A STUDY ON THE IMPACT OF PADDY YIELD WITH WEATHER CONDITIONS IN INDIA USING DATA MINING AND MACHINE LEARNING APPROACHES

J. Karthikeyan[1], Dr. A. Murugan[2]

Article History: Received: 08.05.2023 **Revised:** 20.06.2023 **Accepted:** 15.07.2023

Abstract

Agriculture and accessories contribute to approximately 17% of India's GDP, still the most popular occupation amongst 70% of India's population. The agriculture sector provides different outputs used by diverse segments, including, but not limited to, use as raw materials by various industries, sources of nutrition and businesses, etc. Indian farmer still struggles to pick up the right crop for the proper biological and non-biological factors. Thus, in this case, different machine-learning techniques have been proposed for paddy growth with weather datasets to accelerate the paddy yield of crops. In this paper, we present a summarization of these different approaches: the regression model, random forest, and random tree. These techniques are a part of the paradigm, Precision Agriculture, specifically in paddy data analysis. These algorithms consider and implement external factors, like meteorological data like rainfall and temperature and others like pesticides, to give the best recommendations, which not only lead to better yields but also minimum use of resources and capital.

Index Terms: Precision agriculture, weather conditions, regression model, random tree, and random forest.

[1]Research Scholar, Department of Computer and Information Science, Annamalai University, Annamalainagar – 608 002, Tamil Nadu, India
Email: thalapathik80@gmail.com

[2]Assistant Professor, Department of Computer Science, Periyar Arts College, Cuddalore, (Deputed from Annamalai University, Annamalainagar) Tamil Nadu, India
Email: drmuruganapcs@gmail.com

DOI: 10.31838/ecb/2023.12.6.186

1. Introduction

Agriculture has been the cornerstone of almost all ancient civilizations for the very primary reason of sustenance. Today, agriculture is a $2.4 trillion industry worldwide and is one of the most prominent contributors to developing third-world countries. However, agriculture inherently faces many problems, including highly unpredictable, lack of rain, floods, and drastic weather changes, to name a few. These factors, along with the unmeasured use of insecticides and chemical fertilizers, institutional factors like lack of credit, not enough subsidies by the governing body, and corruption, lead to alienation of the farming body and an increase in the debt, which ultimately leads to suicides and families, laden with more debt. The above reasons necessitate the ushering in of Artificial Intelligence and the Internet of Things [1] in the field of agriculture to use statistical prowess to provide better yields at relatively lower costs. Along these lines, we present various frameworks based on precision agriculture.

Precision Agriculture [2] includes the use of technology in proposing fertilizers, farming techniques, and crops, among other things, to farmers. The frameworks we discuss are focused on a subsection of precision agriculture called crop recommendation systems which use different machine learning algorithms to recommend crops depending on specific rules and data. The correctness of the recommendation depends on the type and the amount of data fed. The statistical nature of these algorithms can lead to a significant increase in yield. A high accuracy measure is desired as the consequences of the otherwise will be immense, which include waste of seeds, time, drastic loss in productivity, etc. Many different predictors can be used for recommendations like temperature, soil properties, humidity, etc. In India, the concept of precision agriculture has yet to be fully embraced, and farmers still use traditional methods of growing crops that do not lead to excellent yield and run a high risk of failure whilst letting a better alternative of technology-based agriculture go unused.

2. Literature survey

Ensemble methods [3] belong to that branch of machine learning that blends several learning methods into one model. The prominent feature of these algorithms is that they give better performance than any of their sub-models. Some examples of these are boosting, bagging, and stacking. Majority Voting is a well-known ensembling technique requiring at least two base learners. Learners are selected in such a way that they are aggressive yet supportive of each other. The primary motive is to achieve a better measure of performance than any of the individual models. This method has been used in [4], [5] for crop prediction. Recommendations are made for a localized area of the Madurai district in Tamil Nadu through a UI-based system [4]. The learners used in this system include Random Tree [6], Naive Bayes [7], KNN [8], and CHAID [9]. The essential features are extracted and introduced to the ensemble model, which generates the induction rules for the recommendation. A very similar approach is adopted in [5], which uses learners Naive Bayes, Linear SVM [10], and Random Forest.

Data mining is acquiring, preparing, and finding patterns in data to gain valuable insights from historical data. It is based on the commonness of machine learning, statistics, and database management. There is a growing need to extrapolate the farmers' existing knowledge by harnessing vast amounts of past information.

Available to us. Such an approach is utilized in [11], which proposes that different types of classification algorithms may behave differently when exposed to

data from other crops. It uses many classification-based algorithms like ANN [10] and KNN (for wheat and potato), Harmonic analysis of NDVI Time-series algorithm (for sugarcane), J48, LAD Tree (for rice), and many others and compares their accuracies. An android application is presented in [12], which leverages the power of ANN and proposes a crop calculator that recommends the best crops based on different values provided, like soil Ph., water availability, temperature, and month of cultivation, in the backend on the web and data server. It also provides a subscription to the farmers, gives personalized information, and takes feedback so the system can be made more user-friendly. A simple comparative study is presented in [13], which compares classification algorithms like J48 [14], LAD Tree [15], LWL, and IBK using the WEKA tool [16] to classify into different seasons, and the preprocessing was done using DSS. The performance is compared based on metrics like RMSE, MAE, and RAE [17].

The RSF Recommender [18] presents a modular approach where different databases are used for storing information related to temperature, seasonal crops, and crop growth rate. The country is divided into regions called Upazilas. It has a UI-based system that detects the location and identifies the Upazilas, which provides information regarding the different thermal, demographic, and physiographic properties. Using this information and the information maintained in the database regarding the seasonal crops, the most similar Upazilas are chosen using Pearson Correlation Similarity, and the top N most similar Upazilas are selected. Finally, these, along with the crop production rates, are used to recommend crops keeping in mind the season by the specialized algorithm.

Big Data analytics [19] studies the examination of considerable datasets to find patterns, inherent collations, and hidden insights. An expert system [20] has a knowledge base that is acquired from domain experts, an inference engine that uses the knowledge and statistical methods to draw interpretations from the database, and an interface that allows input/extraction of commands or rules into/from the knowledge base. An amalgamation of big data analytics and expert systems is presented in [21]; it gives insight into a decision system aided by extensive data analysis for guiding agricultural production. An expert system has a knowledge base that is acquired from domain experts, an inference engine that uses the knowledge and statistical methods to draw interpretations from the database, and an interface that allows input/extraction of knowledge or rules into/from the knowledge base. An intelligent agriculture decision system is presented with different components like knowledge acquisition, knowledge representation and reasoning, knowledge base, and establishing a reasoning program. It is integrated with a Big Data framework which improves the speed and performance of the decision system as a whole. As an instance of the combination of decision system and big data analytics, a wheat agricultural intelligent decision system was presented in the end, which concludes decision-making for a variety of tasks like prevention and control of diseases and pests, green stage decision, selection decision of type, sowing time decision, among others.

A Crop Selection Method (CSM) [22] categorizes the crops as annual (crops that can be grown anytime in the whole year), seasonal (crops that can be grown in a specific season), long-term crops (take a long time to grow) and short-term (take a short time to develop) crops. In many tropical and subtropical countries, including India, the amount of rain received is indicative of the crop yield that can be expected that year. According to the

approach discussed in [22], multiple sequences of crops are possible from the four categories, but only the series which gives the best mean yield is selected. For the selection of the best arrangement of crops, prediction of the yield rate of the crop is imperative and is predicted based on features like water density, weather, soil type, and crop type. Using the expected yields of the crop, multiple sequences are possible; they are filtered, and the best arrangement is selected.

3. Methodologies

In this section, explain the different descriptions of the method is provided. In this paper, consider three tasks were performed, and concepts worked based on various machine learning models and their model accuracy.

3.1 Random Forest

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex issue and improve the performance of the model. As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and, based on the majority votes of forecasts, predicts the final output. The more significant number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

Random Forest works in two phases. The first is to create the random forest by combining the N decision tree, and the second is to make predictions for each tree created in the first phase [23]. The Working process can be explained in the below steps and diagram:

Step-1: Select random K data points from the training set.

Step-2: Build the decision trees associated with the selected data points (Subsets).

Step-3: Choose the number N for the decision trees that you want to build.

Step-4: Repeat Steps 1 & 2.

Step-5: For new data points, find the predictions of each decision tree, and assign the latest data points to the category that wins the majority votes.

3.2 Random Tree

A single decision tree is easy to conceptualize but will typically suffer from high variance, which makes them not competitive in terms of accuracy. One way to overcome this limitation is to produce many variants of a single decision tree by selecting every time a different subset of the same training set in the context of randomization-based ensemble methods [24]. Random Forest Trees (RFT) is a machine learning algorithm based on decision trees. Random Trees (RT) is a class of machine learning algorithms that does ensemble classification. The term ensemble implies a method that makes predictions by averaging over the predictions of several independent base models.

"There are three main choices to be made when constructing a random tree. These are (1) the method for splitting the leaves, (2) the type of predictor to use in each leaf, and (3) the method for injecting randomness into the trees" [25]. A common technique for introducing randomness in a Tree "is to build each tree using a bootstrapped or sub-sampled data set. In this way, each tree in the forest is trained on slightly different data, which introduces differences between the trees" [26]. Randomization can also occur by randomizing "the choice of the best split at

a given node... experiments show however that when noise is important, Bagging usually yields better results" [25].

When optimizing a Random Trees model, "special care must be taken so that the resulting model is neither too simple nor too complex. In the former case, the model is indeed said to underfit the data, i.e., to be not flexible enough the capture the structure between X and Y. In the latter case, the model is said to overfit the data, i.e., to be too flexible and to capture isolated structures (i.e., noise) that are specific to the learning set" [26]. There is then a need to define stropping criteria to stop the growth of a tree before it reaches too many levels to prevent overfitting: "Stopping criteria are defined in terms of user-defined hyper-parameters" [27]. Among those parameters, the most common are:

- The **minimum number of samples** in a terminal node to allow it to split
- The **minimum number of pieces** in a leaf node when the terminal node is split
- The **maximum tree depth**, that is, the maximum number of levels a tree can grow
- Once the **Trees accuracy** (defined by the Gini Impurity index) is less than a fixed threshold

3.3 Linear Regression Model

Simple linear regression is a statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables: (1) One variable, denoted x, is regarded as the predictor, explanatory, or independent variable. (2) The other variable, y, is the response, outcome, or dependent variable. Because the other terms are used less frequently today, we'll use the "predictor" and "response" terms to refer to the variables encountered in this course [28].

$$y=a_x+b \quad \dots (1)$$

3.4 R2 Score

The R2 score or correlation coefficient or the coefficient of determination is a dimension used to explain how important variability of one factor can be caused by its relationship to another affiliated factor. This correlation, called goodness of fit, is represented by values between 0.0 and 1.0. A value of 1.0 indicates a perfect fit and is, therefore, a largely dependable model for future prediction, while a value of 0.0 would indicate that the computation fails to accurately model the data at all. But a value of 0.20, for illustration, suggests that 20% of the dependent variable is predicted using an independent variable. In contrast, a value of 0.50 suggests that 50% of the dependent variables are predicted using the independent variable [28].

$$\mathbf{r} = \frac{\mathbf{n}(\sum \mathbf{xy})-(\sum \mathbf{x})\,(\sum \mathbf{y})}{\sqrt{[\mathbf{n}\sum \mathbf{x^2}-(\sum \mathbf{x})^2]-[\mathbf{n}\sum \mathbf{y^2}-(\sum \mathbf{y})^2]}} \quad \dots (2)$$

3.5 Mean Absolute Error (MAE)

Mean Absolute Error calculates the average disparity between the quantified and actual values. It is also known as scale-dependent precision as it quantifies error in observations taken on an identical scale. It is adopted as an assessment metric for regression patterns in machine learning. It calculates errors between actual values and values foreseen by the model. It is utilized to predict the correctness of the machine learning model.

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} \quad \dots (3)$$

where Σ: Summation, y_i is: Actual value for the i^{th} observation, x_i is the Calculated value for the i^{th} observation, and n is the Total number of observations [29].

3.6 Mean Square Error (MSE)

The mean squared error quantifies how close a regression line is to a set of information points. It is a risk function that corresponds to the expected significance of the squared error loss. The mean squared error is evaluated by taking the average, specifically the mean, of the squared errors from data relevant to a function. A larger MSE shows that the data points are commonly scattered around its central moment (mean), while a smaller MSE suggests the opposite. A smaller MSE is preferred as it demonstrates that your data points are closely spaced around their central moment (mean). It reflects the centralized distribution of your data values, the fact that it is unbiased, and, most importantly, has less error [30].

$$\mathrm{MSE} = \left(\frac{\sum_{i=1}^{n} |y_i - x_i|}{n}\right)^2 \quad \ldots (4)$$

where Σ: for summation, y_i is: the actual value for the i^{th} observation, x_i is: the calculated value for the i_{th} observation, and n is: the total number of observations

3.7 Root Mean Square Error (RMSE)

Root Mean Square Error (RMSE) is the standard deviation of the prediction error. Residuals are a quantification of how far data points are from the regression line; RMSE is a quantification of how distributed these residuals are. It tells you how concentrated the data is around the line of best fit. The mean square error is frequently used in climatology, predicting, and regression model evaluation to verify empirical results [31].

$$\mathrm{RMSE} = \sqrt{\left(\frac{\sum_{i=1}^{n} |y_i - x_i|}{n}\right)^2} \quad \ldots (5)$$

where Σ: for summation, y_i is: the actual value for the i^{th} observation, x_i is: the calculated value for the i_{th} statement, and n is: the total number of observations

3.8 Relative Absolute Error (RAE)

The ratio of the absolute error of the measurement to the specific value is called the relative error by calculating the relative error with the hypothesis of how good the measurement is compared to the actual size. From the relative error, we can ascertain the magnitude of the absolute error. If the specified value is not available, the relative error with regard to the measured significance of the quantity can be calculated. The relative error is dimensionless and has no unit. It is written as a ratio by multiplying it by 100. The relative error is computed from the proportion of the absolute error and the specific value of the quantity. If the total error of the measurement is x, the special deal is x0, the quantified value is x, and the relative error is expressed [32].

$$x_r = \frac{\sum_{i=1}^{n} |y_i - x_i|}{\sum_{i=1}^{n} |x_i - \bar{y}|} \quad \ldots (6)$$

where Σ: for summation, y_i is: the actual value for the i^{th} observation, x_i is: the calculated value for the i_{th} observation and n is: the total number of observations and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

3.9 Root Relative Squared Error (RRSE)

The relative root square error (RRSE) is relative to what it would have been if a simple predictor were used. Exactly, this simple predictor is just the average of the actual values. Thus, the relative squared error takes the total squared error and normalizes it by dividing it by the total squared error of the simple predictor. By taking the square root of the relative squared error, one reduces the error to the identical dimensions as the forecasted magnitude. Mathematically, the root of the relative squared error of a single model is estimated by the following equation: [32]

$$x_r = \frac{\sum_{i=1}^{n} (y_i - x_i)^2}{\sum_{i=1}^{n} (|x_i - \bar{y}|)^2} \quad \ldots (7)$$

where Σ: Summation, y_i: Actual value for the i^{th} observation, x_i: Calculated value for the i^{th} observation, and n: Total number of observations and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

Numerical Illustrations

The science of training machines to learn and produce models for future predictions is widely used, and not for nothing. Agriculture plays a critical role in the global economy. With the continuing expansion of the human population, understanding worldwide crop yield is central to addressing food security challenges and reducing the impacts of climate change. Crop yield prediction is an important agricultural problem. The Agricultural outcome primarily depends on weather conditions (rain, temperature, etc.), pesticides, and accurate information about the history of crop yield is essential for making decisions related to agricultural risk management and future predictions [33]. All datasets (publicly available datasets) here are taken from the Food and Agriculture Organization (FAO) [34], World Data Bank (WDB) [35], and Data retrieved from the Ministry of Earth Sciences and India Metrological Department (IMD) [36]. The corresponding dataset comprises 7 attributes and 4048 instances. In this case, taking into consideration that the only country name is India, and the name of the items is Paddy. We reduce two columns, namely the name of the country and the name of the items, repeatedly in all the instances. Finally, after preprocessing, the dataset contains 5 attributes, and 507 samples are not possible to display, then only display 14 cases between 1990 to 2013. The attribute name year is optional for making data mining and machine learning approaches. In this case, the models fit into the data and perform the analysis with accurate parameters.

Table 1. Rice Paddy Yield with Weather Conditions in India

Year	Yield (hg/ha)	Average Rainfall (mm)	Pesticides (Tonnes)	Average Temp (C)
1990	26125	1401.40	75000.00	25.58
1991	26271	1170.20	72133.00	25.85
1992	26092	1102.50	70791.00	25.69
1993	28303	1207.80	66388.00	25.88
1994	28645	1295.30	61357.00	25.75
1995	26972	1242.50	61257.00	25.86
1996	28226	1182.90	56114.00	25.63
1997	28457	1183.10	52279.00	24.92
1998	28805	1208.80	49157.00	25.91
1999	29782	1116.60	46195.00	26.12
2000	28508	1035.40	44957.52	25.98
2001	31158	1100.70	43720.04	25.76
2002	26163	936.00	42482.56	26.66
2004	29756	1106.50	35113.00	26.11
2005	31537	1208.20	35342.00	25.85
2006	31759	1161.70	37423.00	26.34
2007	32924	1179.30	27422.77	26.00
2008	32509	1118.10	14485.33	25.57

2009	32366	953.80	28707.01	26.55
2010	33587	1215.60	40093.69	26.51
2011	35878	1116.30	55540.00	25.53
2012	36909	1054.70	52980.00	25.86
2013	36070	1242.00	45620.00	26.69

Table 2 Basic Descriptive Statistics

Attributes	Min	Max	Mean	StdDev
Yield (kg/ha)	26092.00	36909.00	30295.73	3228.87
Average rainfall (mm)	936.00	1401.40	1153.88	101.27
Pesticides (Tones)	14485.33	75000.00	48459.04	15010.33
Average Temperature (Celsius)	23.26	28.85	26.01	0.911

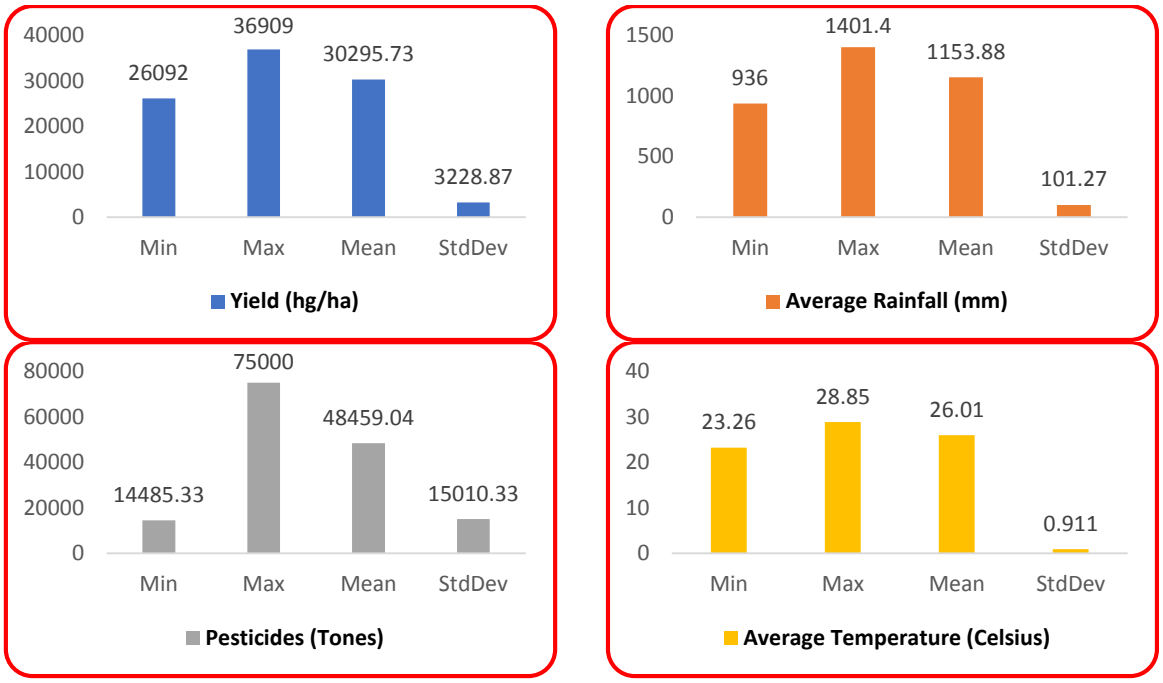

Fig. 1. Descriptive statistics for yield, average rainfall, pesticides, and average temperature

Table 2. Performance Comparison Using R2 Score

Attributes	M5P	Random Forest	Random Tree	REPTree
Yield (kg/ha)	0.9889	1.0000	1.0000	0.9999
Average rainfall (mm)	0.9748	1.0000	1.0000	1.0000
Pesticides (Tones)	0.9850	1.0000	1.0000	1.0000
Average Temperature (Celsius)	0.2251	0.2073	0.2137	0.2431

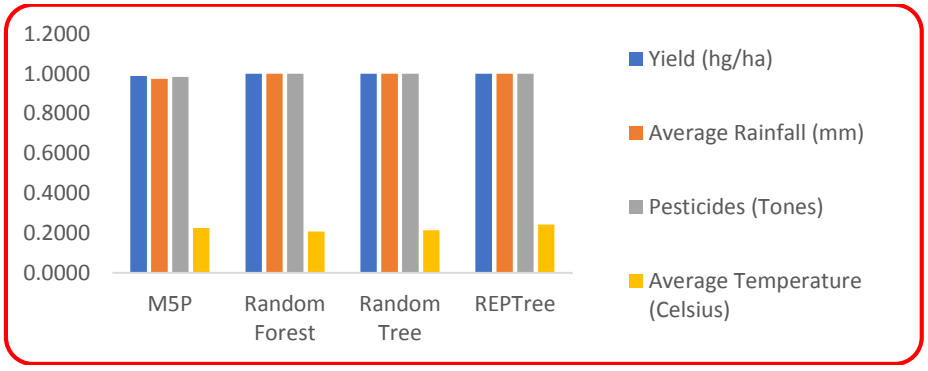

Fig. 2. Performance Comparison Using R2 Score

Table 3. Machine Learning Approaches with Accuracy Performance Using MAE

Attributes	M5P	Random Forest	Random Tree	REPTree
Yield (kg/ha)	493.5799	12.9330	12.3703	15.8671
Average rainfall (mm)	25.1635	0.1756	0.0991	0.0039
Pesticides (Tones)	2190.8100	7.7364	3.5970	4.4073
Average Temperature (Celsius)	0.6992	0.7038	0.7011	0.6930

Table 4. Machine Learning Approaches with Accuracy Performance Using RMSE

Attributes	M5P	Random Forest	Random Tree	REPTree
Yield (kg/ha)	493.5799	12.9330	12.3703	15.8671
Average rainfall (mm)	25.1635	0.1756	0.0991	0.0039
Pesticides (Tones)	2190.8100	7.7364	3.5970	4.4073
Average Temperature (Celsius)	0.6992	0.7038	0.7011	0.6930

Table 5. Machine Learning Approaches with Accuracy Performance Using RAE

Attributes	M5P	Random Forest	Random Tree	REPTree
Yield (kg/ha)	17.8427	0.4675	0.4472	0.5736
Average rainfall (mm)	32.1520	0.2244	0.1267	0.0050
Pesticides (Tones)	17.9543	0.0634	0.0295	0.0361
Average Temperature (Celsius)	96.1931	96.8312	96.4617	95.3436

Table 6. Machine Learning Approaches with Accuracy Performance Using RRSE

Attributes	M5P	Random Forest	Random Tree	REPTree
Yield (kg/ha)	20.1011	0.8793	0.9675	1.1023
Average rainfall (mm)	29.3683	0.8640	0.5201	0.0200
Pesticides (Tones)	18.8683	0.2220	0.0897	0.0994
Average Temperature (Celsius)	97.5097	99.0504	98.7949	97.2824

4. Results and Discussions

Based on the results and discussions, table 1 indicates rice paddy yield with weather conditions in India, which describes the paddy yield, average temperature, pesticides, and average temperature. Based on descriptive statistics like mean and standard deviations, we have summarized the related results shown in Table 2 and Figure 1. This table includes four statistical parameters, namely minimum, maximum, mean, and standard deviations.

This research considers four different machine learning approaches, namely M5P, random forest, random tree, and REP tree approaches which are used to predict the best parameters. In this case, find out the corresponding machine learning approaches and their performance metrics shown in different tables between Table 2 to Table 6 and the related visualization presented in Figure 2. Numerous machine learning techniques are used to recommend precision agriculture, namely crop yield using weather conditions.

Different models can be better for different crops in terms of accuracy, namely R2 score, MAE, RMSE, RAE, and RRSE. Compare the performance of four very commonly used machine learning models, namely M5P, random forest, random tree, and REP tree, to classify results. We have calculated the performance of these models on the dataset of India in terms of Relative Absolute Error (RAE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Root Relative Squared Error (RRSE) as performance matrices in various tables and figures, respectively.

Numerical illustrations in Table 2 and Figure 2 show the machine learning approaches with accuracy parameters, namely R2 score, return very strong positive correlation M5P, random forest, random tree, and REP tree approaches. In this case, for using four parameters except average temperature remaining three parameters, namely paddy yield, average rainfall, and Pesticides, return 0.9889 to 1 and perform only 1% return negative correlation for using average temperature.

ML approaches with accuracy parameters, namely MAE, RMSE, RAE, and RRSE, are used to find the performance. In this case, using random forest, random tree, and REPTree returns a minimum error. M5P returns maximum error compared to others. The related numerical illustrations are shown in Table 3. The RMSE returns a low error rate for using random forest, random tree, and REP Tree, and the related results are shown in Table 4.

Similarly, the performance metrics, namely RAE and RRSE, return a low minimum error rate for using random forest, random tree, and REP Tree with three parameters, namely paddy yield, average rainfall, and pesticides. In this case, the M5P approach returns a maximum error rate compared to the other three decision three systems. The related results and discussion are shown in Table 5 and Table 6.

5. Conclusion and & Future Work

This research clearly indicates 99% to find the future prediction for using three parameters, namely paddy yield, average rainfall, and pesticides. The variable selection process was also performed using ML approaches. In the future, implantation adds some other weather parameters and ML approaches to enhance the model accuracy and variable selections with higher test statistics accuracy.

6. References

1. S. Singh and N. Singh, "Internet of Things (IoT): Security challenges, business opportunities & reference architecture for E-commerce," Proc. 2015 Int. Conf. Green Comput. Internet Things, ICGCIoT 2015, pp. 1577–1581, 2016.

2. S. Babu, “A software model for precision agriculture for small and marginal farmers,” c2013 IEEE Glob. Humanit. Technol. Conf. South Asia Satell. GHTC-SAS 2013, pp. 352–355, 2013.
3. Y. Ren, L. Zhang, and P. N. Suganthan, “Ensemble Classification and Regression-Recent Developments, Applications and Future Directions [Review Article],” IEEE Comput. Intell. Mag., vol. 11, no. 1, pp. 41–53, 2016.
4. S. Pudumalar, E. Ramanujam, R. H. Rajashree, C. Kavya, T. Kiruthika, and J. Nisha, “Crop recommendation system for precision agriculture,” 2016 8th Int. Conf. Adv. Comput. ICoAC 2016, vol. 6, no. V, pp. 32–36, 2017.
5. N. H. Kulkarni, G. N. Srinivasan, B. M. Sagar, and N. K. Cauvery, “Improving Crop Productivity Through A Crop Recommendation System Using Ensembling Technique,” 2018 3rd Int. Conf. Comput. Syst. Inf. Technol. Sustain. Solut., pp. 114–119, 2019.
6. A. Gupta and P. Jain, “A Map Reduce Hadoop Implementation of Random Tree Algorithm based on Correlation Feature Selection,” Int. J. Comput. Appl., vol. 160, no. 5, pp. 41–44, 2017.
7. G. Singh, B. Kumar, L. Gaur, and A. Tyagi, “Comparison between Multinomial and Bernoulli Naïve Bayes for Text Classification,” 2019 Int. Conf. Autom. Comput. Technol. Manag., pp. 593–596, 2019.
8. S. Taneja, C. Gupta, S. Aggarwal, and V. Jindal, “MFZ-KNN-A modified fuzzy based K nearest neighbor algorithm,” Proc. - 2015 Int. Conf. Cogn. Comput. Inf. Process. CCIP 2015, pp. 1–5, 2015.
9. Z. Mingqi and S. Zhijun, "Research on CHAID decision tree model based on the rating of China's small enterprises," 2008 Int. Conf. Wirel. Commun. Netw. Mob. Comput. WiCOM 2008, pp. 1–5, 2008.
10. S. Umadevi and K. S. J. Marceline, "A survey on data mining classification algorithms," Proc. IEEE Int. Conf. Signal Process. Commun. ICSPC 2017, vol. 2018-Janua, no. July, pp. 264–268, 2018.
11. Y. Gandge and Sandhya, “A study on various data mining techniques for crop yield prediction,” Int. Conf. Electr. Electron. Commun. Comput. Technol. Optim. Tech. ICEECCOT 2017, vol. 2018-Janua, pp. 420–423, 2018.
12. R. Shirsath, N. Khadke, D. More, P. Patil, and H. Patil, “Agriculture decision support system using data mining,” Proc. 2017 Int. Conf. Intell. Comput. Control. I2C2 2017, vol. 2018-Janua, pp. 1–5, 2018.
13. S. Mishra, P. Paygude, S. Chaudhary, and S. Idate, “Use of data mining in crop yield prediction,” Proc. 2nd Int. Conf. Inven. Syst. Control. ICISC 2018, no. Icisc, pp. 796–802, 2018.
14. R. Patil and V. M. Barkade, “Class-Specific Features Using J48 Classifier for Text Classification,” Proc. - 2018 4th Int. Conf. Comput. Commun. Control Autom. ICCUBEA 2018, pp. 1–5, 2018.
15. S. R. Kalmegh, “Comparative Analysis of WEKA Data Mining Algorithm RandomForest, RandomTree and LADTree for Classification of Indigenous News Data,” Int. J. Emerg. Technol. Adv. Eng., vol. 9001, no. 1, pp. 507–517, 2008.
16. The University of Waikato, “WEKA: Waikato Environment for Knowledge Analysis,” Proc. New Zeal. Comput. Sci. Res. Students Conf., pp. 57–64, 2014.

17. A. Botchkarev, "Performance Metrics (Error Measures) in Machine Learning Regression, Forecasting, and Prognostics: Properties and Typology," pp. 1–37, 2018.
18. M. J. Mokarrama and M. S. Arefin, "RSF: A recommendation system for farmers," 5th IEEE Reg. 10 Humanit. Technol. Conf. 2017, R10-HTC 2017, vol. 2018-Janua, pp. 843–850, 2018.
19. A. Londhe and P. P. Rao, "Platforms for Big Data Analytics: Trend towards Hybrid Era," 2017 Int. Conf. Energy, Commun. Data Anal. Soft Comput., pp. 3235–3238, 2017.
20. C. Yau and A. Sattar, "Developing expert system with soft systems concept," Proc. IEEE Int. Conf. Expert Syst. Dev., pp. 79–84, 1994.
21. J. C. Zhao and J. X. Guo, "Big data analysis technology application in agricultural intelligence decision system," 2018 3rd IEEE Int. Conf. Cloud Comput. Big Data Anal. ICCCBDA 2018, pp. 209–212, 2018.
22. R. Kumar, M. P. Singh, P. Kumar, and J. P. Singh, "Crop Selection Method to maximize crop yield rate using machine learning technique," 2015 Int. Conf. Smart Technol. Manag. Comput. Commun. Control. Energy Mater. ICSTM 2015 - Proc., no. May, pp. 138–145, 2015.
23. https://www.javatpoint.com/machine-learning-random-forest-algorithm
24. Breiman, Leo. 2001. Random Forests. Machine Learning. Vol-45, p.5-32.
25. Denil, Misha., Matheson, David., de Freitas, Nando. 2014. Narrowing the Gap: Random Forests in Theory and in Practice. Proceedings of the 31st International Conference on Machine Learning, Beijing, China. JMLR: W and P. Vol.32. 9 pages.
26. Louppe, Gilles. 2014. Understanding Random Forests, From Theory to Practice. The University of Liège. Faculty of Applied Sciences. Department of Electrical Engineering and Computer Science. 223 pages.
27. https://catalyst.earth/catalyst-system-files/help/concepts/focus_c/oa_classif_intro_rt.html#:~:text=Random%20Forest%20Trees%20(RFT)%20is,of%20several%20independent%20base%20models.
28. Rajesh, P., and Karthikeyan, M., 2017. A comparative study of data mining algorithms for decision tree approaches using the WEKA tool. Advances in Natural and Applied Sciences, 11(9), pp. 230-243.
29. Albahri, A.S., Hamid, R.A., Al-qays, Z.T., Zaidan, A.A., Zaidan, B.B., Albahri, A.O., AlAmoodi, A.H., Khlaf, J.M., Almahdi, E.M., Thabet, E. and Hadi, S.M., 2020. Role of biological data mining and machine learning techniques in detecting and diagnosing the novel coronavirus (COVID-19): a systematic review. Journal of medical systems, 44(7), pp.1-11.
30. Abdulkareem, N.M., Abdulazeez, A.M., Zeebaree, D.Q. and Hasan, D.A., 2021. COVID-19 world vaccination progress using machine learning classification algorithms. Qubahan Academic Journal, 1(2), pp.100-105.
31. Rustam, F., Reshi, A.A., Mehmood, A., Ullah, S., On, B.W., Aslam, W. and Choi, G.S., 2020. COVID-19 future forecasting using supervised machine learning models. IEEE Access, 8, pp.101489-101499.
32. Sun, N.N., Yang, Y., Tang, L.L., Dai, Y.N., Gao, H.N., Pan, H.Y. and Ju, B., 2020. A prediction model based on machine learning for diagnosing early COVID-19 patients. MedRxiv.
33. https://www.kaggle.com/code/omarkhald/crop-yield-regression/input
34. https://www.fao.org/home/en/
35. https://data.worldbank.org/

36. http://www.indiaenvironmentportal.org.in