزش از بستر Google Colab و نسخه python ۳.۱۰.۱۲ استفاده گردید.

۲. پیش‌پردازش داده‌ها

به دلیل استفاده از بستر Google Colab برای عملیات پیش پردازش و ساخت مدل رگرسیونی ابتدا فایل داده‌ها درGoogle Drive بارگذاری شد. سپس با نصب کردن Drive در بستر Google Colab فایل داده‌ها با استفاده از بسته pandas با نسخه ۲.۰.۳ بارگیری شد و داده‌ها برای عملیات پیش‌پردازش آماده شدند و سایر مراحل در محیط Google Colab با استفاده از CPU این محیط (Intel(R) (Xeon(R) Platinum 8259CL CPU @ 2.50GHz انجام گردید.

به‌منظور جلوگیری از سوگیری آموزش مدل (آموزش تنها برروی یک سری از داده‌ها)، داده‌ها به‌صورت تصادفی مرتب شدند و داده‌های دسته‌ای علامت‌گذاری شدند. ستون‌های نام استان‌ها، نام شهرستان‌ها و نام محصولات که ماهیت دسته‌ای و متنی داشتند، با استفاده از روش علامت-گذاری برچسبی (Label Encoding) به مقادیر عددی تبدیل شدند. این تبدیل به مدل اجازه می‌دهد تا بتواند

مقادیر کیفی را تحلیل کند. برای علامت‌گذاری برچسبی از عملگر LabelEncoder موجود در بسته keras استفاده شد.

نرمال‌سازی داده‌های عددی یکی از مهم‌ترین پیش‌پردازش‌هایی است که سبب افزایش سرعت آموزش مدل می‌شود. این پیش‌پردازش ویژگی‌های عددی مانند سطح زیر کشت (هکتار)، میزان عملکرد را با استفاده از روش‌های MinMaxScaler و Normalize ویژگی‌های عددی در بازه مشخصی نرمال‌سازی شدند تا داده‌ها مقیاس یکسانی داشته باشند و تأثیر ویژگی‌های با مقیاس بزرگ‌تر کاهش یابد. داده‌های میزان عملکرد به صورت دستی Normalize شدند. بدین معنا که تمامی اعداد موجود در ستون داده‌های عملکرد بر مقدار بیشینه آن‌ها (۲۱۶۶۶/۶۷) تقسیم شدند اما عامل سطح زیر کشت با استفاده از عملگر MinMaxScaler موجود در بسته keras پیش‌پردازش گردید.

۳. تقسیم‌بندی داده‌ها

پس از پیش‌پردازش داده‌ها، مجموعه داده‌ها به دو بخش آموزش (۹۰٪) و آزمون (۱۰٪) تقسیم شدند. این تقسیم‌بندی با استفاده از روش train_test_split انجام شد تا امکان ارزیابی مدل‌ها بر روی داده‌های نادیده فراهم شود و بر اساس آن بتوان خطای مدل ساخته شده را بررسی کرد.

۴. مدل‌سازی با استفاده از الگوریتم‌های مختلف

در این پژوهش، چهار الگوریتم یادگیری ماشین مختلف رگرسیونی برای پیش‌بینی عملکرد محصولات گلخانه‌ای استفاده شد. به‌منظور دستیابی به بهترین عملکرد برای هر مدل، تنظیم فراعامل‌ها به‌صورت گسترده و با استفاده از روش آزمون و خطا و به صورت دستی انجام شد. در این فرآیند، ترکیب‌های متعددی از فراعامل‌ها مورد ارزیابی قرار گرفتند و با تحلیل دقیق خطاهای حاصل از هر پیکربندی، مقادیر بهینه فراعامل‌ها انتخاب شدند. جزئیات هر مدل در زیر شرح داده شده است.

۴.۱- مدل رگرسیون بردار پشتیبان (*SVR*)

مدل رگرسیون بردار پشتیبان (SVR[1]) مبتنی بر الگوریتم ماشین‌های بردار پشتیبان (SVM[2]) طراحی شده و برای مسائل رگرسیون استفاده می‌شود. SVR تلاش می‌کند تابعی پیدا کند که داده‌ها را با بیش‌ترین دقت ممکن و کم‌ترین خطا مدل‌سازی کند. این روش برای داده‌های پیچیده که توزیع غیرخطی دارند، با استفاده از هسته‌هایی مانند RBF و Polynomial بسیار مفید است. این مدل برای پیش‌بینی داده‌ها در محیط‌های پر نوفه و با داده‌های پیچیده مناسب است و از مفاهیم حداقل‌سازی خطر ساختاری بهره می‌برد (Alex and Schoelkopf, 2004).

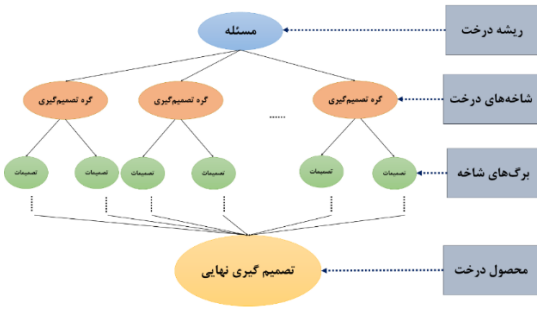
مدل ماشین بردار پشتیبان رگرسیون با استفاده از هسته *RBF* (تابع شعاعی پایه) پیاده‌سازی شد. این مدل برای یادگیری روابط غیرخطی میان ویژگی‌ها و متغیر هدف مناسب است. عامل‌های هسته که به بهینه‌سازی بهتر مدل و آموزش بهتر مدل کمک می‌کند، شامل مقادیر C=۱/۵ و $gamma$=۱/۷ تنظیم شدند. سایر فراعامل‌های مدل برای آموزش بدون تغییر مانند و مقادیر آن‌ها به حالت پیش‌فرض باقی ماندند.

۴.۲- درخت تصمیم (*Decision Tree*)

درخت تصمیم (DT[3]) یک روش یادگیری ماشین تحت نظارت است که برای طبقه‌بندی و رگرسیون استفاده می‌شود. این مدل با تقسیم داده‌ها به زیرمجموعه‌های کوچک‌تر بر اساس ویژگی‌های داده‌های ورودی، ساختاری درختی ایجاد می‌کند که هر گره داخلی نشان‌دهنده یک آزمون روی یک ویژگی، هر شاخه نشان‌دهنده نتیجه آزمون و هر برگ نشان‌دهنده یک طبقه (مسائل طبقه‌بندی) یا مقدار پیش‌بینی (مسائل رگرسیونی) شده است. از مزایای این روش می‌توان به سادگی تفسیر، توانایی کار با داده‌های عددی و دسته‌ای و عدم نیاز به نرمال‌سازی داده‌ها اشاره کرد. با این حال، مدل انعطاف‌پذیر بوده و قابلیت تفسیر بالایی دارد، اما در صورت عدم تنظیم صحیح، ممکن است دچار بیش‌برازش (Overfitting) شود (Everitt, 2005).

برای ساخت مدل درخت تصمیم به صورتی که مدل دچار بیش‌برازش و یا کم برازشی نشود با حداکثر عمق (*max_depth*) ۱۲۸ و حداقل تعداد نمونه‌های لازم برای انشعاب (*min_samples_split*) ۶۴ طراحی شد تا توانایی پیش‌بینی مقادیر را به بهترین شکل داشته باشد و سایر فراعامل‌های مدل برای آموزش به صورت پیش‌فرض باقی ماندند (شکل ۱).

[1]-Support Vector Regression
[2]-Support Vector Machines
[3]- Decision Tree



شکل ۱- ساختار یک مدل نمونه DT که شامل گره‌های تصمیم گیری (شاخه) و تصمیمات مختلف (برگ‌ها)

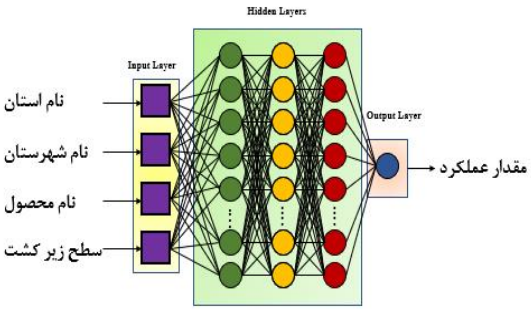
۴.۳- شبکه عصبی چندلایه (*MLP*)

یک نوع شبکه عصبی مصنوعی است که از چندین لایه شامل لایه ورودی، لایه‌های مخفی و لایه خروجی تشکیل شده است. MLP[1] با استفاده از الگوریتم‌های یادگیری مانند بازپخش پسرو (Backpropagation)، روابط غیرخطی پیچیده میان داده‌ها را مدل‌سازی می‌کند و معمولاً در مسائل پیش‌بینی و طبقه‌بندی استفاده می‌شود. شبکه‌های پرسپترون چندلایه به دلیل توانایی در یادگیری روابط غیرخطی، کاربرد گسترده‌ای در مسائل پیش‌بینی و طبقه‌بندی دارند (Taki *et al.*, 2016).

شبکه عصبی چندلایه طراحی‌شده در این پژوهش دارای دو لایه مخفی با تعداد نرون‌های (*hidden_layer_sizes*) ۶۴ و ۳۲ بود. این معماری با استفاده از روش بهینه‌سازی Backpropagation آموزش داده شد. شبکه با تعداد تکرارهای حداکثر (*max_iter*) ۱۰۰۰ و مقدار اول تصادفی برای وزن‌ها پیاده‌سازی شد. برای استفاده از این روش داده‌های نام استان‌ها، شهرستان‌ها، نوع محصول و سطح زیر کشت آن‌ها به عنوان ورودی به مدل داده می‌شود و در مدل آموزش داده می‌شود و یک مقدار را به عنوان خروجی پیش‌بینی می‌کند (شکل ۲).

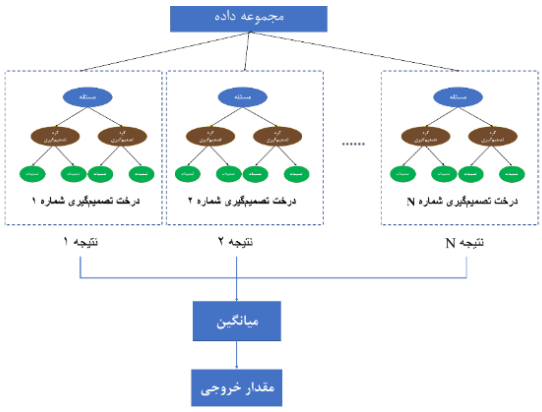
۴.۴- جنگل تصادفی (*Random Forest*)

جنگل تصادفی (RF[2]) یک روش یادگیری ماشین تحت نظارت است که از ترکیب چندین درخت تصمیم برای بهبود دقت و کاهش بیش‌برازش (Overfitting) استفاده می‌کند. این مدل با ایجاد مجموعه‌ای از درخت‌های تصمیم که به صورت مستقل آموزش داده می‌شوند، نتایج را با روش‌هایی مانند رأی اکثریت (برای طبقه‌بندی) یا میانگین‌گیری (برای رگرسیون) ترکیب می‌کند. هر درخت در جنگل تصادفی بر روی یک زیرمجموعه تصادفی از داده‌های آموزشی (با جایگزینی) و یک زیرمجموعه تصادفی از ویژگی‌ها آموزش می‌بینند، که این امر باعث افزایش تنوع و کاهش وابستگی به داده‌های خاص می‌شود (Baştanlar & Özuysal, 2013). این مدل در پیش‌بینی‌های دقیق و پایدار، به‌ویژه در مسائل داده‌های پیچیده و حجیم، عملکرد فوق‌العاده‌ای دارد. این مدل با ترکیب درخت‌های تصمیم به‌شکل تصادفی، پایداری و دقت بالایی در پیش‌بینی ارائه می‌دهد و برای داده‌های پیچیده بسیار مناسب است (Chahidi *et al.*, 2021).



شکل ۲- مدل MLP که شامل چهار ورودی و یک خروجی (رگرسیون)

مدل جنگل تصادفی با ۲۰۰ درخت (*n_estimators*) و حداکثر عمق (*max_depth*) ۶۴ طراحی شد. این مدل با ترکیب پیش‌بینی‌های چندین درخت تصمیم، خطای کلی مدل را کاهش می‌دهد و از بیش‌برازش جلوگیری می‌کند (شکل ۳).



شکل ۳- ساختار مدل RF شامل ۲۰۰ درخت تصمیم‌گیری با عمق ۶۴

[1]-multilayer perceptron

[2]-Random Forest

۵. ارزیابی مدل‌های یادگیری ماشین

برای ارزیابی میزان تطابق خروجی‌های پیش‌بینی‌شده با مقادیر واقعی، از معیار ضریب تعیین (R^2) استفاده شد. این معیار نشان می‌دهد که چه نسبتی از واریانس داده‌های واقعی توسط مدل توضیح داده شده است. مقدار R^2 بین صفر تا یک متغیر است و هرچه به عدد یک نزدیک‌تر باشد، نشان‌دهنده عملکرد بهتر مدل است. برای محاسبه این معیار از تابع آماده r2_score در کتابخانه scikit-learn استفاده شد و مقدار آن از رابطه (۱) به دست می‌آید که در این رابطه y_i مقدار واقعی، $\hat{y}$ مقدار پیش‌بینی، $\bar{y}$ میانگین مقادیر واقعی و n تعداد نمونه‌ها است. برای سنجش دقت پیش‌بینی مدل، از میانگین مربعات خطا (MSE) استفاده شد. این معیار میزان متوسط مجذور اختلاف بین مقادیر واقعی و پیش‌بینی‌شده را اندازه‌گیری می‌کند و مقدار کم‌تر آن نشان‌دهنده دقت بیش‌تر مدل است. محاسبه این معیار با تابع mean_squared_error از کتابخانه scikit-learn انجام گرفت و مقدار آن از رابطه (۲) به دست می‌آید.

$$R^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{\sum_{i=1}^{n}(\bar{y} - y_i)^2} - 1 \quad (۱$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2 \quad (۲$$

برای بررسی خطای مطلق میان مقادیر پیش‌بینی‌شده و واقعی، از میانگین قدر مطلق خطا (MAE) استفاده شد. این معیار میانگین فاصله مطلق بین پیش‌بینی‌ها و واقعیت را بیان می‌کند و درک ساده‌تری نسبت به MSE دارد و مقدار عددی آن با استفاده از رابطه (۳) بدست می‌آید.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - y_i| \quad (۳$$

برای اندازه‌گیری دقت پیش‌بینی مدل به واحد اصلی متغیر هدف، از ریشه میانگین مربعات خطا (RMSE) استفاده شد. این معیار همان MSE است که از آن جذر گرفته شده و بنابراین تفسیر ساده‌تری دارد و از طریق رابطه (۴) محاسبه می‌شود.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - y_i)^2} \quad (۴$$

برای بررسی پراکندگی خطاهای مدل یا مقادیر پیش‌بینی‌شده، از انحراف معیار (Std) استفاده شد. این معیار نشان‌دهنده میزان پراکندگی داده‌ها نسبت به مقدار میانگین است. مقدار کم‌تر Std بیانگر ثبات و دقت بیش‌تر مدل است که با استفاده از رابطه (۵) محاسبه می‌شود.

$$\sigma = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(x_{\hat{i}} - x)^2} \quad (۵$$

این معیارها دقت و توانایی مدل‌ها در پیش‌بینی داده‌ها را به صورت عددی و عینی ارزیابی می‌کنند. علاوه بر این، حجم مدل‌ها نیز به عنوان یکی از فاکتورهای مهم در کارایی محاسباتی و پیچیدگی مدل مورد بررسی قرار می‌گیرد. برای ارزیابی معناداری آماری تفاوت عملکرد مدل‌های یادگیری ماشین، از آزمون تحلیل واریانس یک‌طرفه (One-Way ANOVA) بر روی خطاهای حاصل از پیش‌بینی مدل‌ها استفاده شد. این تحلیل به کمک بسته‌ی scipy.stats در زبان برنامه‌نویسی پایتون پیاده‌سازی گردید. هدف از این آزمون، مقایسه نتایج پیش‌بینی‌شده توسط مدل‌های مختلف با یک‌دیگر و هم‌چنین بررسی تفاوت معنادار بین نتایج پیش‌بینی‌شده و مقادیر واقعی بود.

نتایج و بحث

پس از آموزش و ارزیابی هر مدل، نتایج به دست آمده برای هر مدل به شرح جدول ۱ است:

جدول ۱- ارزیابی دقت و عملکرد مدل‌های ساخته شده

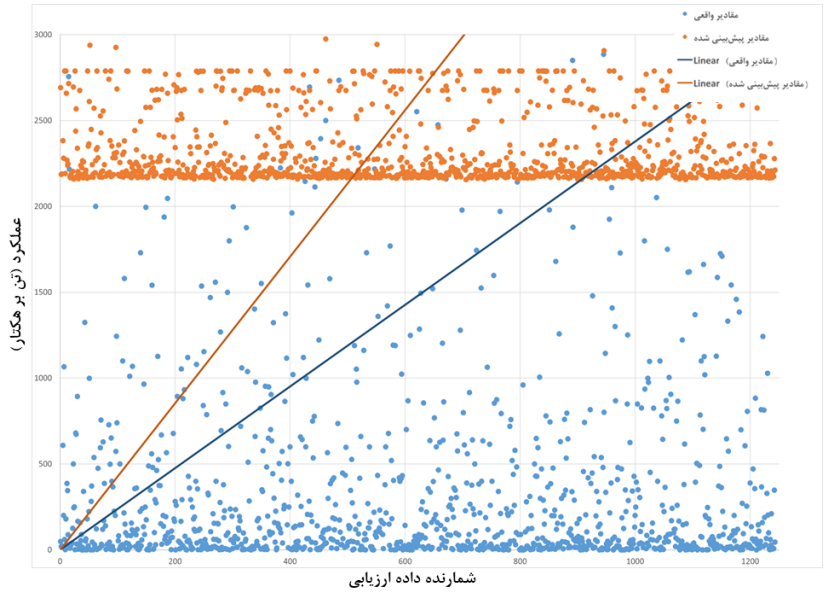
مدل	امتیاز R^2	MSE	حجم مدل (کیلوبایت)
SVR	۰/۸۵۴۶	۳۳/۶ × ۱۰^۶	۱۸۲
MLP	۰/۸۸۷۸	۲۷/۳ × ۱۰^۶	۸۵
RF	۰/۹۵۹۸	۱۵/۶ × ۱۰^۶	۱۸۴۴۸۸
DT	۰/۸۸۶۴	۴۴/۳ × ۱۰^۶	۵۵

مدل SVR در این تحقیق توانست مقدار R^2 برابر با ۰/۸۵۴۶ را کسب کند، هم‌چنین، مقدار MSE برابر با ۱۰^۶× ۳۳/۶ برای این مدل به دست آمد. این نتایج نشان می‌دهند که مدل SVR توانسته پیش‌بینی‌های دقیقی برای داده‌های آزمون ارائه دهد.

اگرچه مدل SVR توانسته پیش‌بینی‌های نسبتاً دقیقی را برای برخی از داده‌ها ارائه دهد، اما هم‌چنان در شبیه‌سازی برخی الگوهای داده‌ای محدودیت‌هایی دارد یکی از دلایل این امر ممکن است به حساسیت بالای مدل SVR به تنظیمات عامل‌ها مانند عامل هسته (kernel)، C و gamma مربوط باشد که در صورت بهینه نبودن، می‌توانند تأثیر منفی بر عملکرد مدل بگذارند. در این پژوهش تلاش شد تا بهترین فراعامل‌ها انتخاب شوند و مدل بهینه‌ترین پیش‌بینی‌ها را ارائه دهد. نمودار مقادیر واقعی و

پیش‌بینی‌شده توسط مدل در شکل ۴ نمایش داده شده است. همان‌طور که در شکل ۴ قابل مشاهده است شیب خطی مقادیر واقعی و شیب خطی مقادیر پیش‌بینی شده متفاوت هستند که می‌تواند نشان دهنده اختلاف زیاد مقادیر پیش‌بینی و واقعی باشد.

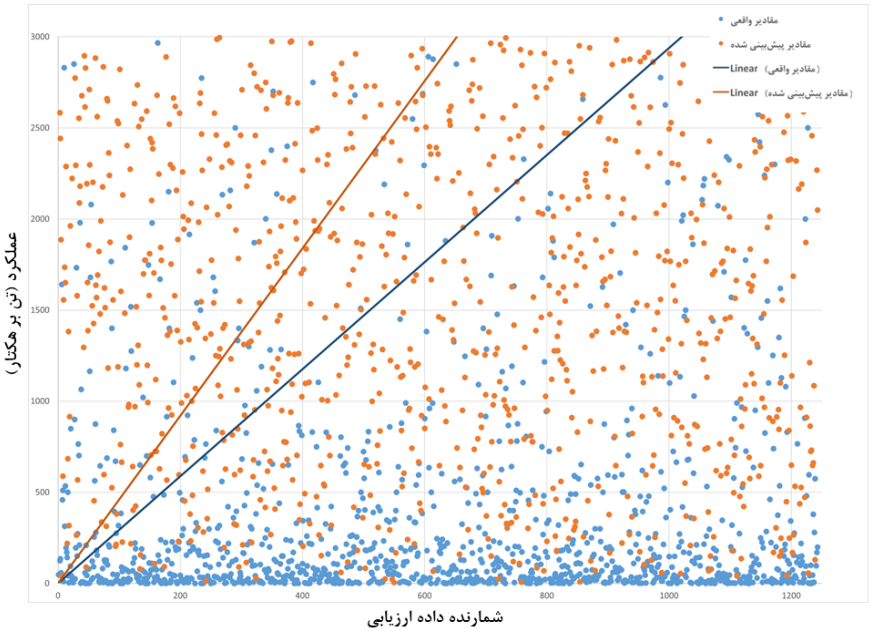
مدل SVR معمولاً به دلیل داشتن الگوریتم‌های پیچیده برای شبیه‌سازی مرزهای تصمیم‌گیری در فضای چندبعدی، حجم نسبتاً بزرگتری نسبت به مدل‌های ساده‌تر دارد. با این حال، حجم ۱۸۲ کیلوبایت برای یک مدل SVR نسبت به برخی مدل‌های دیگر چندان زیاد محسوب نمی‌شود. این مدل برای داده‌هایی که ویژگی‌های غیرخطی دارند، عملکرد خوبی دارد، ولی از لحاظ پیچیدگی محاسباتی معمولاً سنگین‌تر از مدل‌های ساده‌تر است.



شکل ۴- نمودار مقادیر دقیق (نقاط آبی) و پیش‌بینی (نقاط نارنجی) مدل SVR

مدل MLP توانست با R^2 برابر با ۰/۸۸۷۸ عملکرد نسبتاً خوبی را نشان دهد و در مقایسه با سایر مدل‌ها، بیش‌ترین دقت را داشته است. به طور خاص، مقدار MSE برابر با $۲۷/۳ \times ۱۰^{۶}$ نشان‌دهنده این است که مدل MLP توانسته است پیش‌بینی‌هایی با خطای کم‌تری نسبت به مدل SVR ارائه دهد. نمودار مقادیر واقعی و پیش‌بینی‌شده توسط مدل در شکل ۵ نمایش داده شده است. با توجه به شکل ۵ می‌توان دریافت که شیب خطی مقادیر واقعی و شیب خطی مقادیر پیش‌بینی شده متفاوت هستند که می‌تواند نشان دهنده اختلاف زیاد مقادیر پیش‌بینی و واقعی باشد اما در مقایسه با نمودار شکل ۴، می‌توان مشاهده کرد که اختلاف شیب دو خط کاهش یافته است که این امر نشان‌دهنده دقت بالاتر مدل MLP در مقایسه با مدل SVR است. این عملکرد بهتر نسبت به SVR می‌تواند به این دلیل باشد که مدل MLP قادر است روابط پیچیده و غیرخطی بین ویژگی‌های ورودی و عملکرد گلخانه را یاد بگیرد. اما این مدل نیز دارای خطاهایی به مراتب زیادی است و این نیز به دلیل عمق کم‌تر مدل است.

مدل MLP، به دلیل استفاده از چندین لایه از نورون‌ها و وزن‌های متعدد، می‌تواند حجم به‌مراتب بزرگ‌تری نسبت به مدل‌های پایه مانند درخت تصمیم داشته باشد. با این حال، حجم ۸۵ کیلوبایت در مقایسه با مدل‌های پیچیده‌تری مانند جنگل تصادفی یا شبکه‌های عمیق‌تر، مقدار معقولی محسوب می‌شود. مدل MLP به دلیل وجود لایه‌های پنهان و وزن‌های پیچیده، نیاز به محاسبات بیش‌تری دارد، اما در عین حال نسبت به مدل‌های دیگر می‌تواند انعطاف‌پذیری بیش‌تری در یادگیری الگوهای پیچیده‌تری داشته باشد مدل MLP، به‌دلیل داشتن چندین لایه و تعداد زیادی عامل (وزن‌ها)، ممکن است سرعت پردازشی پایین‌تری، به‌ویژه در هنگام آموزش و پیش‌بینی داشته باشد.

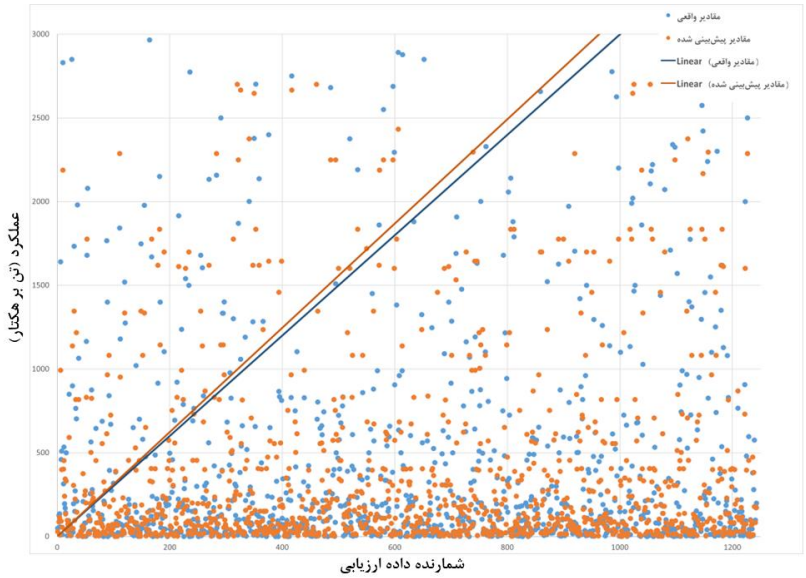


شکل ۵- نمودار مقادیر دقیق (نقاط آبی) و پیش‌بینی (نقاط نارنجی) مدل MLP

مدل DT در این تحقیق با R^2 برابر با ۰/۸۸۶۴ و MSE برابر با $۴۴/۳ \times ۱۰^۶$ عملکرد نسبتاً خوبی داشته است، اما در مقایسه با سایر مدل‌ها مقدار MSE بالاتری دارد. این مسئله به دلیل پیش‌بینی‌های نادرست در برخی داده‌های آزمون رخ داده است که منجر به کاهش مقدار R^2 و افزایش مقدار MSE شده است. هم‌چنین در شکل ۶ مشاهده می‌شود که شیب خطی مقادیر واقعی و پیش‌بینی‌شده نسبتاً به یکدیگر نزدیک هستند، که می‌تواند نشان‌دهنده عملکرد خوب مدل DT در پیش‌بینی است، اما به دلیل چندین پیش‌بینی نادرست در داده‌های ارزیابی، مدل دچار افزایش مقدار MSE و کاهش مقدار R^2 شده است.

مدل درخت تصمیم، به دلیل ساختار ساده خود، حجم کم‌تری نسبت به مدل‌های پیچیده‌تری مانند MLP یا جنگل تصادفی دارد. حجم ۵۵ کیلوبایت برای این مدل منطقی است، زیرا درخت تصمیم یک مدل پایه است که پیش‌بینی‌ها را از طریق تقسیم‌بندی داده‌ها بر اساس ویژگی‌ها و تعیین نقاط تصمیم‌گیری در ساختار درخت انجام می‌دهد. به دلیل ساختار ساده، این مدل معمولاً حجم کمی دارد، اما در مدل‌سازی روابط پیچیده‌تر با محدودیت‌هایی مواجه است، که آن را برای مسائل ساده‌تر مناسب‌تر می‌کند.

مدل‌های درخت تصمیم معمولاً نسبت به مدل‌های پیچیده‌تری مانند MLP و SVR سریع‌تر عمل می‌کنند. به دلیل ساختار سلسله‌مراتبی و طراحی ساده، این مدل‌ها سرعت بالاتری در آموزش و پیش‌بینی دارند. اگرچه مدل درخت تصمیم ممکن است در مجموعه داده‌های پیچیده دقت کم‌تری داشته باشد، اما سرعت بالای آن در پردازش، این مدل را به گزینه‌ای مناسب برای مسائل ساده‌تر تبدیل می‌کند.

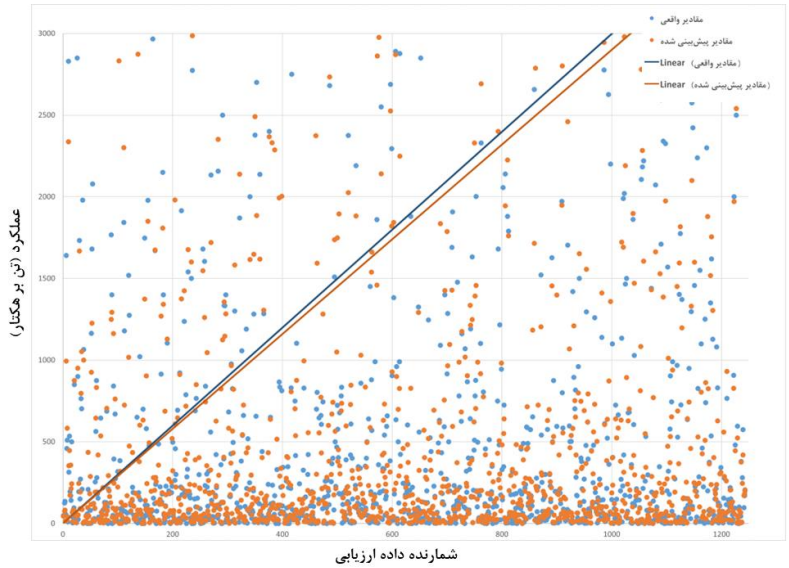


شکل ۶- نمودار مقادیر دقیق (نقاط آبی) و پیش‌بینی (نقاط نارنجی) مدل DT

مدل جنگل تصادفی، به دلیل توانایی در مقابله با داده‌های پیچیده و پرت، در این پژوهش عملکرد موفقی داشت. نتایج این تحقیق نشان داد که مدل جنگل تصادفی با R^2 برابر با ۰/۹۵۹۸ و MSE برابر با $۱۵/۶ \times ۱۰^{۶}$، عملکرد بسیار خوبی داشته است. این مدل توانست پیش‌بینی‌های دقیقی ارائه دهد و میزان خطای آن کم‌تر از سایر مدل‌ها بود. هم‌چنین، شکل ۷ نشان می‌دهد که پوشش مناسب نقاط پیش‌بینی‌شده و تطابق شیب خطی مقادیر واقعی و پیش‌بینی شده، بیان‌گر عملکرد برتر مدل RF نسبت به سایر مدل‌ها است. مدل RF به دلیل برخورداری از ساختار پیچیده و عمق مناسب، توانایی درک بهتر روابط بین داده‌ها را داشته و در نتیجه، عملکرد بهتری نسبت به سایر مدل‌ها ارائه داده است.

مدل جنگل تصادفی با ترکیب چندین درخت تصمیم، پیش‌بینی‌های دقیق‌تری ارائه می‌دهد در نتیجه، حجم این مدل به‌طور طبیعی بیش‌تر از یک درخت تصمیم منفرد است. حجم ۱۸۴۴۸۸ کیلوبایت برای این مدل نشان‌دهنده پیچیدگی بالای آن است. به دلیل وجود چندین درخت تصمیم و نیاز به ذخیره‌سازی عامل‌های متعدد، مدل جنگل تصادفی معمولاً حجمی بیش‌تر از سایر مدل‌ها دارد. علاوه بر این، ترکیب پیش‌بینی‌های مختلف از درخت‌ها برای بهبود دقت مدل، نیاز به محاسبات بیش‌تری دارد که باعث افزایش حجم مدل می‌شود.

مدل‌های جنگل تصادفی معمولاً کندتر از درخت‌های تصمیم واحد عمل می‌کنند. این کاهش سرعت به دلیل نیاز به پردازش تعداد زیادی درخت در طول فرآیند پیش‌بینی است. در این مدل، برای هر ورودی، پیش‌بینی‌های تمامی درخت‌ها محاسبه و سپس با یکدیگر ترکیب می‌شوند. این فرآیند می‌تواند زمان‌بر باشد، به‌ویژه هنگامی که تعداد درخت‌ها زیاد باشد و حجم مدل افزایش یابد.



شکل ۷- نمودار مقادیر دقیق (نقاط آبی) و پیش‌بینی (نقاط نارنجی) مدل RF

نتایج به‌دست‌آمده نشان می‌دهند که مدل RF بهترین عملکرد را در پیش‌بینی عملکرد گلخانه‌ها ارائه داده است. این مدل علاوه بر دستیابی به بالاترین مقدار R^2، کم‌ترین مقدار MSE را نیز ثبت کرده که نشان‌دهنده توانایی بالای آن در مدل‌سازی روابط غیرخطی داده‌ها است. در مقابل، مدل SVR ضعیف‌ترین عملکرد را از خود نشان داد و نتایج آن حاکی از این است که این مدل به تنظیمات بهینه‌تری برای بهبود دقت نیاز دارد.

از نظر حجم، مدل جنگل تصادفی بزرگ‌ترین اندازه را دارد که ناشی از وجود تعداد زیادی درخت تصمیم است. این حجم بالا منجر به افزایش نیازهای پردازشی و محاسباتی شده است. در مقابل، مدل DT به دلیل ساختار ساده خود، کم‌ترین حجم را دارا است. مدل‌های SVR و MLP به دلیل پیچیدگی ساختاری و تعداد عامل‌های بیش‌تر، حجم‌های میانه‌ای دارند.

به‌طور کلی، در شرایطی که هدف پیش‌بینی روابط غیرخطی پیچیده است، مدل RF بهترین گزینه به شمار می‌آید. با این حال، اگر سرعت پردازش و حجم پایین مدل از اهمیت بالایی برخوردار باشد، مدل DT توصیه می‌شود. این مدل هرچند دارای MSE بالا و R^2 پایین‌تری است، اما دقت بیش‌تری نسبت به مدل‌های MLP و SVR دارد. جدول ۲ نشان می‌دهد که انعطاف‌پذیری این مدل‌ها برای پیش‌بینی داده‌های مختلف چگونه است.

یکی از مسائل مهم ارزیابی عملکرد مدل‌ها جهت بررسی، معتبر بودن آن‌ها در واقعیت است. هر چند که در مدل‌های یادگیری ماشین اعتبارسنجی با بخشی از داده‌ها انجام شده است، اما به‌منظور مقایسه دقیق‌تر عملکرد مدل‌های مختلف با یکدیگر، از روش‌های آماری مناسب استفاده شده است تا بتوان تفاوت‌های معنادار میان نتایج حاصل از هر مدل را بررسی و تحلیل نمود.

جهت بررسی دقت پیش‌بینی مدل‌ها، سه شاخص خطای MAE، RMSE وSTD مورد محاسبه قرار گرفتند. نتایج این شاخص‌ها برای هر یک از مدل‌های مورد بررسی در جدول ۳ ارائه شده است.

بر اساس نتایج به‌دست‌آمده، مدل RF کم‌ترین مقدار MAE را ارائه داده است که نشان‌دهنده‌ی دقت بالای این مدل در پیش‌بینی عملکرد محصولات گلخانه‌ای است. هم‌چنین، این مدل دارای کم‌ترین RMSE بوده که نشان از پراکندگی کمتر خطاها نسبت به مقدار واقعی دارد. از نظر انحراف معیار نیز این عملکرد پایدارتری نسبت به سایر مدل‌ها داشته است. مدل DT نیز با وجود داشتن MSE بیش‌تر، در عامل MAE نتیجه‌ای به مراتب بهتر از مدل‌های SVR و MLP داشته است.

انحراف معیار بالاتر مدل DT نسبت به سایر مدل نشان دهنده این است که این مدل در چندین پیش‌بینی مقادیر پرتی را پیش‌بینی کرده است که سبب شده است تا خطای آن مقدار بیش‌تری باشد، درحالی که دقت مدل در اکثر پیش‌بینی‌ها مقدار خوبی بوده است و برای اثبات

جدول ۲- مقایسه مقادیر پیش‌بینی نسبت به مقدار واقعی

استان	شهرستان	محصول	سطح زیر کشت (هکتار)	عملکرد واقعی (تن در هکتار)	SVR (تن در هکتار)	MLP (تن در هکتار)	RF (تن در هکتار)	DT (تن در هکتار)
خراسان رضوی	فیروزه	خیار	۰/۴	۳۳/۲	۲۲۱۱	۲۵۳۸	۹۱/۸۸	۷۲
کرمان	کرمان	سایر سبزیجات برگی	۰/۲	۲۹/۹	۲۱۹۵/۶	۲۲۰۱/۵	۱۳/۱۵	۱۶/۶۹
کرمان	کرمان	گوجه فرنگی	۲	۵۰۰	۲۶۹۴	۱۶۵۷/۷	۴۹۵/۹	۴۵۲/۷
خراسان جنوبی	بیرجند	سبزیجات برگی	۰/۷	۵۹/۹	۲۱۶۷/۷	۲۱۳۶/۶	۱۷/۴۶	۱۴/۶۲
فارس	زرقان	بادمجان	۰/۰۲	۳/۹	۲۱۷۲/۹	۳۱۷/۲	۱۲/۱۵	۹۶/۶۳

این ادعا از یک آزمون تحلیل واریانس یک‌طرفه استفاده شد تا اختلاف معنی‌داری مقادیر پیش‌بینی مدل‌ها را با مقادیر اصلی مورد بررسی قرار دهیم که نتایج آن در جدول ۴ ذکر شده است.

جدول ۳- شاخص‌های ارزیابی دقت مدل‌ها

نام مدل	STD	RMSE	MAE
MLP	۵۱۷۵/۱	۵۲۲۸/۶	۲۳۸۳/۸
SVR	۵۷۹۸/۱	۵۷۹۸/۱	۲۶۳۷/۱
RF	۳۹۵۹	۳۹۵۹	۵۴۰/۵
DT	۶۶۵۵/۹	۶۶۵۶/۲	۹۲۸/۹

جدول ۴- مقادیر تحلیل تجزیه واریانس یک طرفه برای مدل‌های مختلف جهت بررسی معنی‌داری مقادیر پیش‌بینی مدل‌ها و مقدار واقعی داده‌ها

مدل	F-value	P-value
MLP	۱/۶۲۸	۰/۲۰۲
SVR	۸/۱۲۲	۰/۰۰۴
RF	$۷/۴۳۰\times۱۰^{-۷}$	۰/۹۹۹
DT	۰/۰۰۶	۰/۹۳۴

بر اساس نتایج حاصل از ارزیابی مدل‌ها، مشخص گردید که سه مدلRF، DT و SVR اختلاف معناداری با واقعی نداشتند؛ این موضوع نشان می‌دهد که پیش‌بینی‌های این مدل‌ها با مقادیر واقعی تطابق قابل‌قبولی داشته و دقت بالایی در برآورد مقادیر هدف از خود نشان داده‌اند. با این حال، مقدار احتمال (P-value) برای مدل SVR برابر با ۰/۰۰۴ محاسبه شد که بیان‌گر وجود اختلاف معنادار در سطح اطمینان ۹۹ درصد بین مقادیر پیش‌بینی‌شده و مقادیر واقعی است.

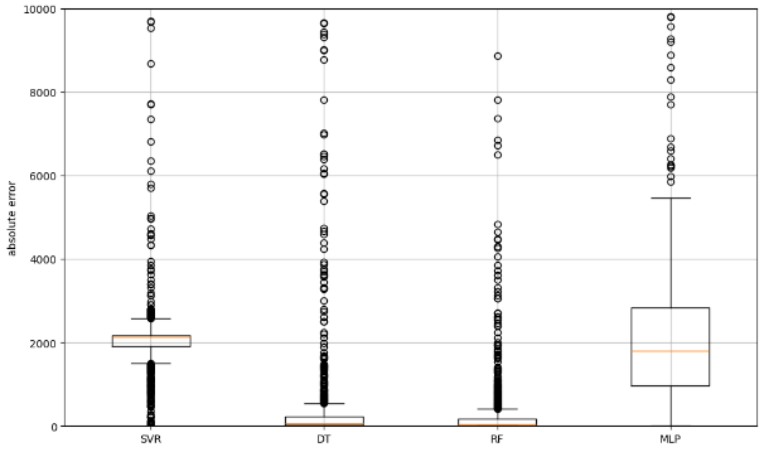
جدول ۵- مقادیر تحلیل تجزیه واریانس دو طرفه جهت بررسی معنی‌داری اختلاف خطای مدل‌ها

مقایسه مدل‌ها	F-value	p-value
SVR vs. MLP	۱/۶۴۹	۰/۱۹۹
SVR vs. RF	۱۲۹/۸۴۲	$۲/۳۵۰\times۱۰^{-۲۹}$
SVR vs. DT	۵۱/۶۹۶	$۸/۵۳۰\times۱۰^{-۱۳}$
MLP vs. RF	۱۱۳/۹۳۰	$۴/۹۰۰\times۱۰^{-۲۶}$
MLP vs. DT	۴۰/۳۸۵	$۲/۴۷۰\times۱۰^{-۱۰}$

این نتیجه حاکی از آن است که عملکرد مدل SVR نسبت به سایر مدل‌ها از دقت کم‌تری برخوردار بوده است. بر این اساس، می‌توان دقت مدل‌ها را به ترتیب RF، DT، MLP و در نهایت SVR طبقه‌بندی نمود. به‌منظور اطمینان بیش‌تر از صحت تفاوت میان دقت پیش‌بینی مدل‌ها، آزمون تحلیل واریانس برای مقایسه‌ی اختلاف میان مقادیر پیش‌بینی‌شده و مقادیر واقعی در بین مدل‌ها انجام گرفت که نتایج این تحلیل در جدول شماره ۵ ارائه شده است. بر اساس نتایج ارائه‌شده در جدول ۵، می‌توان نتیجه‌گیری کرد که بین مقادیر پیش‌بینی‌شده و مقادیر واقعی در چهار مدل

مورد بررسی شامل RF، DT، MLP و SVR تفاوت‌هایی وجود دارد. مقایسه این تفاوت‌ها نشان می‌دهد که مدل‌های SVR و MLP از نظر آماری اختلاف معناداری با سایر مدل‌ها ندارند، که بیان‌گر آن است که این دو مدل در پیش‌بینی داده‌ها عملکرد قابل قبولی نداشته و دقت پیش‌بینی پایینی از خود نشان داده‌اند. در مقابل، اختلاف بین نتایج پیش‌بینی‌شده دو مدل RF و DT با مقادیر واقعی در سطح اطمینان ۹۰ درصد (سطح معنی‌داری ۱۰ درصد) معنادار بوده است؛ این موضوع دلالت بر دقت مناسب این دو مدل در برآورد مقادیر هدف دارد. هم‌چنین، در سایر مقایسه‌ها، تفاوت‌ها در سطح اطمینان ۹۹ درصد (سطح معنی‌داری ۱ درصد) معنادار بوده‌اند، که نشان‌دهنده آن است که دقت پیش‌بینی بین برخی مدل‌ها به‌طور چشم‌گیری متفاوت بوده است. بنابراین، می‌توان نتیجه گرفت که دو مدل RF و DT دارای عملکرد مشابه و دقیق‌تری نسبت به مدل‌های دیگر بوده‌اند و می‌توان آن‌ها را به‌عنوان گزینه‌های برتر برای پیش‌بینی در این مسئله معرفی نمود.

نمودار جعبه‌ای ارائه‌شده در شکل ۸ مقایسه‌ای از خطای مطلق چهار مدل این پژوهش را نمایش می‌دهد. نتایج حاصل از این نمودار نشان می‌دهد که مدل RF کم‌ترین مقدار میانه خطای مطلق و هم‌چنین کم‌ترین پراکندگی خطا را در بین سایر مدل‌ها داراست، که نشان از پایداری و دقت بالای این مدل در پیش‌بینی دارد. مدل DT نیز عملکردی قابل قبول با میانه خطای پایین و پراکندگی محدود ارائه داده است، هرچند احتمال بیش‌برازش در آن وجود دارد. در مقابل، مدل‌های SVR و به‌ویژه MLP دارای میانه خطا و دامنه پراکندگی بیش‌تری هستند، که بیانگر دقت پایین‌تر و نوسان بیش‌تر در پیش‌بینی‌های آن‌ها است. علاوه‌براین، مدل MLP بیش‌ترین تعداد نقاط پرت را به خود اختصاص داده است، که نشان‌دهنده وجود خطاهای پیش‌بینی بسیار بالا در برخی نمونه‌ها است. به‌طور کلی، بر اساس تحلیل خطای مطلق، مدل RF نسبت به سایر مدل‌ها عملکرد بهتری از نظر دقت و پایداری ارائه می‌دهد و می‌تواند گزینه‌ی مناسب‌تری برای مسائل مشابه باشد.



شکل ۸- مقایسه خطای مطلق مدل

نتیجه‌گیری

گلخانه‌ها در بررسی ادبیات حوزه کشاورزی، به عنوان کشاورزی پایدار شناخته می‌شود. یکی از شاخص‌های اصلی کارایی گلخانه‌ها عملکرد تولید است. در این مطالعه، چهار مدل یادگیری ماشین شامل MLP، SVR، DT وRF برای پیش‌بینی عملکرد محصولات گلخانه‌ای در سطح کشور ایران مورد ارزیابی قرار گرفتند. داده‌های ورودی شامل اطلاعات جغرافیایی، نوع محصولات و سطح زیر کشت بودند که پس از پیش‌پردازش و نرمال‌سازی در مدل‌سازی به کار رفتند تا عملکرد گلخانه‌ها در سطح ملی پیش‌بینی شود. مقایسه عملکرد مدل‌ها با استفاده از معیارهای آماری مانند R^2، MSE، MAE، RMSE و Std انجام شد. مدل RF بهترین عملکرد را با R^2 معادل ۰/۹۵۹۸ و MSE برابر $۱۰^۶ \times ۱۵/۶$ ارائه داد و توانست روابط پیچیده بین متغیرها را به خوبی مدل‌سازی کند. هم‌چنین مدل DT که ساختاری مشابه با RF دارد نیز دقت بسیار خوبی را در پیش‌بینی ارائه داد و مقدار R^2 برابر ۰/۸۸۶۴ و MSE معادل $۱۰^۶ \times ۴۴/۳$ را به عنوان خروجی داده‌های آزمون گرفت و از عدم معنی‌داری نتایج اصلی و پیش‌بینی نتیجه گرفته شد که مدل توانایی خوبی در پیش‌بینی مقدار عملکرد داشته است، اما به دلیل چندین خطای فاحش در پیش‌بینی مقدار MSEآن به شدت بالا رفته است. مدلMLP به دلیل

ساختار ساده‌تر در تحلیل روابط داده‌ها، با R^2 برابر ۰/۸۸۷۸ و MSE معادل $۱۰^۶ \times ۲۷/۳$ عملکرد ضعیف‌تری داشت. مدل SVR باوجود تنظیم بهینه هایپرابرعامل‌ها، نتایج نسبتاً ضعیفی با R^2 برابر ۰/۸۵۴۶ وMSE معادل $۱۰^۶ \times ۳۳/۶$ ارائه کرد. هر دو مدل‌هایMLP و SVR اختلاف‌های معنی‌داری نسبت به نتایج اصلی داده‌های ازمون داشتند که نشان دهنده عدم توانایی این دو مدل در پیش‌بینی درست است.

در نهایت، یافته‌های این پژوهش نشان می‌دهد که مدل‌های یادگیری ماشین می‌توانند ابزارهای مؤثری برای سیاست‌گذاران و سرمایه‌گذاران و سازندگان بخش گلخانه در پیش‌بینی مقدار عملکرد محصولات گلخانه‌ای، انتخاب بهینه محصولات و مدیریت کارآمد منابع باشند. مدل RF توسعه‌یافته در این تحقیق، با قابلیت تحلیل داده‌های مکانی و سطح زیر کشت، می‌تواند به‌عنوان ابزاری راهبردی در تصمیم‌گیری‌های کشاورزی در سطوح ملی و محلی برای پیش‌بینی مقدار عملکرد محصولات مورد استفاده قرار گیرد.

با بهره‌گیری از این ابزار می‌توان الگویی بهینه برای کشت محصولات گلخانه‌ای ارائه داد که با استفاده از اطلاعات موقعیت جغرافیایی استان و شهرستان، نوع محصول و سطح زیرکشت، عملکرد را به‌طور دقیق پیش‌بینی می‌نماید. این الگوی پیشنهادی می‌تواند به سیاست‌گذاران بخش کشاورزی در برآورد میزان تولید (تناژ) کمک کرده و مبنایی برای توسعه نظام‌مند الگوی کشت گلخانه‌ای فراهم آورد. هدف این رویکرد، افزایش بهره‌وری تولیدات، جلوگیری از کشت محصولات در مناطق دارای خطر بالا یا عملکرد پایین، و هدایت منابع به‌سوی کشت محصولات پربازده و سازگار با شرایط محیطی هر منطقه است.

مشارکت نویسندگان

نحوه و میزان مشارکت نویسندگان در انجام این پژوهش به صورت زیر است:

خشایار زارع: روش‌شناسی، نگارش

احمد حسین‌نژاد: مفهوم‌سازی، نظارت

دسترسی به داده‌ها

همه اطلاعات و نتایج در متن مقاله ارائه شده است.

اصول اخلاقی

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

تضاد منافع نویسندگان

نویسندگان اعلام می‌کنند که هیچ گونه منافع مالی رقابتی یا روابط شخصی شناخته‌شده‌ای که ممکن است بر کار گزارش‌شده در این مقاله تأثیر گذاشته باشد، ندارند.

حمایت مالی

نویسندگان هیچ حمایت مالی خاصی برای انجام این پژوهش دریافت نکرده‌اند.

منابع

Pifeh, A. (2008). Examining the finished accounting system of agricultural products. 2nd International Conference on Operational Budgeting. Tehran. https://civilica.com/doc/74140

Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. Statistics and computing, 14(3), 199-222. https://doi.org/10.1023/B:STCO.0000035301.49549.88

Baştanlar, Y., & Özuysal, M. (2013). Introduction to machine learning. miRNomics: MicroRNA biology and computational analysis, 105-128. https://doi.org/10.1007/978-1-62703-748-8_7

Ouazzani Chahidi, L., Fossa, M., Priarone, A., & Mechaqrane, A. (2021). Evaluation of supervised learning models in predicting greenhouse energy demand and production for intelligent and sustainable operations. Energies, 14(19), 6297. https://doi.org/10.3390/en14196297

Esmaeli, H., & Roshandel, R. (2020). Optimal design for solar greenhouses based on climate conditions. Renewable energy, 145, 1255-1265. https://doi.org/10.1016/j.renene.2019.06.090

Everitt, B. S. (2005). Classification and regression trees. Encyclopedia of statistics in behavioral science. https://doi.org/10.1002/0470013192.bsa753

Frausto-Solis, J., Gonzalez-Sanchez, A., & Larre, M. (2009, November). A new method for optimal cropping pattern. In Mexican International Conference on Artificial Intelligence (pp. 566-577). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-05258-3_50

Hosseinnejad, A., Saboohi, Y., Zarei, G., & Shayegan, J. (2023). Developing an integrated model for allocating resources and assessing technologies based on the watergy optimal point (water-energy nexus), case study: a greenhouse.

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

Melal, S. R., Aminian, M., & Shekarian, S. M. (2024). A machine learning method based on stacking heterogeneous ensemble learning for prediction of indoor humidity of greenhouse. Journal of Agriculture and Food Research, 16, 101107. https://doi.org/10.1016/j.jafr.2024.101107

Singh, V. K., & Tiwari, K. N. (2017). Prediction of greenhouse micro-climate using artificial neural network. Appl. Ecol. Environ. Res, 15(1), 767-778. http://dx.doi.org/10.15666/aeer/1501_767778

Taki, M., Ajabshirchi, Y., Ranjbar, S. F., & Matloobi, M. (2016). Application of Neural Networks and multiple regression models in greenhouse climate estimation. Agricultural Engineering International: CIGR Journal, 18(3), 29-43.

Van Klompenburg, T., Kassahun, A., & Catal, C. (2020). Crop yield prediction using machine learning: A systematic literature review. Computers and electronics in agriculture, 177, 105709. https://doi.org/10.1016/j.compag.2020.105709

Yang, Y., Gao, P., Sun, Z., Wang, H., Lu, M., Liu, Y., & Hu, J. (2023). Multistep ahead prediction of temperature and humidity in solar greenhouse based on FAM-LSTM model. Computers and Electronics in Agriculture, 213, 108261. https://doi.org/10.1016/j.compag.2023.108261

Zou, W., Yao, F., Zhang, B., He, C., & Guan, Z. (2017). Verification and predicting temperature and humidity in a solar greenhouse based on convex bidirectional extreme learning machine algorithm. Neurocomputing, 249, 72-8 https://doi.org/10.1016/j.neucom.2017.03.023

Journal of Sustainable Development of Energy, Water and Environment Systems, 11(1), 1-32. https://doi.org/10.13044/j.sdewes.d10.0416

Kim, S. Y., Park, K. S., & Ryu, K. H. (2018). Outside Temperature Prediction Based on Artificial Neural Network for Estimating the Heating Load in Greenhouse. KIPS Transactions on Software and Data Engineering, 7(4), 129-134.

Kulyal, M., & Saxena, P. (2022, November). Machine learning approaches for crop yield prediction: A review. In 2022 7th International Conference on Computing, Communication and Security (ICCCS) (pp. 1-7). IEEE. https://doi.org/10.1109/ICCCS55188.2022.10079240

Maraveas, C. (2022). Incorporating artificial intelligence technology in smart greenhouses: Current State of the Art. Applied Sciences, 13(1), 14. https://doi.org/10.3390/app13010014

Gopal, P. M., & Bhargavi, R. (2019). A novel approach for efficient crop yield prediction. Computers and Electronics in Agriculture, 165, 104968. https://doi.org/10.1016/j.compag.2019.104968

Rani, N., Bamel, K., Shukla, A., & Singh, N. (2022). Analysis of Five Mathematical Models for Crop Yield Prediction. South Asian Journal of Experimental Biology, 12(1). https://doi.org/10.38150/sajeb.12(1).p46-54

Palani, H. K., Ilangovan, S., Senthilvel, P. G., Thirupurasundari, D. R., & Kumar, R. (2023, November). AI-powered predictive analysis for pest and disease forecasting in Crops. In 2023 International Conference on Communication, Security and Artificial Intelligence (ICCSAI) (pp. 950-954). IEEE. https://doi.org/10.1109/ICCSAI59793.2023.10421237

Pandorfi, H., Bezerra, A. C., Atarassi, R. T., Vieira, F. M., Barbosa Filho, J. A., & Guiselini, C. (2016). Artificial neural networks employment in the