International Journal of Modern Agriculture, Volume 9, No.4, 2020
ISSN: 2305 - 7246

GENETIC ALGORITHM BASED FEATURE SELECTION AND RANDOM FOREST MODEL FOR RICE YIELD PREDICTION

Avijit Balabantaray[1], Payal Bhadra[2], Rakesh Kumar Ray[3], S. Chakravarty[4], SMIEEE

[1,2,3,4] Centurion University of Technology and Management, Bhubaneswar, Odisha, India
Email: *[1]avijitbalabantaray5@gmail.com, [2]bhadrapayal189@gmail.com, [3]rakesh.ray@cutm.ac.in, [4]sujata.chakravarty@cutm.ac.in*

Abstract

As a major contributor of Rice crop production, India ranks number two after China in whole over the world as a leading country to produce 116.42 million metric tons of Rice in 2018-2019. The production of rice is most suitable for a country like India with favourable climatic and weather condition and majority of population with farmer as occupation. So in order to meet the population demand and growth of country's economy, the major saviour is the rice yield prediction which can predict the amount of Rice by giving the pre-productive information to farmers and agronomist to boost up the production level. In this paper, we come up with an approach to predict the rice yield by making use of collected weather, climatic and agricultural raw materials related data to measure the quantifiable amount of Rice collected from fields. In order to predict the amount of Rice yield, agricultural materials like amount of seeds, pesticides, fertilizers, weather condition like amount of rain and temperature in addition with PH and moisture content of soil are considered which can be correlated with other environmental factors and the yield amount can be predicted by using these determinant factors performing regression analysis by implementing different machine learning models like Linear Regression, SVR, K-Neighbors Regression, Random Forest and Decision tree to compare the value of each regression model with a range analysis of MAPE value collected from each model and genetic algorithm is implemented for feature selection purpose in order to reduce measured MSE and MAPE values up to certain extent.

Key words: Rice Yield Prediction, Regression Analysis, Linear Regression, Support Vector Regression, Random Forest, Genetic Algorithm, Feature Selection

Introduction

Agriculture is the main source of livelihood of Indian people consisting of more than 50% of whole India's population [1] among which Rice is the main cultivar and India itself contributes more than 50% of Rice production in whole over Asia. So Rice yield estimation is one of the significant parts of Agriculture so as to adjust the national grain gracefully and satisfy the populace need [2]. Yet, forecast of yield is tedious and difficult strategy. So we have proposed the necessary yield forecast framework which will anticipate the regular yield of Rice cultivars by analyzing certain ecological variables like land expansion, climate condition, soil dampness level, amount of downpour, pesticides and manures amount through regression analysis using machine learning approach. Certain considerable factors like amount of grains and area of grain particles of rice panicle is one of the important factor that can be taken into consideration for estimating the quantifiable amount of rice yield collected from the paddy fields [3]. Essentially other significant features are certain ecological components like augmentation of land and paddy fields of farmers, measure of provided water, dampness level of the soil, phenotypic element of rice cultivars, use of certain measure of pesticides and manures, bio-synthetic compounds like N_2, P_2O_5 and PH level of soil and so on [4]. Considering the overall elements, it is very conceivable to gauge the measure of rice yield production in every paddy field which will a central point in a nation like India where ranchers can get the exceptional advantages by getting a legitimate perception of speculation of crude materials for future production which will go about as a key job for boosting the economy of a agricultural rich nation like India [5].

So here the primary concerned is the manner by which to anticipate the regular rice yield quantity by utilizing the information gathered from different ecological factors in an increasingly exact and effective

manner. Here in our proposed model, we use the features like amount of raw materials starting from rice cultivars of different phenotypes, amount of water, pesticides, fertilizers, PH, N_2, P_2O_5 and other climatic conditions to predict the yield by performing the regression prediction using different machine learning models like Linear Regression, K-Neighbors Regression, SVR, Decision tree and Random forest. We try to compare each and every model to get the most desirable R- squared (R^2) value by comparing the Mean absolute percentage error (MAPE) of each model and put genetic algorithm technique to reduce further error by feature selection.

2. Literature Survey:

Previously image processing based methods have been proposed for analyzing the plant health condition or phenotypic analysis of different plant cultivars. Similarly image based efficient yield prediction has been proposed in [6] by extracting the grain area of rice panicles by using image segmentation method and comparing the image based model r-squared value with the regression analysis value of weight parameters in order to predict the amount of yield. A tremendous work has been done in [7] by analyzing the rice canopy digital images by considering the changes in red, green and blue light value parameters. Many of the works have been done by making use of satellite images of paddy fields where Yi-ShiangShiuet.al [8] has been proposed a rice prediction system by considering the global and local regression models with the use of the data collected from satellite images. Some of the past works has been analyzed by using the environmental factors such as climatic conditions and soil composure like the ANN based rice yield prediction by [9].

3. Methodology

3.1. Proposed Methods

In the proposed method, data based regression prediction techniques are used to analyzed the performance of predictive models which can give better result with keen comparison among different machine learning models in addition with Genetic algorithm approach for feature selection to get an efficient prediction method for rice yield. Raw data determining the yield of different Rice cultivars consisting of different types of attributes are collected to perform the regression analysis to calculate net surplus amount collected from a particular land area. After performing the required pre-processing on collected data, various machine learning models are implemented upon the collected data like raw materials amount and climatic conditions which give a overview about the performance of each model in accordance with yield amount prediction. Machine Learning models like Linear, logistic regression, SVR, K-Neighbors, Decision trees and Random forest are applied to the collected data to compare each model performance by keeping net amount of yield in view which can be easily analyzed by comparing the R-squared values with MSE and MAPE. In order to reduce the calculated Mean Absolute Percentage Error (MAPE) and Mean Square Error (MSE) from different machine learning models, Genetic algorithm technique is implemented for feature selection purpose by increasing the model performance into some extent by reducing the error. Then an overall comparison is performed among the machine learning models to get that which model outperforms the others and most reliable for the efficient yield production. The whole model implementation will generalized by following the steps as mentioned in Figure 1 where data are collected and processed followed by model implementation and training of the sample inputs with making a quantitative comparison among various model metrics like R-squared value and MAPE with model comparison.

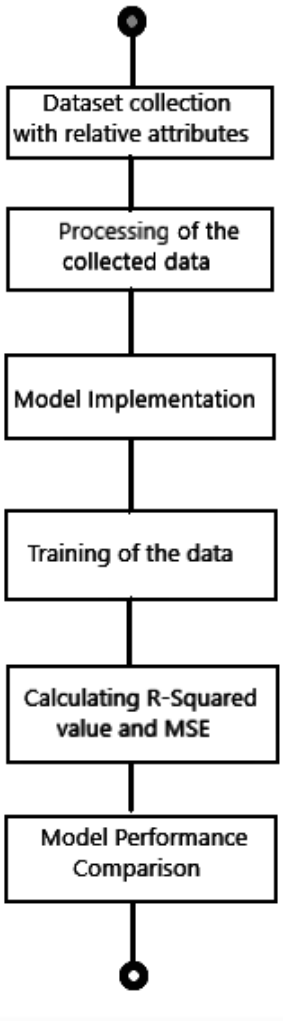

Figure 1. Graphical Representation of Flow diagram of Machine Learning Model implementation

3.2. Concise Model illustration:

3.2.1. Linear Regression: Regression analysis technique is performed in order to decide the connection or correlation between at least two factors having cause impact relations and to make predictions for the theme by utilizing the measure of correlation [10]. Linear Regression is a measurable strategy for figuring the estimation of a dependent variable from an independent variable. It quantifies the relationship between two factors or variables. The principle thought behind it, is to anticipate the dependent variable by utilizing at least two or more independent factors [11]. The regression technique involving only single independent variable is considered as Univariate Regression analysis or simple linear regression whereas regression involving multiple independent variables is termed as Multivariate regression analysis or multiple linear regressions [10, 12].

Simple Linear Regression:

In case of simple linear regression, only single regressor x is compared in contrast with y by establishing a relationship with a straight line as the analysis outcome which can be illustrated by the expression in Eq.1 [12],

$$y = \beta_0 + \beta_1 + \varepsilon \quad (1)$$

Where, the unknown constants are intercept β_0 and the slopeβ_1 and ε is a random error component .

Multiple Linear Regressions:

The regression technique involving multiple number of regressors that is k no of variables likex_1, $x_2 \ldots x_k$, which is linearly related to the variable y and expressed as in Eq.2 [13],

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots + \beta_k x_k + \varepsilon \quad (2)$$

3.2.2. Support Vector Regressor: Support Vector Regression is an extended version of the Support Vector Machine to perform the regression analysis of non-linearly distributed data. So the main idea behind the support vector regression is to unravel a nonlinear regression in a straight manner by mapping nonlinear data collected from the dataset from unique dimensional element space to a higher dimensional component space [14].Suppose the model training values are in the form of different sets like (x_1, y_1), $(x_2, y_2)\ldots\ldots(x_N, y_N)$ where x is the training data factors and y is the corresponding labels with N numbers of samples. So according to Vapnik's ε based SVR theory, the function f(x)as a linear function is represented as in Eq.3,

$$f(x) = (w, x) + b \quad \text{where } w\epsilon \text{ x and b } \epsilon \text{ R} \quad (3)$$

Where X is the space of the input patterns, w is the weight parameters with the dot product relation with each input values in order to achieve flatness by seeking small w values with bias b. The most feasible approach to accomplish the task in order to get f(x) with no more than ε deviation from actual input values where y, is by resolving the optimization problem demonstrating by [15] in Eq.4 and 5,

$$\text{Minimize } \frac{1}{2}\|w\|^2 \quad (4)$$

$$\text{Subjected to constraints } \begin{cases} y_i - wx_i - b \leq \varepsilon \\ wx_i + b - y_i \leq \varepsilon \end{cases} \quad (5)$$

It should be obtained which always maintains at most ϵ deviation from the actual output label value y for all the training samples by maintaining a flatter shape throughout the whole regression analysis [16] which can be illustrated in the Figure 2 adaptable hyper plane of negligible sweep is shaped evenly around the assessed function, such that the absolute errors with minimal value the concerned threshold value ϵ are considered as insignificant throughout the hyperplane [17].

3.2.3. K-Neighbors Regression: K-Neighbors Regression technique where the values of the input data are by taking the average of the K nearest neighbours which contributes more as compared to the distant ones. For example, in case of a data point y having nearest neighbor points y_1 and y_2, the most appropriate solution for predicting y is in Eq.6 [18],

$$y = \frac{y_1 + y_2}{2} \quad (6)$$

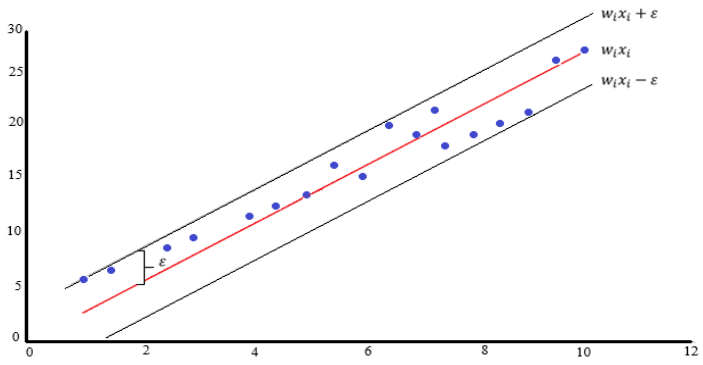

Figure 2. Support Vector Regression Model with Hyperplane and Regression Line

So the predictive value for sample inputs is taken as the average of the K nearest neighbors. In order to measure the nearest distance between the queried points and given training samples, distance measuring formulas like Euclidean distance. Given example like points p=($p_1, p_2, p_3, p_4 \dots, p_n$), in order to calculate K nearest neighbors that are similar to the training samples x=($x_1, x_2, x_3, x_4 \dots, x_n$) the measured Euclidean distance is represented in Eq.7 [19],

$$D(x,p) = \sqrt{(x_1 - p_1)^2 + (x_2 - p_2)^2 + \cdots + (x_n - p_n)^2} \quad (7)$$

For each queried points the nearest K-neighbors of input samples are calculated and the most appropriate predicted value is assigned to each queried point.

3.2.4. Decision Tree Regression: Decision Tree is a regression technique with tree structure where the tree gradually increases by portioning the input samples in each level by creating subsets of samples. The algorithm followed by decision tree is ID3 algorithm approach proposed by J.R.Quinlan based on a greedy choice approach throughout each level of tree avoiding backtracking. The ID3 algorithm focuses on reduction of Standard deviation value instead of information gain in order to analyze the decision tree regression [20].

The whole process of decision tree is based upon building a top-down approach tree by portioning the input samples into different subsets in each level by examining the homogeneity score of each numerical samples by using Eqs.8, 9 and 7,

$$\bar{x} = \frac{\sum x}{n} \quad (8)$$

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n}} \quad (9)$$

$$v = \frac{s}{x} \times 100\% \quad (10)$$

Where x are the input samples and n is the number of samples [21]. The average of the whole inputs is calculated as $\bar{x}$ which is later used for calculating the Standard deviation s. The coefficient of variation v should be calculated to test the homogeneous nature of input samples. If s=0, then samples are homogeneous.

3.2.5. Random Forest Regression: Random Forest is an ensemble model which combines the outcomes for different other models like decision trees by giving the most appropriate prediction by reducing the variation in comparison with single decision trees [22]. The ensemble learning in case of Random forest is totally inspired by the bagging method which is generally used for random feature selection in order to enhance the accuracy and to calculate the generalized error [23]. It aggregates multiple decision trees with multiple predictions where each tree obtain a sample of training sets from the given input data and add some random points to provide the evaluated value.

3.2.6. Genetic Algorithm for feature selection: Genetic Algorithm is an optimization algorithm based upon the concept of genetics, natural selection and biological evolution. It is used to gain the optimal solution by performing the feature selection efficiently by traversing through the supplied data by making using of all kinds of biological evolution theory such as selection, crossover and mutation [24]. GA basically consists of several steps by which it performs the feature selection by choosing the optimum solution by choosing the best fit off-springs which has been illustrated in the Figure 3. It works through repetitive iterative process where each generation tries to select better fit individuals by performing the evolutionary operations like crossover followed by mutation where with each optimal solution is come up with a better model with reduced error than the previously provided model [25]. The crossover operation is performed to crossover the selected parents in order to make new off-springs from parents. Then the mutation operation is carried out to diversify the generation by

changing the features of the off-springs. After this selection operation select the fit optimum solution to pass over the next generation by calculating the fitness value.

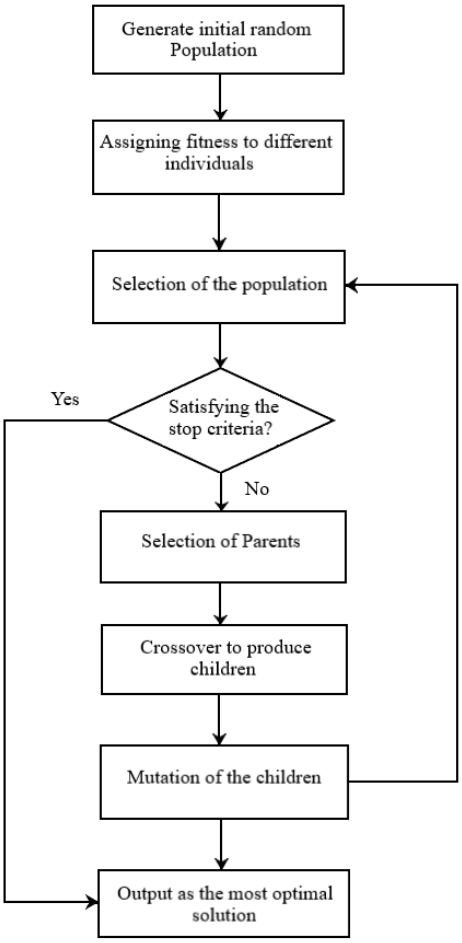

Figure 3. Flow Diagram of different phases of Genetic algorithm to achieve optimal solution

The fitness function is the major quantitative measure to identify which individual is more fit to be inherited to the next level which can be expressed as in Eq.11,

$$f = \sum_{i=1}^{n} f_i \quad (11)$$

Where f_i is the difference of the square values of original weight values and reduced weight values for each response and n is the no of population [26].

3.3. Dataset Preparation and Feature Selection

The data which are utilized for the exploration reason in this paper has been collected from the University resources through academic supervision. The dataset contains the rice production data of different farmers with required information like individual information, for example, Name, Age, Location, Land Extension, Paddy fields expansion and so forth with rural data of a specific season production like measure of seeds, wet/dry/wet-dry land alongside the staples required for the harvesting like composts, for example, PH of the soil, manures particularly Urea amount, Pesticides amount, Nitrogen, Iron and so forth etc. Alongside all these data the measure of all out yield per every rancher are gathered with other required information like work cost, water flexibly cost and other essential data. From the collected dataset, different unnecessary features which shouldn't be considered for our proposed models are eliminated such as Name, age and other unusual

data like labour and work cost. The necessary data which can mostly affect our machine learning models should be prioritized by performing the correlation operations among different contribution features like Land extension, Seeds measure, Soil texture alongside with fertilizers, pesticides and agro-minerals, so that to measure the relationship among different variables that how closely they are linearly related and directly proportional. By performing the correlation operation, the features or the factors by which our models are mostly affected, can be derived easily by eliminating the unnecessary features which will worsen the model prediction [25]. The correlation feature mapping graph can be clearly illustrated in the Figure.4 where the linear relationship among different features can be measured through the 1.0 scale which is the highest value of correlation that how useful that particular feature is.

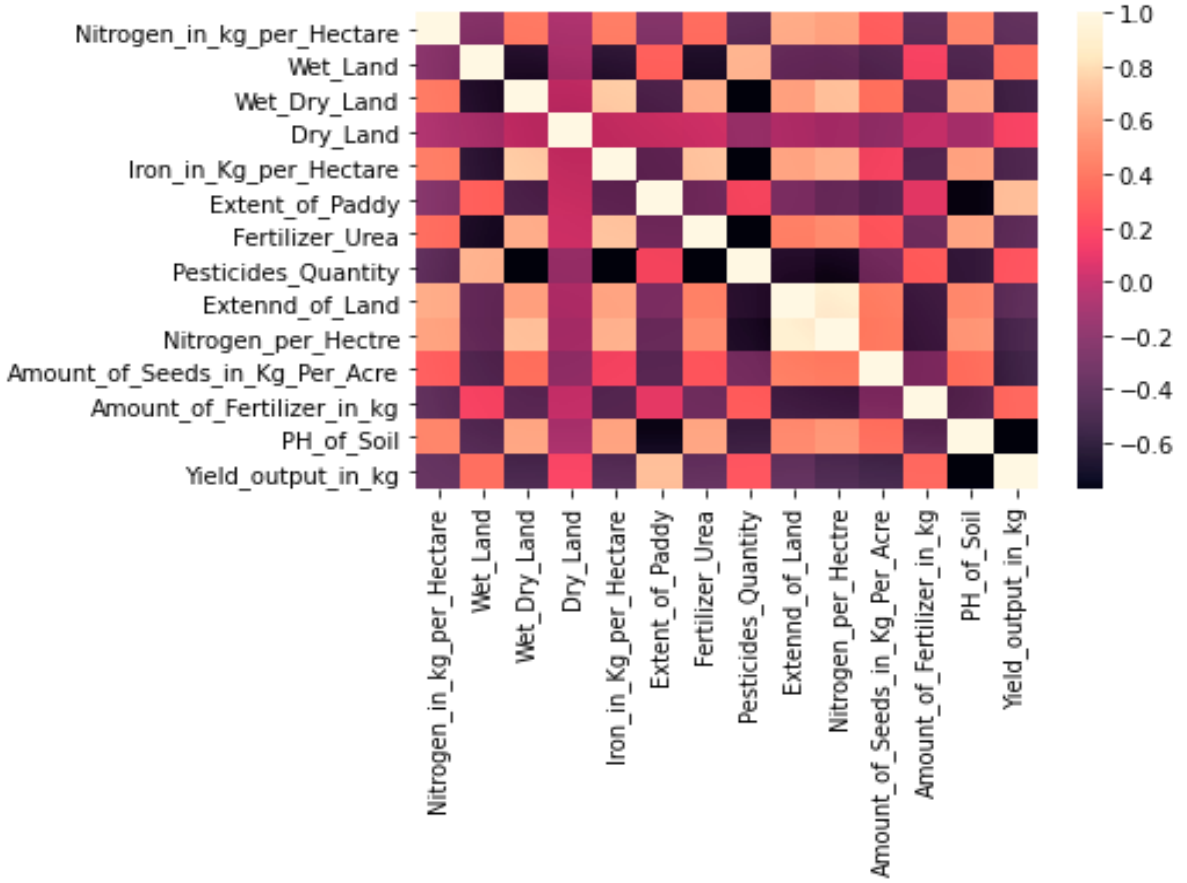

Figure 4. Selecting features based on the correlation mapping graph

In order to choose the appropriate factors, the statistical approach by calculating the P value is performed by comparing each columns of the dataset that which columns are adversely affecting the model performance. The P value helps in feature selection as it calculate the probability of the validation of a particular hypothesis or assumption. We need to test this speculation for each component and choose whether the features hold some centrality in the prediction of the reaction. [26]. The correlation value below 0.8 are taken with the considered hypothesis that the features taken have less impact on the model prediction and P-value is calculated to detect whether the hypothesis is true or not and according to the null hypothesis criteria, the features are selected by taking different combination of columns from the dataset. If the P-value will overcome a certain threshold value, then the feature combination will not considered as the appropriate combination of the model. After feature selection, the data distribution of the selected features is plotted in the Fig. 5 to visualize the linear or non-linear distribution with maximum and minimum elevation points.

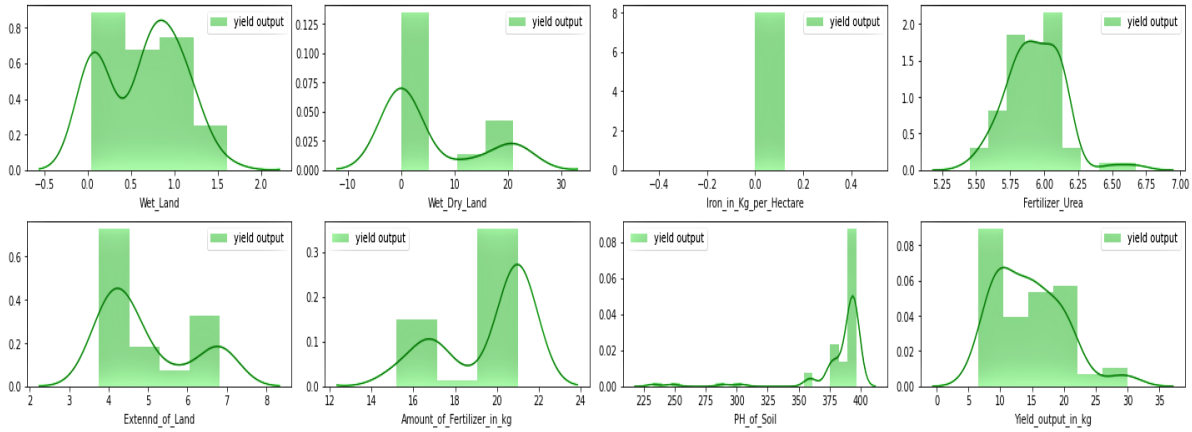

Fig.5. Selecting features based on the correlation mapping graph

Pre-processing of the Collected Data:

The collected data from the dataset with 500 samples are subjected to pre-processing before implementing in the machine learning models. The prerequisite step is the feature scaling operation where the collected data from various ranges are scaled in a range of 0 to 1 by performing the normalization operation which makes the data suitable for model implementation by scaling them into a particular range. After performing the scaling operation, the data is converted into training and testing part by performing the train-test-split operation where the training data and the testing data are distributed in the ratio of 8:2 having 375 samples in Train set and 125 samples for the Test set.

3.4. Model Implementation

The respective model is implemented on the dataset having 500 input samples after performing the preprocessing operations like min max scalar and train-test split. According to the multiple linear regression concepts, the data collected from our dataset are subjected for estimation of regression prediction where the linear regression is expressed as in Eq.12,

$$h(\theta)(x) = \theta_0 + \sum_{i=1}^{m} \left(\theta_i(x^i)\right) \quad (12)$$

Where $h(\theta)(x)$is a speculated value for given contribution for a specific arrangement of parameters θ. The model is implemented by following the Eq.13 of cost function in order to calculate the mean squared error for each of the determinant values of x in our dataset by comparing the predicted value with the actual labels y by generalizing the concept of error reduction at each step,

$$J(\theta) = \frac{\sum_{i=1}^{m}(h_\theta(x^i) - y^i)^2}{2m} \quad (13)$$

The calculated cost function should be optimized by minimizing the error for each inputs which is represented in Eq.14 that how the error is updated for each parameters to obtain the minimum error for model execution,

$$\theta_j = \theta_j - \frac{\alpha \sum_{i=1}^{m}(h_\theta(x^i) - y^i)(x^i)}{m} \quad (14)$$

Where α is the learning rate and m is is the number of training sets. The cost function is updated in each cycle in order to minimize the error till the convergence of the absolute error graph[27].The model is trained to find best regression line with slope and intercept to best fit our data which is presented in the Fig. a with the plotting of the linear regression with the optimized form. The observed R^2 value is 0.62 in 0-1 scale which is

visualized in the percentage form in the Table.1 by measuring the scattering of the data points around the optimized line. The respective Mean Squared Error, Mean absolute error, Mean absolute percentage error and Mean Percentage errors are calculated by using specific equations in order to evaluate the model performance which are represented in Table.1. As well, SVR regressor is implemented on the preprocessed data by taking a linear kernel and a C value of 1E10 in order to fit our model to get the appropriate hyperplane for regression analysis in which the error should not exceed the threshold value. After training of the model, the R^2 value calculated with a performance analysis of 0.61 in 0-1 scale whose percentage value is illustrated in the Table.1. The model metrics such as MSE, MAE, MAPE and MPE are calculated in order to visualize the model performance according our given dataset which are presented in the Table.1. In order to implement the K-Neighbors regression technique on our dataset, the basic steps of algorithm are put forth starting from Euclidean distance calculation followed by nearest neighbors finding with regression prediction [28]. The calculated Euclidean distance is then used further to finding the nearest neighbors for each input samples of the dataset by cross validating unique combination of input sets. After calculation the optimal K value is extracted which is 10 in our case. Then the model is implemented to predict the actual y value in each cases and to analyze the regression technique. The calculated R^2 value in K-Neighbors case is 0.74 in 0-1 scale whereas the value in % is presented in Table. Model metrics like MSE, MAE, MAPE and MPE are also calculated for this regression problem and are demonstrated in the Table.1. Decision tree regression technique is used in order to get the better prediction as compared to other machine learning models which is deployed with the criteria like MSE as the error mapping criterion and levels of tree which is the major criteria for the model functionality. In our proposed decision tree model, 5 levels are taken to determine the regression operation. After training the DT model with the training data, the R^2 value is calculated which is 0.73 and the percentage measure of R^2 value and calculated MSE, MAPE values are showed in the Table 1.

Results and Analysis

4.1. Model Evaluation

We implemented all 5 machine learning models to predict the rice yield. After performing the training of the data by using all the models, the model metrics such as R^2 value, Mean squared error as the measure of squared error for each training model with Mean absolute percentage error, Mean absolute error and Mean Percentage errors [29, 30] are calculated in order to define the efficiency of the test machine learning models that which one is the more reliable one for regression prediction of Rice yield. The model is basically evaluated by the performance analysis quantified by the R^2 value that which particular model outperforms the other. The evaluation of the R^2 value is totally based upon the statistical measure, simply called coefficient of determination followed by the Eq.15,

$$R^2 = 1 - \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} \quad (15)$$

Where the upper value is the square difference of the total error with the lower value is the difference between the actual values. By analyzing the R^2 value the variance between the actual value and calculated value can be determined with the model performance. For each of the models starting from Linear Regression followed by Support Vector Regression, K-neighbors, Decision tree and Random forest models, the respective R^2 value is calculated which is represented in the Table.1 In which it can be clearly visualized that the R^2 value of the RF model is the highest one with 0.96 in 0-1 scale followed by 0.73 and 0.74 for decision tree and random forest showing the most similar variance in the predictive and actual value. The SVR and Linear Regression model has the lowest R^2 value of 0.62 and 0.61 which is calculated by our respective models. The squared error loss for each models are analyzed by using the Mean Squared Error as the criterion evaluating by the Eq.16,

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2 \quad (16)$$

Table 1. Representation of Model metrics of each proposed machine learning models for both Training and Testing data

MODEL	TRAINING DATA					TESTING DATA				
	R^2	MSE	MAE	MAPE	MPE	R^2	MSE	MAE	MAPE	MPE
Linear Regression	62.85	0.1016	0.0819	3.0791	-1.3175	57.08	0.0153	0.0814	3.7815	-2.2323
Random Forest Regression	96.42	0.0019	0.0235	0.9479	-0.5049	79.69	0.0072	0.0551	2.0850	-1.1374
Decision Tree Regression	73.29	0.0019	0.0456	1.9923	-0.4923	60.97	0.0153	0.0799	3.0098	-2.4924
K-Neighbors Regression	74.96	0.0110	0.0614	1.9134	-0.4944	66.10	0.0121	0.0781	4.0199	-2.2368
Support Vector Regression	61.28	0.0170	0.0833	2.964631	-0.1042	55.89	0.0158	0.0873	3.5928	-1.9200

The MSE value calculated for each models are illustrated in the Table where RF model has the smallest squared loss with the value of 0.0019 which is very small that the difference between the actual and predicted Rice yield value are small whereas other models like K-Neighbors and Decision tree have comparing values of 0.013 and 0.017 with a greater difference between the actual and predicted. At the same time SVR and LR have the squared loss of 0.1 which is very high as compared to other models. The MSE value is used to plot the regression lines where the difference between the actual and estimated value can be easily visualized with distance comparison of input samples from the optimal line. The regression lines for each model are plotted in the Figure 6 where scattering of the sample data in the graphs shows the difference between the actual optimal line and the predicted values. As shown in the Figure 7 RF has the most optimal predicted values scattered all over the regression line by performing the best for our Rice Yield dataset whereas the models LR and SVR has the data points scattering far from the optimal line showing poor performance in case of our dataset. In regression prediction Mean Absolute Percentage Error (MAPE) is calculated as the measure of relative error is calculated as the measure of relative error context by evaluating the expression. Random forest has the lowest possible MAPE value with less relative error measured 0.75 followed by Decision tree and K-Neighbors with comparing value ranges between 2.5 to 2.3 with 2% variation in the overall prediction rate showing less amount of deviation. Whereas the relative error for LR and SVR ranges between 2.8 to 2.7 showing more variation as compared to other used models in our proposed methods.

4.2. Model Performance Analysis

The different machine learning models are successfully implemented on the Rice yield dataset with an outstanding result showing efficiency of different algorithms like Random Forest, Decision Tree etc. From the results evaluated from each trained model, Random Forest algorithm outperforms other algorithms which specifically showed worthy performance with 96% model score which is calculated R-Squared value in our regression prediction. The comparison graph of R-Squared value is presented in Figure 8 showing the difference in coefficient of determination in case of each algorithm. Most of the actual labeled Rice yield values are measured as the accurate predicted values. In case of our dataset, the number of samples and degree of correlation is low by making the input data diverse and the degree of variance is more among the data. So algorithms like Linear Regression, SVR, K-Neighbors can't perform efficiently as the algorithms are effective on small dataset with less variance and high correlation coefficient. Decision Tree shows an intermediate performance as the model is prone to overfitting. So in case of our dataset the

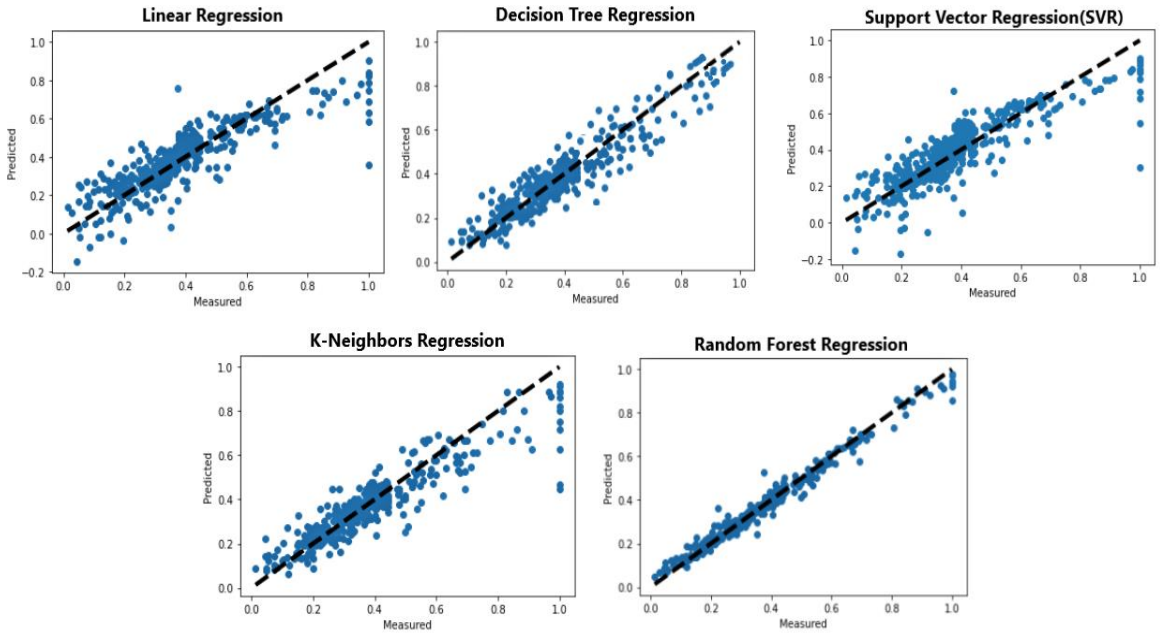

Figure 6. Regression analysis graph with scattered data plotting for each Machine Learning Models

variance in the data and diverse distribution has highly affected. So in order to minimize the error and degree of variance, Random forest is used which make use of different subsets of samples by distributing them throughout the whole tree structures by decreasing the variance and avoiding the overfitting condition. Our collected data are performed extensively well with the Random forest algorithm with the MSE value of 0.0019 and MAPE value of 0.94. The actual and predicted Rice yields are plotted graphically in Figure 9 for each model showing the variation of model predictions. The genetic algorithm is implemented further in order to reduce the calculated error in case of each machine learning models. With each new generation the selected features leads to better offspring giving optimal features by reducing the error in each iteration. The reduced MAPE values for each model implemented with GA are compared to the previous MAPE value in the Fig.10where the reduction of error can be visualized with after successful selection of optimal features. Before GA the MAPE values for models ranges between 2.8 to 0.75 and after GA it got reduced to range between 2.4 to 0.5 which has drastically reduced due to the suitable features selection with each iteration of generation. After the error reduction process, our models has enhanced with the performance showing more accurate predictive value where the Rice yield prediction can be done using sample features in more convenient and reliable way.

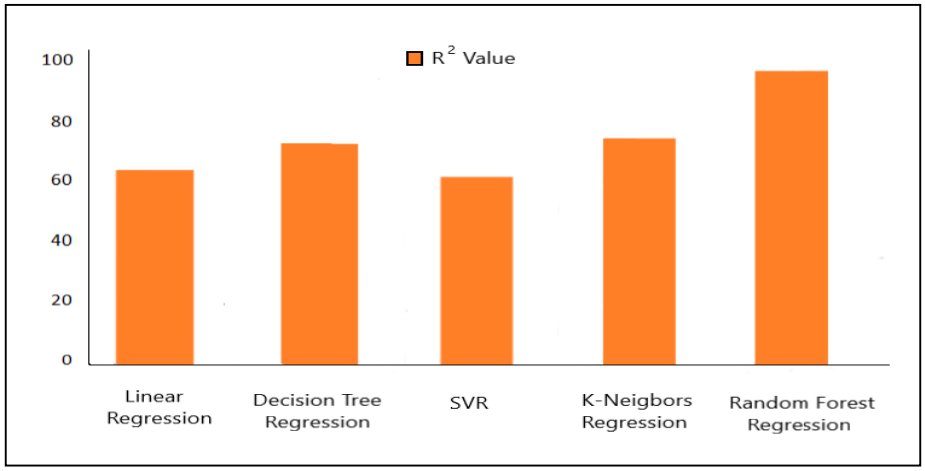

Figure 7. R-Squared Value distribution for every proposed Machine learning algorithms with Comparison

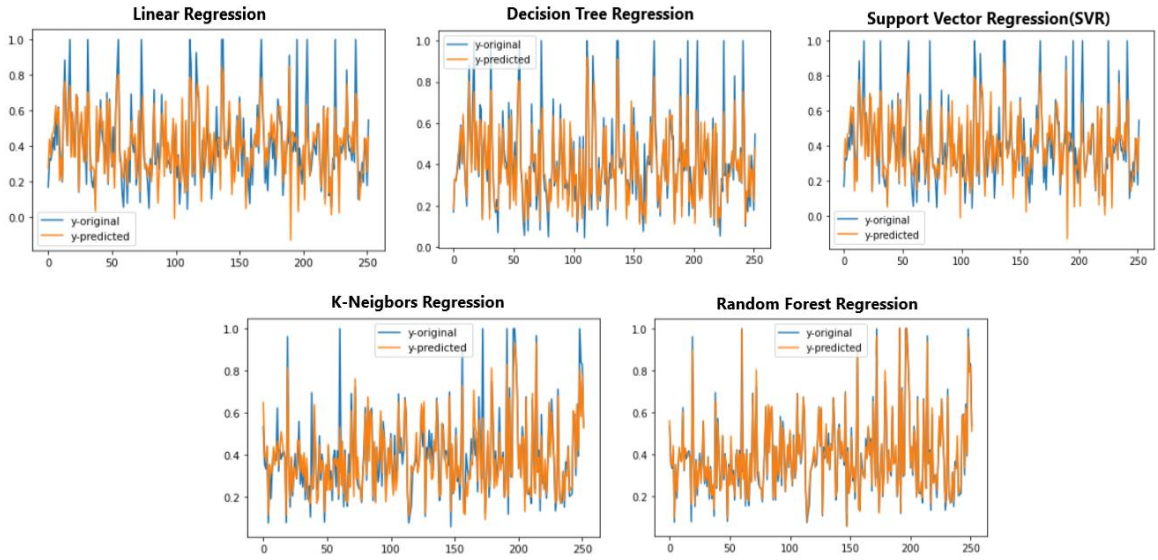

Figure 8. Actual Yield and Predicted Yield value Comparison by plotting prediction graph

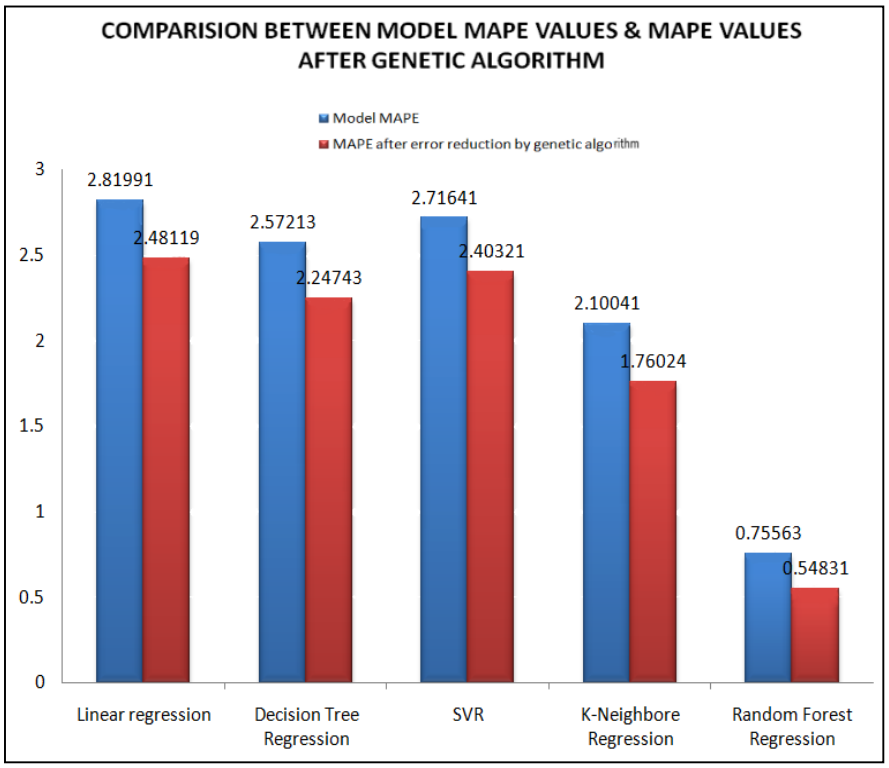

Figure 9. Keen comparison between the MAPE values before GA implementation and after GA feature selection by reducing the error

Conclusion

We successfully implemented different machine learning models on our collected dataset to achieve the most appropriate result by calculating the Rice yield by making use of different dependable factors like

agricultural raw materials manures and pesticides along with certain environmental features. Among the machine learning models Random Forest algorithm performed excellent by giving a regression determination coefficient of 0.96 which is far better as compared other implemented models. As our collected datasets degree of variance and nature of correlation varies greatly as compared to other models collection of complex decision trees that is Random forest gave an outstanding result by predicting the accurate Rice yield values for each sample data making the yield prediction more efficient. In order to boost the performance of the models, feature selection is performed to reduce the further relative error by making use of Genetic algorithm to further advance the degree of yield prediction in more effective way.

References

1. Agriculture in India, https://en.wikipedia.org/wiki/Agriculture_in_India.
2. Sanqin Zhao, Haonan Zheng, Mingmei Chi, Xicun Chai, Yutao Liu, "Rapid yield prediction in paddy fields based on 2D image modelling of rice panicles", Computers and Electronics in Agriculture, Volume 162, **(2019)**, Pages 759-766, ISSN 0168-1699, https://doi.org/10.1016/j.compag.2019.05.020.
3. Wu, W., Liu, T., Zhou, P. et al. "Image analysis-based recognition and quantification of grain number per panicle in rice", Plant Methods 15, 122 **(2019)**. https://doi.org/10.1186/s13007-019-0510-0
4. Rao P.R., Gowda S.P., Prathibha R.J. **(2019)**, "Paddy Yield Predictor Using Temperature, Rainfall, Soil pH, and Nitrogen", In: Sridhar V., Padma M., Rao K. (eds) Emerging Research in Electronics, Computer Science and Technology. Lecture Notes in Electrical Engineering, vol 545. Springer, Singapore.
5. Dhekale, B.S., Nageswararao, M.M., Nair, A. et al. Prediction of kharif rice yield at Kharagpur using disaggregated extended range rainfall forecasts. Theor Appl Climatol 133, 1075–1091 **(2018)**. https://doi.org/10.1007/s00704-017-2232-4.
6. Shiu, Y.-S.; Chuang, Y.-C. Yield Estimation of Paddy Rice Based on Satellite Imagery: Comparison of Global and Local Regression Models. *Remote Sens.* **(2019)**, *11*, 111..
7. Biaojun Ji, Y. Sun, S. Yang, J. Wan, "Artificial neural networks for rice yield prediction in mountainous regions", The Journal of Agricultural Science 145(03), **(2007)**, doi: 10.1017/S0021859606006691
8. Gülden Kaya Uyanık, Neşe Güler, "A Study on Multiple Linear Regression Analysis", Procedia - Social and Behavioral Sciences, Volume 106, **(2013)**, Pages 234-240, ISSN 1877-0428, https://doi.org/10.1016/j.sbspro.2013.12.027.
9. Kumari, Khushbu, and Suniti Yadav. "Linear regression analysis study." Journal of the Practice of Cardiovascular Sciences 4.1 **(2018)**: 33.
10. Sunthornjittanon, Supichaya, "Linear Regression Analysis on Net Income of an Agrochemical Company in Thailand" **(2015)**, University Honors Theses, Paper 131, 10.15760/honors.137.
11. Montgomery, Douglas C., Elizabeth A. Peck, and G. Geoffrey Vining. Introduction to linear regression analysis. Vol. 821. John Wiley & Sons, **(2012)**.
12. Lu Bin, Ni Shaoquan, Washburn Scott S., "A Support Vector Regression Approach for Investigating Multianticipative Driving Behavior", Mathematical Problems in Engineering, **(2015)**, 1024-123X, https://doi.org/10.1155/2015/701926
13. Harris Drucker, Christopher J. C. Burges, Linda Kaufman,Alexander J. Smola, V. Vapnik, "Support vector regression machines", Advances in neural information processing systems, **(1997)**, 28(7):779-784
14. Smola, Alex J., and Bernhard Schölkopf. "A tutorial on support vector regression." Statistics and computing 14.3 **(2004)**: 199-222.
15. Awad, Mariette, and Rahul Khanna. "Support vector regression." Efficient Learning Machines. Apress, Berkeley, CA, **(2015)**, 67-80.
16. Imandoust, SadeghBafandeh, and Mohammad Bolandraftar. "Application of k-nearest neighbor (knn) approach for predicting economic events: Theoretical background." International Journal of Engineering Research and Applications 3.5 **(2013)**: 605-610.
17. Novitasari, H. B., et al. "K-nearest neighbor analysis to predict the accuracy of product delivery using administration of raw material model in the cosmetic industry (PT Cedefindo)." Journal of Physics: Conference Series. Vol. 1367. No. 1. IOP Publishing, **(2019)**.
18. Kuhn, Lisa, et al. "The process and utility of classification and regression tree methodology in nursing research." Journal of advanced nursing 70.6 **(2014)**: 1276-1286.
19. Decision Tree, https://www.saedsayad.com/decision_tree_reg.htm#:~:text=Decision%20tree %20builds%20regression%20or,decision%20tree%20is%20incrementally%20developed.&text=Leaf%20node%20(e.g.%2C%20Hours %20Played,decision%20on%20the%20numerical%20target.

20. Couronné, R., Probst, P. & Boulesteix, A. "Random forest versus logistic regression: a large-scale benchmark experiment", BMC Bioinformatics 19, 270 **(2018)**. https://doi.org/10.1186/s12859-018-2264-5
21. Breiman L, "Random Forests", Machine Learning, 45, 5–32 **(2001)**. https://doi.org/10.1023/A:1010933404324
22. Shrestha Ajay, Mahmood Ausif, "Improving Genetic Algorithm with Fine-Tuned Crossover and Scaled Architecture", Journal of Mathematics, **(2016)**, 2314-4629, https://doi.org/10.1155/2016/4015845
23. Bhawna, Gaurav Kumar, Pradeep Kumar Bhatia, "Software Test Case Reduction using Genetic Algorithm: A Modified Approach", International Journal of Innovative Science, Engineering & Technology, Vol. 3 Issue 5, pp. 349-354, May **(2016)**.
24. A. Adel and K. Salah, "Model order reduction using genetic algorithm," 2016 IEEE 7th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), New York, NY, **(2016)**, pp. 1-6, doi: 10.1109/UEMCON.2016.7777856.
25. Vishal R, "Feature selection — Correlation and P-value", https://towardsdatascience.com/feature-selection-correlation-and-p-value-da8921bfb3cf.
26. PravinDhandre, "Selecting Statistical-based Features in Machine Learning application", https://hub.packtpub.com/selecting-statistical-based-features-in-machine-learning-application/, March 14, 2018.
27. Nishant Agarwal, "Linear Regression", https://www.academia.edu/31830761/Linear_Regression.
28. JasonBrownlee, "Develop k-Nearest Neighbors in Python From Scratch" https://machine learningmastery.com/tutorial-to-implement-k-nearest-neighbors-in-python-from-scratch/, Oct 24, 2019
29. S. Chakravarty, P.K. Dash, (2012), A PSO Based Integrated Functional Link net and Interval Type-2 Fuzzy Logic System for Predicting Stock Market Indices, Applied Soft Computing, Elsevier, vol. 12, pp. 931-941.
30. S. Chakravarty, P.K. Dash, (2011), Dynamic filter weights neural network model integrated with differential evolution for day-ahead price forecasting in energy market, Expert Systems with Applications, Elsevier, vol.38, pp.10974–10982.

Authors

Avijit Balabnataray, the BTech. Student from Centurion University has been actively working on research works related to the field of AI & ML project works like Sixth Sense Robot, Automated Elephant System and Rice yield prediction using Machine learning algorithms.

Payal Bhadra, BTech. Did her from Centurion University. She has been actively working on the field of AI, ML and Computer vision by showing the skills with some of the Achievements like Automated Elephant detection system, Sixth Sense Robot and Rice yield prediction.

Rakesh Kumar Ray is pursuing his Ph.D. in Centurion University of Technology and Management. His research area includes Machine Learning and smart in the field of Agriculture.

Dr. Sujata Chakravarty is a Senior Member of IEEE. Currently she is working as Associate Professor and HoD, Dept. of CSE, Centurion University of Technology & Management, Bhubaneswar, India. She is a reviewer of many International journals like Elsevier (Neurocomputing, Knowledge-Based Systems, Econometrics and Statistics, Karbala International Journal of Modern Science), Inderscience, International Journal of Intelligent and Fuzzy Systems and IEEE. She has published 4 book-chapters and about 90 articles in many International journals and conferences. Her research area includes multidisciplinary fields like Application of Computational Intelligence and Evolutionary Computing Techniques in the field of Financial Engineering, Bio-medical data

classification, Smart Agriculture, Intrusion Detection System in Computer-Network, Analysis and prediction of different financial time series data.