Journal of Agriculture, Biology and Applied Statistics
Vol. 3, No. 2, 2024, pp. 81-90

URL : www.arfjournals.com
https://DOI:10.47509/JABAS.2024.v03i02.04

A Machine Learning Approach for Prediction of Grape Yield in Chikkaballapur District of Karnataka

Manoj B. G.[1],Vasantha Kumari J.[1], Ashalatha K. V.[1], Ashwini. C. B.[1], Mangala C. D.[1], Roshan K. Bhardwaj[2], Hamsa K. R.[3] and Lakshmi Narsimhaiah[4*], Vikas Jain[5] and R.P. Joshi[6]

[1]Department of Agricultural Statistics, University of Agricultural Sciences, Dharwad, India
[2]College of Agriculture and Research Station, Korba, Chhattisgarh, India
[3]College of Agriculture and Research Station, Kurud, Chhattisgarh, India
[4]College of Agriculture and Research Station, Jashpur, Chhattisgarh, India
[5]College of Agriculture, Powarkheda, J.N.K.V.V. (M.P.), India
[6]College of Agriculture,Rewa,J.N.K.V.V.(M.P.), India
[*]Correspondence author Email: lakshmi.narasimhaiah1988@gmail.com

To cite this article

Manoj B.G., & et al. (2024). A Machine Learning Approach for Prediction of Grape Yield in Chikkaballapur District of Karnataka. Vol. 3, No. 2, pp. 81-90. *https://DOI:10.47509/JABAS.2024.v03i02.04*

Abstract: Data mining is an important application to predict grape yield in Chikkaballapur district, Karnataka, focusing on precision agriculture. It emphasizes the pivotal role of climate in crop production, particularly in the context of climate change and extreme weather events. The research analyses various stages of grape development and their sensitivity to weather parameters, underscoring the need for precise interval-based division of crop growth phases. The integration of advanced technology in agriculture, including crop modelling and predictive tools, offers significant potential for improving crop yield predictions. By providing farmers with timely and accurate information based on meteorological, soil, and other relevant data, these tools empower them to make informed decisions and enhance crop productivity while mitigating losses. Machine learning algorithms such as LASSO, Ridge, Elastic Net (ELNET), Support Vector Regression (SVR), and K-Nearest Neighbor (KNN) are employed for predictive modelling, with evaluation metrics like Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) guiding model performance assessment. The research reveals correlations between weather parameters and grape yield, highlighting the significance of these variables in predicting crop outcomes. Support Vector Regression (SVR) emerges as the most effective algorithm, showcasing its potential for handling complex, nonlinear data and improving grape yield predictions. This study contributes valuable insights to precision agriculture and crop yield prediction, benefiting viticulture in Chikkaballapur and potentially revolutionizing agricultural practices worldwide.

Article History Received : 22 March 2024; Revised : 27 April 2024; Accepted : 10 May 2024; Published : 30 September 2024

Introduction

Data mining is the process of extracting useful insights or knowledge from large amount of data. At present, data mining is becoming an extremely important tool to transform available data into information. When the Data mining techniques are used with agriculture data, the term is known as precision agriculture (Shah and Shah, 2018). Machine learning and data mining are two important applications of statistics that are becoming increasingly popular due to the rise of big data. Climate is one of the primary factors beyond human control, which determines the crop yield. In this sense, modelling and prediction of crop yield by considering the climate variables has become an interesting research topic (Wickramasinghe*et al.*, 2021). In recent years, there has been growing interest in using machine learning algorithms for predictive modelling in grapes. Machine learning techniques can help identify the key factors that affect grape yield and quality through feature selection and develop predictive models that can be used to optimize cultivation practices and increase grape yield.

Climate change has posed mammoth challenges for global viticulture, and almost all the growing regions are facing the mounting pressure exerted owing to this unchecked climatic challenge. Pedo-climatic and topographic features largely affect the production and quality of table grapes and wine (Rafique*et al*., 2023). Changes in precipitation patterns and increase in the frequency and intensity of extreme weather events (high temperatures and heat waves) harm crop productivity (Rajath*et al.*, 2023). The effect of weather on grapes varies with the stage of its development. The influence of weather on grapes is found to be dependent on the magnitude of weather variables as well as how the weather is distributed across the various growth stages of this crop. This is because different growth stages of the crop have different sensitivities to weather parameters; some are very sensitive to weather fluctuations, while others are not. Hence, it is necessary to divide the entire crop growth phase into very narrow intervals in order to forecast precisely.

The combination of higher technology and agriculture to improve the production of crop yield is becoming more interesting newly. Due to the rapid development of new high technology, crop models and predictive tools might be predictable to become a crucial element of agriculture (Devika and Ananthi, 2018). Estimating agricultural crop yield prior to harvest is an important issue in agriculture, as the changes in crop yield from year to year influence international business, food supply, and global market prices. Also, early prediction of crop yield provides useful information to policy planners (Palanivel and Surianarayanan, 2019). The provision of accurate and timely information such as meteorological, soil, use of fertilizers, use of pesticides can assist farmers to make the best decision for their cropping situations. This could benefit them to attain greater crop productivity if the conditions are suitable or help them to decrease the loss due to unsuitable conditions for the crop yield (Gandhi *et al*., 2016).

This study aims at understanding machine learning-enabled methodology for grape yield prediction based on the set of weather parameters as predictors for production of

grapes or commonly called as independent variables and production of grapes as label or dependent variable . Supervised machine learning algorithms including LASSO, Ridge, ENET, SVR and K-NN techniques were used to predict the yield based on weather parameters. Accuracy metrics such as MSE, RMSE and MAE were used for evaluation of model performances.

Materials and Methods

Nature of study area

Chikkaballapur is located at a height of 915 meters above sea level. According to the Koppen-Geiger climate classification system, the climate of Chikkaballapur district falls under the category of Aw, which represents a tropical savanna climate. The latitude and longitude of Vijayapura are 13.4355° N and 77.7315° E. It comes under the Eastern Dry Zone of Karnataka. The district is well known for sericulture, mango, grapes, pomegranate, sapota, guava, papaya, banana and citrus and cut flower cultivation.

Nature and sources of data

The present study was based on secondary data on weather parameters from a period of 1981-2021. The climatic scenario dataset for 1981-2021 was downloaded from Coupled Model Intercomparison Project Phase-6 (CMIP-6) with a resolution of 0.5 x 0.5 degree. The data was then analysed and sorted to the specific locations using the FERRET software in the Linux platform. Crop yield data were collected from the District Statistical Office, Chikkaballapur.

Correlation analysis

Correlation measures the degree of closeness or association between two variables and the strength of the relationship between different parameters. Correlation heatmaps are generated in the form of a matrix with radiant colours, if darker colour signifies a stronger correlation, then lighter colour signifies a weak correlation, and vice-versa. The correlation heatmap was generated using the Seaborn library written in Python language.

LASSO regression

It is a method that combines the least-squares loss with an L_1- constraint, or bound on the sum of the absolute values of the coefficients. Relative to the least-squares solution, this constraint has the effect of shrinking the coefficients and even setting some to zero. In this way it provides an automatic way for doing model selection in linear regression. Moreover, unlike some of other criteria for the model selection, the resulting optimization problem is convex and can be solved efficiently for large problems. Given a collection of N predictor-response pairs $\{(x_i, y_i)\}$ from i = 1 approaching to N, the lasso finds the solution (β_0, β_j) to the optimization problem.

$$\text{minimize}\left\{\frac{1}{2N}\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2\right\}$$

$$\textit{Subject to } \sum_{j=1}^{p}\left|\beta_j\right| \le t$$

Ridge Regression

Ridge regression causes the regression coefficients to shrink so that factors that have a negligible impact on the outcome have their coefficients near to zero. The reduction of the coefficients is accomplished by penalizing the regression model with a term known as L_2-norm, which is the sum of the squared coefficients (Zou and Hastie, 2005). Where, the loss is defined as:

$$L_{ridge}\left(\hat{\beta}\right) = \sum\left(y_i - x_i^{'}\hat{\beta}\right)^2 + \lambda\sum_{j=1}^{m}\beta_j^2 = y - X\hat{\beta}^2 + \lambda\hat{\beta}^2$$

where represents the independent variable, β represents the coefficient associated with it, and λ represents the L_2 norm penalty.

Elastic net (ELNET) regression

The ELNET model has features of both LASSO and ridge regressions i.e., it considers both the L_1 and L_2 norms (Hoerl and Kennard, 1970). This causes some coefficients to shrink and some coefficients to be set to zero. Therefore, it reduces the impact of various features without eliminating them completely (Cho *et al*., 2009).

$$L_2 = \sum\left(\hat{Y}_i - Y_i\right)^2 + \lambda\sum\beta^2$$

where, represents the independent variable, β represents the corresponding coefficient and λ represents the penalty.

Support Vector Regression

Support Vector Regression introduced by Vladimir N. Vapnik and Alexey Ya. Chervonenk is in 1963, is the regression model of Support Vector Machine, on a dataset consisting of L samples of form $\{(x_1, y_1), (x_2, y_2), \ldots, (x_L, y_L), x \in R^m, y \in R\}$ is a linear function which can estimate output values based on inputs.

$$y = (w.x) + b$$

where, y is the estimated value, x is the input vector, w is the weight vector and b are the bias.

SVR creates a hyperplane or set of hyperplanes in a high or infinite dimensional space, which is utilized for regression, classification or other tasks. SVR uses linear functions for learning. In case of nonlinear cases, SVR uses a kernel technique to plot the data into a higher dimensional feature space, in which linear functions can be applied (Palanivel*et al*., 2019).

While using the SVM for regression analysis, a margin of tolerance is fixed in approximation to the SVM which has been used for the problem. SVMs represent the hyperplane as optimized hyperplanes with support vectors (Mohapatra and Chaudhary, 2022).

Hyper parameters in SVR

1. Hyperplane: Hyperplanes are decision boundaries for predicting the continuous output. Support Vectors are the data points on either side of the hyperplane that are closest to the hyperplane. These are used to draw the required line that shows the algorithm's predicted outcome.
2. Kernel: A kernel is a collection of mathematical functions that take data and change it into the desired form. These are most commonly used to find a hyperplane in higher-dimensional space. Linear, Non-Linear, Polynomial, Radial Basis Function (RBF) and Sigmoid are the most commonly used kernels. RBF is the kernel that is used by default. Depending on the dataset, each of these kernels are used.
3. Boundary Lines: These are the two lines that are drawn at a distance of ? (epsilon) from the hyperplane. It's used to separate the data points by a margin.
4. Support Vectors: The closest point of the lines from both the classes.

The Support Vector Regressor model fits the hyperplane which has the maximum points and uses a threshold value. The Support Vector Machine algorithm model can also be useful as a regression analysis technique while considering the significant features that can characterize the algorithm.

K-Nearest Neighbour

Cover & Hart introduced K-Nearest Neighbour which is a machine learning method used for regression as well as classification. K-NN considers each data record as a vector in an m-dimensional space (where m is the number of features), and predicts the value of each new sample based on the values of K records that are closest to that point in that space (Enas and Choi, 1986).

The way in which the algorithm decides which of the points from the training set are similar enough to be considered when choosing the class to predict for a new observation is to pick the K closest data points to the new observation, and to take the most common class among these (Karthikeya *et al*., 2020). This is why it is called the K-Nearest Neighbors algorithm. This distance is calculated using various measures such as Euclidean distance, Minkowski distance, Mahalanobis distance. The larger is K; the better is classification. For instance, the closeness of the new point x and the training point xi is measured by a Euclidean distance function in the form of equation as following.

$$d(x, x_i) = \sqrt{\sum_{j=1}^{m}(x_i^j - x^j)^2}$$

i =1, 2,..,n. and j=1,2,..,m.

Where n is the number of training samples and m is the number of input samples.

Samples that are closer to the new sample will have a greater impact on the prediction.

The implementation of algorithm can be noted as below :

1. Load the data
2. Initialize K to your chosen number of neighbors
3. For each example in the data
4. Calculate the distance between the query example and the current example from the data.
5. Add the distance and the index of to an ordered collection.
6. Sort the ordered collection of distances and indices from smallest to largest (in ascending order) by the distances
7. Pick the first K entries from the sorted collection
8. Get the labels of the selected K entries
9. If regression, return the mean of the K labels

Results and Discussion

Correlations between grape yield and weather parameters were useful because they can indicate a predictive relationship that can be exploited in practice. Weather parameters are believed as the significant factors that influence grape yield. Therefore, it is inevitably important to know the correlation between grape yield and weather parameters. The weather parameters considered were Maximum temperature, Minimum temperature, Relative humidity, Rainfall and Wind speed. The correlation matrix between weather parameters and grape yield is represented in the following tables.

Table 1: Correlation of weather parameters and grape production in Chikkaballapur district.

	Yield	*Max T*	*Min T*	*RH*	*RF*	*WS*
Yield	1.000					
Max T	0.288	1.000				
Min T	0.543**	0.421**	1.000			
RH	-0.023	-0.612**	0.054	1.000		
RF	0.087	-0.575**	0.093	0.635**	1.000	
WS	-0.002	0.507**	-0.008	-0.421**	-0.488**	1.000

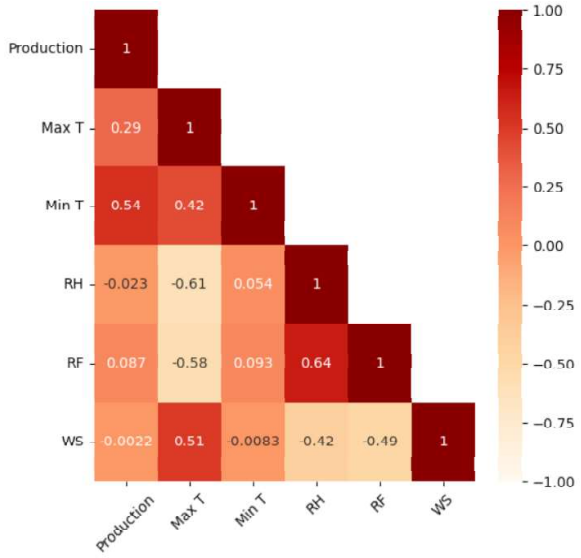

Fig. 1: Correlation heatmap for Chikkaballapur district

For Chikkaballapur district, as presented in the form of heatmap (Figure), a positive association was observed between production of grapes and maximum temperature. While there was positive association between production of grapes and minimum temperature which was highly significant indicating that increase in the minimum temperature increases the production of grapes, a negative association was observed in between production of grapes and relative humidity and also with windspeed, which was found out to be non-significant indicating that increase in the relative humidity and windspeed decreases the production of grapes. Multicollinearity was observed with several weather parameters which hinders the performance of a model as in case of maximum temperature and minimum temperature.These results were on par with the study conducted by Bhagat*et al*. (2021), reported that the production of cotton in Jalgaon district of Maharashtra was highly dependent on relative humidity at evening and bright sunshine hours. However, wind velocity was adversely affecting the productivity of cotton in the Jalgaon district.

The study aimed to predict grape yield using weather-related data. The dataset was divided into two parts: a training set (80% of the data) and a testing set (20% of the data). Several machine learning algorithms were employed to create predictive models. These models were then assessed using common evaluation metrics, which included Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and Median Absolute Error (MeAE). The model with the lowest values for these metrics was considered the most suitable for predicting groundnut yield based on the weather-related variables.

Table 2: Comparison of different machine learning algorithms for prediction of grapes yield in Chikkaballapur district.

Models	*MSE*	*RMSE*	*MAE*	*MeAE*
LASSO	374.518	19.352	15.905	13.013
Ridge Regression	364.452	19.091	15.667	13.138
ELNET	357.204	18.900	15.601	13.171
SVR	185.485	13.619	11.767	8.266
KNN	208.805	14.450	12.597	13.001

Least values of MSE, RMSE, MAE and MeAE values was obtained in the case of SVR algorithm when compared to other algorithms, which were found to be 185.485, 13.619, 11.767 and 8.266 respectively. However, MSE, RMSE, MAE and MeAE values were obtained in the case of LASSO regression were found to be high with values 374.518, 19.352, 15.905 and 13.013 respectively. It is evident that SVR and K-NN outperformed the other three algorithms with comparatively less MSE, RMSE, MAE and MeAE values. Possibly, the effectiveness of Support Vector Regression (SVR) can be attributed to its utilization of the 'Kernel' feature, which simplifies the handling of complex and non-linear data. The Kernel function facilitates the transformation of a lower-dimensional dataset into a higher-dimensional one, thereby enabling the creation of hyperplanes that are instrumental for predictive purposes.

It has been observed that SVR performs well with smaller test datasets, in contrast to K-Nearest Neighbor, which demonstrates suboptimal performance with such datasets. This finding aligns with the results obtained by Bondre and Mahagaonkar (2019), who examined various machine learning techniques for forecasting future crop production. In their study, SVR outperformed the Random Forest algorithm. Similarly, the efficacy of SVR was affirmed in a study conducted by Khosla *et al.*(2020), where they explored crop yield prediction using aggregated rainfall-based modular Artificial Neural Networks and Support Vector Regression. In this instance as well, SVR outperformed other techniques.

Conclusion

The study explored the application of machine learning techniques to predict grape yield based on weather parameters, shedding light on the intricate relationship between climate and agriculture. The analysis revealed that maximum and minimum temperatures positively impact grape production, while relative humidity and wind speed exhibit negative correlations. These findings align with the broader understanding of how weather influences crop yield. The study compared several machine learning algorithms, with Support Vector Regression (SVR) and K-Nearest Neighbor (K-NN) emerging as the top performers. SVR, in particular, demonstrated exceptional predictive capabilities, reflected in its lower Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Median Absolute Error (MeAE) values. This underscores SVR's efficacy in handling complex and

non-linear data, making it a valuable tool in precision agriculture for grape yield estimation. The implications of this research extend beyond viticulture. It underscores the potential of machine learning and data-driven insights to optimize agricultural practices, adapt to changing climate conditions, and enhance crop productivity. As climate change continues to pose challenges to global agriculture, the adoption of such predictive models becomes increasingly important, offering a promising avenue for sustainable and efficient food production.

References

Bhagat A A, Shivgaje A J and Badgujar C D, 2021, A study on the variability in rainfall and relationship of weather parameters with cotton crop in Jalgaon district of Maharashtra. *International Journal of Farm Sciences*, 11(4): 62-65.

Bondre D A and Mahagaonkar S, 2019, Prediction of crop yield and fertilizer recommendation using machine learning algorithms. *International Journal of Engineering Applied Sciences and Technology*, 4(5):371-376.

Cho S, Kim H, Oh S, Kim K and Park T, 2009, Elastic-net regularization approaches for genome-wide association studies of rheumatoid arthritis. *BMC Proceed*, 3(7): 1-6

Devika B and Ananthi B, 2018, Analysis of crop yield prediction using data mining technique to predict annual yield of major crops. *International Research Journal of Engineering and Technology*, *5*(12): 1460-1465.

Enas G G and Choi S C, 1986, Choice of the smoothing parameter and efficiency of k-nearest neighbor classification. Statistical Methods of Discrimination and Classification, 2(3): 235-244.

Gandhi N, Armstrong L J, Petkar O and TripathyA K, 2016, Rice crop yield prediction in India using support vector machines. 13th International Joint Conference on Computer Science and Software Engineering (JCSSE), 1-5 (IEEE).

Hoerl A E and Kennard R W, 1970, Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55-67.

Karthikeya H K, Sudarshan K and Shetty D S, 2020, Prediction of agricultural crops using KNN algorithm. *International Journal of Innovative Science and Research Technology*, 5(5): 1422-1424.

Khosla E, Dharavath R and Priya R, 2020, Crop yield prediction using aggregated rainfall-based modular artificial neural networks and support vector regression. *Environment, Development and Sustainability*, 22:5687-5708.

Mohapatra S and Chaudhary N, 2022, Statistical Analysis and Evaluation of Feature Selection Techniques and implementing Machine Learning Algorithms to predict the Crop Yield using Accuracy Metrics *Engineered Science*, 21: 787.

Palanivel K and Surianarayanan C, 2019, An approach for prediction of crop yield using machine learning and big data techniques. *International Journal of Computer Engineering and Technology*, *10*(3): 110-118.

Palanivel K and Suriyanarayanan C, 2019, An approach for prediction of crop yield using machine learning and big data techniques *International Journal of Computer Engineering and Technology*, 10(3): 110-118.

Rafique R, Ahmad T, Kalsoom, T, Khan M A and Ahmed M, 2023, Climatic Challenge for Global Viticulture and Adaptation Strategies. *Springer International Publishing*, 611-634.

Rajath H P, Kumar C, Hanumanthappa M, Bhanuprakash H R, Yogesh G S and Chandrakala H, 2023, Impact of weather parameters on maize agroecosystem and adaptation strategies under changing climatic conditions: A review on Sustainable and climate-resilient adaptation strategies in maize agroecosystem. *Plant Science Today*, 10: 1-10.

Shah V and Shah P, 2018, Groundnut crop yield prediction using machine learning techniques. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 3(5): 1093-1097.

Wickramasinghe L, Weliwatta R, Ekanayake P and Jayasinghe J, 2021, Modeling the relationship between rice yield and climate variables using statistical and machine learning techniques. *Journal of Mathematics*, *2021*: 1-9.

Zou H and Hastie T, 2005, Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society*, 67 (2): 301–302.