An Intelligent Framework for Rice Yield Prediction using Machine Learning based Models

N. Gnanasankaran, E. Ramaraj, T. Manikumar

Abstract - Rice is considered as a major crop productivity in India. It is an essential staple food for the significant proportion of the population in the country. Rice productivity may be predicted, by the use of climate (seasonal) forecast to a crop yield model, at the start of the season or even long earlier the season begins. Several researchers have developed a set of automated rice plant productivity prediction models. This paper devises an intelligent machine learning (ML) based regression models for the prediction of rice plant productivity. The proposed model involves different ML models namely K-star, logistic regression (LR), bagging, multilayer perceptron (MLP), radial basis function (RBF) network, additive regression and gaussian process. The performance of the proposed model has been validated using a dataset collected from Tamil Nadu. The proposed model has been implemented using WEKA tool. The experimental results indicated that the K-star model has offered maximum outcome over the compared methods with the maximum correlation coefficient of 0.954, minimum mean absolute error (MAE) of 223.43, root mean square error (RMSE) of 365.22, Relative Absolute Error(RAE) of 26.65 and Root Relative Squared Error (RRSE) of 32.55.

Index Terms - Machine learning, Regression model, Rice plant, Crop productivity, Predictive model

1 Introduction

Nowadays, India is one of the major rice producers globally. Agricultural productivity depends on the core segments such as soil, attributes, seasonal changes and climatic changes says Wu Fan et al. [1]. Kathkar et al. [2] compares India with other countries and says, India holds second position in the Rice production and consumer account for 22.3% of global production. Rice is cultivated throughout the country as well as it gives more than 40% in entire food grains productivity. India is the leading originator of the Basmati Rice to the global markets. India generates around 4.25 million metric loads of basmati rice which is roughly 70% of the total worldwide production. Basmati Rice is an important commodity which is massively exported to different countries such as Iran and Saudi Arabia. At the same time, the top basmati rice importers are America and England.

At present, major suppliers of Rice in India and its exporters are largely planted in this area namely Tamil Nadu, Andhra Pradesh, West Bengal, Punjab, Chhattisgarh, Uttar Pradesh, Bihar, and Orissa, as shown in Fig. 1. These are the massive in Rice production reports to 72% and hold more than 75% of shares in entire Rice production in India. The harvesting of crop is linked to a variety of collected data and its relationship with harvest. These collected data are applied to the ML algorithms to instruct them. This helps the farmer to assume the profit of the crop. With this knowledge, they have a clear view in the quantity of crop which they can invest and also do modifications before harvesting often locking an additional cost than staying till the reap.

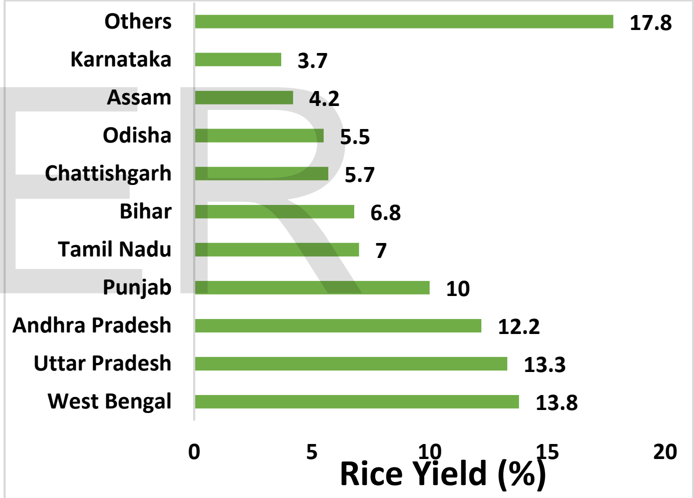


Fig. 1 Statistics of Rice Yield in India

A good tendency of analyzing and forecasting the production of essential crop area make more impact on many bank accounts. The key analysis of creation and production is a requirement for correct data based on the ecology and suitable research works for producing utmost profit. Analysing the tendency of the crops such as garlic and papaya are the challenging ones. An unpredicted decline in production trims down extra profit and revenue for the farmers and direct to hike in price. Likewise, boosting the production could reflect sharp price reduction and has unfavourable cause on farmer's incomes. Bouncing the cost of a necessary product plays important function in deciding the cost of salary, wages, inflation, and other strategies in an economy. The appropriate prediction will give technique for good profit and debit management in balancing the cost and assure income for the farmers.

Some approaches such as simulation designing and remote sensing are mainly employed to predict the harvest rate and land selection. Although, estimation is essential in early at once the crop yielding is done or the crop sown. It can be realized only through crafting the data history and obtaining the predictions. Data mining approach executes a method to calculate the yield with old records. This was attained by relating association rule mining on agriculture records. Data was gathered by K-means clustering algorithm and analysed the short calculation of harvest of the crop was experimented.

Dahikar et al. [3] measured a range of condition of climatological phenomena disturbing neighbouring climate environment in different side of the world. It has sturdy cause on yielding the crop. Benefits of Artificial Neural Networks (ANN) have been demonstrated as great tool in predicting and modelling, to enhance its effectiveness. The methodology of calculating the usage of Crop was done to sense the appropriate crop in sensing different factor of soil and factors allied to atmosphere. It has employed different models like Multi Linear Regression, ANN, Fuzzy Logic and Adaptive NeuroFuzzy Inference System and examined to identify the finest technique in predicting the crop yield.

Gnanasankaran and Ramaraj [4] predicted that Multiple Liner Regression Model is best suited to predict Rainfall Data using Indian Meteorological data using WEKA tool.

Several techniques for predicting the yield of crop are balanced by its factors namely Mean Square Error (MSE), RMSE, correlation coefficient and *R*2 to verify Adaptive Neuro Fuzzy Inference System (ANFIS) prediction technique is greater than other technique. Kumar et al. [5] carried out graphic analytics on sugarcane crop database. Supervised ML algorithm related to discover the real estimated price and LS-SVM algorithm shows its efficiency and mean squared error at cross-validation phase.

Yimit et al. [6] make use of geostatistical technique for revising historical and geographical deviations of ground-water salinity in the Ili River Irrigation region in the West China. It initiates that the ground-water height and salinity with spatial dependence, spherical and exponential models which is mean to be optimal semivariogram technique. In Mazandaran Province, Iran, the interpolation method in ArcGIS were observed for mapping the sulphate concentration (SO4 2−), Cl− ,EC, SAR, TDS, pH, TH and TH. Datasets were gathered in the spring of 2004 by illustrating from twenty-three wells and the region with minimum water feature are summarized once the plotting operation has been completed.

Dash et al. [7] employed the OK and IK approach in plotting and acquiring possibility charts, and they concluded that sixty-nine percentage of the revise that the region had salinities and 24% of the region had the maximum possibility of ground-water salinity to beat yield in controlling maximum thresholds. Seyed mohammadi et al. [8] calculated various interpolation processes to notice the accuracy for examining the spatial deviation of ground-water EC quantity in central region of Guilan Province. It states that, OK approach is an effective estimation technique and the Gaussian representation was considered to be an optimal empirical semivariogram of variable dataset in OK. The complicated region in Guilan Province is that, the limited harvest which happens because of salinity in Sefidrud River basin.

Ahmadpour et al. [9] studied the connection among ground-water EC and distance from the Caspian Sea in Guilan Province and the Sefidrud River. They finalized that there is a significant negative relationship from groundwater EC and the distance from the Sefidrud River. It is clear that, the salinity circulation would be in a triangle shape whose bottom is the opening of sink and tangent to Caspian Sea. Chandrasekharan et al. [10] studied the uptrend and downtrend of ground-water's EC in the seaside band in dehydrated and rainy months and declared in which uptrend of EC in the station cause because of seawater movements into the water level of the seaside band. Rezaei et al. [11] assured that sea-water intrusion on the shores which receives minimum rainfall throughout the year enhanced by maximizing the ground-water consumption. They described that in the case of constant excessive usage of ground-water, this can have a pessimistic result on the crop yield in the Guilan area. In all research study, precipitation plays main aspect in ground-water quality; it results updating constantly.

Ashrafzadeh et al. [12] considered that the sufficient ground-water quality for Rice irrigation in the paddy field of Guilan region by way of 2 geostatistical method, OK and ordinary cokriging (OCK). The spatial plotting of the sum of main cations and anions (SCA) and EC during 2010–2014 classify into 4 qualities: unsuitable, risky, excellent, and good. They have done that ground-water in the areas of Eastern parts was risky whereas the area in the Western parts have excellent ground-water quality for Rice irrigation. Finally, this research study projected that farmers utilize mutually ground-water and surface water at the same time. On the other hand, the research study only measured the current situation with respect to ground-water salinity and non-tendency of ground-water level modifies over upcoming years.

Though several models have been available in the literature, there is still a need to develop rice crop prediction model. This paper introduces an intelligent ML based regression models to predict the productivity of the rice plants. The proposed model involves different ML models namely K-star, logistic regression (LR), bagging, multilayer perceptron (MLP), radial basis function (RBF) network, additive regression and gaussian process. The performance of the proposed model has been validated using a dataset collected from Tamil Nadu.

2 The Proposed Rice Crop Productivity Prediction Model

The complete workflow of the projected method is shown in Fig. 2. As shown, the input data undergo preprocessing to remove the unwanted data exist in the dataset. Then, the preprocessed data will be applied to the predictive models and the rice crop productivity can be forecasted. Finally, the performance of these ML models can be analyzed interms of several aspects.

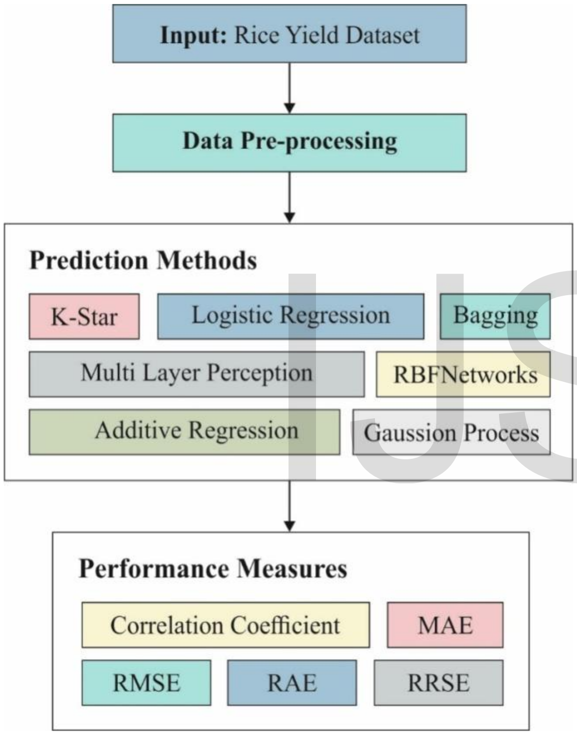


Fig. 2 Block diagram of proposed model

2.1 Preprocessing

The input rice yield dataset undergoes preprocessing to remove the redundant data and enhance the classification outcome. In this case, the state name Tamil Nadu is found to be common in all the instances, therefore, the state name has been removed in the preprocessing step. Then, the preprocessing data is sent to the classification model for further processing.

2.2 K-Star Model

K-star or K* is defined as an instance-oriented classification method. The class of a sample instance depends upon the training instance which is evaluated by similarity function. It is varied from alternate instance based learners where entropy-based distance function is applied. Instance-based learners divide the instance by relating a database of pre-classified instances. The basic consideration of this classifier is that, same instances would have identical classification. The “similar instance” and “similar classification” are related units of an instance-based learner as distance function calculates the similarity of these instances, as well as classification function shows the resultant accuracy. As per Ashrafzadeh et al. [12] the K-star method applies entropic measure, which depends upon probability of converting an instance into another feasible transformation. Under the help of entropy, a meter of instance distance is advantageous and data theory guides in processing the distance of instances. The complications of a transformation of single instance into actual distance from several instances. It is accomplished in 2 phases. Initially, a finite set of conversions are mapped from one instance to another. Secondly, transform instance using the program in a definite sequence of conversions. The provided collection of infinite points as well as pre determined transformations T, assume t be a value of set T. The t maps t: I → I. These instances were mapped with o in T(σ(a) = a), where σ terminates P and the collection of prefix codes from T^*. Participants of T^* and of P describes a conversion on I. Here, P means a probability function on T^*. It meets the following features:

$$0 \leq \frac{p(\bar{t}u)}{p(\bar{t})} \leq 1 \qquad (1)$$

$$\sum_{u} p(\bar{t}u) = p(\bar{t}) \qquad (2)$$

$$p(\Lambda) = 1 \qquad (3)$$

Finally, the following function is satisfied:

$$\sum_{t\,\in p} p(\bar{t}) = 1 \qquad (4)$$

The probability function P^* is described as probability of paths from instance a, to instance b:

$$p * \left(\frac{b}{a}\right) = \sum_{t\in p:t(a)=b} P(t) \qquad (5)$$

It is evident that P *meets the given characteristics.

$$\sum_{b} P * \left(\frac{b}{a}\right) = 1\,, 0 \leq P * \left(\frac{b}{a}\right) \leq 1 \qquad (6)$$

The K^* function is represented by:

$$K * \left(\frac{b}{a}\right) = -\log_2 P * \left(\frac{b}{a}\right) \qquad (7)$$

2.3 Logistic regression (LR)

The variations among logistic and linear regression models were resembled in parametric methods. When these differences are considered, the techniques applied in examining task with the help of LR method which employs the similar strategy in linear regression. According to Mahmood et al. [13], the approaches utilized in linear regression analysis are also employed for LR investigation. Assume Y is the response variable with the values of 0 or 1. The required response is the possibility of variable with value 1, say (x).The logistic response function is expressed as

$$\mathrm{E(Y|x)} = \pi(\mathrm{x}) = \exp\frac{(\beta_0 + \beta_1 \mathrm{x})}{\{1 + \exp(\beta_0 + \beta_1 \mathrm{x})\}} \qquad (8)$$

The logistic conversion of $\pi(\mathrm{x})$ is provided by

$$\eta = \beta_0 + \beta_1 \mathrm{x} = \ln[\pi(\mathrm{x})/\{1 - \pi(\mathrm{x})\}] \qquad (9)$$

It is facilitated as link function among probability and linear expression on RHS. The proportion $\pi(\mathrm{x})/\{1 - \pi(\mathrm{x})\}$ is named as odds. Higher likelihood model is applied for identifying estimates of β_0 and β_1, as provided by $\hat{\beta}_0$ and $\hat{\beta}_1$, correspondingly. Once the coefficients are determined, the related hypothesis testing were processed either the model is sufficient or the autonomous variables are connected to resultant variable. Odds ratio is expressed by

$$\mathrm{Odds}_{\mathrm{xi+1}}/\mathrm{Odds}_{\mathrm{xi=}} w^{\beta_1} \qquad (10)$$

The odds ratio is interpreted as an estimated improvement in odds of one-unit modification in the measure of predicted estimates.

2.4 Bagging Model

A learning set $\mathcal{L}$ is comprised with data $\{(y_{\mathrm{n}}, x_n), n = 1 \dots., N\}$ where y's indicate class labels or mathematical response. Consider the steps of applying learning set for developing a predictor $\varphi(x, \mathcal{L})$ — when input is x and it predicts y by $\varphi(x, \mathcal{L})$. Then, sequence of learning sets is provided by$\{\mathcal{L}_k\}$with Nautonomous observations from similar distribution as $\mathcal{L}$. The main aim of this model is to exploit the $\mathcal{L}$ which results in best prediction when compared with individual learning set predictor $(x, \mathcal{L})$. The limitation is that, it is enabled to be operated with sequence of predictors $\{\varphi(x, \mathcal{L}_k)\}$.When y is arithmetical, the better step is to exchange $\varphi(x, \mathcal{L})$ by maximum $\varphi(x, \mathcal{L}_k)$. By k, in which $\varphi_A(x) = E_{\mathcal{L}}\varphi(x, \mathcal{L})$ where $E_{\mathcal{L}}$is the requirement across $\mathcal{L}$, and subscript A in $\varphi_A$implies an aggregation. While $\varphi(x, \mathcal{L})$detects a class $j \in$ $\{1 \dots..J\}$, then aggregating the $\varphi(x, \mathcal{L}_k)$is processed by voting system. Assume $N_j = nr\{k; \varphi(x. \mathcal{L}_k) = j\}$ and consider $\varphi A(x) = argmax_j N_j$, where, the j for $N_j$is higher.In general, the single learning set $\mathcal{L}$with no luxury of replicates of $\mathcal{L}$ are replaced.Yet, the resemblance of the task which leads to $\varphi_A$can be accomplished. The bootstrap samples$\{\mathcal{L}^{(B)}\}$ from $\mathcal{L}$, and results in $\{\varphi(x, \mathcal{L}^{(B)})\}$.When y is numerical, considerφ_B as

$$\varphi_{\mathrm{B}}(\mathrm{x}) = \mathrm{a}v_B \varphi(x, \mathcal{L}^{(B)}). \qquad (11)$$

If y is a class label, and the $\{\varphi(x, \mathcal{L}^{(B)})\}$ is a vote and develops $\varphi_{\mathrm{B}}(\mathrm{x})$. This is named as bootstrap aggregating and applies the acronym bagging.The $\{\mathcal{L}^{(B)}\}$ norm refers data sets, with N cases, obtained randomly; however, by replacement, from $\mathcal{L}$. All (y_n, x_n) functions are displayed many times not specifically as $\mathcal{L}^{(B)}$. The $\{\mathcal{L}^{(B)}\}$ is an replicated data set acquired from bootstrap distribution approximation using $\mathcal{L}$. For background on bootstrapping, a significant factor in bagging enhances the accuracy with a stable procedure for developing φ. When any changes occur in $\mathcal{L}$, then it generates minimum alterations in φ, and $\varphi_B$would be closed to φ. Followed by, enhancing task has been carried out for unreliable procedures in which small change in £ leads to maximum changes in φ. Instability is examined from Ananthakumar et al. [14] that is pointed from neural nets, classification and regression trees, and subset selection in linear regression are unstable, whereas k-nearest neighbor (kNN) model is stable.

2.5 Multilayer perceptron (MLP)

MLP is defined as a category of feedforward ANN with least number of 3 layers. In this model, n, h and m shows the count of input, hidden and output nodes correspondingly. Here, weights of an MLP are saved in the matrix W and these biases were saved in matrix B. The MLP resultant evaluation is provided in the following equation. Initially, the weighted sums of inputs are determined by

$$s_j = \sum_{i=1}^{n} \left(W\left((j-1)n + i\right) * I_i\right) + B(j), j = 1, 2, \cdots, h \qquad (12)$$

$W\left((j-1)n + i\right)$ defines weight from ith input node to jth hidden node, $B(j)$ means bias of jth hidden node, and $I_i$shows ith input. The weights and biases are projected in the form of matrix where it defines the storage of matrix.Followed by, result for all hidden nodes are provided by

$$S_j = sigmoid\,(s_j) = \frac{1}{\left(1 + \exp\left(-s_j\right)\right)}, j = 1,2, \cdots, h \qquad (13)$$

Finally, the simulation outcome is estimated by,

$$o_k = \sum_{j=1}^{h}\left(W(nh+(k-1)h+j)*S_j\right)+B(h+k), k = 1,2,\dots,m \qquad (14)$$

$$O_k = sigmoid(o_k) = \frac{1}{(1+\exp(-o_k))}, k = 1,2,\cdots,m \qquad (15)$$

where $W(nh+(k-1)h+j)$refers weight from jth node in hidden layer to kth node in output layer, and $B(h+k)$ depicts the bias of kth output node. The major task of MLP training is to acquire the best combination of weight and bias that results in better simulation outcome for the provided input.

In problem formulation, every weight and biases are combined to develop a candidate in population where weights and biases are assumed to be the attribute of a candidate solution.

$$Candidate\ Solution = \{W(1),\dots,W(nh+mh),B(1),\dots\dots,B(h+m)\} \qquad (16)$$

Overall weights ought to be optimized is $nh+mh$ while entire biases to be optimized is $h+m$. Hence, parameters should be optimized for MLP training as $(n+1)h+(h+1)m$. Basically, the infrastructure of MLP is indicated as $n-h-m$. The inputs are referred as features. Thus, numbers of input layer nodes in the classification problem as mentioned by Breiman et al. [15]. Since the standard rules are lagging for the selection of hidden nodes, the rule is applied

$$H = 2XN + 1 \qquad (17)$$

where N is the number of input nodes, and H is the number of hidden nodes. It is apparent that overall weights and biases ought to be optimized which must be directly proportional to feature values. As the complications are enhanced periodically, the MLP is accelerated, and make the searching process a complex task.

2.6 Radial basis function (RBF) Network

Basically, ANN applies RBF an activation functions which has been expressed mathematically. The result attained from this method is a type of linear integration of inputs and neuron parameters. This RBF is applied for many applications such as process approximation, time series analysis, classification, and system management. RBF is composed of 3 layers namely, input layer, hidden layer along with a non-linear RBF activation function as well as linear output layer. Here, input is assumed to be vector of real numbers $x \in \mathbb{R}^n$ while result of the system is scalar function of input vector,$\varphi: \mathbb{R}^n \to \mathbb{R}$, which is expressed by,

$$\varphi(x) = \sum_{i=1}^{N} a_i\,\rho(||x-c_i||) \qquad (18)$$

where Ndenotes the count of neurons in hidden layer, $c_i$implies the middle vector for neuron i, and $a_i$are weights of neuron $i$from linear output neuron. The process which is based on the distance from a middle vector is symmetric as quoted by Battacharjee et al. [16]. Usually, inputs are linked with hidden neuron. It is considered to be Euclidean distance while RBF is assumed to be Gaussian function

$$\rho(\|x-c_i\|) = \exp[-\beta\|x-c_i\|^2] \qquad (19)$$

The Gaussian basis functions were local to middle vector from,

$$\lim_{\|x\|\to\infty} \rho(\| x - c_i\|) = 0 \qquad (20)$$

where adjustable parameters of single neuron have minimum impact for input values which has been away from middle neuron.Based on the specific conditions regarding the structure of activation function, RBF networks are global approximators on compact subset of $\mathbb{R}^n$. It refers that, an RBF network which has sufficient hidden neurons approximates the constraints on closed, bounded set with random precision.The parameters a_i, c_i and $\beta_i$were computed by optimizing the fitness among φ.

2.7 Additive regression Model

Additive Model (AM) is defined as non-parametric regression approach. It was defined is a basic portion of ACE scheme. The AM applies 1D smoother for developing limited class of non-parametric regression models. Due to this factor, it has minimum influence when compared with p-dimensional smoother. Additionally, AM is highly reliable than a standard linear model, which has been interpretable than a normal regression surface interms of approximation errors. Some of the problems involved in AM are, model selection, over-fitting, and multi-collinearity. Provided a data set $\{y_i, x_{i1},, x_{ip}\}_{i=1}^{n}$of n statistical units, where $\{x_{i1},\dots,x_{ip}\}_{i=1}^{n}$shows predictors and $y_i$means the result, hence, AM is derived by:

$$E[y_i|x_{i1},\dots,x_{ip}] = \beta_0 + \sum_{j=1}^{p} f_j\,(x_{ij}) \qquad (21)$$

or

$$\mathrm{Y} = \beta_0 + \sum_{j=1}^{p} f_j\,(X_j) + \varepsilon \qquad (22)$$

where $E[\varepsilon] = 0$, $Var(\varepsilon) = \sigma^2$ and $E[f_j(X_j)] = 0$. The functions $f_j(x_{ij})$ are unknown smooth functions that fit for the data [17].

Fitting the AM, which refers that, functions $f_j(x_{ij})$is processed using back-fitting algorithm.

2.8 Gaussian process (GP)

The GP is one of the familiar approaches in MLwhich has been applied extensively in time series examination. A GP presents a set of arbitrary variables, and finite number of joint Gaussian distribution [18]. The GP is employed for characterizing the probability distribution across functions by using 2 functions such as mean function m(u) and the covariance function. The mean function $k(u_l, u_2)$. The real process f(u)is defined as GP, and expressed by

$$f(u) \sim \mathcal{GF}\big(m(u), k(u_l, u_2)\big), \qquad (23)$$

where,

$$m(u) \ = \ \mathbb{E}[f(u)] \qquad (24)$$

$$k(u_l, u_2) = \ E\big[\big(f(u_1) - \ m(u_1)\big)\big(f(u_2) - \ m(u_2)\big)\big] \qquad (25)$$

In regression, a data set D of N observations are provided; $D = \{(u_i, u_i)|i = \ 1, , N\}$, with $u_i \in \mathbb{R}^D$and $v_i \in \mathbb{R}$, the main aim is to detect new v_* given $u_*$by applying f(u)where: $v_i = f(u_i) + \delta_i$ in which δ_i denotes Gaussian noise with mean 0 and variance σ^2. Hence, the closing prices in stock market are effective as actual prices are computed at closing time. The advance distribution of required target v described by,

$$v \sim \mathcal{N}\big(0, K(U, U)\big), \qquad (26)$$

where, K(U, U)implies covariance matrix among every pairs of training points and U denotes $(n \times m)$matrix of input. Here, RBF kernel is applied as,

$$k(u_l, u_2) = \ \exp\left(-\sigma||u_l - u_2||^2\right) \qquad (27)$$

The predictive distribution of v ∗has been determined by conditioning the training data which results in $p(f(u_*)|u_*, D)$. The joint distribution of v and predictions of u_* is demonstrated by:

$$\begin{bmatrix} v \\ f(u_*) \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K(U,|U) & K(U,|u_*) \\ K(u_{*l}|U) & K(u_{*l}|u_*) \end{bmatrix}\right) \qquad (28)$$

The conditional distribution enables to obtain the predictive distribution of v_*using mean and covariance:

$$\bar{f}(u_*) = K(U, u_*)^T(K + \sigma_n^2 I)^{-1}v, \qquad (29)$$

$$V_f(u_*) = K(u_*, u_*) - K(U, u_*)^T(K + \sigma_n^2 I)^{-1}K(U, u_*) \qquad (30)$$

3 Performance Validation

The proposed model has been simulated using WEKA tool. For experimental analysis, the data has been collected from various districts in Tamil Nadu. Fig. 3 shows the frequency distribution of the attributes present in the dataset. Besides, a sample visualization of the dataset is shown in Fig. 4.

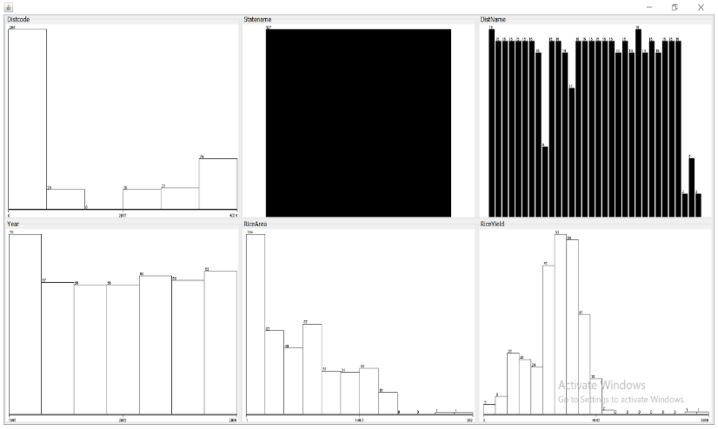
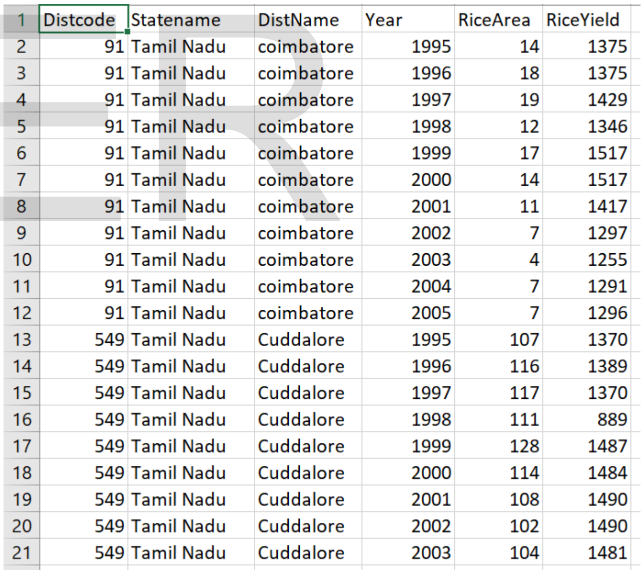


Fig. 3 Frequency Distribution of Applied Dataset



1	Distcode	Statename	DistName	Year	RiceArea	RiceYield
2	91	Tamil Nadu	coimbatore	1995	14	1375
3	91	Tamil Nadu	coimbatore	1996	18	1375
4	91	Tamil Nadu	coimbatore	1997	19	1429
5	91	Tamil Nadu	coimbatore	1998	12	1346
6	91	Tamil Nadu	coimbatore	1999	17	1517
7	91	Tamil Nadu	coimbatore	2000	14	1517
8	91	Tamil Nadu	coimbatore	2001	11	1417
9	91	Tamil Nadu	coimbatore	2002	7	1297
10	91	Tamil Nadu	coimbatore	2003	4	1255
11	91	Tamil Nadu	coimbatore	2004	7	1291
12	91	Tamil Nadu	coimbatore	2005	7	1296
13	549	Tamil Nadu	Cuddalore	1995	107	1370
14	549	Tamil Nadu	Cuddalore	1996	116	1389
15	549	Tamil Nadu	Cuddalore	1997	117	1370
16	549	Tamil Nadu	Cuddalore	1998	111	889
17	549	Tamil Nadu	Cuddalore	1999	128	1487
18	549	Tamil Nadu	Cuddalore	2000	114	1484
19	549	Tamil Nadu	Cuddalore	2001	108	1490
20	549	Tamil Nadu	Cuddalore	2002	102	1490
21	549	Tamil Nadu	Cuddalore	2003	104	1481

Fig. 4 Sample Data

Table 1 shows the detailed comparison of the results offered by the proposed model with existing models on the test dataset.

Table 1 Performance of Proposed Method with the Existing Methods

Methods	Correlation Coefficient	MAE	RMSE	RAE	RRSE
K-Star	0.954	223.43	365.22	26.65	32.55
LR	0.5504	666.96	936.57	79.57	83.48

Gaussian Process	0.7251	548.63	790.54	65.64	70.47
MLP	0.767	572.48	760.77	68.29	67.81
RBF	0.092	839.01	1117.03	100.09	99.57
Bagging	0.791	464.89	700.11	55.46	62.41
Additive Regression	0.534	680.735	949.21	81.21	84.61

Fig. 5 shows the correlation coefficient analysis of the proposed model on the applied data. The figure displayed that the RBF model is the worst performer, which has attained minimal correlation coefficient value of 0.092. At the same time, the LR and additive regression models have tried to show better results over the RBF model by attaining near identical correlation coefficient of 0.5504 and 0.534 respectively.

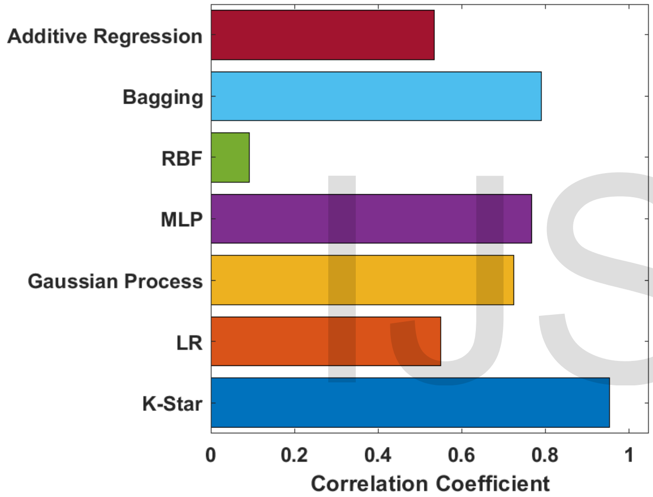


Fig. 5 Correlation Coefficient analysis of various models

Besides, the Gaussian processes have shown somewhat manageable results over the other methods by attaining moderate correlation coefficient of 0.7251. Also, the MLP and Bagging model leads to competitive and closer correlation coefficient values of 0.767 and 0.791 respectively. At last, the presented K-Star model has showcased effective performance by attaining maximum correlation coefficient of 0.954.

Fig. 6 showcases the MAE analysis of the projected model on the given data. The figure portrays that the RBF method is an ineffective performer, which has reached higher MAE value of 839.01. Simultaneously, the LR and additive regression methodologies have attempted to demonstrate moderate results than the RBF model by achieving closer identical MAE of 666.96 and 680.735 correspondingly. On the other side, the MLP have depicted reasonable results when compared with alternate models by accomplishing gradual MAE of 572.48. Additionally, the Gaussian process and Bagging approaches results competing and nearby MAE values of 548.63 and 464.89 respectively. Finally, the projected K-Star technology has illustrated superior performance by obtaining maximum MAE of 223.43.

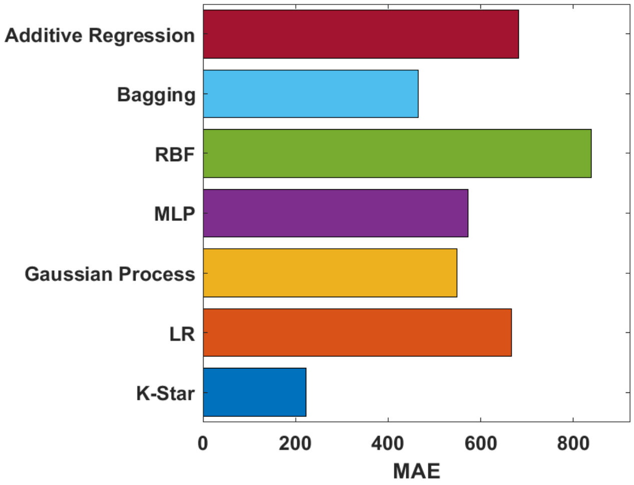


Fig. 6 MAE analysis of various models

Fig. 7 provides the RMSE analysis of the proposed approach on the applied data. The figure states that the RBF model is the poor performer, which has accomplished maximum RMSE value of 1117.03.

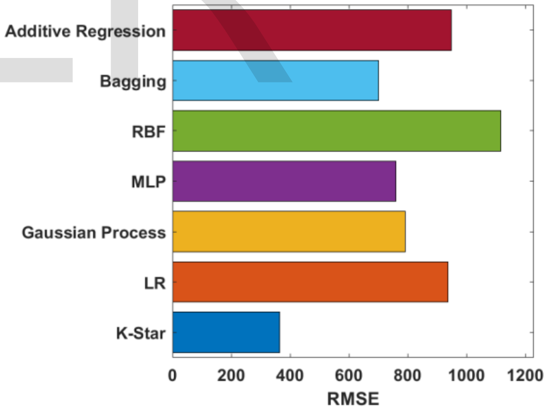


Fig. 7 RMSE analysis of various models

Fig. 8 offers the RAE analysis of the developed method on the applied data. The figure depicted that the RBF approach is an unfit performer, which has obtained higher RAE value of 100.09. Concurrently, the LR and additive regression frameworks attempted to demonstrate good results over the RBF model by accomplishing closer identical RAE of 79.57 and 81.21 correspondingly.

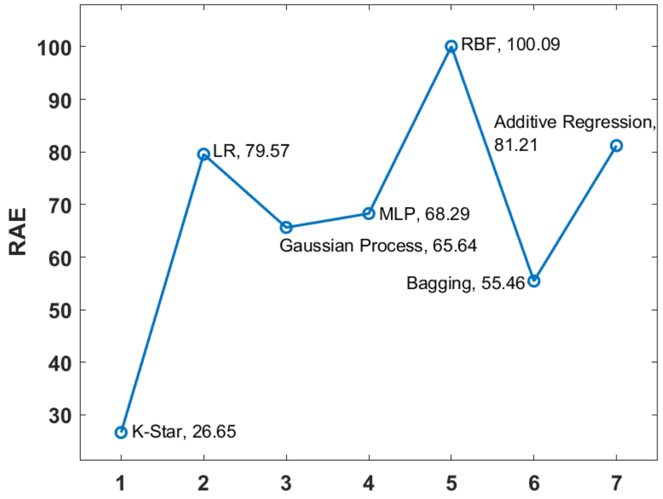


Fig. 8 RAE analysis of various models

Next, the MLP have illustrated appreciable results over the compared models by reaching gradual RAE of 68.29. In addition, the Gaussian process as well as Bagging schemes leads to competing and closer RAE values of 65.64 and 55.46 respectively. Consequently, the presented K-Star approach has demonstrated efficient performance by achieving best RAE of 26.65.

Fig. 9 provides the RRSE analysis of the presented model on the applied data. The figure revealed that the RBF technique is an inferior one, which has reached maximum RRSE value of 99.57. Meanwhile, the LR and additive regression models have attempted to imply good results over the RBF model by accomplishing closer identical RRSE of 83.48 and 84.61 respectively. On the other hand, the Gaussian processes have demonstrated better results over the earlier methods by reaching gradual RRSE of 70.47.

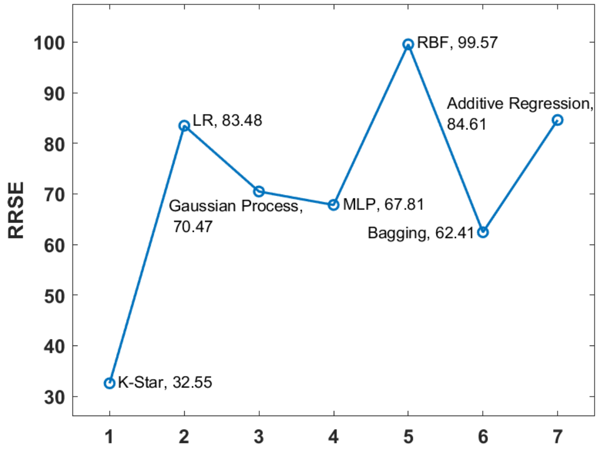


Fig. 9 RRSE analysis of various models

Furthermore, the MLP and Bagging methodologies produces more or less equal and closer RRSE values of 67.81 and 62.41 correspondingly. Finally, the proposed K-Star model has displayed productive performance by obtaining maximum RRSE of 32.55.

4 Conclusion

This paper has developed an intelligent ML based regression models for the prediction of rice plant productivity. The proposed model involves different ML models namely K-star, LR, bagging, MLP, RBFNetwork, additive regression and gaussian process. The performance of the proposed model has been validated using a dataset collected from Tamil Nadu. The experimental results indicated that the K-star model has offered maximum outcome over the compared methods with the maximum correlation coefficient of 0.954, minimum MAE of 223.43, RMSE of 365.22, RAE of 26.65 and RRSE of 32.55. These values ensured that the K-star model is found to be effective model for proper rice crop productivity. In future, the performance of these models can be improvised by the use of deep learning based predictive models.

Acknowledgements

This research work was carried out with financial support of RUSA–Phase 2.0 grant sanctioned vide Letter No. F. 24-51/2014-U, Policy (TNMulti-Gen), Dept. of Edn. Govt. of India, Dt. 09.10.2018.

References

[1] Wu Fan, Chen Chong, Guo Xiaoling, Yu Hua, Wang Juyun. Prediction of crop yield using big data. 8th International Symposium on Computational Intelligence and Design (ISCID).2015;1, 255-260.

[2] Khatkar, B.S., Chaudhary, N. and Dangi, P., 2016. Production and consumption of grains in India.

[3] Dahikar, S.S. and Rode, S.V., 2014. Agricultural crop yield prediction using artificial neural network approach. *International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering*, *2*(1), pp.683-686.

[4] N. Gnanasankaran and E. Ramaraj., 2020. A Multiple Linear Regression Model to predict Rainfall using Indian Meteorological Data. *International Journal of Advanced Science & Technology, 29(8), 746-758.*

[5] Kumar, A., Kumar, N., Kumar, V.V., Kumar, N. and Vats, V., 2018. Efficient Crop Yield Prediction Using Machine Learning Algorithms. *International Research Journal of Engineering and Technology*, *5*(06).

[6] Yimit, H., Eziz, M., Mamat, M., & Tohti, G. (2011). Variations in groundwater levels and salinity in the Ili River Irrigation Area, Xinjiang, Northwest China: a geostatistical approach. International Journal of Sustainable Development & World Ecology, 18(1), 55–64. https://doi.org/10.1080 /13504509.2011.544871.

[7] Dash, J. P., Sarangi, A., & Singh, D. K. (2010). Spatial variability of groundwater depth and quality parameters in the National Capital Territory of Delhi. Environmental Management, 45(3), 640–650. https://doi.org/10.1007/s00267-010-9436-z.

[8] Seyed mohammadi, J., Esmaeelnejad, L., &Shabanpour, M. (2016). Spatial variation modelling of groundwater electrical conductivity using geostatistics and GIS. Modeling Earth Systems and Environment, 2(4), 1–10. https://doi. org/10.1007/s40808-016-0226-3.

[9] Ahmadpour, H., Khaledian, M. R., & Ashrafzadeh, A. (2015). Examine the relationship between groundwater salinity with Sefidrud River and the Caspian Sea. In Third International Symposium on Environment and Water Resources Engineering. 2–3 June. Tehran, Iran.

[10] Chandrasekharan, H., Sarangi, A., Nagarajan, M., Singh, V. P., Rao, D. U. M., Stalin, P., & Anbazhagan, S. (2008). Variability of soil–water quality due to Tsunami-2004 in the coastal belt of Nagapattinam District, Tamilnadu. Journal of Environmental Management, 89(1), 63–72.

[11] Rezaei, M., Davatgar, N., Tajdari, K. H., & Aboulpour, B. (2010). Investigation the spatial variability of some important groundwater quality factors in Guilan, Iran. Journal of Water and Soil, 24(5), 32–41.

[12] Ashrafzadeh, A., Roshandel, F., Khaledian, M., Vazifedoust, M., & Rezaei, M. (2016). Assessment of groundwater salinity risk using kriging methods: a case study in northern Iran. Agricultural Water Management, 178, 215–224. https://doi. org/10.1016/j.agwat.2016.09.028.

[13] Mahmood, D.Y. and Hussein, M.A., 2013. Intrusion detection system based on K-star classifier and feature set reduction. International Organization of Scientific Research Journal of Computer Engineering (IOSR-JCE) Vol, 15(5), pp.107-112.

[14] Ananthakumar, U. and Sarkar, R., 2017, November. Application of logistic regression in assessing stock performances. In 2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech) (pp. 1242-1247). IEEE.

[15] Breiman, L., 1996. Bagging predictors. Machine learning, 24(2), pp.123-140.

[16] Bhattacharjee, K. and Pant, M., 2019. Hybrid particle swarm optimization-genetic algorithm trained multi-layer perceptron for classification of human glioma from molecular brain neoplasia data. Cognitive Systems Research, 58, pp.173-194.

[17] https://en.wikipedia.org/wiki/Radial_basis_function_network

[18] https://en.wikipedia.org/wiki/Additive_model

[19] Alamaniotis, M., Ikonomopoulos, A. and Tsoukalas, L.H., 2011, September. A Pareto optimization approach of a Gaussian process ensemble for short-term load forecasting. In 2011 16th International Conference on Intelligent System Applications to Power Systems (pp. 1-6). IEEE.

Author Details:

1] Dr.N. Gnanasankaran, Assistant Professor, Department of Computer Science, Thaigarajar College, Madurai, Tamilnadu, India. Mail id: sankarn.iisc@gmail.com.

2] Dr.E. Ramaraj, Professor and Head, Department of Computer Science, Alagappa University. Karaikudi, Tamilnadu, India. Mail id: eramaraj@rediffmail.com.

3] Dr.T. Manikumar, Assistant Professor. Department of Computer Applications, Thiagarajar College, Madurai, Tamilnadu, India. Mail id: tmanikumar.mku@gmail.com.