International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2024: 7.101

Comparative Performance of Support Vector Machine and Random Forest Algorithms for Mango Yield Prediction in Dharwad and Kolar Districts of Karnataka

Keerthi P[1], Vasantha Kumari J[2], Lalitha V. M.[3]

[1]MSc Scholar, Department of Agricultural Statistics, COA, Dharwad, UAS, Dharwad
Email: *keerthi998k[at]gmail.com*

[2]Assistant Professor, Department of Agricultural Statistics, COA, Dharwad, UAS, Dharwad

[3]MSc Scholar, Department of Agricultural Statistics, COA, Dharwad, UAS, Dharwad

Abstract: ***Aims: The study aimed to develop and compare the performance of two machine learning models-Support Vector Machine (SVM) regression and Random Forest (RF)-for predicting mango (Mangifera indica L.) yields in Dharwad and Kolar districts of Karnataka, representing distinct agro-climatic zones. It further assessed the predictive capability of these models under varying climatic conditions. Study Design: A retrospective analytical study was conducted using machine learning-based regression modelling. Place and Duration of Study: The study was carried out in the Department of Agricultural Statistics, University of Agricultural Sciences, Dharwad, using secondary data on mango yield and weather parameters spanning 44 years (1980-2023). Methodology: A dataset comprising mango yield statistics and meteorological variables-rainfall, maximum and minimum temperatures, and wind speed was used. Both SVM and RF models were trained on 80% of the data and tested on the remaining 20%. Model performance was evaluated using the coefficient of determination (R^2), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). Scatter plots were utilized to visualize relationships between actual and predicted yields. Results: The Random Forest model exhibited superior predictive accuracy compared to SVM in both districts. In Dharwad, RF achieved an R^2 of 0.513 versus 0.433 for SVM, while in Kolar, RF attained 0.760 compared to 0.079 for SVM. Scatter plots indicated that RF predictions aligned more closely with observed yields, particularly in Kolar. Conclusion: Ensemble-based models such as Random Forest outperform kernel-based SVM for mango yield prediction. Integrating long-term meteorological data with machine learning techniques enhances yield forecasting accuracy and supports climate-resilient agricultural planning.***

Keywords: Mango yield, Machine learning, Support Vector Machine, Random Forest, Weather parameters

1. Introduction

Mango cultivation holds a central place in Karnataka's horticultural sector, contributing significantly to farm income and trade. However, yield fluctuations are common due to high sensitivity to weather conditions, particularly temperature variations, irregular rainfall, and wind speed. Traditional statistical tools often oversimplify such nonlinear and interactive influences, which limits their predictive reliability.

Machine learning (ML) methods provide alternatives capable of modelling complex, nonlinear, and high-dimensional datasets. Among ML methods, Support Vector Machine (SVM) regression and Random Forest (RF) are widely applied for yield forecasting. SVM, a kernel-based algorithm, is known for its ability to handle nonlinear regression problems with high generalization capacity. RF, an ensemble learning method, combines multiple decision trees to improve predictive accuracy and minimize overfitting (Shahhosseini *et al.,* 2020).

This study compares the prediction efficiency of SVM and RF for mango yields in two contrasting agro-climatic regions: Dharwad, representing the Northern Transition Zone, and Kolar, representing the Eastern Dry Zone of Karnataka.

2. Methodology

2.1 Study Area

- **Dharwad District**: Located at 15.45°N latitude and 75.00°E longitude, at an altitude of 750 m above sea level. It receives an annual average rainfall of 1,200 mm, with a tropical monsoon climate. Mango is cultivated but not as extensively as in Kolar.
- **Kolar District**: Located at 13.13°N latitude and 78.12°E longitude, at an altitude of 849 m. It receives an average annual rainfall of 750 mm and has a dry climate suitable for large-scale mango orchards, making it one of Karnataka's leading mango-producing districts.

2.2 Data Sources

- **Mango production data** (1980–2023): Directorate of Economics and Statistics and Department of Horticulture, Government of Karnataka.
- **Meteorological data**: Rainfall (mm), maximum temperature (°C), minimum temperature (°C), and wind speed (m/s). Data were collected from the Department of Agrometeorology, UAS Dharwad, and NASA POWER database.

2.3 Pre-processing of Data

The dataset was checked for missing values and outliers. Missing values were imputed using interpolation methods. All input variables were standardized using StandardScaler to ensure that parameters measured in different units (e.g., °C vs mm) were placed on a common scale, preventing bias in model training.

2.4 Support Vector Machine Regression (SVM)

SVM (Vapnik *et al.,*1997) is based on the principle of mapping input data into a high-dimensional feature space using kernel functions and finding a hyperplane that minimizes prediction error within a defined tolerance (Dang *et al.,* 2021)

1) **Kernel function**: Radial Basis Function (RBF) was used, as it effectively captures nonlinear patterns (Saruta *et al.,* 2013).
2) **Hyperparameters**:
   a) **C (regularization parameter)**: Controls the trade-off between minimizing training error and maximizing generalization.
   b) **ε (epsilon)**: Defines the margin of tolerance within which predictions are not penalized.
   c) **γ (gamma)**: Determines the influence of individual training points.
3) **Optimization:** Hyperparameters were tuned using grid search and k-fold cross-validation to achieve the best fit (Nitze *et al.,* 2012).

2.5 Random Forest Regression (RF)

RF is an ensemble technique that builds multiple decision trees using bootstrap sampling and random feature selection, and then averages their predictions (Breiman, 2001). This reduces variance and prevents overfitting.

1) **Key Features**:
   a) **Bootstrap Aggregation (Bagging)**: Each tree is trained on a random sample with replacement.
   b) **Random Feature Selection**: At each node, a random subset of features is considered for splitting, improving model diversity (Champaneri *et al*., 2016).
   c) **Aggregation**: Predictions from all trees are averaged to produce the final output (Everingham *et al*., 2016).

2) **Hyperparameters**:
   a) **Number of trees (n_estimators)**: Determines ensemble size.
   b) **Maximum depth**: Controls complexity of trees.
   c) **Minimum samples per split**: Prevents overfitting by requiring a minimum number of observations at each split.

3) **Tuning**: Hyperparameters were optimized through grid search and cross-validation.

2.6 Model Training and Evaluation

1) **Training**: 80% of the dataset was used for training and 20% for validation.
2) **Evaluation Metrics**:
   - **Coefficient of Determination (R^2)**: Measures goodness-of-fit.
   - **Mean Absolute Error (MAE)**: Indicates average prediction deviation.
   - **Mean Absolute Percentage Error (MAPE)**: Provides error relative to actual values.
   - **Root Mean Squared Error (RMSE)**: Penalizes larger errors more heavily.

3. Results and Discussion

The comparative analysis of SVM and RF models revealed significant differences in predictive performance. In Dharwad, the RF model achieved an R^2 value of 0.513, while SVM lagged behind at 0.433. Similarly, in Kolar, RF demonstrated superior performance with an R^2 of 0.760, compared to 0.079 for SVM. Lower RMSE, MAE, and MAPE values in RF further confirmed its predictive efficiency as shown in Tables 1 to 4. RF predictions closely followed the line of equality, which was shown in Figures 1 and 2, while SVM predictions deviated considerably, especially for extreme yield values.

These results suggest that RF can effectively model the non-linear influence of meteorological parameters on mango yield. The ensemble nature of RF allowed it to capture complex interactions and minimize prediction errors, whereas SVM, though capable, was less effective in handling the variability inherent in long-term agricultural datasets. The better performance of RF in Kolar may be attributed to the more consistent weather–yield relationships in this region compared to Dharwad, which exhibits greater climatic fluctuations.

Overall, RF offered a reliable approach for mango yield forecasting in Karnataka, providing valuable insights for farmers, researchers, and policymakers to make data-driven decisions.

3.1 Model Performance in Dharwad

Table 1: Performance of SVM and RF in Dharwad

Model	R^2	RMSE (MT)	MAE (MT)	MAPE (%)
SVM	0.433	26,071.25	17,689.11	36.16
RF	0.513	24,151.62	17,499.39	25.10

Table 2: Comparative results of actual values and predicted values of mango production for SVR and Random Forest methods in Dharwad district

Year	Actual value	Predicted value	
		SVR	RF
2019	77459	49452.38	62103.44
2020	64676	49846.75	43653.81
2021	75437	68344.34	79552.97
2022	78298	67852.82	80090.08
2023	107135	75707.64	97367.38

3.2 Model Performance in Kolar

Table 3: Performance of SVM and RF in Kolar

Model	R^2	RMSE (MT)	MAE (MT)	MAPE (%)
SVM	0.079	1,09,049.33	96,608.06	34.9
RF	0.76	35,347.24	32,211.59	9.24

Table 4: Comparative results of actual values and predicted values of Mango production for SVR and Random Forest methods in Kolar district

Year	Actual value	Predicted value	
		SVR	RF
2019	422218	378837.9	417270.3
2020	429185	379984	396482.3
2021	403884	403883.9	420004.7
2022	395310	379698	403041.3
2023	405519	385723.4	397309.9

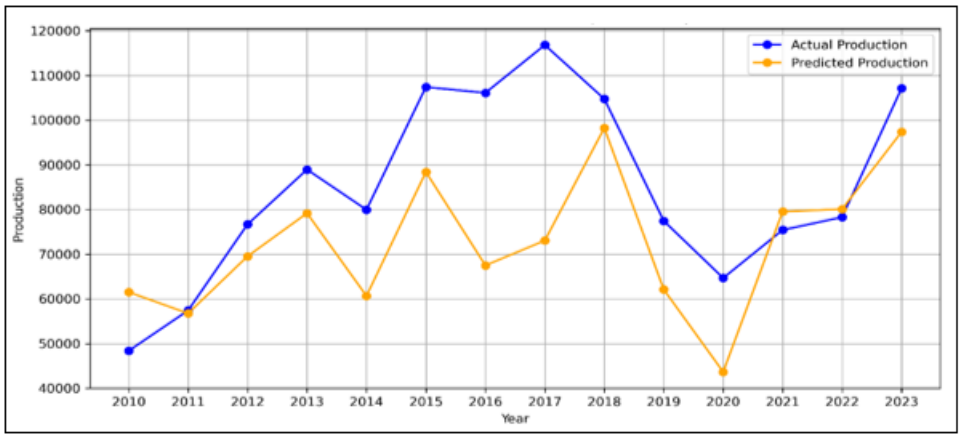

Figure 1: Performance evaluation of the best-performing model (RF) for mango yield prediction in Dharwad

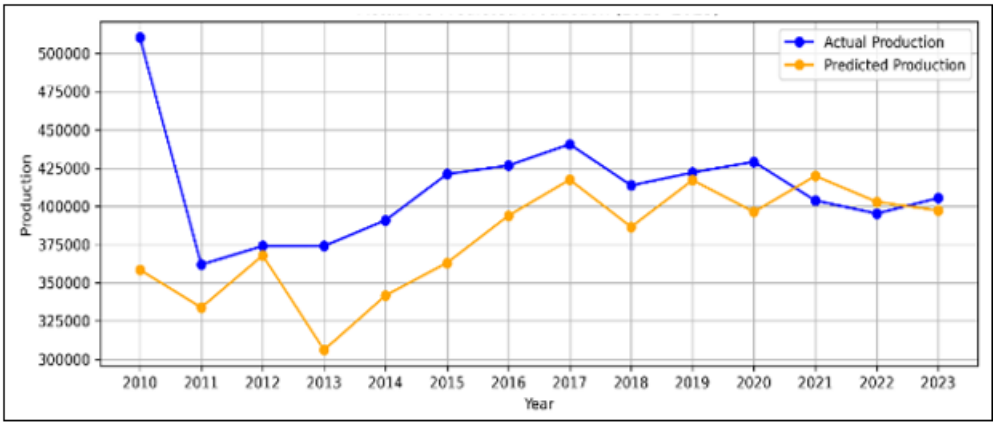

Figure 2: Performance evaluation of the best-performing model (RF) for mango yield prediction in Kolar

4. Conclusion

This comparative analysis highlights the superiority of Random Forest over Support Vector Machine for mango yield prediction in Dharwad and Kolar districts of Karnataka. While SVM could model nonlinear patterns to some extent, its predictive capacity was relatively weak, especially in Kolar. RF, by leveraging ensemble learning, captured variability more effectively, providing robust and accurate yield forecasts in both districts.

Such predictive modelling can assist farmers, planners, and policymakers in decision-making related to crop management, resource allocation, and adaptation strategies under climate variability. Future work may integrate deep learning methods or hybrid models for further accuracy enhancement.

Disclaimer (Artificial Intelligence) Author(s) hereby declares that NO generative AI technologies such as Large Language Models (ChatGPT, COPILOT, etc) and text-to-image generators have been used during the writing or editing of this manuscript.

Competing Interests
The authors have declared that no competing interests exist.

References

[1] Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32.
[2] Champaneri, M., Chachpara, D., Chandvidkar, C., and Rathod, M. 2016. Crop yield prediction using machine learning Technology. *International Journal of Science and Research,* 9(38).
[3] Dang, C., Liu, Y., Yue, H., Qian, J., and Zhu R. 2021. Autumn crop yield prediction using data-driven approaches: Support vector machines, random forest, and deep neural network methods. *Canadian Journal of Remote Sensing*, 47(2),162-181.

[4] Das, S. P., and Padhy, S. 2012. Support vector machines for prediction of futures prices in Indian stock market. *International Journal of Computer Applications*, 41(3).

[5] Everingham, Y., Sexton, J., Skocaj, D., and Inman-Bamber, G. 2016. Accurate prediction of sugarcane yield using a random forest algorithm. *Agronomy for sustainable development*, 36(1),1-9.

[6] Nitze, I., Schulthess, U., and Asche, H. 2012. Comparison of machine learning algorithms random forest, artificial neural network, and support vector machine to maximum likelihood for supervised crop type classification. *Proceedings of the 4th GEOBIA, Rio de Janeiro, Brazil*, 79(1), 3540.

[7] Saruta, K., Hirai, Y., Tanaka, K., Inoue, E., Okayasu, T., and Mitsuoka, M., 2013. Predictive models for yield and protein content of brown rice using support vector machine. *Computers and Electronics in Agriculture*, 99(1), 93-100.

[8] Shahhosseini, M., Hu, G., and Archontoulis, S. V. 2020. Forecasting corn yield with machine learning ensembles. *Frontiers in Plant Science*, 11(1), 1120.

[9] Suresh, N., Ramesh, N. V. K., Inthiyaz, S., Priya, P. P., Nagasowmika, K., Kumar, K. V. H., Shaik, M., and Reddy, B. N. K. 2021. Crop yield prediction using a random forest algorithm. In *2021 7th international conference on advanced computing and communication systems (ICACCS),* 1, 279-282.

[10] Vapnik, V., Golowich, S. E., and Smola, A. J. 1997. Support vector method for function approximation, regression estimation, and signal processing. *Advances in Neural Information Processing Systems*, 9, 281-287.

[11] Zhang, Q., Zhao, X., Han, Y., Yang, F., Pan, S., Liu, Z., Wang, K., and Zhao, C. 2023. Maize yield prediction using federated random forest. *Computers and Electronics in Agriculture*, 210, 107930.