ISSN: 2456-1452
Maths 2023; 8(3): 141-149

https://www.mathsjournal.com
Received: 04-01-2023
Accepted: 08-02-2023

Ajith S
Ph.D. Scholar, Department of Agricultural Statistics, Uttar Banga Krishi Viswavidyalaya (UBKV), West Bengal, India

Manoj Kanti Debnath
Assistant Professor, Department of Agricultural Statistics, Uttar Banga Krishi Viswavidyalaya (UBKV), West Bengal, India

Deb Sankar Gupta
Professor, Department of Agricultural Statistics, Uttar Banga Krishi Viswavidyalaya (UBKV), West Bengal, India

Pradip Basak
Assistant Professor, Department of Agricultural Statistics, Uttar Banga Krishi Viswavidyalaya (UBKV), West Bengal, India

Corresponding Author:
Ajith S
Ph.D. Scholar, Department of Agricultural Statistics, Uttar Banga Krishi Viswavidyalaya (UBKV), West Bengal, India

Application of statistical and machine learning models in combination with stepwise regression for predicting rapeseed-mustard yield in Northern districts of West Bengal

Ajith S, Manoj Kanti Debnath, Deb Sankar Gupta and Pradip Basak

DOI: https://dx.doi.org/10.22271/maths.2023.v8.i3b.1004

Abstract
Rapeseed-mustard crop is an important oilseed crop in India. District-wise yield prediction is essential for various location specific decision making. The performance of two machine learning models namely Support Vector Regression (SVR) and Artificial Neural Network (ANN) were compared with basic linear regression model for district-wise yield prediction of rapeseed-mustard crop. The study area for the present investigation were Cooch Behar, Malda, Jalpaiguri and Uttar Dinajpur districts of West Bengal. Yearly unweighted and weighted weather indices were calculated from weekly weather parameters. The indices that significantly affecting yield were selected using stepwise regression for fitting the models. The ANN model was fitted using backpropagation algorithm. The optimum number of neurons in hidden layer for ANN were ranging between two to four. The Tangent hyperbolic function was found to be suitable hidden layer activation function. The nonlinear Radial Basis Function kernel was the best kernel for Support Vector Regression. While evaluating the performance of fitted models in both calibration and validation stages, the ANN model was the best fitted model for Cooch Behar and Malda and SVR was the best fitted model for Jalpaiguri and Uttar Dinajpur districts. It was concluded that the machine learning models outperformed multiple linear regression model for district-wise yield prediction of rapeseed-mustard crop.

Keywords: Rapeseed-mustard, weather indices, stepwise regression, multiple linear regression, artificial neural network, support vector regression

1. Introduction
Rapeseed-mustard crop is an important oilseed crop in India. India is the second largest cultivator of the crop with 6.86 million hectares of cultivational area with productivity of 1331 kg/ha that yields 9.12 million tonnes of oilseeds. The rapeseed–mustard crop is grown in diverse agroclimatic conditions ranging from north-western/north-eastern hills to down south. Madhya Pradesh, Rajasthan, Haryana, Uttar Pradesh West Bengal and Assam states were leading in rapeseed–mustard production in India [1].

The rapeseed-mustard crop is grown in sub-tropical regions of West Bengal. The cold weather condition prevailing in the northern part of West Bengal is favourable condition for cultivating rapeseed-mustard crop [2]. The crop is mostly grown in rabi season as a cold weather crop in West Bengal [3].

According to the Economic Survey of India-2021-22, due to increasing population growth and urbanization, oil consumption is expected to remain high. The advanced yield prediction provides an insight regarding the quantity of oilseeds will be produced. This will be useful to make import policies to ensure adequate supply of oil.

The productivity of rapeseed-mustard is more prone to vulnerable due to highly dynamic temperature and greater uncertainties in rainfall. The Rapeseed-Mustard crop yield is profoundly influenced by the weather particularly temperature affects various phenological stages. The growth and development of rapeseed-mustard crop differs in different environmental conditions [4].

crop establishment and cold spell and intermittent rains during crop growth stage cause considerable yield losses by physiological disorder and appearance and proliferation of diseases such as white rust, downy mildew and Sclerotinia stem rot and aphid pest [5, 6]. Hence the prediction models using weather parameters can provide more accurate performance.

Fisher (1924) [7] was the first person who identified that the effect of change in weather condition on yield in successive weeks will be an orderly one that follows some mathematical law. The rapeseed-mustard crop requires different climate conditions in different stages of crops [8]. In order to give weightage to weekly weather conditions, weighted weather indices were first developed by Jain *et al*. (1980) [9]. The statistical models using correlation coefficient based weighted indices can be effectively used to predict crop yield [10].

The regression model is a standard linear statistical method to predict one response variable using one or more regressor variables [11]. The performance of advanced machine learning methods is more accurate as it tries to find the pattern of crop response to varying climatic conditions [12]. Artificial Neural Network (ANN) is a machine learning model that able understand the nonlinear effect of input variable on yield [13]. Support Vector Regression (SVR) is another machine learning model that maps the input data into high dimensional space using nonlinear kernel function [14]. With this background, the present study made an attempt to compare two nonlinear machine learning models namely ANN and SVR with Linear Regression for predicting rapeseed-mustard yield using weather indices. The performance of ANN model under four activation functions and performance of SVR under three kernel functions were evaluated.

2. Materials and Methods

Description of Study area and data

Cooch Behar, Malda, Jalpaiguri and Uttar Dinajpur districts that were present in northern part of West Bengal were considered for the present study. Yield data of Rapeseed-mustard crop in the selected districts from 1997-98 to 2020-21 were collected from Directorate of Economics and Statistics under Government of India. Weekly weather data of these districts were collected from Regional Meteorological Centre (RMC), Kolkata.

The Maximum Temperature (T_{Max}), Minimum Temperature (T_{Min}), Rainfall (RF), Relative Humidity (RH) and Windspeed were the weather parameters considered for the study.

Calculation of different weather indices

Weekly weather data of the weeks in which rapeseed-mustard crop is grown were used to develop yearly weather indices. As the rapeseed-mustard crop is grown in rabi season, the weather parameters of 47th Standard Meteorological Week (SMW) to 11th SMW of next year were considered.

Unweighted indices were calculated for each weather parameters as simple average of weather data of the weeks in which rapeseed-mustard crop was grown.

The correlation coefficient based weighted index of jth weather parameter for ith year (C_{ij}) was calculated as follows,

$$C_{ij} = \frac{\sum_{k=1}^{m} r_{jk} \cdot X_{ijk}}{\sum_{k=1}^{m} r_{jk}} \quad (1)$$

Similarly path correlation coefficient based weighted index of jth weather parameter for ith year (P_{ij}) was calculated as follows,

$$P_{ij} = \frac{\sum_{k=1}^{m} p_{jk} \cdot X_{ijk}}{\sum_{k=1}^{m} p_{jk}} \quad (2)$$

Where, X_{ijk} is jth weather parameter in kth week of ith year and r_{jk} and p_{jk} are the correlation coefficient and path coefficient between detrended crop yield and jth weather parameter at kth week respectively.

Indices selection using Stepwise Regression

Five unweighted, five correlations based and five path coefficient based weighted indices were calculated from five weather parameters. Hence there were fifteen weather indices. Inclusion of all independent variables into the model leads to complex model and many parameters need to be estimated. Inclusion of many variables prone to the multicollinearity problem that leads to unstable coefficients [15]. Hence, Stepwise Regression (SR) was employed to select the indices that were significantly influenced the yield among the fifteen developed indices.

Stepwise regression is a classical variable selection methodology which is used to identify and select a useful subset of the important explanatory variables set [16, 17, 18]. Stepwise regression selects explanatory variables based on their statistical significance [19]. It is employed in stepwise manner on choosing the variables that give the best predictions by addition or deletion of variables at each step. The stepwise regression is the compromise between forward selection and backward elimination procedures in managing the limitations of both the methods [20, 21].

Multiple Linear Regression (MLR)

Multiple Linear Regression (MLR) model uses more than one explanatory variable to predict a response or dependent variable [22].

$$Y = A + \sum \beta_i . X_i + \varepsilon \quad (3)$$

Where, Y is the dependent variable, X_i is ith independent variable, β_i's are regression coefficients and ε is the error. The Ordinary Least Square method was used to estimate the regression coefficients [23].

Artificial Neural Network (ANN)

Multilayer Perceptron (MLP) is the most commonly used type of Feedforward Neural Network topology [24]. A typical MLP architecture consists of three layers namely, input layer, hidden layer and an output layer. The input layer obtains input data (x_i) into the neural network system and an appropriate weight (w_i) is to be multiplied with each input. A bias term (b) also incorporated as weight of an edge using a bias neuron. The sum of this product will be transmitted to the hidden layer. In hidden layer, a nonlinear transformation is to be applied through a nonlinear activation function. This computed value transmitted to the output layer as a linear combination of all the neurons of hidden layer [25].

The relationship between output and inputs can be mathematically represented as follows:

$$Y_t = f\left\{\sum_{j=1}^{q} w_j . g\left(\sum_{i=1}^{k} (w_i . x_i) + b . w_0\right)\right\} + e_t \quad (4)$$

The Back Propagation (BP) is a straightforward algorithm to train MLP [26]. The BP algorithm starts with training the input data with random weights and the weights are adjusted in successive steps to reduce error. Learning rate is an important factor in BP algorithm which determines the rate at which the weights were updated in each step.

The critical aspect of neural network model were the number of hidden layer neurons and choice of activation function. The 10-fold cross validation procedure was used to optimise the number of neurons in the hidden layer [27]. The performance MLP neural network model with logistic, Tangent hyperbolic (Tanh) Softmax and Restricted Linear Unit (ReLU) activation functions were evaluated.

Support Vector Machine Regression (SVR)

Support vector machines (SVM) was developed by Vapnik for classification purposes. As a generalization of SVM, Support Vector Regression (SVR) can be used for regression context by introducing ϵ-insensitive tube [28].

The SVR model is given as

$$y = f(x) + \varepsilon \quad (5)$$

Where, f(x) is estimated as

$$f(x) = \sum_{i=1}^{n} x_i . w_i . K(x_i, x_k) + b \quad (6)$$

Where, $K(x_i, x_k)$ is Kernel function. The performance three kernels namely Linear, Radial Basis Function (RBF) and polynomial kernel were evaluated. Linear kernel function can be used when the data is linearly separated. Nonlinear kernel such as Radial Basis Function (RBF) and polynomial kernel were appropriate when the data is not linearly separated [29].

The learning process of SVR is controlled by hyperparameter. Cost (C) and Epsilon (ε) are the hyperparameters to be optimized while training SVR model using linear kernel. Along with Cost and Epsilon, an additional parameter Gamma (γ) has to be optimized in nonlinear RBF kernel function. Another parameter, degree of polynomial (d) has to be optimized in polynomial kernel function. Grid search using 10-fold cross validation was used to tune the hyperparameters [30].

Selection of best fitted models

The MLR model and two machine learning models ANN and SVR were fitted using the indices selected by Stepwise Regression (SR). 80% data were used for model fitting (calibration) and remaining 20% data were used for validation of the fitted models. The following evaluation criteria were used to select the best fitted model for each district.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \quad (7)$$

$$\text{Adj. } R^2 = 1 - \frac{n-p}{n-p-1}(1 - R^2) \quad (8)$$

$$MAE = \frac{\sum_{i=1}^{n}|(Y_i - \hat{Y}_i)|}{n} \quad (9)$$

$$RMSE = \sqrt{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 / n} \quad (10)$$

Where $\hat{Y}_i$ and Y_i are the predicted and observed yield.

3. Results and Discussion

Calculation of weather indices

Weather indices were calculated using the weather parameters of the weeks of the year in which the crop is grown. The yield was detrended before calculating the weighted indices, as there was as a significant trend in the rapeseed-mustard yield. The weighted indices developed using detrended yield was precise [31]. By detrending the yield, the effect of trend causes such as improved varieties and other cultivational practices *etc.*, were removed. It was expected that detrended yield represents only the actual effect of weather factors on yield by removing trend causes [32]. The correlation coefficient based weighted indices were prefixed with CC and path coefficient based weighted indices were prefixed with PC for differentiating different indices of same weather parameters.

Stepwise Regression (SR)

Stepwise regression analysis was applied by taking Rapeseed-Mustard yield as dependent variable and 15 weather indices as explanatory variables for each district separately. The results of stepwise regression were given in the table 1. For Cooch Behar district, correlation coefficient-based index of Relative Humidity (CC_RH) was included in the first step which was significant at 5% Level of Significant (LoS). CC_RH alone explaining 50% variation in the yield. The path coefficient-based index of Maximum Temperature (PC_T_{Max}) was included in the second step which further increased the adjusted R^2 to 0.65. There was a decrease in AIC value. The correlation coefficient-based index of Wind Speed (CC_Wind. Speed) was included in the third step. There was a further increase in adjusted R^2 and decrease in AIC values. All the three includes indices and intercept were significant at 5% LoS. These three indices were cumulatively explaining 72% of variation in the yield of Rapeseed-mustard crop. Inclusion of other indices to the stepwise regression were not increase the adjusted R^2 to a significant level and their respective regression coefficients were not significant at 5% LoS.

Table 1: Summary of Stepwise Regression analysis

District	Step	Predictor	Coefficient	p value	Adj. R^2	AIC
Cooch Behar	1	(Intercept)	344.08	0.00	0.50	282.45
		CC_RH	5.21	0.00		
	2	(Intercept)	-606.19	0.05	0.65	274.79
		CC_RH	4.19	0.00		
		PC_T_{Max}	38.78	0.00		
	3	(Intercept)	-828.76	0.01	0.72	270.53
		CC_RH	2.38	0.05		
		PC_T_{Max}	37.10	0.00		
		CC_Windspeed	220.67	0.02		
Jalpaiguri	1	(Intercept)	1814.86	0.00	0.48	279.01
		PC_T_{Min}	-86.51	0.00		
	2	(Intercept)	733.21	0.20	0.56	276.22
		PC_T_{Min}	-88.75	0.00		

		T_{Max}	42.44	0.04		
	3	(Intercept)	972.14	0.08	0.61	273.81
		PC_ T_{Min}	-69.37	0.00		
		T_{Max}	54.36	0.01		
		T_{Min}	-61.49	0.06		
	4	(Intercept)	-463.01	0.62	0.66	271.61
		PC_T_{Min}	-54.59	0.01		
		T_{Max}	68.69	0.00		
		T_{Min}	-66.30	0.03		
		RH	12.08	0.05		
Malda	1	(Intercept)	1160.28	0.00	0.29	305.45
		RF	-1.58	0.00		
	2	(Intercept)	433.36	0.04	0.56	295.03
		RF	-1.69	0.00		
		PC_RH	9.93	0.00		
	3	(Intercept)	486.04	0.01	0.68	288.98
		RF	-1.49	0.00		
		PC_RH	9.09	0.00		
		CC_RF	2.38	0.01		
Uttar Dinajpur	1	(Intercept)	-363.17	0.24	0.39	306.65
		CC_T_{Min}	74.73	0.00		
	2	(Intercept)	-1548.04	0.00	0.59	299.99
		CC_T_{Min}	70.56	0.00		
		PC_RH	15.56	0.01		
	3	(Intercept)	-1836.89	0.00	0.61	298.17
		CC_T_{Min}	40.92	0.01		
		PC_RH	17.59	0.00		
		PC_T_{Min}	47.35	0.05		

There were four steps for Jalpaiguri district in which path coefficient-based index of Maximum Temperature (PC_T_{Max}), unweighted indices of Maximum Temperature (T_{Max}), Minimum Temperature (T_{Min}) and Relative Humidity (RH) were added in successive steps. These four indices were significant at 5% LoS. There was a decline in the AIC value in each step and it was low when these four indices were included. These four indices were together explaining 66% of the variation present in the yield of Rapeseed-Mustard crop in Jalpaiguri district.

Unweighted index of Rainfall (RF), correlation coefficient-based index of Relative Humidity (CC_RH) and path coefficient-based index of Rainfall (PC_RF) were included in stepwise regression for Malda district. All the three includes indices and intercept were significant at 5% LoS. AIC value was low when these three indices were included in the model. These three indices were cumulatively explaining 68% of variation in the yield of Rapeseed-mustard crop in Malda district.

The correlation coefficient-based index of Minimum Temperature (CC_T_{Min}), path coefficient-based index of Relative Humidity (PC_RH) and Minimum Temperature (PC_T_{Min}) were included in three steps for Uttar Dinajpur district. These three indices along with intercept were significant at 5% LoS and AIC was low when these three indices included in the model. These three indices were cumulatively explaining 61% of variation in the yield of Rapeseed-mustard crop in Uttar Dinajpur district.

The Multiple Linear Regression (MLR) models were fitted for each district by using the indices selected for respected district as independent variables and the yield as dependent variable. The fitted MLR models for each district were given below.

Cooch Behar: Yield = -828.76+2.38xCC_RH+37.10xPC_T_{Max}+220.67xCC_Windspeed

Jalpaiguri: Yield = -463.01-54.59xPC_T_{Min}+68.69x T_{Max}-66.30x T_{Min}+12.08xRH

Malda: Yield = 486.04-1.49xRF+9.09xPC_RH+2.38xCC_RF

Uttar Dinajpur: Yield = -1836.89+40.92xCC_T_{Min}+17.59x PC_RH+47.35x PC_T_{Min}

Artificial Neural Network (ANN) model

The number of input layer neurons were the number of indices selected from stepwise regression. Hence there were three input neurons for Cooch Behar, Malda and Uttar Dinajpur districts and it was four for Jalpaiguri. From the Fig. 1 it can be seen that RMSECV was low for two hidden layer neurons for Cooch Behar and Uttar Dinajpur and the optimum number of hidden neurons were four and three for Jalpaiguri and Malda districts respectively.

The results of ANN model using BP algorithm using four different activation function were given in the table 2. For Cooch Behar district, the ANN model using Tanh activation function was converged to lowest possible Sum of Squared Error (SSE). Learning rate of Tanh activation function was comparatively low level of 0.07. Similarly, the learning rate was low in tanh activation function (0.03) and it converged to lowest possible error of 0.07 in Jalpaiguri district also. For both Malda and Uttar Dinajpur districts, the optimum learning rate of Tanh activation function were at low level of 0.04 and it converged to lowest possible error. The learning rate of other activation functions were high level of 0.10 and converged to high SSE. Due to low learning rate, the Tangent Hyperbolic (Tanh) activation function took many iterations to converge to the global minimum. But it converged to lowest possible error. Similar learning rate were also obtained by [33, 34].

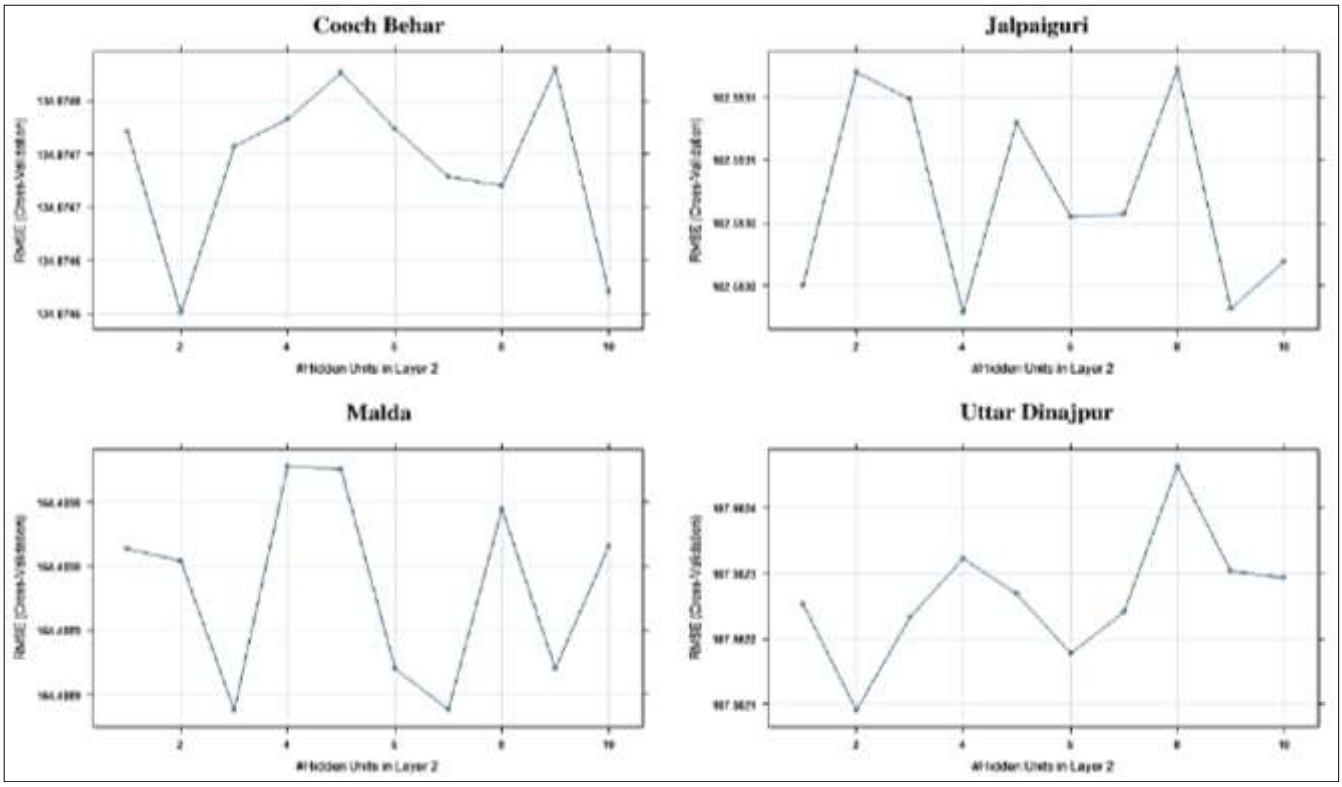

Fig. 1 Cross-validation plot for optimum number of hidden layer neurons

Table 2 Summary of Artificial Neural Network (ANN) model

District	Number of Hidden layer Neurons	Activation Function	Learning Rate	SSE	Steps	AIC	BIC
Cooch Behar	2	Logistic	0.08	0.14	768	22.28	31.44
		Tangent Hyperbolic (Tanh)	0.07	0.12	658	22.25	31.41
		ReLU	0.10	0.68	29	23.35	32.52
		Softmax	0.10	0.68	727	23.37	32.53
Jalpaiguri	4	Logistic	0.10	0.53	33	51.05	71.88
		Tangent Hyperbolic (Tanh)	0.03	0.07	683	50.15	70.98
		ReLU	0.10	0.52	29	51.04	71.87
		Softmax	0.10	0.12	936	50.25	71.07
Malda	3	Logistic	0.09	0.39	19	32.77	46.10
		Tangent Hyperbolic (Tanh)	0.04	0.06	1308	32.13	45.46
		ReLU	0.10	0.39	32	32.78	46.11
		Softmax	0.10	0.39	35	32.78	46.11
Uttar Dinajpur	2	Logistic	0.08	0.15	569	22.32	31.48
		Tangent Hyperbolic (Tanh)	0.04	0.06	2400	22.11	31.28
		ReLU	0.10	0.62	25	23.23	32.39
		Softmax	0.05	0.14	1543	22.29	31.45

The AIC and BIC values were also comparatively low in case of Tanh activation function for all districts. Hence, it can be concluded that MLP architecture of ANN model using Backpropagation learning algorithm that had Tangent Hyperbolic (Tanh) function as activation function using weather indices as input performs better for predicting Rapeseed-mustard yield. The main advantage of Tanh activation function is its output were zero-centered [35].

Support Vector Regression (SVR) model

SVR models were fitted to each district to predict Rapeseed-Mustard yield using the respective weather indices selected from Stepwise Regression as input. The best combination of these hyperparameters were selected for each kernels using grid search algorithm.

The performance of various kernels and their hyperparameter of SVR were given in the table 3. The optimum cost (C) was lowest level of one and margin of the hyperplane (ϵ) was also at lowest level of 0.01 in linear kernel for all districts. The degree of polynomial (d) of polynomial kernel was one for all districts except Uttar Dinajpur which is similar to linear kernel. The value of d for Uttar Dinajpur was four. But the cost in polynomial kernel was increased to 4 in all districts except Uttar Dinajpur due to which there was a decline in number of support vectors. Further gamma was also high for polynomial kernel. In RBF kernel, the optimum cost (C) was four for all districts except for Jalpaiguri it was 32. There was a considerable reduction in number of support vectors in RBF kernel as the hyperplane margin (ϵ) was neither too high nor too low. For Uttar Dinajpur, the margin of the hyperplane (ϵ) was decreased to zero that leads to many support vectors in both nonlinear kernels. As the gamma was low in RBF, there was low curvature in RBF function than polynomial function. Due to appropriate choices of hyperparameters, the nonlinear RBF kernel based SVR model was performed better than linear and polynomial kernels with lowest RMSE for all districts. Similar results were obtained by [36, 37].

Table 3 Summary of Support Vector Regression (SVR) model

District	Kernel Function	Cost (C)	Epsilon (ϵ)	Gamma (Y)	Degree (d)	Number of Support Vectors	RMSE
Cooch Behar	Linear	1	0.10	NA	NA	15	72.69
	Polynomial	4	0.50	0.70	1	9	67.82
	RBF	4	0.30	0.10	NA	7	59.56
Jalpaiguri	Linear	1	0.10	NA	NA	17	66.18
	Polynomial	4	0.80	0.30	1	15	76.16
	RBF	32	0.30	0.10	NA	3	39.73
Malda	Linear	1	0.10	NA	NA	16	125.36
	Polynomial	16	0.80	0.20	1	10	108.74
	RBF	4	0.40	0.10	NA	4	85.93
Uttar Dinajpur	Linear	1	0.10	NA	NA	17	119.1
	Polynomial	4	0.00	0.10	1	17	119.38
	RBF	4	0.00	0.10	NA	17	80.06

Evaluating performance of fitted models

The R^2, Adjusted R^2, MAE and RMSE were used to examine goodness of fit of the models. The performance of models for testing data were validated using MAE and RMSE. The model evaluation parameters were given in the table 4.

Table 4 Evaluation parameters of fitted models

District	Model	During calibration				During validation	
		R^2	Adj. R^2	MAE	RMSE	MAE	RMSE
Cooch Behar	MLR	0.77	0.72	57.14	67.24	63.62	80.64
	ANN	0.82	0.78	50.32	59.56	72.99	77.46
	SVR	0.82	0.78	51.91	60.17	66.82	89.52
Jalpaiguri	MLR	0.77	0.66	58.22	65.25	74.88	89.26
	ANN	0.86	0.85	41.97	50.93	68.69	94.90
	SVR	0.92	0.89	37.78	39.73	67.37	79.49
Malda	MLR	0.77	0.68	84.17	95.32	121.39	138.33
	ANN	0.83	0.79	69.24	80.15	63.52	85.93
	SVR	0.81	0.77	77.01	85.93	79.11	106.15
Uttar Dinajpur	MLR	0.71	0.61	88.16	116.00	124.42	168.62
	ANN	0.91	0.89	55.99	66.16	181.07	245.24
	SVR	0.87	0.84	43.16	80.06	134.58	159.90

For Cooch Behar district, the R^2 and Adjusted R^2 of both ANN and SVR were 0.82 and 0.78 respectively. But MAE and RMSE were comparatively low in ANN in both calibration and validation stage of the model. Hence the ANN model was the best fitted model for Cooch Behar. The best fitted ANN model was graphically given in the Fig 2.

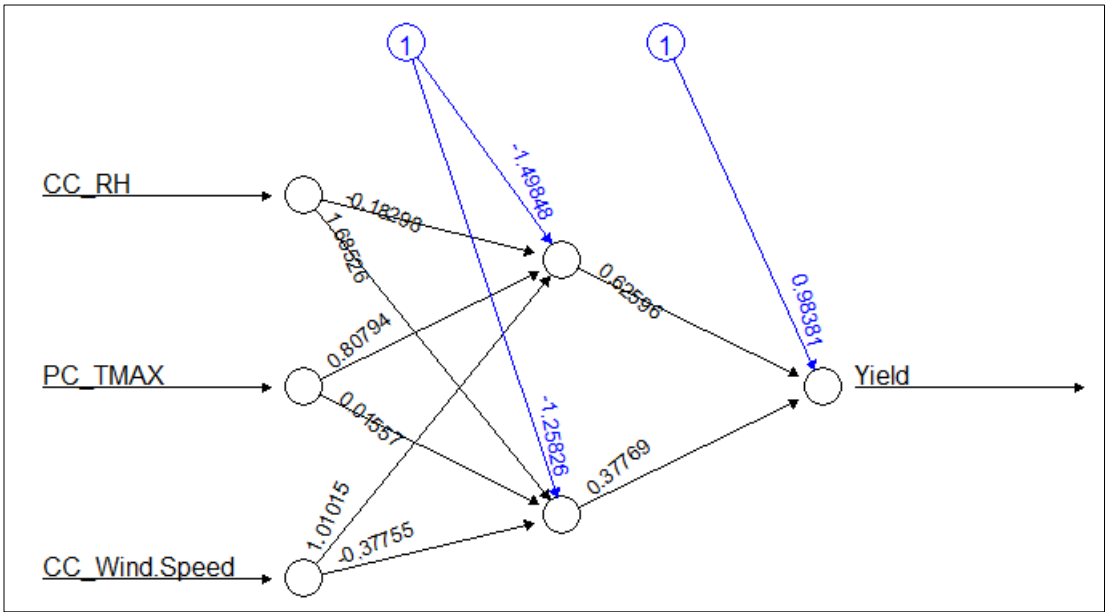

Fig. 2 The best fitted ANN model for Cooch Behar district.

For Jalpaiguri, the R^2 and Adjusted R^2 were high for SVR model which were 0.92 and 0.89 respectively. The MAE and RMSE were comparatively low in SVR in both calibration and validation stages. Hence, the SVR model was the best fitted model for Jalpaiguri. For Malda, the R^2 and Adjusted R^2 were high for ANN model which were 0.83 and 0.79 respectively. The MAE and RMSE were comparatively low in ANN in both calibration and validation stages. Hence, the ANN model was the best fitted model for Malda. The best fitted ANN model was graphically in the Fig 3.

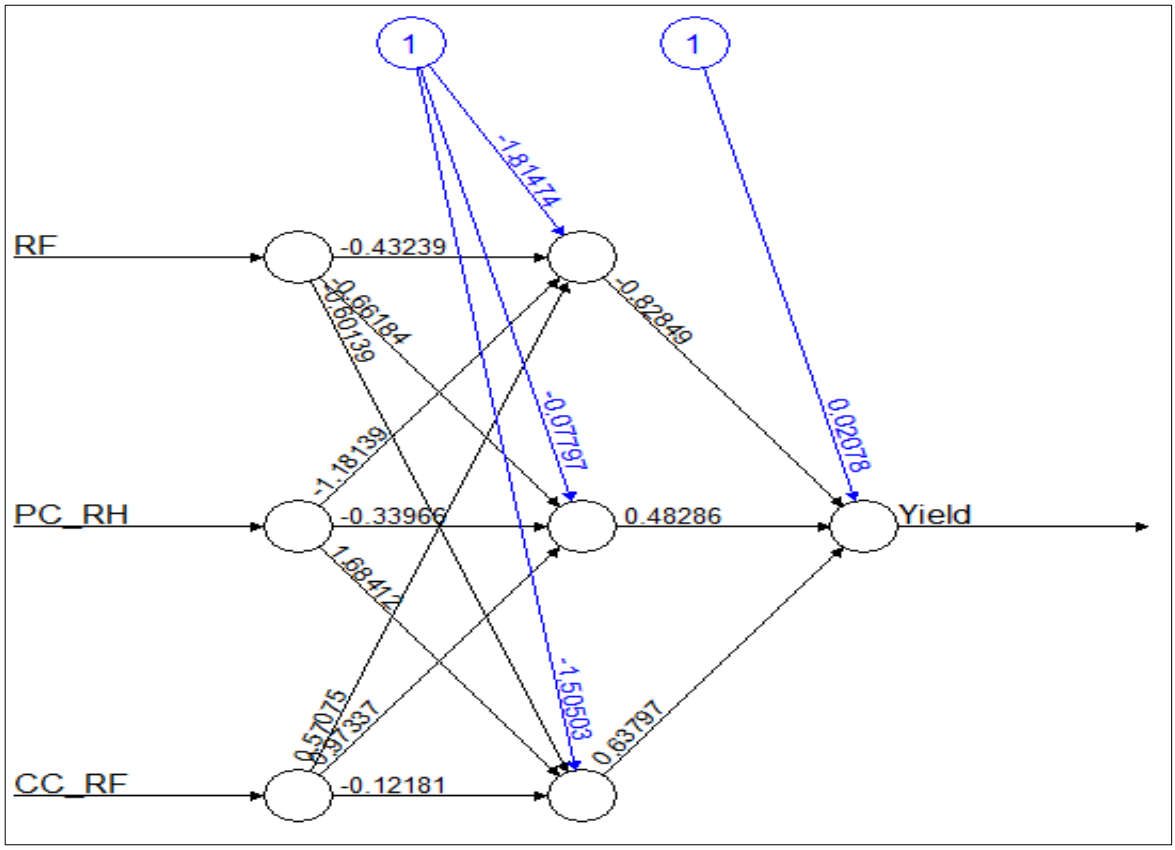

Fig 3: The best fitted ANN model for Malda district.

For Uttar Dinajpur district the R^2 and Adjusted R^2 were high for ANN model but the MAE and RMSE were comparatively high due to the possibility of over fitting the model. But the SVR model that had R^2 and Adjusted R^2 of 0.87 and 0.84 were having comparatively low MAE and RMSE in both calibration and validation of the models. Hence, SVR model was the best fitted model for Uttar Dinajpur district.

The role of variable selection is important in fitting a model as it removed the variables that were redundant and not significantly influenced the yield [38]. The variable selection reduces the issue of overfitting as well as makes algorithm to work fast [39]. Using the stepwise regression, only three or four variables that affect the yield significantly were selected from fifteen independent indices. The performance of nonlinear machine learning models was better than the linear regression model to predict crop yield [40, 41].

4. Conclusion

The district-wise yield prediction is necessary for location specific decision making. The weather indices-based prediction models were effective for location specific models. Variable selection for model fitting is important to keep only significant input variables in the model and it avoid unnecessary complexity in the model. Only three or four indices that significantly affect the yield were selected using stepwise regression. The optimum number of neurons in the hidden layer for ANN were ranging between two to four. The Tangent hyperbolic function was found to be the suitable hidden layer activation function for Multilayer Perceptron using back propagation algorithm. The nonlinear Radial Basis Function kernel was the best kernel for Support Vector Regression. The ANN model was the best fitted model for Cooch Behar and Malda and SVR was the best fitted model for Jalpaiguri and Uttar Dinajpur districts. In overall, the machine learning models performed better than the multiple linear regression model for district-wise yield prediction of rapeseed-mustard crop.

5. Conflict of Interests

The author(s) declare(s) that there is no conflict of interest related to this article.

6. References

1. Kumar V, Tiwari A. Sparking yellow revolution in India again. Rural Pulse. 2020;34:1-4.
2. Banerjee H, Chatterjee S, Sarkar S, Gantait S, Samanta S. Evaluation of rapeseed-mustard cultivars under late sown condition in coastal ecosystem of West Bengal. Journal of Applied and Natural Science. 2017;9(2):940-949.
3. Shekhawat K, Rathore SS, Premi OP, Kandpal BK, Chauhan JS. Advances in agronomic management of Indian mustard (*Brassica juncea* (L.) Czernj. Cosson): An overview. International Journal of Agronomy. 2012;2012:1-14.
4. Jain G, Sandhu SK. Agroclimatic indices and yield of mustard under different thermal regimes. Journal of Agricultural Physics. 2018;18(2):232-239.
5. Punia R, Kumari P, Kumar A, Rathi AS, Avtar R. Impact of weather parameters on alternaria blight of Indian mustard [(*Brassica juncea* (L.) Czern. & Coss.)]. Bangladesh Journal of Botany. 2021;50(1):15-19.
6. Paliwal HB, Sharma R, Samota SK. Effect of Different Sowing Dates and Varieties on Mustard Growth and Yield in Prayagraj Conditions. International Journal of Environment and Climate Change. 2022;12(11):2316-2322.
7. Fisher RA. Studies in crop variation III-The influence of rainfall on the yield of wheat at Rothamsted. Philosophical Transactions of the Royal Society of London. 1924;213:89-142.

8. Kaur B, Gill KK Development of Weather based Weekly Thumb Rules for Potential Productivity of Mustard Crop in Punjab Vayu Mandal. 2017;43(1):72-81.
9. Jain G, Sandhu SK. Agroclimatic indices and yield of mustard under different thermal regimes. Journal of Agricultural Physics. 2018;18(2):232-239.
10. Pandey KK, Maurya D, Gupta G, Mishra SV. Yield forecasting models based on weather parameters for eastern UP. Vegetos. 2016;29(1):22-24.
11. Sagar BM, Cauvery NK. Agriculture data analytics in crop yield estimation: A critical review. Indonesian Journal of Electrical Engineering and Computer Science. 2018;12(3):1087-1093.
12. Kang Y, Ozdogan M, Zhu X, Ye Z, Hain C, Anderson M. Comparative assessment of environmental variables and machine learning algorithms for maize yield prediction in the US Midwest. Environmental Research Letters. 2020;15(6):064005.
13. Aubakirova G, Ivel V, Gerassimova Y, Moldakhmetov S, Petrov P. Application of artificial neural network for wheat yield forecasting. Eastern-European Journal of Enterprise Technologies. 2022;3(4):31-39.
14. Paidipati KK, Chesneau C, Nayana BM, Kumar KR, Polisetty K, Kurangi C. Prediction of rice cultivation in India—support vector regression approach with various kernels for non-linear patterns. Agricultural Engineering. 2021;3(2):182-198.
15. Liu B, Zhao Q, Jin Y, Shen J, Li C. Application of combined model of stepwise regression analysis and artificial neural network in data calibration of miniature air quality detector. Scientific Reports. 2021;11(1):3247.
16. Lewis-Beck MS. Stepwise regression: A caution. Political Methodology. 1978;5(2):213-240.
17. Johnsson T. A procedure for stepwise regression analysis. Statistical Papers. 1992;33(1):21-29.
18. Aljarrah M, Al-Jarrah Y, Galvão RK, Araújo MC, Fragoso WD, Silva EC, *et al*. Using stepwise regression to investigate c ustomers' propensity to change cellular phone providers. Global Journal of Pure and Applied Mathematics. 2017;13(9):5013-5020.
19. Serrone DG, Moretti L. A stepwise regression to identify relevant variables affecting the environmental impacts of clinker production. Journal of Cleaner Production. 2023;398:136564.
20. Huberty CJ. Problems with stepwise methods-better alternatives. Advances in Social Science Methodology. 1989;1:43-70.
21. Ghani IM, Ahmad S. Stepwise multiple regression method to forecast fish landing. Procedia-Social and Behavioral Sciences. 2010;8:549-554.
22. Uyanık GK, Güler N. A study on multiple linear regression analysis. Procedia-Social and Behavioral Sciences. 2013;106:234-240.
23. Lukman AF, Ayinde K, Siok Kun S, Adewuyi ET. A modified new two-parameter estimator in a linear regression model. Modelling and Simulation in Engineering; c2019.
24. Kittichotsatsawat Y, Tippayawong N, Tippayawong KY. Prediction of arabica coffee production using artificial neural network and multiple linear regression techniques. Scientific Reports. 2022;12(1):14488.
25. Fernandes PO, Teixeira JP. Applying the artificial neural network methodology for forecasting the tourism time series. 2008;1:26-38.
26. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. nature. 1986;323(6088):533-536.
27. Srinivasan K, Cherukuri AK, Vincent DR, Garg A, Chen BY. An efficient implementation of artificial neural networks with K-fold cross-validation for process optimization. Journal of Internet Technology. 2019;20(4):1213-1225.
28. Nti IK, Nyarko-Boateng O, Adekoya FA, Weyori BA. An empirical assessment of different kernel functions on the performance of support vector machines. Bulletin of Electrical Engineering and Informatics. 2021;10(6):3403-11.
29. Al Azies H, Trishnanti D, PH EM. Comparison of kernel support vector machine (SVM) in classification of human development index (HDI). IPTEK Journal of Proceedings Series. 2019;30(6);53-57.
30. Pence I, Kumaş K, Siseci MC, Akyüz A. Modeling of energy and emissions from animal manure using machine learning methods: the case of the Western Mediterranean Region, Turkey. Environmental Science and Pollution Research. 2023;30(9):22631-52.
31. Agrawal R, Jain RC, Jha MP. Models for studying rice crop-weather relationship. Mausam. 1986;37(1):67-70.
32. Huzsvai L, Zsembeli J, Kovács E, Juhász C. Response of winter wheat (*Triticum aestivum* L.) yield to the increasing weather fluctuations in a continental region of four-season climate. Agronomy. 2022;12(2):1-19.
33. Nawi NM, Ghazali R, Salleh MN. The development of improved back-propagation neural networks algorithm for predicting patients with heart disease. In Information Computing and Applications: First International Conference, ICICA 2010, Tangshan, China, October 15-18, 2010. Proceedings 1. Springer Berlin Heidelberg; c2010 p. 317-324.
34. Igiri CP, Anyama OU, Silas AI. Effect of learning rate on artificial neural network in machine learning. International Journal of Engineering Research & Technology. 2015;4:395-363.
35. Nwankpa C, Ijomah W, Gachagan A, Marshall S. Activation functions: Comparison of trends in practice and research for deep learning. arXiv preprint arXiv:1811.03378; c2018.
36. Truong VH, Pham HA. Support vector machine for regression of ultimate strength of trusses: A comparative study. Engineering Journal. 2021;25(7):157-166.
37. Shafiee S, Lied LM, Burud I, Dieseth JA, Alsheikh M, Lillemo M. Sequential forward selection and support vector regression in comparison to LASSO regression for spring wheat yield prediction based on UAV imagery. Computers and Electronics in Agriculture. 2021;183:106036.
38. Shahhosseini M, Hu G, Archontoulis SV. Forecasting corn yield with machine learning ensembles. Frontiers in Plant Science. 2020;11:1-16.
39. Kumar S, Attri SD, Singh KK. Comparison of Lasso and stepwise regression technique for wheat yield prediction. Journal of Agrometeorology. 2019;21(2):188-192.
40. Zhou W, Liu Y, Ata-Ul-Karim ST, Ge Q, Li X, Xiao J. Integrating climate and satellite remote sensing data for predicting county-level wheat yield in China using machine learning methods. International Journal of Applied Earth Observation and Geoinformation. 2022;111:102861.

41. Sridhara S, Manoj KN, Gopakkali P, Kashyap GR, Das B, Singh KK, *et al*. Evaluation of machine learning approaches for prediction of pigeon pea yield based on weather parameters in India. International Journal of Biometeorology. 2023;67(1):165-180.