Original Article
ISSN (Online): 2582-7472

ShodhKosh: Journal of Visual and Performing Arts
June 2024 5(6), 1714–1720

ENHANCING TOMATO VARIETY SELECTION AND YIELD PREDICTION MODEL USING ADVANCED MACHINE LEARNING TECHNIQUES

R. Usha Devi [1], Dr. N. A. Sheela Selvakumari [2]

[1] Research Scholar, Department of Computer Science, Sri Krishna Arts and Science College, Coimbatore, Assistant Professor, Department of Data Science, Nirmala College for Women, Coimbatore, India
[2] Associate Professor, Department of Computer Science, Sri Krishna Arts and Science College, Coimbatore, India

ABSTRACT

One important agricultural practice that greatly contributes to the world's food supply is the growing of tomatoes. However, choosing the right tomato varieties and accurately estimating their yields are difficult undertakings that depend on several variables, such as crop statistics, climate, and geography. Advanced machine learning approaches can be used to improve prediction accuracy, feature selection, and preprocessing to overcome these problems. To optimize tomato variety selection and yield prediction, this work investigates an integrated approach that uses improved preprocessing approaches for outlier and missing information, an advanced feature selection method, and a hybrid algorithm.

Keywords: Crop Yield Prediction, Spatial Data, Climate Data, Missing Value, PCA

Corresponding Author

R. Usha Devi, ushaofficial16@gmail.com

DOI
10.29121/shodhkosh.v5.i6.2024.2371

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

1. INTRODUCTION

In Tamilnadu, India, tomatoes are one of the most extensively grown and significant crops from an economic standpoint. As seen in Fig. 1.1 [1], precise and timely crop production forecasts are essential for efficient planning and decision-making processes in the agricultural industry. Nonetheless, illnesses, pests, and climate change can all have a big influence on tomato production [2]. Researchers have looked into using several machine-learning algorithms for tomato yield prediction to overcome this difficulty. Planning requires accurate crop yield prediction [3]. The choice of tomato varieties is a major factor in crop productivity. Studies have indicated that choosing tomato breeding lines according to characteristics such as fruit weight and quantity of fruits per plant can result in high yield potential [4].

Figure:1.1 Technology Advancement in Agriculture

Furthermore, high-yielding tomato hybrid varieties that are ideal for growth under particular conditions can be identified by evaluating traits including plant height, leaf surface area, root weight, and fruit output per bush [5]. To increase agricultural production and guarantee food security, weather-based crop output prediction using machine learning models is essential [6]. Planning for agriculture, allocating resources, and ensuring food security all depend on accurate tomato yield predictions [7]. Promising strategies to improve the precision of these forecasts are provided by machine learning. To help farmers choose the best tomato varieties for maximum yield, a Tomato Crop Variety Recommendation System can be created with machine learning techniques and data analytics [8]. Through the incorporation of variables such as soil properties (nitrogen, phosphorus, and potassium), meteorological conditions, past crop performance, and market trends, this system can generate customized recommendations that are appropriate for certain agricultural situations [9]. By using sophisticated preprocessing and feature selection strategies, this work attempts to address typical issues in agricultural datasets and assess the accuracy of several machine learning models in forecasting tomato yield. The work is structured as follows: Section II addresses previous research on yield prediction and crop recommendation, with a focus on data handling challenges in agriculture. A thorough explanation of the suggested model for the aforementioned problems is provided in Section III. Comparing the outcomes of the suggested model with the previous research models is covered in Section IV. The paper in Section V finally came to an end.

2. RELATED WORKS

Crop recommender system that boosts farming productivity and sustainability by using machine learning to suggest appropriate crops, including tomato varieties, based on soil and meteorological data [10]. A machine learning-based decision support system is shown in [11] that advises farmers on the best crop to plant depending on weather and soil parameters such as temperature, humidity, rainfall, nitrogen, potassium, phosphorus, and pH of the soil. As a workable solution to balancing covariates whose distribution is anticipated to differ if the values are not missing entirely at random, stable balancing weights are shown in [12]. Discrepancy statistics can be employed to compare the density of imputed and observed values. For preprocessing, a modified imputation approach is proposed that outperforms popular imputation schemes such as mean imputation, regression, and Markov Chain Monte Carlo [13]. To improve agricultural practices and optimize crop selection, feature selection in crop data entails utilizing machine learning algorithms to detect important features such as weather and soil composition [14]. The importance of the best feature selection methods in machine learning algorithms for crop-related applications has been emphasized by several research. For example, a unique hybrid feature selection method that combined Grey Wolf Optimisation with Mutual Information and ReliefF filters showed improved results in crop mapping [15]. Furthermore, the accuracy of crop classification was greatly improved by the employment of spectral, index, and texture features in classification schemes in addition to random forest feature optimization [16]. This highlights the significance of feature selection in crop mapping applications.

3. PROPOSED METHODOLOGY

The model that is suggested is shown in Fig. 3.1. The investigation starts with a thorough pretreatment of the data, which involves numerous imputations to handle missing values. Correlation-based feature selection is then used to determine which elements have the greatest impact on tomato yields. Lastly, the program forecasts the expected production and recommends an effective tomato variety that is appropriate for the specified soil and environment.

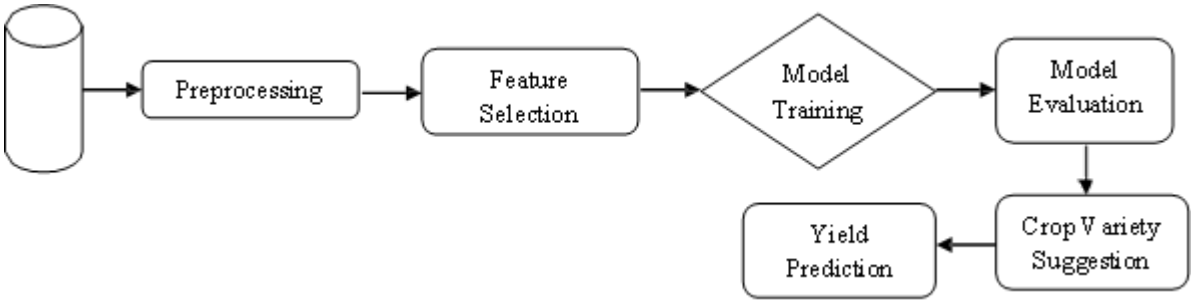

Figure: 3.1. Proposed Model

3.1. DATASET

The dataset contains attributes like soil properties, irrigation information, temperature, precipitation, and humidity, as well as historical yield data. gathered historical data on Tamilnadu tomato yield for the previous ten years. Temperature, precipitation, humidity, and solar radiation are all included in weather data. Data on soil consist of pH, organic matter, texture, and nutrient concentration. Crop data include the occurrence of pests and diseases as well as a range of performance.

3.2. DATA PREPROCESSING

The gathered agricultural data was preprocessed using a multi-stage method. Recursive Imputation and the outlier segregate algorithm (RIOS), which is based on prior estimations, are used to handle missing data. A statistical method called recursive imputations is used to deal with missing data. It involves splitting up a dataset into numerous complete datasets, analyzing each dataset independently, and then combining the results to get estimates that take the uncertainty caused by the missing data into account. This preprocessing step can be very important in the context of crop yield prediction to ensure the robustness of the subsequent analysis, particularly when working with big agricultural datasets that may have missing values for a variety of causes, such as unrecorded observations or failures of the soil type. The Filter is a technique that, given measurements taken over time that include statistical noise and other errors, yields estimates of some unknown variables. It changes estimates depending on fresh data and earlier estimations since it functions recursively. Moreover, imputation can assist in locating and fixing measurement mistakes, resulting in more accurate and dependable crop yield projections. Imputation can also help locate and resolve problems with the quality of the data, which can raise the analysis's overall accuracy and lower the possibility that false conclusions will be made.

3.3. FEATURE SELECTION

The hybrid technique has the potential to greatly improve machine learning model performance by identifying the most pertinent features from the data. By combining the advantages of both approaches, swarm intelligence with correlation-based feature selection (SICBFS) allows for effective and efficient feature selection. Algorithms with swarm intelligence are excellent at sifting through the search space and making use of the best results. This guarantees an exhaustive and efficient search for the ideal feature subset. Redundancy among features and relevance to the target variable are the two criteria used by CFS to assess feature subsets. This guarantees that the features chosen are non-redundant and informative. This approach evaluates each feature's capacity for prediction as well as the degree of feature redundancy. The technique determines the value of a subset of features by taking into account each feature's unique predictive capacity as well as the degree of redundancy among them. By altering the fitness function (CFS score) and the

settings of the swarm intelligence algorithm, the method can be tailored to various data kinds and assessment standards. To determine the most pertinent tomato variety and yield predictors, SICBFS was used.

3.4. CROP RECOMMENDATION WITH YIELD PREDICTION

An enhanced variant of conventional multiple linear regressions is called Enhanced Multiple Linear Regression (EMLR). A statistical method for simulating the relationship between a scalar dependent variable and one or more independent variables is called enhanced multiple linear regression (EMLR). It has been demonstrated to offer notable accuracy in agricultural yield prediction and has been applied extensively. The ability to include several independent variables in multiple linear regression (MLR) makes it an excellent fit for the intricate, multidimensional nature of agricultural decision-making.

In the context of tomato variety recommendation, the EMLR model can be represented as:

$Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + ... + \beta_nX_n + \varepsilon$-- Equ 3.1 In Equ. 3.1: Y is the dependent variable (e.g., expected yield or a performance score) X_1, X_2... X are independent variables (environmental and agronomic factors) β_0, β_1, β_2... β_n are the regression coefficients ε is the error term.

Interpreting the effects of different elements on varied performance is made simple by EMLR models. Include categorical and continuous variables (using dummy coding). Because EMLR models are easy to compute, real-time applications can benefit from using them. Apply k-fold cross-validation, which has k-folds (usually 5 or 10). A similar technique called K-fold cross- validation is applied to assess machine learning models on a small sample of data. It's especially helpful in situations like predicting tomato yield, where there may not be as much data available and it's critical to utilize what is [17]. When predicting tomato yield—where environmental factors can vary greatly—this approach is especially helpful. to manage the variation in tomato yields caused by various growth environments or seasons as depicted in several folds. For tomato production prediction, k-fold cross-validation in EMLR can be used to create a more dependable and robust model that takes into account the inherent unpredictability of agricultural data.

4. RESULT AND DISCUSSION

The effectiveness of the deployed machine learning models is then carefully assessed using a range of common metrics, which offer a detailed evaluation of the prediction accuracy of the model. Examples of these metrics are Root Mean Square Error, Mean Absolute Error, and Coefficient of Determination.

4.1. ROOT MEAN SQUARE ERROR (RMSE)

The RMSE is a widely used metric to quantify the discrepancies between observed and predicted values in a model. The square root of the average squared discrepancies between the observed (actual) and anticipated values is measured by RMSE. A smaller RMSE denotes a better fit, and it indicates how closely the predicted values match the actual values [18].

$$RMSE=\frac{\sqrt{\sum(yi-\hat{y}i)^2}}{n} \quad \text{-- Equ 4.1}$$

In Equ. 4.1: yi = observed values ŷi = predicted values n = number of observations

4.2. MEAN ABSOLUTE ERROR (MAE)

The mean absolute difference (MAE) between the observed (actual) and expected values is computed. Without taking into account the direction of the errors, MAE calculates the average magnitude of errors in a set of forecasts. A lower MAE indicates more accurate predictions, and it provides an estimate of the size of the errors in a set of forecasts [19].

$$MAE=\frac{\sqrt{(\sum|yi-\hat{y}i|)}}{n} \quad \text{-- Equ 4.2}$$

In Equ. 4.2 : yi = observed values ŷi = predicted values n = number of observations

4.3. COEFFICIENT OF DETERMINATION

The percentage of the dependent variable's variation that can be predicted from the independent variables is expressed as R^2. Higher values suggest a better fit. The range is 0 to 1. Regression predictions that have an R2 of 1 fully fit the data [20].

$$R^2 = 1 - \left(\frac{SSres}{SStot}\right) \quad \text{-- Equ 4.3}$$

In Equ. 4.3: SSres = $\Sigma(yi - \hat{y}i)^2$ (Sum of squares of residuals) SStot = $\Sigma(yi - \bar{y})^2$ (Total sum of squares) $\bar{y}$ = mean of observed values.

Table: 4.1 Comparison of Models based on the above Metrics

Metrics	EMLR	RF	DT	LR
RMSE	8.2	10.5	12.1	10.8
MAE	6.7	8.4	9.5	8.6
R2	0.85	0.78	0.75	0.77

The comparison of the suggested models and the current models using the metrics of RMSE, MAE, and R2 is displayed in Table 4.1. The table clearly shows that EMLR performs better than the other models in each of the three metrics: it has the greatest R2, the lowest RMSE, and MAE. This suggests that out of the four models, EMLR offers the most precise and trustworthy forecasts.

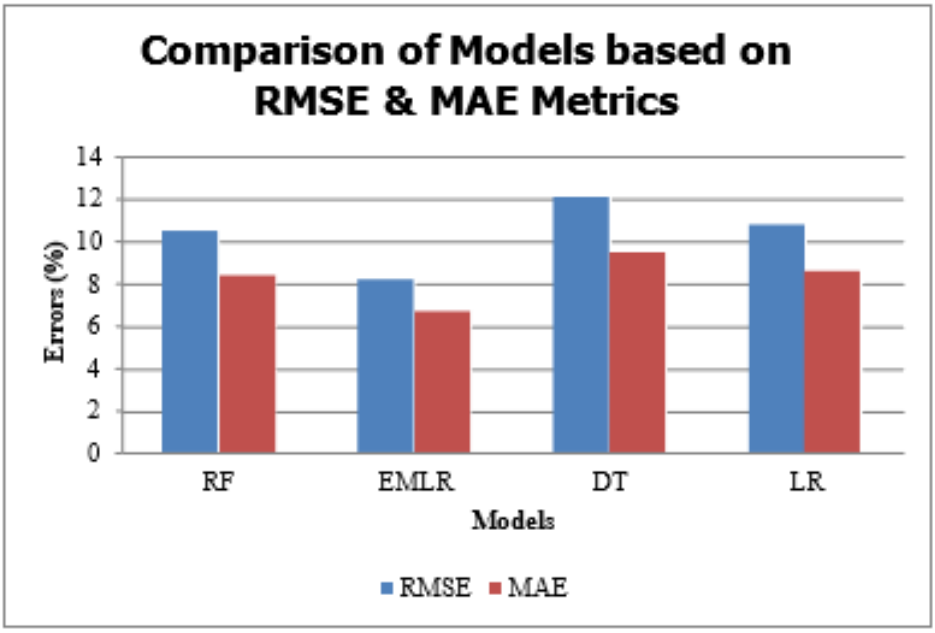

Figure: 4.1 Comparison of Models based on RMSE & MAE Metrics

Fig. 4.1 shows a pictorial representation of model comparison based on RMSE and MAE. From the figure, it is clear that EMLR outperforms the other models in the above two metrics with the lowest RMSE and MAE of 8.2 and 6.7 respectively whereas RMSE of RF, DT, and LR are 10.5, 12.1, and 10.8 respectively, and MAE of 8.4, 9.5 and 8.6 respectively which is higher than the proposed model.

Fig. 4.2 shows a pictorial representation of model comparison based on R2. From the figure, it is clear that EMLR outperforms the other models with the highest R2 of 0.85 whereas RF, DT, and LR are 0.78, 0.75, and 0.77 respectively which is lower than the proposed model.

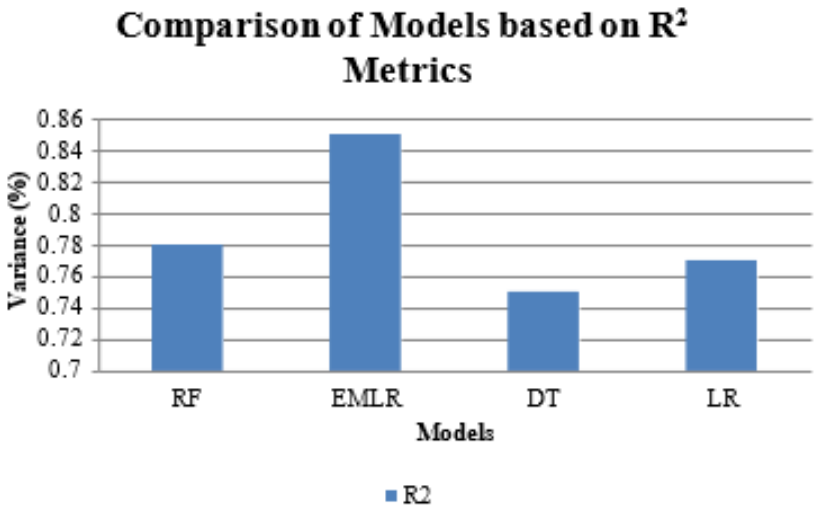

Figure: 4.1 Comparison of Models based on R2 Metrics

5. CONCLUSION

The findings could help academics, farmers, and legislators create more sensible plans for managing and producing tomatoes sustainably. A potent method for maximizing tomato variety selection and yield prediction is the combination of improved preprocessing techniques, sophisticated feature selection, and hybrid prediction algorithms. The results of this study show how sophisticated machine learning methods can be used to increase the precision and dependability of tomato yield predictions. This approach helps farmers make more informed decisions and increases prediction accuracy, which in turn leads to improved crop management and higher yields. The use of machine learning in agriculture will surely result in more effective and sustainable farming methods as it continues to advance. In the future, better insights into the genetic determinants impacting production and disease resistance can be obtained by combining genomic data of many tomato types. To suggest the most robust and productive cultivars, machine learning models can evaluate this genomic data in combination with climate and geographic information.

CONFLICT OF INTERESTS

None.

ACKNOWLEDGMENTS

None.

REFERENCES

Banshilal, Patidar, -., Samarth, Sinha., Anurag, Jain. (2024). Crop yield prediction using automated analytical and predictive techniques. doi: 10.1201/9781032700502-70

Shalini, Bhadola., Kavita, Rathi. (2024). Application of machine learning for crop yield prediction challenges and future directions. doi: 10.58532/v3bdai2p3ch3

Ramakrishnan, Raman., Hanumaji, Kantari., Atul, A, Gokhale., K., Elangovan., B., Meenakshi., S., Srinivasan. (2024). Agriculture Yield Estimation Using Machine Learning Algorithms. doi: 10.1109/autocom60220.2024.10486107

Witri, Nur, Anisa., Enik, Nurlaili, Afifah., Rudi, Hari, Murti. (2022). Selection of tomato breeding lines based on morphological traits associated with high yield potential in double- cross population. Biodiversitas, doi: 10.13057/biodiv/d230624

Toshtemir, Ostonakulov., Gulshoda, Anvar, qizi, Saidova., Anvar, Akbarovich, Shamsiev. (2023). Isolation of adaptive varieties of tomato and innovative technologies for their cultivation on slightly saline soils of the Bukhara region. E3S web of conferences, doi: 10.1051/e3sconf/202338903012

S., Vijayalakshmi., Aadhila, Begum, A, -., K., Preethi. (2024). Weather and Crop Yield Prediction by Machine Learning Model and Spiking Neural Network. International Journal For Multidisciplinary Research, doi: 10.36948/ijfmr.2024.v06i01.14350

Murali, Krishna, Gumma., Ramavenkata, Mahesh, Nukala., Pranay, Panjala., Pavan, Bellam., Snigdha, Gajjala., S., K., Dubey., Vinay, Kumar, Sehgal., Ismail, Mohammed., Kumara, Charyulu, Deevi. (2024). Optimizing Crop Yield Estimation through Geospatial Technology: A Comparative Analysis of a Semi-Physical Model, Crop Simulation, and Machine Learning Algorithms. AgriEngineering, doi: 10.3390/agriengineering6010045.

Jie, Zhang., Qingqing, Zhao., Xiaohong, Li., Hecang, Zang., Guoqiang, Li., Gu, Zheng. (2022). Design and Application of Crop Variety Recommendation and Production Decision System in Agricultural Production. doi: 10.1109/iccsmt58129.2022.00012.

Shivanoori, Sai, Samhith., T., V., Rajinikanth., Burma, Kavya., Alley, Yashwanth, Sai, Krishna. (2024). Crop Recommender System. doi: 10.1007/978-981-99-7954-7_55.

Yogesh, Gupta., Geetika, Srivastava. (2024). Crop recommendation system for growing the best suitable crop. International Journal of Science and Research Archive, doi: 10.30574/ijsra.2024.12.1.1111.

D., Kavitha. (2023). Crop Recommendation System using ML. doi: 10.55041/isjem00457

Moritz, Marbach. (2021). Choosing Imputation Models. arXiv: Methodology

Elenita, T., Caparino., Ariel, M., Sison., Ruji, P., Medina. (2018). A Modified Imputation Method to Missing Data as a Preprocessing Technique. doi: 10.1109/HNICEM.2018.8666394

Suhas, Kakade., Ravi, S., Kulkarni., Somesh, Dhawale., M, J.C.. (2023). Utilization of Machine Learning Algorithms for Precision Agriculture: Enhancing Crop Selection. Green Intelligent Systems and Applications, doi: 10.53623/gisa.v3i2.313

Marwa, S., Moustafa., Amira, S., Mahmoud., Eslam, Farg., Mohsen, Nabil., S.M., Yasir, Arafat. (2024). Bi-stage feature selection for crop mapping using grey wolf metaheuristic optimization. Advances in Space Research, doi: 10.1016/j.asr.2024.02.037

Zhanqing, Zhao., Guofei, Shang., Yubo, Liu., Shizhuo, Liu., Xiao, Zhang. (2023). Optimal feature selection and crop extraction using random forest based on GF-6 WFV data. International Journal of Remote Sensing, doi: 10.1080/01431161.2023.2216856.

Jiang, Y., et al. (2021). Evaluating Machine Learning Models for Predicting Building Energy Consumption: A Comparative Study. Energy and Buildings, 232, 110677. doi:10.1016/j.enbuild.2020.110677.

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? Geoscientific Model Development, 7(3), 1247-1250.

Kumar, A., & Singh, S. K. (2020). An Integrated Framework of Improved Mean Absolute Error Model for Image Denoising. Multimedia Tools and Applications, 79, 9935-9952. doi:10.1007/s11042-019-08354-5.

Najafabadi, A. H. Z., & Moosavi, V. (2018). Spatial modeling of wetland vegetation using remote sensing data: A comparison of four modeling approaches in the Anzali Wetland. Journal of Environmental Management, 206, 286-300. doi:10.1016/j.jenvman.2017.10.056.