Research Article

Agricultural Engineering, 48(2) (2025) 315 - 341 ISSN (E): 2588-526X

DOI: 10.22055/agen.2025.49302.1773 ISSN (P): 2588-5944

Diagnosis and yield estimation of sugarcane fields using support vector machine and random forest algorithms (A case study: Dehkhoda Sugarcane Agro-industry Company)

E. Noori[1] and N. Monjezi [2]*

1. MSc., Biosystems Engineering Dept., Faculty of Agriculture, Shahid Chamran University of Ahvaz, Ahvaz, Iran
2. Associate Professor, Biosystems Engineering Dept., Faculty of Agriculture, Shahid Chamran University of Ahvaz, Ahvaz, Iran

Received: 23 March 2025 Accepted: 16 June 2025 *Corresponding Author: n.monjezi@scu.ac.ir

Abstract

Introduction: Sugarcane is cultivated on 135,000 hectares of agricultural land in Khuzestan Province. Some farms operated by the Sugarcane and By-Products Development Company achieve an average yield of 125 tons per hectare annually, highlighting the region's strong potential for sugarcane cultivation. However, in recent years, the yield per unit area has declined, necessitating extensive studies to investigate the causes of this decrease and to develop strategies for improving crop performance. Sugarcane yield is influenced by various factors, but it is challenging to simultaneously assess the impacts of management practices, water availability, soil conditions, and climate in field studies. Therefore, modeling provides a viable solution. By utilizing data mining and machine learning tools, we can analyze data to identify the factors affecting sugarcane performance. This research aims to pinpoint the issues and estimate the performance of sugarcane farms in Dehkhoda Agro-Industry using support vector machine and random forest algorithms.

Materials and Methods: Data and information were collected from the Sugarcane and By Products Development Company and the Dehkhoda Sugarcane Agro-Industry Company for the crop years 2019-2023. This included both statistical data and field measurements, resulting in a total of 2,032 records across four categories: meteorological data, crop data, soil data, and farm management data. The study focused on various variables analyzed using two algorithms: support vector machine and random forest. These variables included crop variety, harvest month, nitrogen and phosphorus fertilizer application, plant age (first crop or ratoon), irrigation frequency during the growing season, the ratio of sprayed area to total field area, soil texture, soil electrical conductivity, water consumption per hectare, drainage rate, field management practices, length of the growing season, field area, sugarcane yield, and climatic conditions. To evaluate model accuracy, the study employed three indicators: root mean square error, average absolute error, and correlation coefficient. Modeling was conducted using Python software version 3.12.4. In this study, 70% of the data (1,626 samples) was designated for the training set, while the remaining 30% (406 samples) was used for the testing set.

Results and Discussion: The scores of the support vector machine and random forest models, with sugarcane yield as the target variable, were 84.5% and 85.6% on the testing datasets, respectively, and 86.9% and 87.1% on the training datasets, respectively. In the experimental phase, the random forest model demonstrated significant performance, achieving a correlation coefficient (R^2) of 0.92, a mean absolute error (MAE) of 0.34, and a root mean square error (RMSE) of 62.81. Various metrics and methods were employed

© 2025 The Author(s). Published by Shahid Chamran University of Ahvaz. This is an open-access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

to evaluate the models' accuracy, stability, and reliability in handling new data. The constructed models were assessed using cross-validation, confusion matrices, performance characteristic curves, and sensitivity accuracy curves. For sugarcane yield prediction, the hybrid classifier with soft voting achieved accuracies of 93% on training data and 89% on test data. The accuracy-frequency for the sugarcane yield prediction model using the random forest was 96%, representing a significant achievement in this research. This high accuracy indicates that the model effectively identified and modeled the complex, nonlinear patterns present in the sugarcane data. Furthermore, it suggests that the selected input variables had strong predictive power. The random forest output curve also demonstrates that the algorithm performed acceptably in predicting sugarcane yield. The upward trend followed by the stabilization of the curve indicates that the model converged well and reached its optimal point.

Conclusion: The results demonstrated significant success in using machine learning techniques to predict the performance of sugarcane fields. The random forest model achieved 97% accuracy, confirmed by 10-layer cross-validation, indicating its reliability as a tool for sugarcane plantation management. Complementary results from the support vector machine algorithm further support these findings. The balanced data distribution in the box plot, strong performance in the confusion matrix, and favorable ROC curve all suggest that these models can significantly influence decision-making in sugarcane plantation management. Additionally, these results indicate that employing the random forest technique in agriculture can enhance productivity and improve resource management.

Keywords: Sugarcane, Support Vector Machine, Random Forest, Crop Yield Prediction.

مهندسی زراعی (مجله علمی کشاورزی) جلد ۴۸. شماره ۲، ۱۴۰۴

مقاله پژوهشی

عارضه‌یابی و تخمین عملکرد مزارع نیشکر با استفاده از الگوریتم ماشین بردار پشتیبان (SVM) و جنگل تصادفی (RF) (مطالعه موردی: شرکت کشت و صنعت نیشکر دهخدا)

الهام نوری[۱] و نسیم منجزی[۲*]

۱- دانشجوی کارشناسی ارشد مکانیزاسیون کشاورزی، گروه مهندسی بیوسیستم، دانشکده کشاورزی، دانشگاه شهید چمران اهواز، ایران

۲- دانشیار گروه مهندسی بیوسیستم، دانشکده کشاورزی، دانشگاه شهید چمران اهواز، ایران

تاریخچه مقاله

دریافت: ۱۴۰۴/۰۱/۰۳

پذیرش نهایی: ۱۴۰۴/۰۳/۲۶

کلمات کلیدی:

نیشکر،

ماشین بردار پشتیبان،

جنگل تصادفی،

پیش بینی عملکرد محصول

* عهده دار مکاتبات

Email: n.monjezi@scu.ac.ir

چکیده

کشت نیشکر، ۱۳۵ هزار هکتار از زمین‌های کشاورزی استان خوزستان را به خود اختصاص داده است. هدف از این پژوهش، عارضه‌یابی و تخمین عملکرد مزارع نیشکر کشت و صنعت دهخدا با استفاده از الگوریتم ماشین بردار پشتیبان و جنگل تصادفی است. برای جمع‌آوری داده‌ها و اطلاعات، از آمار شرکت توسعه نیشکر و صنایع جانبی خوزستان و شرکت کشت و صنعت نیشکر دهخدا در سال‌های زراعی ۱۴۰۲-۱۳۹۷ و همچنین جمع‌آوری میدانی داده‌ها و اندازه‌گیری آن‌ها استفاده شد. داده‌های استفاده شده شامل ۲۰۳۲ داده است که از چهار دسته مجموعه داده شامل داده‌های هواشناسی، داده‌های محصول، داده‌های خاک و داده‌های مدیریت مزرعه تشکیل شده است. شاخص‌های مورد استفاده به‌منظور ارزیابی دقت مدل شامل: خطای ریشه میانگین مربعات خطا، متوسط قدر مطلق خطا و ضریب همبستگی بود. برای مدل‌سازی از نرم‌افزار پایتون نسخه 3.12.4 استفاده شد. نتایج بررسی عارضه‌ها و علل کاهش عملکرد نیشکر نشان داد که عوامل اثرگذار در این زمینه عبارتند از: آبیاری و زهکشی، حاصلخیزی خاک و تغذیه گیاه، واریته محصول، آفات و بیماری‌ها و علف‌های هرز، عملیات زراعی خاک‌ورزی، کاشت و برداشت، عوامل اقلیمی و تنش‌های محیطی، تراکم خاک، شوری آب و خاک و نیروی انسانی و مدیریتی. در مدل‌سازی نیز امتیاز مدل ماشین بردار پشتیبان و جنگل تصادفی با در نظر گرفتن عملکرد نیشکر به عنوان متغیر هدف در دو مجموعه داده آموزش و آزمایش، بر روی داده‌های آزمایش به ترتیب ۸٤/٥ و ۸٥/٦ درصد و بر روی داده‌های آموزش به ترتیب ۸٦/۹ و ۸۷/۱ درصد به دست آمد. در مرحله آزمایش نیز عملکرد مدل جنگل تصادفی با ضریب همبستگی (R^2) برابر ۰/۹۲، میانگین خطای مطلق (MAE) برابر ۰/۳٤ و ریشه میانگین مربعات خطا (RMSE) برابر ٦۲/۸۱ قابل توجه بود. دقت-فراخوانی برای مدل پیش‌بینی عملکرد نیشکر با استفاده از جنگل تصادفی ۹٦ درصد به دست آمد. این دقت بالا نشان‌دهنده این است که مدل توانسته به خوبی الگوهای پیچیده و غیرخطی موجود در داده‌های نیشکر را شناسایی و مدل‌سازی کند. همچنین، این دقت بالا نشان می‌دهد که متغیرهای ورودی انتخاب شده برای مدل از قدرت پیش بینی کنندگی بالایی برخوردار بوده‌اند. همچنین، بر اساس نتایج مدل‌سازی، در دو مدل جنگل

تصادفی و ماشین بردار پشتیبان مهم‌ترین متغیرهای اثرگذار بر میزان عملکرد محصول نیشکر، واریته محصول، میزان کود شیمیایی (نیتروژن) و سن گیاه است.

مقدمه

کشت نیشکر، ۱۳۵ هزار هکتار از زمین‌های کشاورزی استان خوزستان را به خود اختصاص داده است (۲۷). هشت واحد از کشت و صنعت‌های نیشکری ایران در قالب شرکت توسعه نیشکر و صنایع جانبی خوزستان در مساحتی بالغ بر ۱۱۰ هزار هکتار می‌باشند که سالیانه بیش از ۸ میلیون تن نیشکر و حدود ۸۰۰ هزار تن شکر تولید می‌کنند (۱۷). بنابراین این شرکت به عنوان یکی از قطب‌های کشاورزی-اقتصادی ایران و منطقه شناخته می‌شود. در برخی از مزارع شرکت توسعه نیشکر و صنایع جانبی خوزستان، سالانه به طور میانگین ۱۲۵ تن در هکتار، محصول برداشت می‌شود که نشانگر استعداد مطلوب استان خوزستان برای کشت نیشکر است (۲۹). کشت نیشکر در شرکت توسعه نیشکر و صنایع جانبی در استان خوزستان از دهه ۱۳۸۰ شروع شد. با وجود مشکلات فراوان در سال‌های اولیه شروع بهره‌برداری از اراضی (مثل شوری خاک و لم یزرع بودن اراضی) عملکرد در واحد در سطح مطلوب بود. در ادامه، متأسفانه به دلایل متعددی از جمله عدم توجه به تولید پایدار، آلودگی‌های زیست محیطی، فشردگی خاک مزارع به دلیل تردد ماشین‌های سنگین زراعی، بالا رفتن شوری آب رود خانه کارون و دز، کمبود آب و خشکسالی، وقوع پدیده گرد و غبار، آتش زدن مزارع نیشکر و از بین رفتن مواد آلی و میکروارگانسیم‌ها و .... عملکرد نیشکر در واحد سطح نسبت به سال‌های اولیه کاهش پیدا کرد و تا پتانسیل اسمی تولید (۱۲۰ تن در هکتار) هنوز فاصله‌ی زیاد دارد. هر چند ممکن است از سالی به سال دیگر به صورت جزئی عملکرد افزایش داشته باشد ولی روند کلی آن کاهشی بوده است. علاوه بر این، در حال حاضر تفاوت چشمگیر در میانگین عملکرد محصول نیشکر در بین کشت و صنعت‌های نیشکری و حتی تفاوت بین عملکرد مزارع مختلف در یک کشت و صنعت، مشهود است. بنابراین برای بررسی علل کاهش عملکرد مزارع نیشکر و ارائه راهکارهایی به منظور بهبود عملکرد محصول نیاز به مطالعات وسیعی در این زمینه است. در این پژوهش یکی از اهداف اصلی، تخمین عملکرد مزارع نیشکر در استان خوزستان است. اما در وهله اول با توجه به نیاز به شناسایی عوامل تأثیرگذار بر عملکرد محصول نیشکر، عارضه‌یابی علل مؤثر در این زمینه، لازم و ضروری است. عارضه‌یابی دلالت بر شناسایی نقطه ضعف‌هایی دارد که باعث شده است تا یک کشت و صنعت از منظر تولید بر واحد سطح از مسیر رشد باز بماند و یا سرعت رشد آن کمتر و کمتر شود.

عملکرد نیشکر تحت تأثیر پارامترهای مختلفی است؛ با توجه به اینکه بررسی همزمان تأثیرات مدیریتی، آب، خاک و اقلیم روی گیاه نیشکر در روش میدانی وجود ندارد؛ بنابراین مدل‌سازی، راه‌حلی برای این مشکل است (۱۶).

ابزارهای داده‌کاوی و یادگیری ماشین می‌توانند داده‌ها را تحلیل و عوامل مؤثر بر عملکرد محصول را شناسایی کنند. داده‌کاوی، فرآیند شناسایی و استخراج اطلاعات مهم و مفید و الگوهای پنهان از مجموعه داده‌ها و تبدیل آن‌ها به اطلاعات قابل فهم است (۲۱). الگوریتم‌های یادگیری ماشین در زمینه تخمین عملکرد محصولات کشاورزی تا کنون کاربرد گسترده‌ای داشته‌اند.

کاوسی گیوشاد و همکاران[1] (۱۱) برای پیش‌بینی عملکرد زعفران از مدل ماشین بردارپشتیبان و برنامه‌ریزی بیان ژن استفاده کردند. هدف از این تحقیق، توانایی مدل شبیه‌سازی ماشین بردار پشتیبان[2] و مدل برنامه‌ریزی بیان ژن در پیش‌بینی عملکرد زعفران بر اساس داده‌های هواشناسی (حداقل دما، حداکثر دما، بارش، تبخیر و رطوبت نسبی و عملکرد یکسال قبل) بوده است. نتایج نشان داد مدل برنامه‌ریزی بیان ژن نتایج دقیق‌تری را در تخمین عملکرد زعفران نسبت به مدل ماشین بردار پشتیبان دارا است.

حافظی و همکاران[3] (۶) هیبرید شبکه عصبی مصنوعی را با الگوریتم‌های فرا-ابتکاری در پیش‌بینی عملکرد نیشکر استفاده کردند. در این تحقیق به منظور پیش‌بینی عملکرد نیشکر از

1- Kavoosi-Givshad *et al.*
2- Support Vector Machine
3- Hafezi *et al.*

عامل‌های مؤثری نظیر مشخصات گیاه، هدایت الکتریکی آب و خاک، حداکثر درجه حرارت و میانگین ساعات آفتابی در طول دوره رشد محصول و در مقیاس زمانی هفت ساله استفاده گردید. نتایج بدست آمده از تحقیق، حاکی از آن است که استفاده از الگوریتم‌های فرااکتشافی سبب بهبود عملکرد شبکه عصبی شده و قدرت پیش‌بینی را به طور محسوسی افزایش داده است.

ماهاجان و همکاران[1] (۱۵) پیش‌بینی عملکرد محصولات را با استفاده از الگوریتم‌های یادگیری ماشین (درخت‌های تصمیم‌گیری، جنگل تصادفی[2]، رگرسیون بردار پشتیبان و تقویت گرادیان) ارائه دادند که می‌تواند برای پیش‌بینی بازده محصولات عمده تولید شده در مناطق جامو و کشمیر در هند استفاده شود. نتایج نشان داد در بین مدل‌های مورد استفاده، رگرسیون درخت تصمیم، تولید محصول را با بیشترین دقت (۹۶٪) پیش‌بینی می‌کند؛ در حالی که رگرسیون بردار پشتیبان، پیش‌بینی تولید محصول را با ضعیف‌ترین دقت (۸۹٪) انجام داد.

کورادسنگ و همکاران[3] (۱۳) در پیش‌بینی عملکرد محصول از مدل‌های یادگیری ماشین استفاده کردند. در این تحقیق از روش‌های یادگیری ماشینی برای پیش‌بینی برداشت محصول بر اساس داده‌های آب و هوا و انتقال اطلاعات در مورد روند تولید استفاده شد. داده‌های آب، هوا و عملکرد محصول برای سیب زمینی ایرلندی و ذرت از منابع مختلف جمع‌آوری شد. داده‌های جمع‌آوری شده از طریق جنگل تصادفی، رگرسیون چند جمله‌ای و رگرسیون بردار پشتیبان مورد تجزیه و تحلیل قرار گرفت. نتایج نشان داد که جنگل تصادفی بهترین مدل با ریشه میانگین مربعات خطای ۸/۵۱ و ۹/۱۱۳ برای سیب زمینی و ذرت است، در حالی که ضریب همبستگی برای مجموعه داده‌های محصولات مشابه ۰/۸۷۵ و ۰/۸۱۷ بود. شرایط آب و هوایی بهینه برای عملکرد بهینه هر دو محصول شناسایی شد.

ساریجالو و همکاران[4] (۲۴) برای تخمین عملکرد هیبریدهای ذرت از الگوریتم‌های یادگیری ماشینی استفاده کردند. در این تحقیق از الگوریتم‌های مختلف یادگیری ماشین از جمله درخت تصمیم، ماشین تقویت گرادیان، جنگل تصادفی، تقویت تطبیقی، XGBoost و شبکه عصبی برای پیش‌بینی عملکرد هیبریدهای ذرت استفاده شده است. نتایج نشان می‌دهد پیش‌بینی به‌دست‌آمده توسط XGBoost دقیق‌تر از مدل‌های دیگر با ریشه میانگین مربع خطا برابر ۰/۰۵۲۴ است.

پانیراهی و همکاران[5] (۲۰) از یک رویکرد مقایسه‌ای مبتنی بر یادگیری ماشین برای پیش‌بینی عملکرد محصول با استفاده از یادگیری نظارت شده با مدل‌های رگرسیون استفاده کردند که هدف اصلی این پروژه تحقیقاتی توسعه یک مدل یادگیری ماشینی (ML) برای پیش‌بینی عملکرد بود. برای تخمین عملکرد محصول، داده‌ها با استفاده از یادگیری ماشینی نظارت شده با شش مدل رگرسیون مجزا جمع‌آوری و آموزش داده شد. با میانگین خطای مطلق ۴۶۸/۱۶، جنگل تصادفی از سایر مدل‌ها بهتر عمل کرد.

فگاد و پاوار[6] (۴) برای پیش‌بینی محصول از شبکه عصبی مصنوعی و ماشین بردار پشتیبان استفاده کردند. در این تحقیق، طبقه‌بندی‌کننده‌های مورد استفاده، ماشین بردار پشتیبان و شبکه‌های عصبی مصنوعی بودند. پیش‌بینی محصول با در نظر گرفتن پارامترهایی مانند میزان بارندگی، دمای حداقل و حداکثر، نوع خاک، رطوبت و مقدار pH خاک انجام شد. نتایج نشان داد شبکه عصبی ۸۶/۸۰ درصد از دقت پیش‌بینی را ارائه می‌دهد.

باتول و همکاران[7] (۲) از یک رویکرد ترکیبی برای پیش‌بینی عملکرد محصول چای از مدل‌های شبیه‌سازی و یادگیری ماشینی استفاده کردند. این مطالعه تحقیقاتی با هدف ارائه یک مطالعه تطبیقی از روش‌های پیش‌بینی عملکرد چای با استفاده از مدل شبیه‌سازی آکواکراپ[8] سازمان ملل متحد و روش‌های یادگیری ماشینی است. نتایج نشان داد که الگوریتم رگرسیون یادگیری ماشین در پیش‌بینی بازده با استفاده از داده‌های کمتر نسبت به مدل شبیه‌سازی بهتر عمل می‌کند.

1- Mahajan *et al.*
2- Random Forest
3- Kuradusenge *et al.*
4- Sarijaloo *et al.*
5- Panigrahi *et al.*
6- Fegade and Pawar
7- Batool *et al.*
8- AquaCrop

رضایی و همکاران[1] (۲۲) به بررسی تشخیص تیپ اراضی کشاورزی دشت عباس با استفاده از تحلیل سری زمانی تصاویر ماهواره سنتینل ۲[2] پرداختند. در این مطالعه، از شاخص NDVI و تصاویر سری زمانی سنجش از دور ماهواره سنتینل ۲ برای بررسی پویایی پوشش گیاهی در بخش‌هایی از دشت عباس دهلران در سال زراعی ۱۴۰۰-۱۴۰۱ استفاده شد. طبقه‌بندی کاربری زمین با استفاده از الگوریتم ماشین بردار پشتیبان انجام شد و صحت نتایج با استفاده از داده‌های کاداستر سنجیده شد. نتایج نشان داد که دقت طبقه‌بندی برای اراضی آبی ۹۸/۴ درصد و برای اراضی دیم ۸۶/۷ درصد بود. این نتایج نشان دهنده دقت بالای نقشه‌های کاداستر در تشخیص و تفکیک نوع اراضی کشاورزی است.

در پژوهش حاضر از روش‌های داده‌کاوی استفاده شده است. داده‌کاوی تئوری‌های پایگاه داده‌ها، هوش مصنوعی، یادگیری ماشین و علم آمار را در هم می‌آمیزند تا زمینه کاربردی فراهم شود در صورتی‌که در روش‌های کلاسیک داده‌ها امکان بررسی و کشف الگوها در حجم انبوهی از داده‌ها به راحتی میسر نیست و نیازمند صرف زمان و هزینه است. همچنین در روش‌های داده‌کاوی امکان تحلیل متغیرهای کیفی مانند مدیریت تولید، بافت خاک، واریته محصول، زمان برداشت، تعداد دفعات آبیاری و .... و وارد کردن این متغیرها به مدل به راحتی امکان‌پذیر است. بنابراین از آنجا که الگوریتم‌های کاربردی در زمینه‌ی پیش‌بینی و تخمین عملکرد محصولات کشاورزی متنوع و زیادند، در پژوهش حاضر باتوجه به ماهیت متغیرهای تأثیرگذار بر عملکرد محصول نیشکر که ماهیتی کمی و کیفی دارند، الگوریتم ماشین بردار پشتیبان و جنگل تصادفی انتخاب شده است. بنابراین در این تحقیق به عارضه‌یابی و تخمین عملکرد مزارع نیشکر کشت و صنعت دهخدا با استفاده از الگوریتم ماشین بردار پشتیبان و جنگل تصادفی به‌صورت خاص پرداخته شد.

مواد و روش‌ها

منطقه مورد مطالعه

کشت و صنعت نیشکر دهخدا در شمال شهرستان اهواز، حد فاصل جاده اندیمشک تا رودخانه کارون، کیلومتر ۲۳ از جاده دغاغله واقع گردیده است. ارتفاع از سطح دریا اراضی زراعی شرکت دهخدا، ۲۰ متر و طول و عرض جغرافیایی آن به ترتیب ۴۸ درجه و ۴۰ دقیقه شرقی و ۳۱ درجه و ۲۵ دقیقه شمالی است. خالص ارضی این کشت و صنعت ۱۱۳۶۹ هکتار و ظرفیت اسمی تولید در آن در سال به ترتیب ۱۰۰ هزار تن شکر خام، ۱۷۵ هزار تن شکر سفید، ۳۰ الی ۳۸ هزار تن ملاس و فرآورده‌های آن و ۳۳۰ هزار تن باگاس است.

جمع‌آوری و پیش پردازش داده‌ها

برای جمع‌آوری داده‌ها و اطلاعاتی که در این پژوهش به کار گرفته‌شد، از آمار شرکت توسعه نیشکر و صنایع جانبی خوزستان، موسسه تحقیقات و آموزش توسعه نیشکر و صنایع جانبی خوزستان و شرکت کشت و صنعت نیشکر دهخدا در سال‌های زراعی ۱۴۰۲-۱۳۹۷ و همچنین جمع‌آوری میدانی داده‌ها و اندازه‌گیری آن‌ها استفاده شد. متغیرهای مورد بررسی در دو الگوریتم ماشین بردار پشتیبان و جنگل تصادفی شامل: واریته محصول، زمان برداشت، کود شیمیایی نیتروژن، کود شیمیایی فسفر، سن گیاه (کشت اول یا راتون)، تعداد دفعات آبیاری مزرعه در طی فصل زراعی، نسبت سطح سم‌پاشی مزرعه (نسبت سطح سم‌پاشی شده به مساحت فم کل مزرعه)، بافت خاک، هدایت الکتریکی خاک، مقدار مصرف آب در هکتار، میزان زهکشی، مدیریت مزرعه (مدیریت تولید)، طول فصل زراعی، مساحت مزرعه، عملکرد نیشکر و متغیرهای آب و هوایی بود.

داده‌های استفاده شده شامل ۲۰۳۲ داده است که از چهار دسته مجموعه داده شامل داده‌های هواشناسی، داده‌های محصول، داده‌های خاک و داده‌های مدیریت مزرعه تشکیل شده است. با توجه به اینکه متغیرهای مورد استفاده برای مدل‌سازی مقیاس یکسانی نداشتند، در این پژوهش از روش نرمال‌سازی مین-ماکس[3] استفاده شده است (۷). همچنین به منظور بررسی همبستگی بین متغیرهای مستقل با متغیر وابسته (عملکرد نیشکر) از همبستگی پیرسون استفاده شد (شکل ۱). جدول ۱ توصیف متغیرهای گسسته و جدول ۲ توصیف متغیرهای پیوسته مورد استفاده را نشان می‌دهد.

1- Rezaei *et al.*
2- Sentinel-2
3- Min-Max

تحلیل توصیفی داده‌ها

نمودار شکل ۲ (الف)، میانگین عملکرد نیشکر را در بازه زمانی ۱۴۰۲ - ۱۳۹۷ نشان می‌دهد. از سال ۱۳۹۷ تا ۱۴۰۲ عملکرد نیشکر روند صعودی داشته و در سال ۱۴۰۲ به حداکثر مقدار خود رسیده است. همان‌طور که نمودار نشان می‌دهد در سال زراعی ۱۳۹۹ عملکرد نیشکر افت چشمگیری داشته است که می‌تواند ناشی از عوامل مختلفی باشد. یکی از این عوامل میزان شوری خاک است که در شکل ۲ (ب)، میانگین شوری خاک در سال‌های ۱۴۰۲-۱۳۹۷ محاسبه شده و نشان می‌دهد که در سال زراعی ۱۳۹۹ میزان شوری بیشتر از سایر سال‌ها بوده است.

در شکل ۳، عملکرد نیشکر به تفکیک واریته‌های مختلف و سال زراعی ارائه شده است. با توجه به نمودار واریته IRC99-02 بیشترین عملکرد را در سال‌های ۱۳۹۷ تا ۱۴۰۲ دارد.

به طور کلی می‌توان نتیجه گرفت که مجموعه‌ای از متغیرها در تغییرات عملکرد مزارع نقش دارند و دقیقاً به خاطر همین پیچیدگی روابط است که تحلیل‌های آماری و رگرسیون‌های آماری قادر به استخراج الگوهایی از میان حجم عظیم داده‌ها با دقت بالا نیستند و در این باره باید از روش‌های داده‌کاوی و یادگیری ماشین استفاده شود. لازم به ذکر است، هدف از ارائه این تحلیل توصیفی در این پژوهش، نشان دادن این نکته است که نتیجه‌گیری صرفاً بر اساس یک عامل (مانند شوری خاک)، نمی‌تواند منطبق بر واقعیت باشد و عوامل مختلفی در افزایش یا کاهش عملکرد تأثیر دارند. بنابراین باید تا آنجایی که امکان‌پذیر است تمامی متغیرهای اثرگذار بر عملکرد شناسایی شوند و در مدل‌سازی به کار گرفته شوند.

اجرای روش‌های RF و SVM

ماشین بردار پشتیبان (SVM)

ماشین بردار پشتیبان یک الگوریتم یادگیری ماشینی[1] (ML) است که ممکن است برای دستیابی به راه‌حل‌هایی برای مدیریت بهتر محصول استفاده شود. ماشین بردار پشتیبان در مسائل طبقه‌بندی دو یا چند گروهه و رگرسیون مورد استفاده قرار می‌گیرد. مانند دیگر روش‌های یادگیری ماشینی، در این روش نیز فرآیند ساخت مدل، به دو مرحلهٔ آموزش و آزمایش تقسیم می‌شود. همچنین داده‌ها به صورت تصادفی و با حفظ نسبت کلاس‌ها در هر مجموعه، با استفاده از روش نمونه‌گیری تصادفی طبقه‌بندی شده[2] به مجموعه‌های آموزش و آزمایش تقسیم شده‌اند تا از لحاظ آماری، توزیع کلاس‌ها در هر مجموعه، مشابه کل داده‌ها باشد. سهم داده‌های آموزشی ۷۰ درصد داده‌ها و سهم داده‌های آزمایشی ۳۰ درصد داده‌ها است و بر مبنای الگوریتم داده‌کاوی عمل می‌کند. ماشین بردار پشتیبان یکی از روش‌های دسته‌بندی نظارت شده خطی داده‌ها است (۲۵).

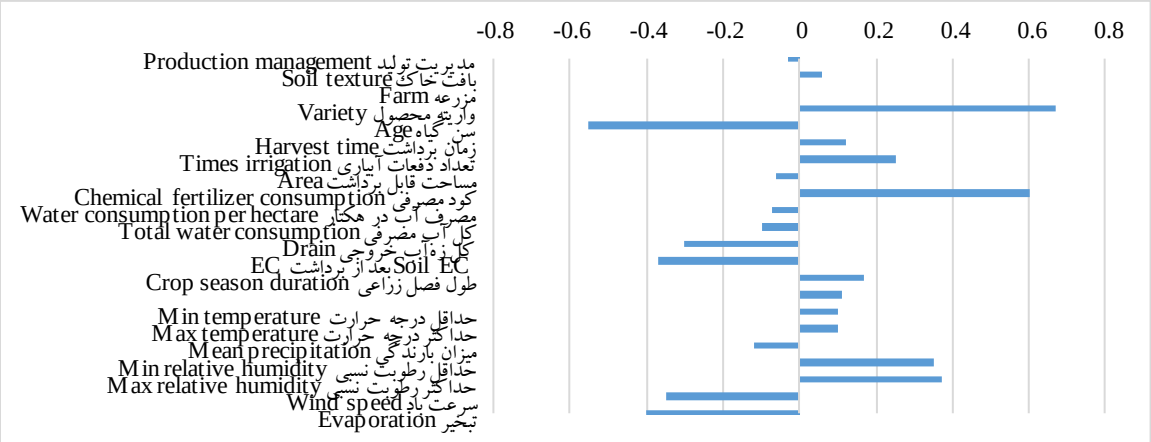
ایده اصلی ماشین بردار پشتیبان، یافتن یک ابر صفحه مطلوب به‌عنوان سطح تصمیم‌گیری است، به طوری که فاصله بین دو کلاس را حداکثر کند. داده‌هایی که به ابر صفحه نزدیک‌اند، بردارهای پشتیبان نامیده می‌شوند. درصورتی که داده‌ها در مسائلی به‌صورت خطی جداپذیر نباشند، داده‌ها با یک کرنل[3] غیرخطی به فضای با ابعاد بیشتر منتقل می‌شوند و در فضای جدید به‌صورت خطی جدا می‌شوند تا بتوان ابر صفحه مطلوب را تعیین کرد (۱۲ و ۲۳). کرنل تابعی است که ضرب داخلی[4] بین نگاشت‌های ویژگی را در فضای ویژگی‌های پنهان محاسبه می‌کند. در واقع کرنل، ضرب داخلی دو بردار

1- Machine Learning
2- stratified random sampling
3- Kernel
4- Inner product

در فضای ویژگی‌ها را بدون نیاز به محاسبه‌ی مستقیم نگاشت آن‌ها با بعد بالاتر محاسبه می‌کند.

روش کار ماشین بردار پشتیبان به این صورت است که فرض کنید I داده‌های آموزشی موجود می‌باشد که هر یک با (y_i, x_i) نشان داده می‌شود. x_i بردار ویژگی n بعدی و $y_i \in \{1, -1\}$ برحسب آن است. هدف، یافتن ابر صفحه است که دو کلاس را برحسب ۱ و ۱- با بیشترین حاشیه از هم جدا کند. این ابر صفحه را می‌توان با رابطه ۱ بیان کرد.

$$W^T\varphi(X) + b = 0 \qquad (۱)$$



شکل (۱) همبستگی بین متغیرهای مستقل با متغیر وابسته (عملکرد نیشکر)

Figure (1) Features correlation with the dependent variable (sugarcane yield)

جدول (۱) توصیف متغیرهای گسسته

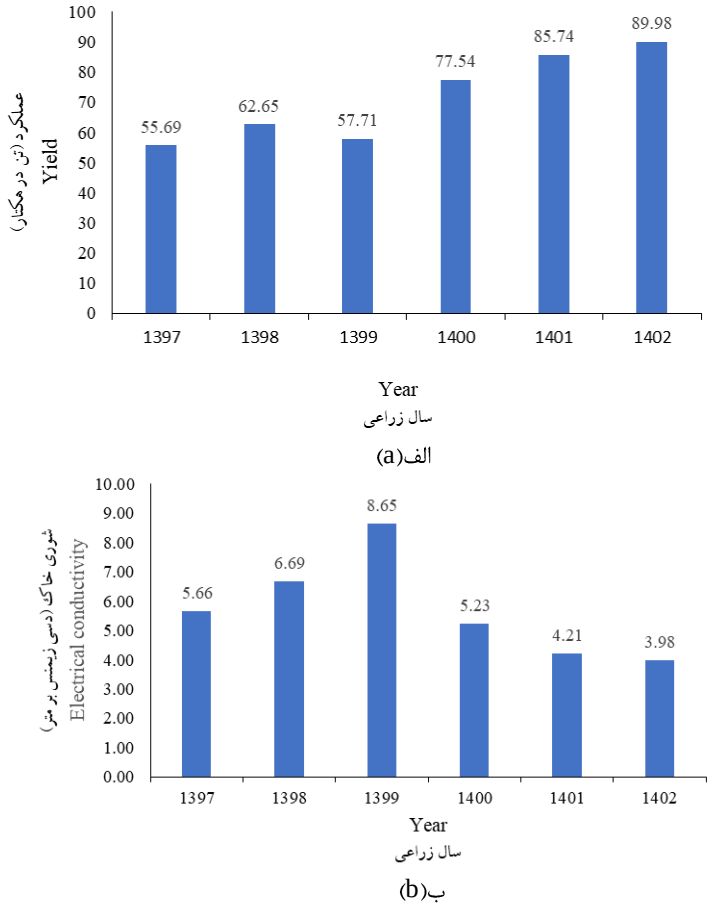
Table (1) Description of discrete variables

ردیف Row	نام متغیر Variable name	نوع داده Data type	توصیف متغیر Variable description
1	مدیریت تولید Production management	اسمی Nominal	1, 2, 3, 4
2	بافت خاک Soil texture	اسمی Nominal	شنی رسی لومی، شنی لومی، سیلتی رسی ، سیلتی رسی لومی، سیلتی لوم، رسی لومی، لومی sandy clay loam, sandy loam, silty clay, silty clay loam, silty loam, clay loam, loam
3	مزرعه Farm	اسمی Nominal	440 مزرعه (440 Farms)
4	واریته محصول Variety	اسمی Nominal	SP70-1143 ،IRC99-02، CP69-1062، CP57-614 ،CP48-103
5	سن گیاه Age	ترتیبی Ordinal	PCمزارع تازه کشت شده و R مزارع بازرویی یا کشت راتون
6	زمان برداشت Harvesting time	اسمی Nominal	فروردین، اردیبهشت، مهر، آبان، آذر، دی، بهمن، اسفند April, May, October, November, December, January, February, March
7	تعداد دفعات آبیاری Number of irrigation times	ترتیبی Ordinal	12-34 نوبت (12 to 34 times)

جدول (۲) توصیف متغیرهای پیوسته

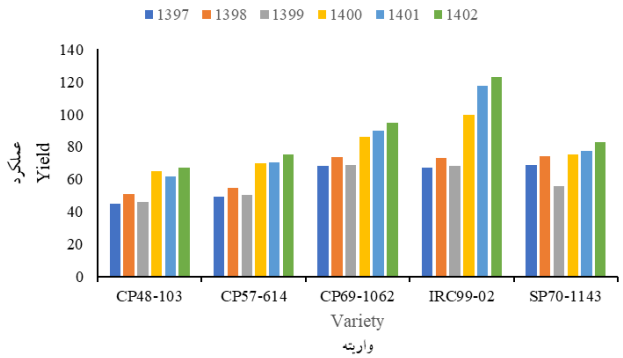
Table (2) Description of continuous variables

ردیف Row	نام متغیر (Variable name)	نوع داده Data type	توصیف متغیر (Variable description)	واحد (Unit)
1	مساحت قابل برداشتArea	عددی Numerical	مساحت مزرعه	هکتار (ha)
2	کود مصرفی Chemical fertilizer consumption	عددی Numerical	میزان کل کود مصرفی برای مزارع	کیلوگرم بر هکتار (Kg/ha)
3	مصرف آب در هکتار Water consumption per hectare	عددی Numerical	میزان آب داده شده به مزرعه در هر نوبت آبیاری در هر هکتار	مترمکعب بر هکتار M^3/ha
4	کل آب مصرفی Total water consumption	عددی Numerical	میزان کل مصرف آب در یک فصل زراعی در هر هکتار	مترمکعب بر هکتار M^3/ha
5	کل زه‌آب خروجی Drain	عددی Numerical	میزان آب خارج شده از سیستم زهکشی مزرعه	مترمکعب بر هکتار M^3/ha
6	عملکرد Yield	عددی Numerical	مقدار نیشکر برداشت شده در واحد سطح	تن بر هکتار (ton/ha)
7	EC بعد از برداشت Soil EC	عددی Numerical	هدایت الکتریکی شوری خاک پس از عملیات برداشت محصول	دسی زیمنس بر متر dS/m
8	طول فصل زراعی Crop season duration	عددی Numerical	برای مزارع تازه کشت زمان کشت تا برداشت و برای مزارع بازرویی فاصله زمانی برداشت قبل تا برداشت بعدی	روز (Day)
9	فاصله زمانی قطع آب تا برداشت محصول Time interval of drying off to crop harvest	عددی Numerical	مدت زمان بین قطع عملیات آخرین آبیاری مزرعه تا عملیات برداشت محصول (تنش خشکی به گیاه جهت رسیدگی مطلوب)	روز (Day)
10	حداقل درجه حرارت Min temperature	عددی Numerical	میانگین حداقل درجه حرارت در بازه ۴ ماهه اردیبهشت- شهریور در هر سال زراعی که به‌صورت روزانه اندازه گیری شده است.	سلسیوس Celsius
11	حداکثر درجه حرارت Maximmum temperature	عددی Numerical	میانگین حداکثر درجه حرارت در بازه ۴ ماهه اردیبهشت - شهریور در هر سال زراعی که به‌صورت روزانه اندازه گیری شده است.	سلسیوس Celsius
12	میزان بارندگی Mean precipitation	عددی Numerical	میانگین میزان بارش در بازه ۷ ماهه مهر -فروردین در هر سال زراعی که به‌صورت روزانه اندازه گیری شده است.	میلی متر (mm)
13	حداقل رطوبت نسبی Minimum relative humidity	عددی Numerical	میانگین حداقل رطوبت نسبی در بازه ۴ ماهه اردیبهشت- شهریور در هر سال زراعی که به‌صورت روزانه اندازه گیری شده است.	درصد Percentage
14	حداکثر رطوبت نسبی Maximum relative humidity	عددی Numerical	میانگین حداکثر رطوبت نسبی در بازه ۴ ماهه اردیبهشت- شهریور در هر سال زراعی که به‌صورت روزانه اندازه گیری شده است.	درصد Percentage
15	سرعت باد (Wind speed)	عددی Numerical	میانگین سرعت باد در بازه ۴ ماهه اردیبهشت - شهریور در هر سال زراعی که به‌صورت روزانه اندازه گیری شده است.	متر بر ثانیه (m/s)
16	تبخیر Evaporation	عددی Numerical	میانگین تبخیر در بازه ۴ ماهه اردیبهشت - شهریور در هر سال زراعی که به‌صورت روزانه اندازه گیری شده است.	میلی متر mm



شکل (۲) میانگین عملکرد نیشکر (الف) و میانگین شوری خاک (ب) در بازه زمانی ۱۳۹۷-۱۴۰۲

Figure (2) Average sugarcane yield (a) and Average soil salinity (b) in the period 2018-2023



شکل (۳) عملکرد نیشکر به تفکیک واریته‌های مختلف و سال زراعی

Figure (3) Sugarcane yield by different varieties and growing season

جنگل تصادفی (RF)

جنگل تصادفی یکی از انواع الگوریتم‌های یادگیری ماشین است و به طور وسیعی در تحقیقات، به ویژه برای پیش‌بینی محصولات کشاورزی استفاده می‌شود. همچنین دارای تعداد زیادی درخت تصمیم برای طبقه‌بندی یا رگرسیون است (۱ و ۹). روش جنگل تصادفی که جزء روش‌های یادگیری ماشین است، قابلیت

استفاده برای پیش‌بینی داده‌های کمی و رده‌بندی داده‌های کیفی را داراست. RF یکی از بهترین الگوریتم‌های یادگیری ماشین است که بر اساس درخت تصمیم‌گیری است که در آن درختان زیادی برای تخمین، ترکیب می‌شوند (۵، ۸). این مدل طبقه‌بندی با دقت بالا را برای تعداد زیادی از مجموعه‌های داده استفاده می‌کند (۱۴). مانند دیگر روش‌های یادگیری ماشینی، در این روش نیز فرآیند ساخت مدل به دو مرحلهٔ آموزش و آزمایش تقسیم می‌شود. سهم داده‌های آموزشی ۷۰ درصد داده‌ها و سهم داده‌های آزمایشی ۳۰ درصد داده‌ها است. در این الگوریتم، یکی از پارامترهای اصلی که باید تنظیم شود، متغیر k برای هر گره است. راحت‌ترین راه برای تعیین k محاسبه ریشه مربع تعداد کل متغیرهای مستقل است. به‌عنوان مثال رابطه ۲، $n \leq \sqrt{k}$، که n تعداد متغیرهای ورودی است.

$$Y'_n = (fX_n) = \frac{1}{k}\sum_{K}^{K} = 1\, T_{L(\theta\pi)} \qquad (۲)$$

هر درخت برآورده کننده $T_{L\theta}$ به یک بردار یا خط تصادفی θوابسته بوده که نشان دهنده دسته‌بندی نمونه از مجموعه آموزشی L است. برآورد کننده نهایی f، اکثریت رأی یا میانگین برای همه درخت‌ها است.

بهینه‌سازی پارامترهای مدل و ثبت نتایج مدل بهینه

همان‌طور که اشاره شد هر یک از روش‌های RF و SVM در ساختار خود دارای پارامترهای متعددی هستند که تعیین مقدار این پارامترها در اختیار کاربر است. یکی از مهمترین مراحل تحقیق، انتخاب صحیح پارامترهای مدل برای رسیدن به نتایج مطلوب است. معمولاً در فرآیند بهینه‌سازی پارامترهای مدل، از انواع نمودارهای روند تغییرات و الگوریتم‌های بهینه‌سازی استفاده می‌شود. در این پژوهش نیز هر دو روش مذکور به کار گرفته شده است. توجه شود که در برخی موارد در بهینه کردن یک پارامتر، یک رابطه‌ی مستقیم بین پیچیدگی مدل و عملکرد آن وجود دارد. به طوری که بهبود عملکرد مدل نیازمند پیچیدگی بیشتر در مدل است. بنابراین کاربر انتخاب می‌کند که حاضر است چه مقدار پیچیدگی مدل را بپذیرد تا به دقت مدنظرش برسد. در واقع ممکن است کاربر از مدل مطلوب‌تر صرف‌نظر کند تا مدلی ساده‌تر و کم‌هزینه‌تر (در اینجا هزینه می‌تواند زمان اجرای مدل باشد) داشته باشد. بهینه‌سازی پارامترهای مدل RF با استفاده از نمودارهای تغییرات انجام شده است. در بهینه‌سازی پارامترهای مدل SVM از الگوریتم هوک و جیوز[1] استفاده شده است. پس از آنکه در یک پایگاه داده، مدل بهینه پیدا شد، نتایج مورد نیاز این مدل (معیارهای دقت و خطا و مقادیر پیش بینی شده‌ی متغیر پاسخ) ثبت می‌شود. همچنین در زمینه یادگیری ماشین، دو مفهوم اساسی به نام‌های بیش‌برازش و کم‌برازش وجود دارد که بر عملکرد و پایداری مدل‌ها تأثیرگذار هستند. وقتی یک مدل آن‌قدر به داده‌های اموزش حساس می‌شود که الگوهای تصادفی یا نویز را هم یاد می گیرد، به طوری که روی داده‌های جدید عملکرد ضعیفی دارد، مدل دچار بیش‌برازش شده است و زمانی که مدل به قدری ساده است که نمی‌تواند حتی الگوهای اساسی موجود در داده‌های آموزشی را یاد بگیرد، مدل دچار کم‌برازش شده است. همچنین بیش‌برازش به معنای دقت دسته‌بندی داده‌های آموزشی است. در این حالت، دقت مدل بر روی داده‌های آموزشی بسیار بالا و خطای آن بسیار پایین است. اما، ممکن است مدل به دلیل پیچیدگی بالا، داده‌های آموزشی را به طور دقیق یاد بگیرد و دارای بایاس کمی باشد اما در برابر داده‌های آزمایشی حساسیت زیادی دارد و نمی‌تواند آن‌ها را به خوبی تشخیص دهد که باعث افزایش واریانس می‌شود. به طور کلی، وقتی بیش‌برازش رخ می‌دهد، مدل‌های پیچیده‌ای ایجاد می‌شوند که اما قابلیت تعمیم‌پذیری ندارند و عملکرد آن‌ها قابل‌قبول نیست. یکی از راه‌های کنترل بیش‌برازش، تنظیم و انتخاب بهینه هایپرپارامترها است. به عنوان مثال در الگوریتم جنگل تصادفی، تنظیم حداکثر عمق درخت و تعداد درخت‌ها پارامترهای مهمی هستند که با تعداد کمتر درخت و عمق کمتر، بیش‌برازش را کاهش می‌دهند. در مقابل، در کم‌برازش، دسته‌بندی غیردقیقی از داده‌ها انجام می‌شود و

1- Hook and Jeeves method

مدل ساخته شده ساده‌تر و با دقت پایین‌تر و خطا بالاتر است. مدل در این حالت دارای واریانس پایین و بایاس بالایی است. اما، عملکرد آن بر روی داده‌های آزمایشی و آموزشی پایین است و عملکرد آن قابل قبول نیست.

برای کنترل بیش‌برازش و کم‌برازش، ابتدا هایپرپارامترهای مدل با استفاده از جستجوی شبکه‌ای تنظیم شد و سپس صحت مدل‌ها بر روی داده‌های آموزشی و آزمایشی بررسی شد. جستجوی شبکه‌ای به عنوان یک روش تنظیم هایپرپارامتر مورد استفاده قرار گرفت که بهینه‌سازی عملکرد مدل را هدف قرار می‌دهد. هایپرپارامترها یا فراپارامترها، پارامترهایی هستند که مقادیر آن‌ها فرآیند یادگیری را کنترل می‌کند و مقادیر پارامترهای مدلی را تعیین می‌کند که یک الگوریتم یادگیری به یادگیری آن‌ها ختم می‌شود.

جستجوی شبکه‌ای اساساً یک الگوریتم بهینه‌سازی است که این امکان را می‌دهد بهترین پارامترها برای مسئله بهینه‌سازی از فهرستی از گزینه‌های پارامتری که ارائه می‌شود، انتخاب کرد، بنابراین روش آزمون و خطا را خودکار می‌کند.

ارزیابی مدل‌ها

اعتبارسنجی متقابل

اعتبارسنجی متقابل، یکی از روش‌های ارزیابی مدل است که تعیین می‌کند نتایج مدل ساخته شده بر روی یک مجموعه داده تا چه اندازه قابل تعمیم و مستقل از داده‌های آموزشی است. در اعتبارسنجی متقابل که همراه با پارامتر k استفاده می‌شود، داده‌های اولیه به‌صورت تصادفی به k زیرمجموعه‌ی $kD1$، $kD2$، ... که دو به دو ناسازگار هستند و اندازه تقریباً یکسانی دارند، تقسیم می‌شوند. فرآیند آموزش و آزمایش k دفعه اجرا می‌شود. در تکرار i ام، بخش iD به عنوان مجموعه آزمایشی و بخش‌های دیگر به عنوان مجموعه آموزشی مدل در نظر گرفته می‌شود. بنابراین در تکرار اول با کمک زیرمجموعه‌های $kD1$، $kD2$، $kD3$ ... مدل ساخته شده و با زیرمجموعه $1D$ آزمایش و ارزیابی می‌شود و در تکرار دوم زیرمجموعه‌های $kD1$، $kD2$، $kD3$ ... برای ساخت مدل و $2D$ برای ارزشیابی مدل استفاده می‌شود و کار به همین ترتیب ادامه پیدا می‌کند.

به طور معمول، اعتبارسنجی متقابل با ۱۰ لایه برای ارزیابی مدل‌ها استفاده می‌شود. در این روش اعتبارسنجی متقابل طبقه‌بندی شده، مجموعه داده‌های پژوهش که شامل ۲۰۳۲ نمونه است، به ۱۰ لایه (زیرمجموعه) تقسیم می‌شود. در هر مرحله، مدل با ۹ لایه آموزش می‌بیند و با یک لایه برای آزمایش و اعتبارسنجی استفاده می‌شود. امتیاز هر مرحله ثبت می‌شود و این فرآیند ادامه می‌یابد تا تمامی لایه‌ها یک‌بار اعتبارسنجی شوند. در نهایت، امتیازهای به‌دست‌آمده از همه مراحل میانگین‌گیری شده و به عنوان امتیاز نهایی مدل اعلام می‌شود. در این روش، معیار صحت برای ارزیابی استفاده شده است.

ماتریس درهم ریختگی

استفاده از ماتریس درهم‌ریختگی به‌عنوان یک ابزار مفید برای ارزیابی دقت دسته‌بندی مدل‌ها مورد توصیه قرار می‌گیرد. این ابزار این امکان را می‌دهد تا عملکرد دسته‌بندی برای تشخیص نمونه‌ها در کلاس‌های مختلف مشاهده شود.

در وظایف طبقه‌بندی در یادگیری ماشین، یک ماتریس درهم‌ریختگی به عنوان جدولی برای خلاصه کردن عملکرد مدل عمل می‌کند که شامل تعداد پیش‌بینی‌های درست مثبت[1]، درست منفی[2]، مثبت کاذب[3] و منفی کاذب[4] است. برای طبقه‌بندی چندکلاسه، این ماتریس به شکل مربعی با ابعادی برابر با تعداد کلاس‌ها می‌شود که هر سلول نمایانگر تعداد نمونه‌ها برای ترکیب برچسب‌ها و کلاس‌های پیش‌بینی‌شده است. ردیف‌ها نشان‌دهنده کلاس‌های واقعی[5] هستند، در حالی که ستون‌ها نشان‌دهنده کلاس‌های پیش‌بینی‌شده می‌باشند. تحلیل این ماتریس دیدگاهی نسبت به نقاط قوت و ضعف مدل در تمایز بین کلاس‌های خاص ارائه می‌دهد.

1- True positive
2- True negative
3- False positive
4- False negative
5- Ground Truth

از آنجا که استفاده از داده‌های آموزشی برای ارزیابی مدل ممکن است به تخمین‌های خوش‌بینانه‌ای منجر شود، بنابراین برای ارزیابی مدل، توصیه می‌شود از داده‌های آزمایشی که در فرآیند آموزش به کار نبرده شده‌اند، استفاده شود.

در مرحله بعدی، مدل‌های ساخته شده با استفاده از سنجه‌های معرفی شده مورد بررسی قرار گرفته و ماتریس درهم‌ریختگی نتایج ارزیابی نمایش داده می‌شود. تمامی سنجه‌ها بر روی مجموعه داده‌های آزمایشی که شامل ۴۰۶ نمونه است محاسبه شد و نتایج ارزیابی اعلام گردید.

منحنی مشخصه عملکرد سیستم

در مسائل طبقه‌بندی و پیش‌بینی احتمال تعلق یک نمونه به یک کلاس خاص، می‌توان با استفاده از مقادیر مختلف آستانه[1] احتمال عضویت در کلاس‌ها را تفسیر کرد. مقدار بهینه آستانه در این مسائل می‌تواند انعطاف‌پذیرتر از پیش‌بینی مستقیم کلاس نمونه باشد. این انعطاف‌پذیری ناشی از توانایی مدل در ایجاد توازن بین خطاهای نادرست مثبت و منفی است، به خصوص در شرایطی که هزینه یک خطا بیشتر از دیگر خطاها باشد. به‌عنوان مثال، ممکن است مقدار آستانه به‌صورت پیش‌فرض برابر با ۰/۵ در نظر گرفته شود، به این معنی که اگر احتمال تعلق کمتر از ۰/۵ باشد، نمونه به کلاس منفی تخصیص می‌یابد و اگر برابر یا بیشتر از ۰/۵ باشد، به کلاس مثبت تخصیص داده می‌شود. در این پژوهش از شاخص یودن[2] به منظور یافتن مقدار بهینه آستانه استفاده شد. این روش به دنبال یافتن آستانه‌هایی است که حداکثر تفاوت بین حساسیت[3] و ویژگی[4] را به دست آورد. به عبارت دیگر، این روش تلاش می‌کند تا بهترین تعادل را بین تشخیص درست نمونه‌های مثبت و تشخیص درست نمونه‌های منفی ایجاد کند. همچنین، دو ابزار تشخیصی که برای تفسیر احتمالات پیش‌بینی در مسائل طبقه‌بندی استفاده می‌شوند، عبارتند از منحنی مشخصه عملکرد و منحنی دقت-فراخوانی. منحنی (ROC[5]) ابزار بصری مفیدی برای مقایسه دو مدل در نظر گرفته می‌شود.

تجزیه و تحلیل داده‌ها

در این تحقیق برای مدل‌سازی با ماشین‌های بردارپشتیبان (SVM) و جنگل تصادفی (RF) از نرم‌افزار پایتون نسخه 3.12.4 استفاده شد.

نتایج و بحث

شناسایی و عارضه‌یابی عوامل مؤثر بر عملکرد در مزارع نیشکر

یکی از اهداف تحقیق حاضر، برآورد دقیقی از میزان تأثیر عوامل مختلف تحت بررسی بر میزان عملکرد تن در هکتار مزارع نیشکر است. در این تحقیق، نیاز به بکارگیری روش های نوین در ساماندهی پایگاه‌های بزرگ اطلاعات عملکرد نیشکر و نیز سودمندی روش‌های داده‌کاوی به خصوص الگوریتم‌های یادگیری ماشین در تخمین عملکرد با حداکثر استفاده از دیگر پارامترها و اطلاعات برداشت شده، مورد بحث و بررسی قرار گرفت. بنابراین با توجه به یکی از اهداف تحقیق مبنی بر عارضه‌یابی و کشف علل منجر به کاهش در عملکرد مزارع تحت پوشش شرکت کشت و صنعت نیشکر دهخدا نکات کلی زیر در سطح کلان مدل‌سازی در نظر گرفته شده است:

- بر اساس داده‌های قابل دسترس و همچنین سطح کیفیت آن‌ها، الگوهای بدست آمده سعی در تبیین تفاوت‌های عملکردی مزارع در سطوح کلی دارد. بنابراین تمرکز مدل‌ها و انتخاب متغیرها با هدف توضیح روند گذشته صورت پذیرفته است.

- ارائه راهکارهای اجرایی بهبود فرآیند تولید نیشکر بایستی در سطح جزئیات صورت پذیرد که گاهاً با توجه به عدم کیفیت مناسب داده‌های در دسترس، عدم دسترسی به برخی منابع تاثیرگذار در عملکرد و نبود وضوح مورد نیاز در

1- Threshold
2- Youden's J statistic
3- True Positive Rate
4- True Negative Rate
5- Receiver Operating Characteristic

داده‌ها این موضوع جزو اهداف و بالطبع دستاوردهای این پژوهش نبود. بر اساس تجربیات و مطالعات پیشین (۱۸ و ۲۸) شاخص‌های اجرایی و مدیریتی بر تغییر سطح عملکرد محصول نیشکر تأثیرگذار است. چگونگی تأثیر پذیری سطح عملکرد وابسته به ترکیب‌های خاصی از شاخص‌های اجرایی و مدیریتی است که در قالب مدل‌های ماشین بردار پشتیبان و جنگل تصادفی استخراج شده است. اطلاعات حاصل از این پردازش را می‌توان در اختیار مدیران تولید در کشت و صنعت قرار داد و نتایج به دست آمده می‌تواند در برنامه‌ریزی و آماده‌سازی شرایط مناسب برای رسیدن به اهداف تعیین شده میزان تولید کشت و صنعت نیشکر دهخدا کمک نماید. به طور کلی، نتایج بررسی عارضه‌ها و علل کاهش عملکرد نیشکر در شرکت کشت و صنعت دهخدا نشان داد که عوامل اثرگذار در این زمینه عبارتند از: آبیاری و زهکشی،حاصلخیزی خاک و تغذیه گیاه، واریته محصول، آفات و بیماری‌ها و علف‌های هرز، عملیات زراعی خاک‌ورزی، کاشت و برداشت، عوامل اقلیمی و تنش‌های محیطی، تراکم خاک، شوری آب و خاک و نیروی انسانی و مدیریتی.

مدل‌سازی

برای ساخت مدل، داده‌ها با روش هلداوت[1] به دو بخش تقسیم شده‌اند. در این روش داده‌ها به‌صورت تصادفی به دو مجموعه‌ی مستقل آموزشی و آزمایشی تقسیم می‌شوند. در این پژوهش نیز ۷۰ درصد از داده‌ها که شامل ۱۶۲۶ نمونه است برای مجموعه آموزشی و ۳۰ درصد از داده‌ها که شامل ۴۰۶ نمونه است برای مجموعه آزمایشی انتخاب شدند، شکل ۴، نمودار مجموعه داده‌های آموزشی و آزمایشی را به تفکیک متغیر هدف (عملکرد نیشکر) نشان می‌دهد. مدل‌سازی در دو مرحله انجام شده است. در مرحله اول دو الگوریتم مورد استفاده در پژوهش، با هایپرپارامترهای[2] پیشفرض کتابخانه سایکیت لرن[3] ساخته شده و با معیار صحت ارزیابی شده‌اند. هایپرپارامترهای مدل، مقادیری هستند که میزان پیچیدگی و فرآیند یادگیری مدل را کنترل می‌کنند، همانند عمق درخت تصمیم. جدول ۳، امتیاز مدل‌ها را در مجموعه داده‌های آزمایشی و آموزشی با در نظر گرفتن عملکرد نیشکر به عنوان متغیر هدف نشان می‌دهد. در این پژوهش، ویژگی هدف مسئله برای تعیین کلاس به دو بازه تقسیم شده است و ویژگی عملکرد محصول نیشکر (مقدار نیشکر تولید شده بر حسب تن در هکتار) به دو کلاس مطلوب (۱) و نامطلوب (۰) تقسیم شده است. به طوری که مقادیر بیشتر مساوی ۷۵ برای عملکرد محصول نیشکر در کلاس یک و مقادیر پایین تر از ۷۵ در کلاس صفر قرار دارد. در مرحله پیش پردازش داده‌ها، تعادل توزیع کلاس‌ها[4] به منظور جلوگیری از سوگیری مدل به سمت کلاس غالب، انجام شد. تعداد نمونه‌های کلاس ۰ و ۱ به ترتیب ۱۰۰۵ و ۱۰۲۷ بودند که نشان‌دهنده‌ی تعادل آماری مناسب بین کلاس‌هاست.

همانطور که جدول ۳ نشان می‌دهد، امتیاز مدل ماشین بردار پشتیبان در داده‌های آزمایش و آموزش به ترتیب ۸۴/۱ ٪ و ۸۵/۴ ٪ است. در تحقیقی مشابه، ذکی دیزجی و همکاران[5] (۲۸) در مدل‌سازی پیش‌بینی عملکرد محصول نیشکر با استفاده از درخت‌های تصمیم نیز به دقتی حدود ۸۰ درصد دست یافتند. همچنین بر اساس نتایج جدول ۵، مدل جنگل تصادفی با دقت ۸۵/۹ ٪ عملکرد بهتری نسبت به مدل مدل ماشین بردار پشتیبان را ارائه می‌دهد. اما، دقت این مدل بر روی داده‌های آموزشی ۹۸/۲٪ بوده و این اختلاف بزرگ بین دقت مدل بر روی داده‌های آزمایشی و آموزشی نشان‌دهنده وجود بیش‌برازش[6] است.

جدول ۴ برخی از هایپرپارامترهای مدل‌های استفاده شده در این پژوهش به‌همراه مقدار بهینه انتخاب شده آن‌ها برای مدلسازی عملکرد نیشکر را نشان می‌دهد. برای پیاده‌سازی روش جستوجوی شبکه‌ای از کتابخانه سایکیتلرن، ماژول انتخاب مدل و تابع جستجوی شبکه‌ای استفاده شده است.

1- Holdout
2- Hyperparameters
3- Scikit learn
4- class balance
5- Zaki dizaji *et al.*
6- Overfitting

در ادامه فرآیند آموزش مدل‌ها با پارامترهای بهینه انتخاب شده با روش جستجوی شبکه‌ای انجام شده است، ۷۰٪ از داده‌ها برای آموزش و ۳۰٪ برای آزمایش استفاده شدند.

جدول ۵، امتیاز مدل‌ها را با درنظر گرفتن عملکرد نیشکر به عنوان متغیر هدف در دو مجموعه داده آموزش و آزمایش، بر اساس معیار صحت نشان می‌دهد. همانطور که جدول نشان می‌دهد مدل جنگل تصادفی با بیشترین میزان صحت بر روی مجموعه داده آزمایش، بهترین عملکرد را دارد و بر اساس این معیار، بهترین مدل برای پیش‌بینی عملکرد نیشکر است. در تحقیقی مشابه نادر نژاد و همکاران[1] (۱۹) برای پیش‌بینی عملکرد شکر استحصالی از نیشکر از الگوریتم‌های جنگل تصادفی، آدابوست، تقویت گرادیان حداکثری و ماشین بردار پشتیبان استفاده کردند. مدل جنگل تصادفی با صحت ۹۲/۲ ٪ برای پیش‌بینی شکر استحصالی در بین مدل‌های ارائه شده بهترین عملکرد را داشت.

دسته‌بند ترکیبی رأی گیری[2]

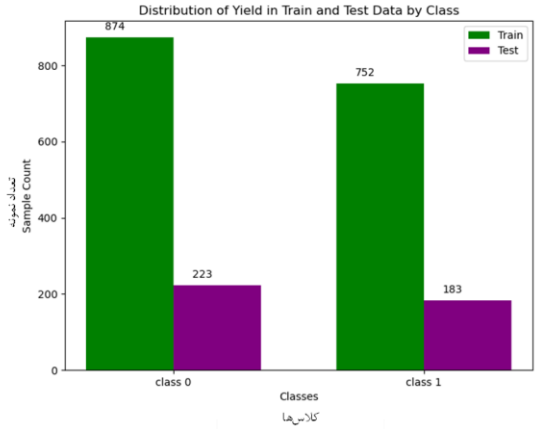
برای پیش‌بینی عملکرد نیشکر، صحت دسته‌بند ترکیبی با رأی‌گیری نرم روی داده‌های آموزشی و آزمایشی محاسبه و در جدول ۶ ارائه شده است. دسته‌بند ترکیبی رأی‌گیری، یک فراطبقه‌بند[3] است که برای ترکیب چندین الگوریتم یادگیری ماشین از دو رویکرد رأی‌گیری سخت[4] (پیش‌بینی نهایی کلاس بر اساس رأی اکثریت دسته‌بندها) و رأی‌گیری نرم[5] (پیش‌بینی نهایی بر اساس میانگین احتمال پیش‌بینی کلاس‌ها) استفاده می‌کند. نمودار حاصل، پیش‌بینی احتمال کلاس‌ها را برای اولین نمونه در مجموعه داده‌های آزمایشی نشان می‌دهد که براساس آن، این نمونه در کلاس یک قرار می‌گیرد (شکل ۵).

ارزیابی مدل‌های طبقه‌بندی و تحلیل نتایج

در این پژوهش، انواع مختلف مدل‌ها روی مجموعه داده‌های مورد مطالعه ایجاد شدند و بر اساس آن‌ها عملکرد نیشکر به دو دسته مطلوب و نامطلوب دسته‌بندی شد. حالا نیاز است تا دقت این مدل‌ها روی داده‌هایی که مدل‌ها بر اساس آن‌ها آموزش نیافته‌اند، بررسی شود. برای ارزیابی دقت مدل‌ها و میزان پایداری و قابل اعتماد آن‌ها در برخورد با داده‌های جدید، از سنجه‌ها و روش‌های مختلفی استفاده می‌شود. در این بخش، مدل‌های ساخته شده با استفاده از روش اعتبارسنجی متقابل داده‌ها[6]، ماتریس درهم‌ریختگی[7]، منحنی مشخصه عملکرد[8] و منحنی دقت فراخوانی[9] ارزیابی شدند. نتایج اعتبارسنجی متقابل برای مدل‌های پیش‌بینی عملکرد نیشکر در جدول ۷ نمایش داده شده است. همچنین، نمودار شکل ۶، نشان‌دهنده دامنه تغییرات امتیاز مدل‌ها در اعتبارسنجی متقابل با ۱۰ لایه و ۲ تکرار برای پیش‌بینی عملکرد نیشکر را نمایش می‌دهد.

همانطور که جدول ۷ نشان می‌دهند، مدل‌های ساخته شده بر اساس معیار صحت و با روش اعتبار سنجی متقابل داده‌ها ارزشیابی شدند. بر این اساس مدل جنگل تصادفی با بیشترین میزان صحت ۸۲/۴ درصد بهترین مدل برای پیش‌بینی عملکرد نیشکر است. همچنین با توجه به نمودار جعبه‌ای به وضوح مشاهده می‌شود که این مدل به طور متوسط عملکرد بهتری نسبت به مدل دیگر دارد. در تحقیقی مشابه، یامپارلا و همکاران[10] (۲۶) برای پیش‌بینی عملکرد محصول از الگوریتم جنگل تصادفی استفاده کردند. تمرکز این تحقیق بر استفاده از چندین الگوریتم برای پیش‌بینی عملکرد محصولات کشاورزی بود. مقایسه روش‌های یادگیری ماشین متعدد، از جمله، رگرسیون درخت تصمیم، رگرسیون خطی، تقویت گرادیان، K نزدیک‌ترین همسایه و جنگل تصادفی انجام شد. در میان آن‌ها، جنگل تصادفی دقیق‌ترین روش بود که دقت ۹۵ درصد را ارائه داد.

1- Nader nejad *et al.*
2- Ensemble Vote Classifier
3- Meta-classifier
4- Hard Voting
5- Soft Voting
6- Cross-validation
7- Confusion matrix
8- Receiver Operating Characteristic Curve
9- Stratified Cross-validation
10- Yamparla *et al.*



شکل (٤) توزیع عملکرد نیشکر در داده‌های آموزشی و آزمایشی به تفکیک کلاس‌ها
Figure (4) Distribution of sugarcane yield in training and experimental data by class separation

جدول (۳) نتایج امتیاز مدل‌های متغیر هدف عملکرد محصول نیشکر
Table (3) Score results of target variable models for sugarcane yield

صحت مجموعه داده‌های آموزش Accuracy of training datasets	صحت مجموعه داده‌های آزمایش Accuracy of the test dataset	مدل (Model)
0.982	0.859	جنگل تصادفی Random Forest
0.854	0.841	ماشین بردار پشتیبان Support Vector Machine

جدول (٤) فراپارامترهای الگوریتم‌های پژوهش
Table (4) Hyperparameters of research algorithms

مقدار بهینه فراپارامترها با روش جستجوی شبکه‌ای به‌همراه اعتبارسنجی متقابل Optimal hyperparameter values s obtained through network search method with cross-validation	الگوریتم Algorithm
تعداد درختان: تعداد درختان تصمیم در جنگل، ۱۰۰ Number of trees: Number of decision trees in the forest, 100 حداکثر عمق درخت: ۷ Maximum depth of tree: 7 حداقل نمونه در گره برگ: کمترین نمونه لازم در گره که برگ محسوب شود،۱۲ Minimum sample at leaf node: The minimum sample required at a node to be considered a leaf, 12 حداکثر ویژگی‌ها: بیشترین تعداد ویژگی‌هایی که هنگام جستجوی بهترین تقسیم باید در نظر گرفته شود، ۷ Max Features: The maximum number of features to consider when searching for the optimal split, 7	جنگل تصادفی (Random Forest)
کرنل: هسته، جداسازی کلاس‌ها در فضای داده‌ها. هسته تابع پایه شعاعی. rbf Kernel: kernel, class separation in data space. Radial basis function kernel. rbf فراپارامتر جریمه C: پارامتر تنظیم میزان جریمه‌ای که به داده‌هایی که اشتباه دسته‌بندی شده‌اند داده می‌شود، ۱۵ Hyperparameter Penalty C: Parameter that adjusts the penalty assigned to misclassified data, 15 گاما: ضریب کرنل برای چندجمله‌ای، سیگموید و پایه شعاعی می‌باشد. ۰,۳ Gamma: Kernel coefficient for polynomial, sigmoid and radial basis. 0.3 درجه: درجه کرنل چند جمله‌ای Degree: Polynomial kernel degree	ماشین بردار پشتیبان (Support Vector Machine)

جدول (۵) نتایج مدل‌های ساخته شده با پارامترهای بهینه متغیر هدف عملکرد نیشکر

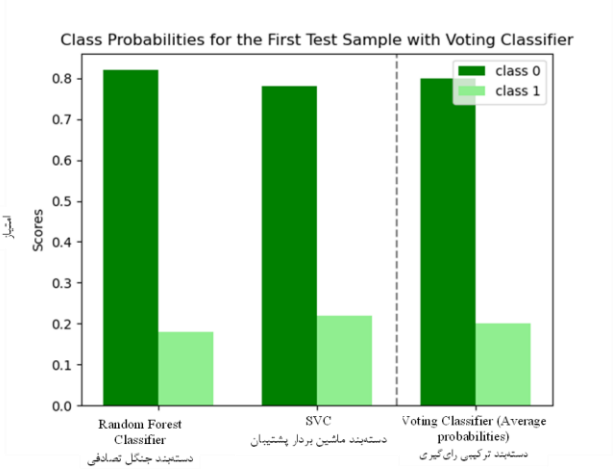
Table (5) Results of models built with optimal parameters of the target variable of sugarcane yield

صحت مجموعه داده آموزش Accuracy of training datasets	صحت مجموعه داده آزمایش Accuracy of the test dataset	مدل (Model)
0.871	0.856	جنگل تصادفی (Random Forest)
0.869	0.845	ماشین بردار پشتیبان (Support Vector Machine)

جدول (٦) نتایج دسته‌بند ترکیبی رأی‌گیری در پیش‌بینی عملکرد نیشکر

Table (6) Results of the combined voting classification in sugarcane yield prediction

صحت روی مجموعه داده آزمایش Accuracy of the test dataset	صحت روی مجموعه داده آموزش Accuracy of training datasets	مدل(Model)
0.89	0.93	دسته‌بند ترکیبی برای پیش‌بینی عملکرد نیشکر Hybrid classifier for sugarcane yield prediction



شکل (۵) احتمال کلاس‌ها برای نمونه اول مجموعه آزمایشی با دسته‌بند ترکیبی رأی‌گیری

Figure (5) Class probabilities for the first sample of the test set with the combined voting classifier

در مرحله بعدی، ماتریس درهم‌ریختگی نتایج ارزیابی نمایش داده می‌شود. در پیش بینی عملکرد نیشکر دو کلاس مطلوب (کلاس مثبت) و نامطلوب (کلاس منفی) وجود دارد. تمامی سنجه‌های استفاده شده بر روی مجموعه داده‌های آزمایشی محاسبه شدند. از ۴۰۶ نمونه آزمایشی، ۲۰۴ نمونه متعلق به کلاس منفی (صفر) و ۱۹۶ نمونه متعلق به کلاس مثبت (یک) است. شکل ۷ ماتریس درهم ریختگی مدل‌های ساخته شده برای پیش‌بینی عملکرد نیشکر را نشان می‌دهد.

در ماتریس درهم ریختگی هر چه مقادیر روی قطر اصلی یا به عبارتی تعداد پیش بینی های درست بیشتر باشند و مقادیر روی قطر فرعی یا پیش بینی های نادرست کمتر باشند عملکرد آن مدل بهتر میباشد. با توجه به شکل ۷، مدل جنگل

تصادفی برای پیش بینی کلاس منفی عملکرد بهتری داشته و است. در تحقیقی مشابه، در پیش‌بینی عملکرد محصول نیشکر در کشت و صنعت نیشکر امیرکبیر با استفاده از الگوریتم داده‌کاوی درخت تصمیم، نتایج نشان داد درخت‌های تصمیم توانستند ۷۳ درصد پیش‌بینی درست و ۲۳ درصد پیش‌بینی نادرست داشته باشند (۲۹). همچنین جدول ۸ و ۹ امتیاز مدل‌ها را به تفکیک کلاس‌ها بر اساس سنجه‌های دقت[1]، حساسیت[2] و امتیاز F_1 نشان می‌دهند. بر اساس جدول ۸، مدل جنگل تصادفی ۸۱ درصد از نمونه هایی را که در کلاس مثبت دسته‌بندی کرده است، واقعا در کلاس مثبت بوده‌اند. همچنین ۸۲ درصد از نمونه‌های متعلق به کلاس مثبت به درستی دسته‌بندی شده‌اند. همچنین این مدل ۸۷ درصد از نمونه‌هایی را که در کلاس منفی دسته‌بندی کرده است، واقعا در کلاس منفی بوده‌اند و ۸۵ درصد از نمونه‌های کلاس منفی را به درستی پیش‌بینی کرده است. معیار F_1 برای جنگل تصادفی برابر با ۸۲ درصد است. همچنین بر اساس جدول ۹، سنجه‌های اندازه‌گیری شده برای مدل ماشین بردار پشتیبان را به تفکیک کلاس‌ها نشان می‌دهد. بر این اساس، مدل ۸۱ درصد از نمونه‌هایی را که در کلاس مثبت دسته‌بندی کرده است، واقعاً در کلاس مثبت بوده‌اند. همچنین ۸۲ درصد از نمونه‌های متعلق به کلاس مثبت به درستی دسته‌بندی شده‌اند. معیار F_1 برای مدل ماشین بردار پشتیبان برابر ۸۲ درصد است. همچنین این مدل ۸۷ درصد از نمونه‌هایی را که در کلاس منفی دسته‌بندی کرده است، واقعاً در کلاس منفی بوده‌اند و ۸۷ درصد از نمونه‌های کلاس منفی را به درستی دسته‌بندی کرده است.

شکل ۸ منحنی مشخصه عملکرد مدل‌های پژوهش برای پیش‌بینی عملکرد نیشکر را نشان می‌دهد. همانطور که شکل ۸ نشان می‌دهد، مدل جنگل تصادفی با بیشترین مساحت زیر نمودار بهترین عملکرد را دارد. خط چین قرمز در نمودار بیانگر حالتی است که هیچگونه مهارت یادگیری ماشینی استفاده نشده و بر اساس شانس یا پیش‌بینی نمونه‌ها در کلاس اکثریت است. هرچه مدل‌ها از این خط فاصله بیشتری داشته و بالای آن قرار بگیرند، عملکرد بهتری دارند.

از مجموع ۵۷۲ نمونه ۱۹۰ نمونه را به درستی پیش بینی کرده

منحنی مشخصه عملکرد سیستم (ROC) ارائه شده در این پژوهش، یکی از مهم‌ترین شاخص‌های ارزیابی مدل‌های طبقه‌بندی را نمایش می‌دهد. این منحنی که رابطه بین نرخ مثبت واقعی و نرخ مثبت کاذب را نشان می‌دهد، به وضوح نزدیک به گوشه بالا–چپ نمودار است که نشان‌دهنده عملکرد عالی مدل است. سطح زیر منحنی[3] (AUC) که معیاری کمّی برای ارزیابی عملکرد کلی مدل است، مقدار بالایی را نشان می‌دهد. این منحنی همچنین نشان می‌دهد که مدل در آستانه‌های مختلف تصمیم‌گیری نیز عملکرد پایداری دارد.

منحنی خروجی جنگل تصادفی که در شکل ۹ نمایش داده شده، روند یادگیری و عملکرد این الگوریتم را به تصویر می‌کشد. این منحنی نشان می‌دهد که الگوریتم جنگل تصادفی نیز توانسته عملکرد قابل قبولی در پیش‌بینی عملکرد نیشکر داشته باشد. روند صعودی و سپس تثبیت منحنی نشان می‌دهد که مدل به خوبی همگرا شده و به نقطه بهینه رسیده است.

منحنی دقت- فراخوانی

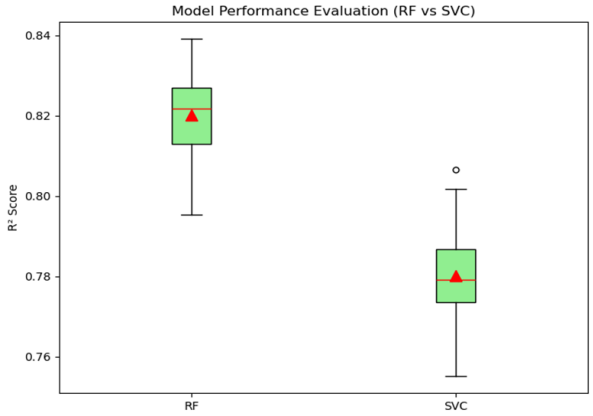
نمودار پیش‌بینی مدل جنگل تصادفی دقت قابل توجه ۰/۹۶ را نشان می‌دهد (شکل ۱۰). این دقت بالا نشان‌دهنده چندین نکته کلیدی است: اولاً، مدل جنگل تصادفی توانسته به خوبی الگوهای پیچیده و غیرخطی موجود در داده‌های نیشکر را شناسایی و مدل‌سازی کند. ثانیاً، پارامترهای مدل به درستی تنظیم شده‌اند که منجر به این عملکرد بالا شده است. همچنین، این دقت بالا نشان می‌دهد که متغیرهای ورودی انتخاب شده برای مدل از قدرت پیش‌بینی‌کنندگی بالایی برخوردار بوده‌اند. نحوه پراکنش نقاط در این نمودار نشان می‌دهد که مدل در طیف وسیعی از مقادیر، پیش بینی‌های دقیقی ارائه می‌دهد.

1- Precision
2- Sensitivity
3- Area under curve

جدول (۷) نتایج اعتبارسنجی متقابل داده‌ها عملکرد نیشکر

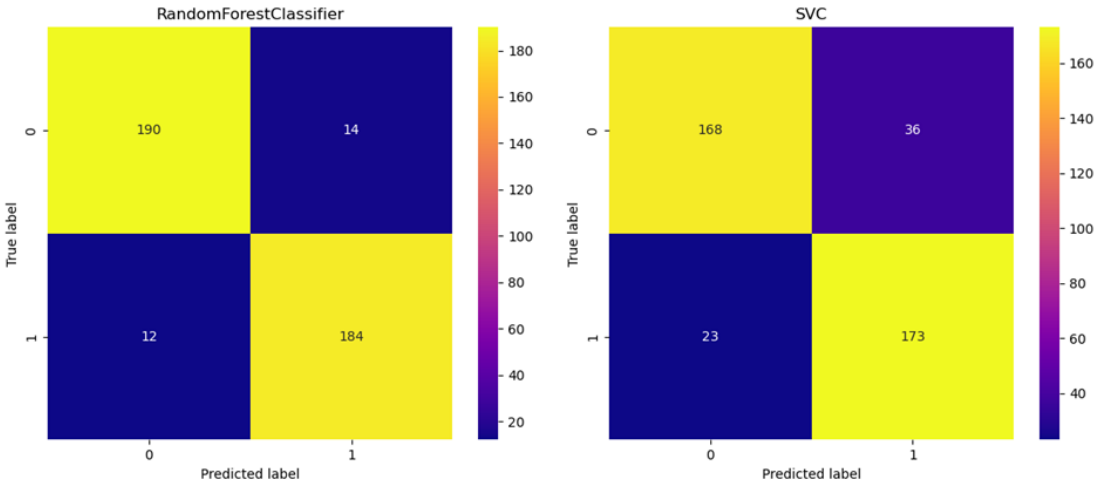
Table (7) Cross-validation Results for sugarcane yield data

صحتAccuracy	مدل (Model)
0.824	جنگل تصادفی (Random Forest)
0.778	ماشین بردار پشتیبان (Supprt Vector Machine)



شکل (٦) نمودار جعبه ای ارزیابی مدل‌ها با روش اعتبارسنجی متقابل برای پیش‌بینی عملکرد نیشکر

Figure (6) Box plot of model evaluation using cross-validation method for sugarcane yield prediction



شکل (۷) ماتریس‌های درهم ریختگی متغیر هدف عملکرد نیشکر

Figure (7) Confusion matrices for the target variable of sugarcane yield

جدول (۸) نتایج سنجه‌های ارزشیابی مدل جنگل تصادفی
Table (8) Results of evaluation metrics for the random forest model

امتیاز F_1 (F_1 Score)	حساسیت (Sensitivity)	دقت (Precision)	نام کلاس (Class name)
0.87	0.87	0.87	نامطلوب (منفی، ۰) Unfavorable (negative, 0)
0.82	0.82	0.81	مطلوب (مثبت، ۱) Favorable (positive, 1)
0.84	0.84	0.84	میانگین Average

جدول (۹) نتایج سنجه‌های ارزشیابی مدل ماشین بردار پشتیبان
Table (9) Results of evaluation metrics for the support vector machine model

امتیاز F_1 (F_1 Score)	حساسیت (Sensitivity)	دقت (Precision)	نام کلاس (Class name)
0.85	0.84	0.85	نامطلوب (منفی، ۰) Unfavorable (negative, 0)
0.79	0.79	0.79	مطلوب (مثبت، ۱) Favorable (positive, 1)
0.82	0.82	0.82	میانگین (Average)

.

دقت و خطای مدل‌ها

با توجه به جدول ۱۰، نتایج دقت مدل‌های مختلف در پیش‌بینی عملکرد نیشکر در مراحل آموزش و آزمایش نشان می‌دهد که الگوریتم‌های متفاوت هر یک عملکرد متفاوتی داشته‌اند. بر اساس داده‌های جدول ۱۰، الگوریتم جنگل تصادفی در داده‌های آموزش با ضریب همبستگی (R^2) برابر ۰/۹۵، میانگین خطای مطلق (MAE) برابر با ۰/۲۱ و ریشه میانگین مربعات خطا (RMSE) برابر ۵/۷۱ توانسته است به دقت بالایی دست یابد. این مقایسه نشان‌دهنده آن است که مدل جنگل تصادفی توانسته به خوبی داده‌های آموزش را به نتایج پیش‌بینی نزدیک کند. در مرحله آزمایش عملکرد مدل جنگل تصادفی همچنان قابل توجه بوده و با ضریب همبستگی (R^2) برابر ۰/۹۲، میانگین خطای مطلق (MAE) برابر ۰/۳۴ و ریشه میانگین مربعات خطا (RMSE) برابر ۶/۲۸ نشان می‌دهد که این مدل در پیش‌بینی داده‌های جدید و نادیده نیز بسیار موثر است.

این نتایج حاکی از آن است که مدل جنگل تصادفی دارای قابلیت تعمیم‌دهی بالایی است و می‌تواند به خوبی از داده‌های آموزش فرا گرفته و به داده‌های جدید پاسخ دهد. به طور کلی این الگوریتم در مقایسه با الگوریتم ماشین بردار پشتیبان عملکرد بهتری از خود نشان داده است. این موضوع به دلیل توانایی بالای الگوریتم در مدل‌سازی روابط پیچیده و غیرخطی بین ویژگی‌ها و متغیر هدف است. همچنین، استفاده از مدل جنگل تصادفی باعث افزایش توانایی مدل در یادگیری الگوهای پیچیده‌تر است که نتیجه آن دقت بالا در هر دو مرحله آموزش و آزمایش است. بنابراین می‌توان نتیجه گرفت که مدل جنگل تصادفی انتخاب بهتری برای پیش‌بینی دقیق عملکرد محصول نیشکر است.

در تحقیق مشابه، جاجاریا و همکاران[1] (۱۰) نتایج استفاده از الگوریتم‌های یادگیری ماشین برای پیش‌بینی عملکرد محصول در ایالت راجستان هند در پنج محصول نشان داد که جنگل تصادفی با ضریب همبستگی ۰/۹۶۳، ریشه میانگین مربعات خطا ۰/۰۳۵ و میانگین مطلق خطا ۰/۰۲۵۱ بهتر از سایرین عمل کرده است.

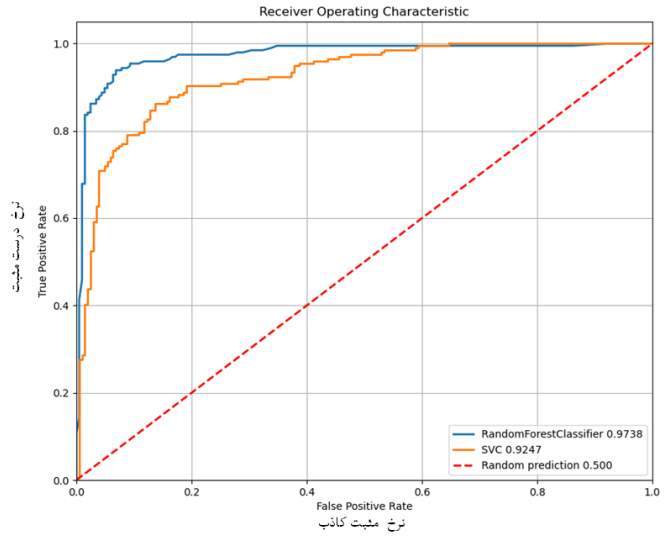
در پژوهشی دیگر، اد-دائودی و همکاران[2] (۳) برای بهبود پیش‌بینی عملکرد محصول از الگوریتم‌های یادگیری ماشین استفاده کردند. در این پژوهش، پتانسیل الگوریتم‌های یادگیری ماشین برای بهبود دقت پیش‌بینی عملکرد محصولات کشاورزی در مراکش و اثر عوامل مختلف بر عملکرد محصول، از جمله الگوهای آب و هوا، سطح رطوبت خاک، و بارندگی بررسی شد. نتایج نشان می‌دهد که الگوریتم‌های یادگیری ماشینی می‌توانند به طور قابل توجهی دقت پیش‌بینی عملکرد محصول را بهبود بخشند که با نتایج حاصل از پژوهش حاضر، هم راستا است؛ اما لازم به ذکر است که در پژوهش حاضر، با توجه به ماهیت متغیرهای تأثیرگذار بر عملکرد محصول نیشکر که ماهیتی کمی و کیفی دارند، الگوریتم‌های ماشین بردار پشتیبان و جنگل تصادفی انتخاب شده است. نتایج این تحقیق ممکن است با نتایج تحقیقات دیگر به دلیل تفاوت در داده‌ها (نوع، کیفیت و حجم داده‌ها)، متغیرهای ورودی (انتخاب و نوع متغیرها)، روش‌ها (الگوریتم‌های یادگیری ماشین، تنظیم پارامترها، و روش‌های ارزیابی) متفاوت باشد.

اهمیت متغیرها در پیش‌بینی عملکرد نیشکر

در الگوریتم یادگیری ماشین با استفاده از روش‌های انتخاب متغیر می‌توان متغیرهای کم اهمیت‌تر را کنار گذاشت و از ابعاد زیاد مدل که منجر به پیچیدگی بیش از حد آن می‌شود، جلوگیری کرد. انتخاب متغیرهای اثرگذار باعث می‌شود که توانایی تعمیم‌پذیری مدل افزایش یابد. با بررسی اهمیت متغیرها و میزان تأثیر آن‌ها بر ویژگی هدف مسئله، می‌توان زیرمجموعه‌ای از متغیرها که مرتبط هستند را انتخاب نمود و با استفاده از آن‌ها به نتیجه‌ای نزدیک یا مشابه به زمانی که همه متغیرها در ساخت مدل استفاده می‌شوند، دست یافت. در چنین حالتی درک و تفسیر مدل راحت‌تر خواهد بود.

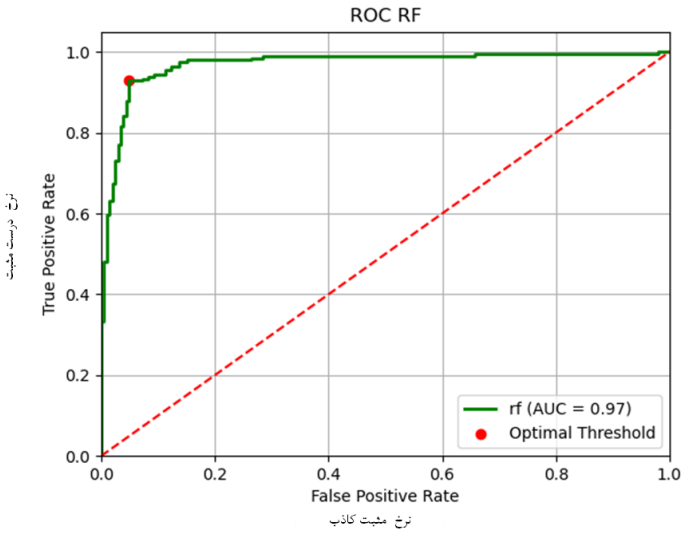
بر اساس نتایج حاصل از مدل‌سازی، متغیرهای واریته محصول، کود شیمیایی (نیتروژن)، سن گیاه، تعداد دفعات آبیاری، هدایت الکتریکی خاک، مقدار آب مصرفی در هکتار، طول فصل زراعی و زمان برداشت از مهم‌ترین و تأثیرگذارترین متغیرها در مدل جنگل تصادفی می‌باشند (شکل ۱۱، الف). در مدل ماشین بردار پشتیبان نیز متغیرهای واریته محصول، کود شیمیایی (نیتروژن)، سن گیاه، هدایت الکتریکی خاک، مقدار آب مصرفی در هکتار و کود شیمیایی (فسفر) از اهمیت بیشتری برخوردار بودند (شکل ۱۱، ب). بنابراین در دو مدل جنگل تصادفی و ماشین بردار پشتیبان مهم‌ترین متغیرهای اثرگذار بر میزان عملکرد محصول نیشکر، واریته محصول، میزان کود شیمیایی (نیتروژن) و سن گیاه می‌باشند. این نتایج با نتایج به دست آمده از تحقیقات ذکی دیزجی و همکاران (۲۸) و ذکی دیزجی و همکاران (۲۹) مطابقت دارد.

1- Jhajharia *et al.*
2- Ed-Daoudi *et al.*



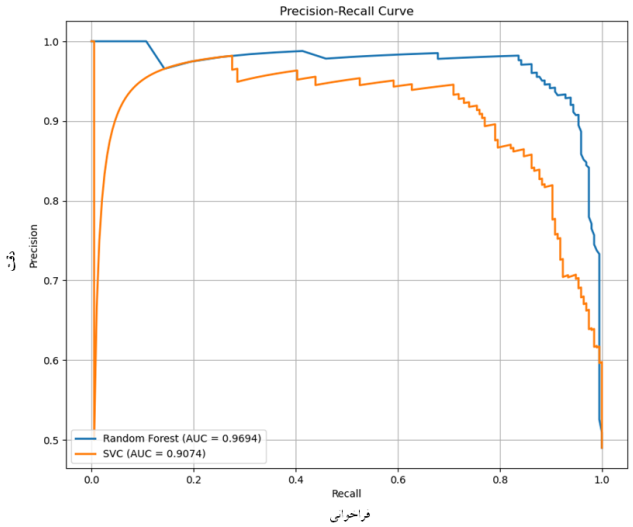
شکل (۸) منحنی مشخصه عملکرد سیستم برای مدل‌ها

Figure (8) System performance characteristic curve for models



شکل (۹) منحنی خروجی مدل جنگل تصادفی

Figure (9) Output curve of the random forest model



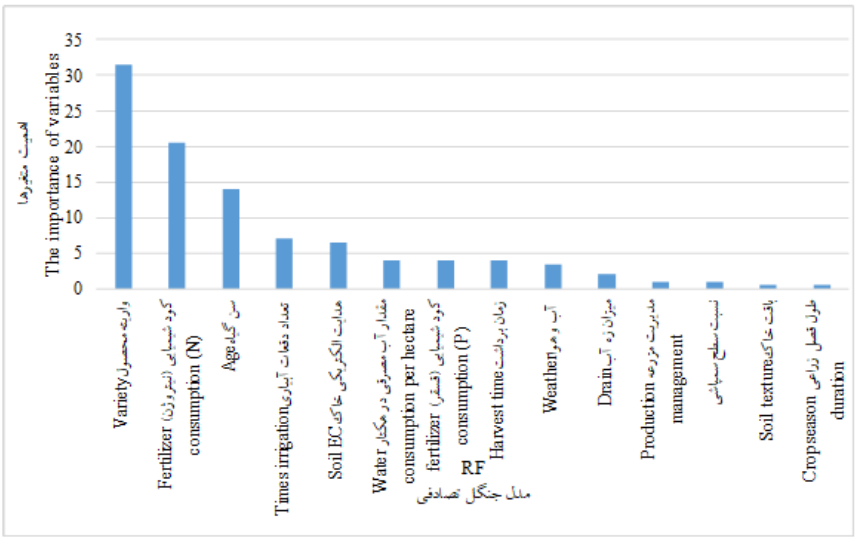
شکل (۱۰) منحنی دقت- فراخوانی مدل‌ها

Figure (10) Accuracy-frequency curve of models

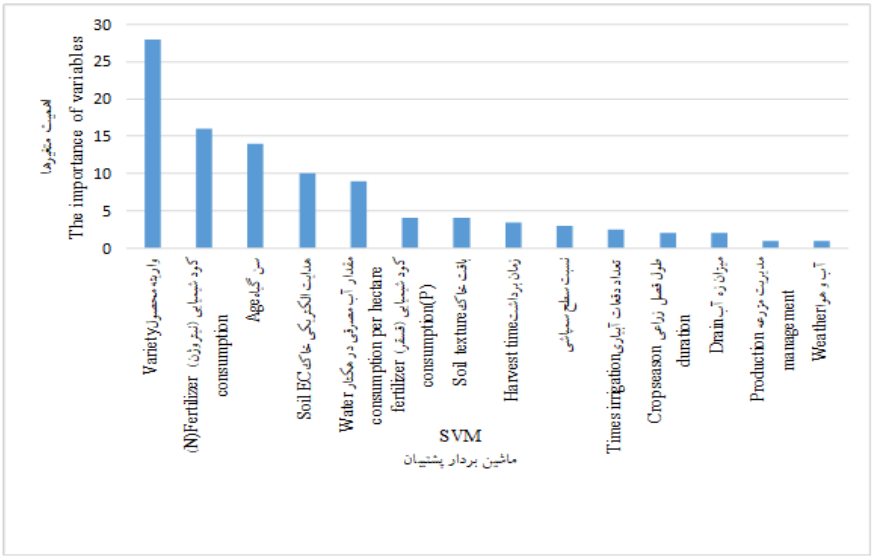
جدول (۱۰) دقت مدل‌های مختلف در پیش‌بینی عملکرد نیشکر

Table (10) Accuracy of various models in predicting sugarcane yield

آزمایش (Test)			آموزش (Training)			مدلModel
MAE	RMSE	R^2	MAE	RMSE	R^2	
0.34	6.28	0.92	0.21	5.71	0.95	جنگل تصادفی (Random Forest)
0.55	7.50	0.86	0.26	6.30	0.91	ماشین بردار پشتیبان (Support Vector Machine)



الف (a)



ب (b)

شکل (۱۱) اهمیت متغیرها در مدل جنگل تصادفی (الف) و در مدل ماشین بردار پشتیبان (ب)

Figure (11) Importance of variables in the random forest model (a) and in the support vector machine model (b)

نتیجه‌گیری

نتایج پژوهش، نشان‌دهنده‌ی موفقیت قابل توجه در استفاده از روش‌های یادگیری ماشین برای پیش‌بینی عملکرد محصول نیشکر است. دقت فراخوانی ۹۶/۹ درصدی مدل جنگل تصادفی که با اعتبارسنجی متقابل ۱۰ لایه‌ای تأیید شده، نشان می‌دهد که این مدل می‌تواند به عنوان ابزاری قابل اعتماد در مدیریت مزارع نیشکر استفاده شود. نتایج مکمل حاصل از الگوریتم ماشین بردار پشتیبان نیز این یافته‌ها را تأیید می‌کند. توزیع متعادل داده‌ها در نمودار جعبه‌ای، عملکرد خوب در ماتریس درهم‌ریختگی و منحنی ROC مطلوب، همگی نشان می‌دهند که این مدل‌ها می‌توانند در تصمیم‌گیری‌های مدیریتی مزارع نیشکر نقش مهمی ایفا کنند. این نتایج همچنین نشان می‌دهند که استفاده از روش جنگل تصادفی به علت بالا بودن میزان دقت مدل در هر دو مرحله آزمایش و آزمون نسبت به روش مدل ماشین بردار پشتیبان می‌تواند به بهبود مدیریت مزرعه، بهبود بهره‌وری و مدیریت بهتر منابع در مزارع نیشکر کمک کند. همچنین با استفاده از این مدل،

می‌توان عملکرد مزارع نیشکر را با دقت بالا پیش‌بینی کرده و تصمیمات بهتری در مورد کوددهی، آبیاری، و سایر عملیات زراعی اتخاذ کرد. پیاده‌سازی مدل در شرایط واقعی مزرعه و گرفتن بازخورد جزء اهداف پژوهش حاضر نبوده است ولی در پژوهش‌های آتی می‌توان بازخورد مدل در واقعیت را سنجید. در ارتباط با محدودیت‌های مدل نیز وابستگی به داده‌های تاریخی یکی از محدودیت‌های مدل‌سازی است. این مدل بر اساس داده‌های جمع‌آوری شده از سال‌های زراعی ۱۳۹۷-۱۴۰۲ در استان خوزستان در کشت و صنعت دهخدا آموزش داده شده است. بنابراین، دقت آن ممکن است در شرایطی که با داده‌های تاریخی تفاوت زیادی داشته باشد، کاهش یابد. همچنین از نظر تعمیم‌پذیری محدود، تعمیم‌پذیری این مدل به مناطق دیگر نیازمند بررسی دقیق شرایط محیطی و مدیریتی آن مناطق است. از نظر عدم قطعیت در شرایط غیرمعمول هم مدل ممکن است در شرایط غیرمعمول (مانند تغییرات شدید آب و هوایی، شیوع آفات و بیماری‌ها، و غیره) دقت کمتری داشته باشد. به طور خلاصه این مدل‌ها می‌توانند در مدیریت مزارع نیشکر مفید باشند، اما محدودیت‌های آن‌ها باید در نظر گرفته شود. استفاده از الگوریتم‌های دیگر داده‌کاوی مانند الگوریتم‌های ترکیبی[1] (XGBoost) جهت استفاده در تحقیقات آتی و مقایسه نتایج با پژوهش حاضر پیشنهاد می‌گردد.

سپاس گزاری

نویسندگان از معاونت پژوهشی دانشگاه شهید چمران اهواز بابت تأمین هزینه‌های این پژوهش سپاسگزاری می‌نمایند.

1- Ensemble methods

References

1. Abbas, F., Afzaal, H., Farooque, A.A. and Tang, S. .2020. Crop yield prediction through proximal sensing and machine learning algorithms. Agronomy, 10 (7): 1046. https://doi.org/10.3390/agronomy10071046

2. Batool, D., Shahbaz, M., Shahzad Asif, H., Shaukat, K., Alam, T. M., Hameed, I. A., and Luo, S. 2022. A Hybrid Approach to Tea Crop Yield Prediction Using Simulation Models and Machine Learning. Plants, 11 (15): 1925. https://doi.org/10.3390/plants11151925

3. Ed-Daoudi, R., Alaoui, A., Ettaki, B. and Zerouaoui, J. 2023. Improving Crop Yield Predictions in Morocco Using Machine Learning Algorithms. Journal of Ecological Engineering, 24 (6): 392-400. https://doi.org/10.12911/22998993/162769

4. Fegade, T. K., and Pawar, B. V. 2020. Crop prediction using artificial neural network and support vector machine. In Data Management, Analytics and Innovation: Proceedings of ICDMAI 2019, 2: 311-324. https://doi.org/10.1007/978-981-13-9364-8_23

5. Ghaffari, H. and Amerikhah, H. 2023. Simulation of daily suspended sediment in the Rud Zard River using different machine learning techniques. Agricultural Engineering, 46 (1): 102-119. (In Persian with English abstract)

6. Hafezi, N., Bahrami, H., Sheikh Davoudi, M.J. and Alavi, S.A. 2021. Hybrid artificial neural network with metaheuristic algorithms in sugarcane yield prediction. Iranian Journal of Biosystem Engineering, 51 (3): 515-526. (In Persian with English abstract)

7. Han, J., Kamber, M., and Pei, J. 2019. Data mining: concepts and techniques, 3rd Ed. Niaze danesh. Tehran. (In Persian)

8. Jafarian, Z. and Kargar, M. 2017. Comparison of Random Forest (RF) and Boosted Regression Tree (BRT) models in predicting the presence of dominant rangeland species in the rangelands of Pulur, Mazandaran. Applied Ecology, 6 (1): 41-55. (In Persian with English abstract)

9. Jeong, J.H., Resop, J.P., Mueller, N.D., Fleisher, D.H., Yun, K., Butler, E.E., Timlin, D.J., Shim, K.M., Gerber, J.S., Reddy, V.R. and Kim, S.H. 2016. Random forests for global and regional crop yield predictions. PLoS One, 11 (6): 0156571. https://doi.org/10.1371/journal.pone.0156571

10. Jhajharia, K., Mathur, P., Jain, S., and Nijhawan, S. 2023. Crop yield prediction using machine learning and deep learning techniques. Procedia Computer Science, 218: 406-417. https://doi.org/10.1016/j.procs.2023.01.023

11. Kavoosi-Givshad, M., Khashaei-Siouki, A. and Behdani, M.A. 2021. The effectiveness of support vector machine models and gene expression programming in predicting saffron crop yield. Saffron Research, 8 (2): 343-358. (In Persian with English abstract)

12. Kavzoglu, T., and Colkesen, I. 2009. A kernel functions analysis for support vector machines for land cover classification, International Journal of Applied Earth Observation and Geoformation, 11: 352-359. https://doi.org/10.1016/j.jag.2009.06.002

13. Kuradusenge, M., Hitimana, E., Hanyurwimfura, D., Rukundo, P., Mtonga, K., Mukasine, A., and Uwamahoro, A. 2023. Crop yield prediction using machine learning models: case of Irish potato and maize. Agriculture, 13 (1): 225. https://doi.org/10.3390/agriculture13010225

14. Lee Sunmin Kim, J. C., Jung, H. S., Lee, M. J., and Lee, S. 2017. Spatial prediction of flood susceptibility using random-forest and boosted-tree models in Seoul metropolitan city, Korea. Geomatics, Natural Hazards and Risk, 8: 1185–1203. https://doi.org/10.1080/19475705.2017.1308971

15. Mahajan, J., Banal, K., and Mahajan, S. 2021. Estimation of crop production using machine learning techniques: a case study of J&K. International Journal of Information Technology, 13 (4): 1441-1448. https://doi.org/10.1007/s41870-021-00653-7

16. Medar, R.A., Rajpurohit, V.S. and Ambekar, A.M. 2019. Sugarcane Crop Yield Forecasting Model Using Supervised Machine Learning. International Journal of Intelligent Systems and Applications, 11 (8): 11-20. https://doi.org/10.5815/ijisa.2019.08.02

17. Monjezi, N. 2023. Sugarcane Technology and Mechanization in Iran. Shahid Chamran University of Ahvaz Publications, 792 p. (In Persian with English abstract)

18. Monjezi, N., Zaki Dizji, H., Sheikh Davoudi, M. J., Marzban, A. and Shamili, M. 2017. Application of fuzzy GERT method in scheduling sugarcane production operations. Agricultural Engineering, 40 (1): 125-139. (In Persian with English abstract) https://doi.org/10.22055/agen.2017.1308

19. Nader nejad, F., Imani, D. M. and Rasouli, M.R. 2022. Presenting a data-driven model for predicting sugar yield from sugarcane. Journal of Agricultural Machinery, 12 (4): 558-543. (In Persian with English abstract) https://doi.org/10.22067/jam.2021.69805.1034

20. Panigrahi, B., Kathala, K. C. R., and Sujatha, M. 2023. A machine learning-based comparative approach to predict the crop yield using supervised learning with regression models. Procedia Computer Science, 218: 2684-2693. https://doi.org/10.1016/j.procs.2023.01.241

21. Ramesh, D., and Vardhan, B. V. 2013. Data mining techniques and applications to agricultural yield data. International Journal of Advanced Research in Computer and Communication Engineering, 2: 3477-3480.

22. Rezaei, M., Aghamohammadi Zanjirabad, H., Azizi, Z., Vafainejad, A. and Behzadi, S. 2023. Identification of agricultural land types in the Abbas Plain using time series analysis of Sentinel 2 satellite images. Sustainability, Development and Environment, 4 (4): 1-13. (In Persian with English abstract) https://doi.org/10.22034/jsde.2023.846486

23. Samadzadegan, F. and Hassani, H. 2012. Determining optimal support vector machines in hyperspectral image classification based on genetic algorithm. Journal of Information and Communication Technology, 13: 9-24. (In Persian with English abstract) https://doi.org/10.4028/www.scientific.net/KEM.500.792

24. Sarijaloo, F. B., Porta, M., Taslimi, B., and Pardalos, P. M. 2021. Yield performance estimation of corn hybrids using machine learning algorithms. Artificial Intelligence in Agriculture, 5: 82-89.

25. Scholkopf, B. and Smola, A. J. 2000. Statistical learning and kernel methods. Cambridge, MA: MIT Press, pp: 29.

26. Yamparla, R., Shaik, H. S., Guntaka, N. S. P., Marri, P., and Nallamothu, S. 2022. Crop yield prediction using random forest algorithm. In 2022 7th International Conference on Communication and Electronics Systems (ICCES), 1538-1543. https://doi.org/10.1109/ICCES54183.2022.9835756

27. Zaki Dizaji, H. and Monjezi, N. 2018. Evaluation of waste sources during the sugarcane production process and providing solutions to reduce waste. Journal of Agricultural Machinery, 8 (1): 63-73. (In Persian with English abstract) https://doi.org/10.22067/jam.v8i1.59027

28. Zaki-Dizaji, H., Monjezi, N. and Sheikh Davoudi, M.J. 2018. Investigating factors affecting sugarcane production yield with the aim of increasing production using a data mining approach. Iranian Journal of Biosystem Engineering, 49 (3): 501-511. (In Persian with English abstract)

29. Zaki Dizaji, H. Bahrami, H. Monjezi, N. and Sheikh Davoudi, M.J. 2019. Modeling variables affecting sugarcane yield using 0.05C and QUEST decision tree algorithms. Journal of Agricultural Machinery, 9 (18): 469-484. (In Persian with English abstract) https://doi.org/10.22067/jam.v9i2.69712