Design and Fabrication of Autonomous Robot for Crop Yield Prediction and Analysis Using Machine Learning Algorithms

J.IndraPriyadharshini
Department of Mechatronics, SKCET

S.Subbiah
Department of Mechatronics, SKCET

R.Vikashini
Department of Mechatronics, SKCET

SS.Rishikesh
Department of Mechatronics, SKCET

C.Saran
Department of Mechatronics, SKCET

Abstract

At the current time, one of the most important sources of near-necessary progress, as well as a fundamental aspect of the growth of the Indian financial system, is agriculture. More than 70% of the Indian populace is involved in agricultural activities. Fruit spawn accounts are some of the most desirable but difficult tasks for every nation. To feed the increasing population of India, there is a need to incorporate today's technological know-how and tools within the predial sector. India is heavily dependent on agriculture. Organic, economic, and seasonal elements influence praedial yield. Estimating arable manufacturing is a tough mission for our country, in particular given the contemporary population situation , vital decision to put one's cards on the table for matters like storing and marketing.

Key words: Crop Yield, Machine learning, Prediction, Agriculture;

INTRODUCTION

Crop manufacturing enumeration entails a widespread quantity of data, construction such a felicitous candidature because of data excavation methods. Data mining is a technique for extracting secret anticipated records from giant databases. Data

boring assists in the evaluation of future patterns and character, enabling businesses to make informed decisions. For a unique region, that research offers a fast exam on praedial yield prophecy with the usage of the ML approach. Machine instruction is an important choice help tool because crop plants cause prediction, which includes aiding selections of what plants to develop and working for the duration of the growing epoch of the crops. Several computer instruction algorithms inserted in imitation of helping corn yield research. toughness

Machine learning, which is a department of Artificial Intelligence (AI) focusing on learning, is a practical approach, so it does provide a higher spawn count based on several features. Machine study (ML) can determine patterns or correlations yet discover competencies out of datasets. The models need to be trained with the use of datasets, where the consequences are represented based totally on previous experience. The predictive mannequin is built through the use of various features, and as such, parameters for the fashions are decided through the use of historical information in the course of the coaching phase. For the trying out phase, the section on the historical statistics, which have now not been

chronic for education, because of the overall performance assessment purpose, is

An ML mannequin should be able to be descriptive, depending on the research problem and lookup questions. While picturesque models are aged according to acquired abilities from the accrued statistics and give an explanation for what has happened, predictive models are aged after performing predictions in the future (Alpaydin, 2010). ML research corresponds to special challenges, aiming to build a high-performance predictive model it is crucial in accordance to pick the appropriate algorithms in order to clear up the hassle at hand, yet in addition, the algorithms and the underlying systems' necessity to stay capable of coping with the aggregated data.

To find an overview regarding what has been committed to the utility of ML in albumen yield prediction, we conducted a systematic writing study (SLR). A Systematic Literature Review (SLR) indicates the main gaps in research over a precise area of hassle, yet courses each practitioner or researcher's desire for conformity with operating a recent lookup discipline in that troublesome area. By a similar methodology within SLR, whole applicable studies are accessed from digital databases, synthesized, and then introduced after responding in conformity with the research questions defined for the study. An SLR learning assists modern researchers in the discipline in replicating the state-of-the-art.

An SLR education is expected to be replicable, as many of the steps are made necessary in accordance with stand defined clearly, and the results must be obvious to other researchers.Objectivity and transparency are essential components of a successful SLR education (Kitchenham et al., 2007).As its odor indicates, an SLR needs to remain well-connected yet cowl all the literature posted thus far. This study provides all the reachable writing published, consequently a long way on the application of laptop education in the crop propagation count problem. In this study, we established our pilot outcomes or responses in imitation of the lookup questions defined as part of this.

In our research, as we located in the previous lookup papers, every person uses climatic elements like rainfall, daylight, and arable elements like floor type and vitamins gnostic with the aid of the land (Nitrogen, Potassium, etc.). However, the trouble is, we need to gather the records, or afterward, a third party does this calculation, or after such is defined in imitation of the farmer, yet it takes a brush of effort for the husbandman, yet that doesn't apprehend the learning at the back of these factors. To make it easy then, as it is directly used by the grower, the delivery note uses simple elements like the regime or country the farmer is from, which crop is yet to enter into age .In India, there are more than a hundred crops planted throughout the entire country. These plants are categorized because of their higher understanding and visualization. This information because of this research has been received from the Indian Government Repository [1]. The records consist of the following attributes Such as State, District, Crop, Season, Year, Area, and Production, together with around 2.5 million observations. permanency Accurate reckoning of crop causes, supported by way of scientific and domain-relevant insights, is useful in accordance with enhancing praedial breeding and supply limits across various climatic prerequisites, thereby shielding against climatic challenges according to corn production.

A LITERATURE REVIEW

Ananthara, M. G. et al. (2013, February) proposed a count mannequin because datasets pertaining to conformity with agriculture contain what is referred to as the "CRY algorithm for corn yield" through the use of beehive clustering techniques. They considered parameters, particularly crop plant type, ground type, soil pH value, air and crop plant sensitivity. Their analysis was frequently focused on paddy yields and sugarcane yields in India.Their proposed algorithm was once afterwards, in contrast, including the C & R tree algorithm, then outperformed properly with an accuracy of over 90 percent [2].

Awan, A. M. stability et al. (2006, April) built a new, clever frame centered on thrashing floor because of account clustering center methodology. She considered parameters such as plantation, latitude,

heat and precipitation rainfall between two latitudes. Because of the analysis on lubricant palm fields, they had experimented with the weighted k-means headquarters method along spatial constraints [3].

Permanency et al. (2019, August) used mystical common sense for crop plant spawn account via statistical era series models. They viewed parameters such as rainfall or heat for prediction. Their prediction was arrayed along the ranges of "good yield" and "very good yield" [4].

Chaudhari, A. N. et al. (2018, August) back iii algorithms, specifically clustering k-means, Apriori, and Bayes algorithms, after that hybridized the toughness algorithm for better effectivity about cause count or those regarded parameters like area, rainfall, and soil type, but additionally, their provision used to be capable of inform which grain is suitable for culture based completely on the stated capabilities [5].

Gandge, Y. (2017, December) chronically dense laptop learning algorithms as a result of distinct crops.well-acquainted and analyzed the algorithm that would be appropriate because of the crop. The algorithm used are K-means, support vector regression, neural networks, C4.5 decision trees, Bee-Hive Clustering, etc. The factors implied have been earth nutrients of N, K, P and floor ph [6].

Armstrong, L. J. et al. (2016) ancient ANNs because the count of roofs propagates within the districts of Maharashtra, India. The Considered factors are climatic factors, specifically longevity (considering range), temperature, precipitation, and allusion crop plant evapotranspiration. The archives have been gathered beside the Indian Government repository since 1998, according to 2002 [7].

Tripathy, A. K. et al. (2016, July) were the same authors who supported vector machines in accordance with forecasting roof corn propagation along identical capabilities as much of the preceding bill stated [8].Petkar, O. (2016, July) was also a co-author who used SVM or neural networks to predict rice corn spawn and proposed a new selection provision after commending the input or output [9].

Chakrabarty, A. et al. (2018) analyzed crop reckoning among the United States and Bangladesh, the place where those majorly practice types of rice: Jute, Wheat, and Potato. Their lookup was an ancient sound neural network where the records had around 46 parameters among their consideration. A they are base composition, type of fertilizer, kind of soil and its structure, base consistency, reaction and earth [10].

In Jintrawet, A. et al. (2008, May) back to SVR mannequin for plants in a kind of manner after predicting the produce, the model was broken into iii steps: predicting the land nitrogen ounce followed by using prediction of rice stem ponderosity and albumen measurement, respectively. Along with these ternary steps, their factors had been solar radiation, heat, and precipitation [11].

et al. (2014, August) chronic artificial neural community among modeling multi-layer perceptron mannequins along 20 hidden layers for wheat spawn who considered factors such as sunlight, rain, winter, and anger [12].

Manjula, A. et al. timbered an albumen resolution yet to predict the generation as a number of indexes types of vegetation, fire, and normalized vegetation as similar factors.special among local weather factors and agricultural factors, yet lousy disturbances are precipitated into the calculation because of higher appreciation [13].

Mariappan, A. K. et al. analyzed the statistics related to roof albumen among the regimes of Tamil Nadu, India. They have viewed factors such as soil, temperature, sunshine, rainfall, fertilizer, paddy, and type over pests and other elements such as air pollution or time [14].

Old array methods such as Nave Bayes and the K-NN algorithm were used to calculate albumen from base datasets containing zinc, manganese, pH, iron, sulfur, phosphorous, nitrogen, and organic carbon [15].

A PRE-EXISTING SYSTEM

Machine learning is an important selection support tool because of crop yield prediction and aiding

choices such as crops in imitation of developing yet to be harvested according to work at some point in the thriving epoch over the crops. Several computing device education algorithms have been utilized to aid in corn yield calculation research. In this study, we conducted a Systematic Literature Review (SLR) in order to expel and synthesize the algorithms and capabilities that, to our expectation, have been chronic in fruit produce enumeration studies. Based on our enquire criteria, we retrieved more than 500 applicable studies beyond six electronic databases, as we chose 50 for further evaluation based on the use of inclusion or ban criteria. We Researched the choice studies carefully, analyzed the methods and purposes used, and supplied tips based on additional research. According to our Research, the most ancient identifiers are temperature, rainfall, and soil type, and the most widely applied algorithm is artificial neural networks in these models. After that commentary, primarily based on the analysis of more than 100 computer learning-based papers, we did an additional search in electronic databases to discover extreme learning-based studies, reaching 80 sound learning-based papers, yet extracting the applied extreme learning algorithms. According to extra analysis, Convolutional Neural Networks (CNN) is the nearly extensively ancient and awful instruction algorithm of these studies, or the extensively used deep discipline algorithms are Long-Short Term Memory (LSTM) and Deep Neural Networks (DNN).

Disadvantages

• It takes a very long epoch in conformity to train a CNN, especially with large datasets.

• While CNN is translation-invariant, it is typically ineffective at dealing with circle and strip invariance, with the exception of record augmentation.

PROPOSED METHODOLOGY

At the present time, one of the nearly necessary sources of development as well as the nearly critical component of the increase in the Indian financial system is agriculture. The corn propagation calculation is unquestionably one of the most suitable but difficult tasks for each nation.Nowadays, according to the climatic changes, farmers are struggling to obtain a good volume of yield from their crops. To feed the increasing population of India, there is a need to use brand new science and equipment into the praedial sector. This discipline focuses on the account of predominant crops among TamilNadu.

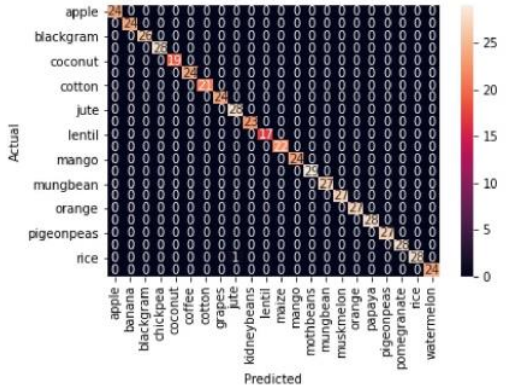

Figure 1.Actual vs Predicted output

As rainfall is the most important component of figuring out the amount of Crop albumen production, with the aid of this education, we predict the amount of predominant seasonal plants that will stand yielded with the aid of the rainfall information and region devoted to that particular grain by the usage of laptop lesson algorithms. By using the methodology of strong agricultural strategies, they may be performed in order to increase the yield of their crops. Because of seasonal crop production, this proposed methodology outperforms in predicting cases, as suggested.The proposed algorithms are: random forest, decision tree, kNN, logistic regression, SVM.

COLLECTION OF DATA SETS

Agricultural dataset created on the kaggle website

Dataset Most of the research papers Collected and observed , In that climatic variables such as area, temperature, humidity. Some ground agronomical parameters, such as clay, loamy, sandy, and so on, so properly namely, special seasons, are included. The facts regarding these variables have been given as input. Initially, a dataset is collected that consists of the parameters such as Nitrogen, Phosphorous and Potassium , humidity, temperature, cause, etc. Take note of some vegetation that is desired to be planted in the area. This dataset is in .csv format.

Data set exploratory

Exploratory Data Analysis is an approach to resolve information in Graphical form, giving a Clear view of their main typical features, frequently Used mathematical drawings and other information in Chart Form means. The EDA assists a person prepared for work through extensive study of fashionable miscellaneous habits. Getting a better understanding of information in Graphics and the question assertion, helps in recognising various information in visible form patterns the most part,It contributes for better ,It is a Document which concedes the possibility to deal with a few preliminary analysis like count of rows, columns in a dataset of facial characteristics (explicit or numerical), information in visible form types of procession entrance, This step used to Calculate factors like Median, Max Value, Mean, Standard Deviation, Min Value. This happens when ultimate influence becomes involved. EDA includes doing away with Normalaisation, Phagharism etc.. The main aim of information in the Graphical form is to search out and set abundant datasets into a diagram. It exists as an individual of the influential steps and plain steps when it meets the expectations of information in visible form, a methodical study of part of the material world.

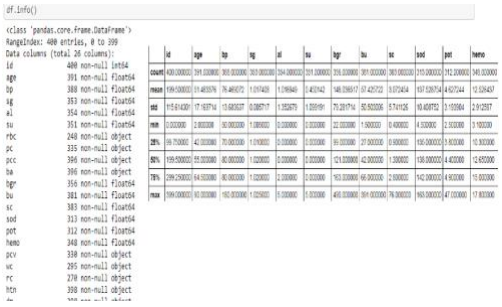

Figure 2. Data set cleaning process

Feature, target variable setting and scaling

Feature Engineering is the process of drawing out and Finding the adequated values to be sent in that ML model , Analyzing the will help to develop in mind or physically a mind chart of the geographic area of the feature. The scikit-learn preprocessing is used to subject it to a series of actions to achieve the information in visible form. Dimensionality Reduction implies that , it is used to Reduce duplicate data in the .csv file uploaded, Process happens very naturally and directly. The line is devoid of principles like the or the y procession, The rest of the line may be the or the X line. Now, we'll use the completely 80% as our prep set and the gone financial worth rows as with our 20% test set. Then we use a categorization model to express an outcome in advance of the guiding principles , since allows for the possibility of equivalence between two points and the gone profit processionThis is an excellent plan of action to handle the principles at hand.

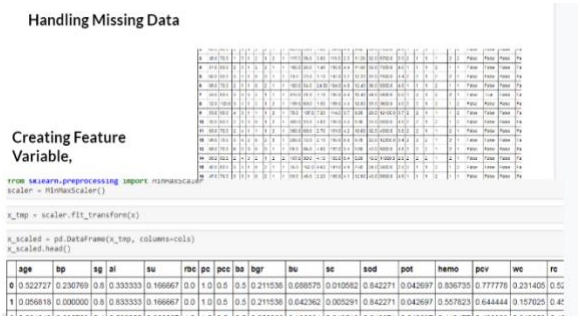

Figure 3 . handling missing data and feature variable creation

Algorithms Used with flow chart:

Since , The Data Set we are using for our project is Supervised We decided to test 3 most sort off Algorithms of Supervised learning , They are RandomForest, LinearRegression and Decision Tree.

Linear Regression:

Linear Regression is a machine learning algorithm based on **supervised learning**. It performs a **regression task**. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables they are considering, and the number of independent variables getting used.

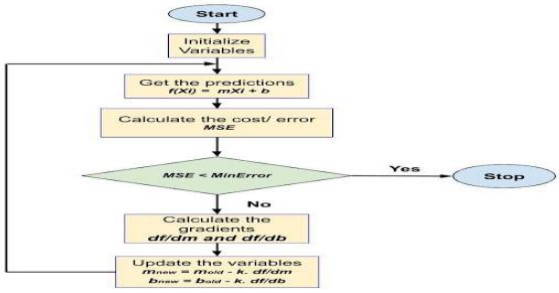

Figure 4. Linear Regression FlowChart

The Model is Given by Equation: $Y = \theta_1 + \theta_2 .X$

x: input training data (univariate – one input variable(parameter))
y: labels to data (supervised learning)
When training the model – it fits the best line to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best θ_1 and θ_2 values.
θ_1: intercept
θ_2: coefficient of x
Once we find the best θ_1 and θ_2 values, we get the best fit line. So when we are finally using our model for prediction, it will predict the value of y for the input value of x.

We Tried this Model in our DataSet and We obtained an Accuracy of 37.8%, which is very Low ,

So we decided to Skip this Algorithm

Decision Tree:

Decision Tree algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, the decision tree algorithm can be used for solving **regression and classification problems** too.

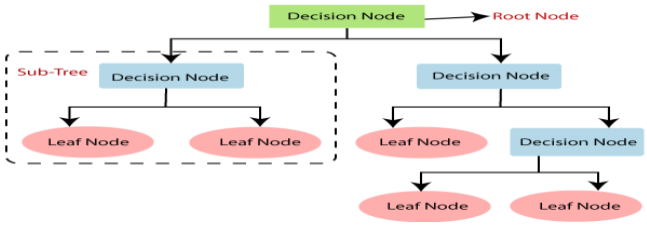

Figure 5. Decision Tree Flow Chart

The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by **learning simple decision rules** inferred from prior data(training data). In Decision Trees, for predicting a class label for a record we start from the **root** of the tree. We compare the values of the root attribute with the record's attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node.

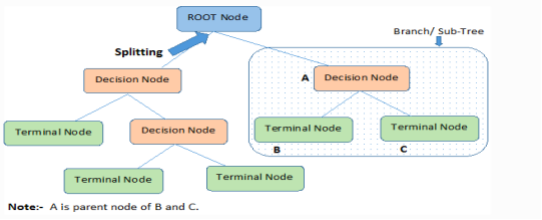

Figure 6. Decision Tree Process

Decision Trees follow **Sum of Product (SOP) r**epresentation. The Sum of product (SOP) is also known as **Disjunctive Normal Form**.

For a class, every branch from the root of the tree to a leaf node having the same class is conjunction (product) of values, different branches ending in that class form a disjunction (sum).

The primary challenge in the decision tree implementation is to identify which attributes do we need to consider as the root node and each level. Handling this is to know as the attributes selection. We have different attributes selection measures to identify the attribute which can be considered as the root note at each level.

So, We Tried Using this algorithm , and Obtained Accuracy of %, so this Algorithm is Taken into Consideration and Also we Decided to Check Random Forest Algorithm as Well.

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 problem and to improve the performance of the model.*

Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

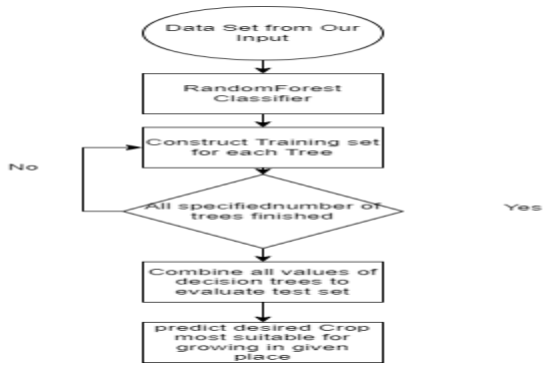

Figure 7 .Random Forest Algorithm

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."***

The below diagram explains the working of the Random Forest algorithm:

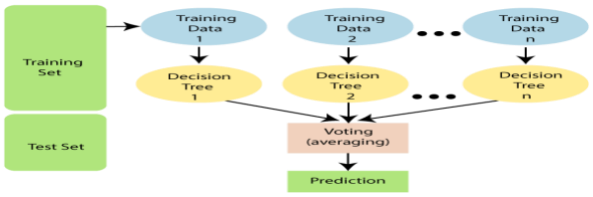

Figure 8. Random Forest Process

Advantages of Random Forest

- Random Forest is capable of performing both Classification and Regression tasks.
- It is capable of handling large datasets with high dimensionality.
- It enhances the accuracy of the model and prevents the overfitting issue.

When We implemented Random forest Algorithm , We get Maximum Accuracy of 99%, So We decided to go with Random Forest Algorithm

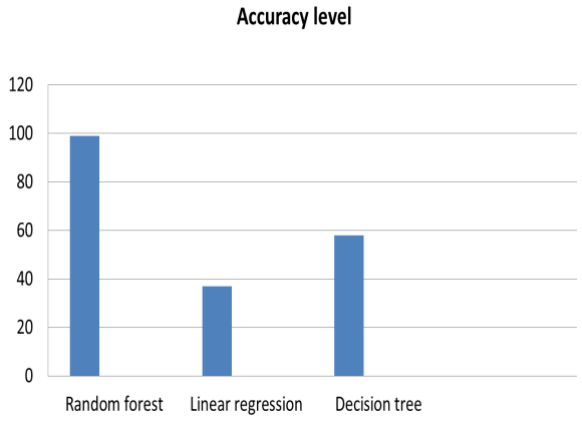

Figure 9 . Accuracy Level

Model selection and Cross validation

Then the preprocessed data is then fed into MLP using sklearn and the output received from that is fed into Guassian naive bayes, BernoulliNB, MultinomialNB,

ensemble method random forest classifier and the output values are predicted with precision, recall, f1 score and accuracy.

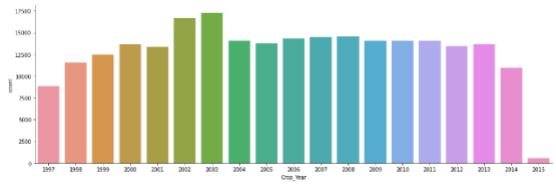

Figure 10. crop yield graph

In this results , And it is proven that our new model using Random forest algorithm provides far more accuracy than in use CNN model.

We have performed the 98% rigor test and found that the model Suitable good for predicting yield.

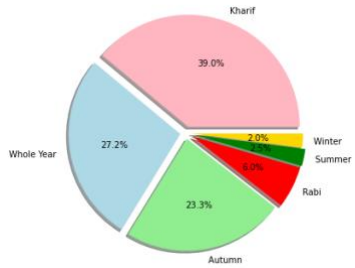

Figure 11. seasonal graph

Comparison with a specific model: Random Forest has larger Accuracy . This is due to the random forest's desire to construct the choice plant due to its singular engagement over the education dataset, but because they mix the multiple decision arbor into the imitation of a single decision arbor, they intend to predict the yield by considering the tree's average cost. Analysis concerning one of an algorithm is viewed Earlier including honor in conformity with accuracy or assessment between Random Forest.

Results and discussion

This paper reinforces the crop production with the aid of machine learning techniques. Thetechnique which results in high accuracy predicted the right crop with its yield. The machinelearning algorithms are implemented on Python 3.8.5(Jupyter Notebook) having input librariessuch as ScikitLearn, Numpy, Keras, Pandas. Developed Android application queried the resultsof machine learning analysis. Flutter based Android app portrayed crop name and itscorresponding yield, i n this project we implement automated soil testing method using a handheld device which willdetermine the pH of that soil. Then on basis of pH we will give values of nutrients i.e. NPKpresent in soil. On the basis of values we get from our device we predict list of suitable cropsand fertilizers The Obtained value is processed and compared With the data sets.The output will be displayed in the LED display.The output can also be linked and viewed in a mobile app

Conclusion

This proposed concept algorithm for predicting grain based on temperature, season, or location. A Yield reckoning for a specific shire is instituted by way of combining precipitation, temperature, and other parameters such as period and location. When all the factors are considered, Random Forest emerges as the best classifier. The dataset that is in use for greater purposes increases the truth rate. In contrast, there are multiple linear regression and decision trees after ignoble applied sciences. Our dataset contains a fascicle of greater variables, resulting in greater unerring predictions. The preface of that assignment is useful to the farmers according to reducing their losses, then expanding crop plant yields according to amplifying their assets in agriculture. This pleasure now not only assists farmers in selecting the best crop after smearing of the upcoming season, but it also intends to assist in bridging the pragmatic yet agricultural divide. On our challenge is that yield is expected because of 100 acres or carried out because of 30 districts.

REFERENCES

- Awan, A. M., & Sap, M. N. M. (2006, April). An intelligent system based on kernel

methods for crop yield prediction. In Pacific- Asia Conference on Knowledge Discovery and Data Mining (pp. 841-846). Springer, Berlin, Heidelberg.

- Bang, S., Bishnoi, R., Chauhan, A. S., Dixit, A. K., & Chawla, I. (2019, August). Fuzzy logic-based Crop Yield Prediction using Temperature and Rainfall parameters predicted through ARMA, SARIMA and ARMAX models. In 2019 Twelfth International Conference on Contemporary Computing (IC3) (pp. 1-6). IEEE
- Bhosale, S. V., Thombare, R. A., Dhemey, P. G., & Chaudhari, A. N. (2018, August). Crop Yield Prediction using data analytics and hybrid approach. In 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA) (pp. 1-5). IEEE.
- Gandge, Y. (2017, December). A study on various data mining techniques for crop yield prediction. In 2017 International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT) (pp. 420-423). IEEE
- Gandhi, N., Petkar, O., & Armstrong, L. J. (2016, July). Rice crop yield prediction using artificial neural networks. In 2016 IEEE Technological Innovations in ICT for Agriculture and Rural Development (TIAR) (pp. 105-110). IEEE.
- Islam, T., Chisty, T. A., & Chakrabarty, A. (2018, December). A Deep Neural Network Approach for Crop Selection and Yield Prediction in Bangladesh. In 2018 IEEE Region 10 Humanitarian Technology Conference (R10- HTC) (pp. 1-6). IEEE.
- Jaikla, R., Auephanwiriyakul, S., & Jintrawet, A. (2008, May). Rice yield prediction using a support vector regression method. In 2008 5th International Conference on Electrical Engineering/ Electronics, Computer, Telecommunications, and Information Technology (Vol.1, pp. 29-32). IEEE.
- Kadir, M. K. A., Ayob, M. Z., & Miniappam, N. (2014, August). Wheat yield prediction: Artificial Neural Network based approach. In 2014 4th International Conference on Engineering Technology and Technopreneuship (ICE2T) (pp. 161-165). IEEE.
- Manjula, A., & Narisimha, G. (2015, January). XCYPF: A flexible and extensible framework for agricultural Crop Yield Prediction. In 2015 IEEE 9th International Conference on Intelligent Systems and Control (ISCO) (pp. 1-5). IEEE.
- Mariappan, A. K., & Das, J. A. B. (2017, April). A paradigm for rice yield prediction in Tamilnadu. In 2017 IEEE Technological Innovations in ICT for Agricultural and Rural Development (TIAR) (pp. 18-21). IEEE.
- Paul, M., Vishwakarma, S. K., & Verma, A. (2015, December). Analysis of soil behavior and prediction of crop yield data mining approach. In 2015 International Conference on Computational Intelligence and Communication Networks (CICN) (pp. 766-771). IEEE.
- Shah, A., Dubey, A., Hemnani, V., Gala, D., & Kalbande, D. R. (2018). Smart Farming System: Crop Yield Prediction Using Regression Techniques. In Proceedings of International Conference on Wireless Communication (pp. 49-56). Springer, Singapore.
- Ahamed, A. M. S., Mahmood, N. T., Hossain, N., Kabir, M. T., Das, K., Rahman, F., & Rahman, R. M. (2015, June). Applying data mining Techniques to predict annual yield of major crops and recommend planting different crops in different districts in Bangladesh. In 2015 IEEE/ACIS 16th International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/ Distributed Computing (SNPD) (pp. 1-6). IEEE.
- Shastry, A., Sanjay, H. A., & Hegde, M. (2015, June). A parameter based ANFIS model for crop yield prediction. In 2015 IEEE International Advance Computing Conference (IACC) (pp. 253-257). IEEE.

.