International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

OPTIMIZING CROP PREDICTION WITH AGRICULTURAL ENVIRONMENTAL CHARACTERISTICS USING FEATURE SELECTION AND CLASSIFICATION MODELS

[1]M. Mounika, [2]Kothi.Mounika

[1]Assistant Professor, Department of MCA Student, Sree Chaitanya College of Engineering, Karimnagar

[2]MCA Student, Department of MCA Student, Sree Chaitanya College of Engineering, Karimnagar

ABSTRACT

Research on agriculture is expanding. In agriculture, crop prediction is especially important and mostly depends on soil and environmental factors including temperature, humidity, and rainfall. Farmers used to be able to choose which crop to raise, keep track of its progress, and select when to harvest it. However, the agricultural community now finds it challenging to continue doing so due to the quick changes in the environment. As a result, machine learning methods have supplanted prediction in recent years, and this study has employed a number of these to calculate crop production. Effective feature selection techniques must be used to preprocess the raw data into a dataset that is easily calculable and machine learning (ML) friendly in order to guarantee that a particular ML model operates with a high degree of precision. Only data characteristics that are highly relevant to defining the model's final output should be used in order to cut down on redundancy and improve the accuracy of the ML model. In order to guarantee that only the most pertinent features are included in the model, optimum feature selection is necessary. Our model will become needlessly complex if we aggregate every single attribute from the raw data without considering how each item contributes to the model-building process. Moreover, extra characteristics that don't add much to the ML model will make it more complex in terms of time and space and have an impact on how accurate the model's output is. The findings show that an ensemble approach outperforms the current classification method in terms of prediction accuracy.

1. INTRODUCTION

Crop prediction in agriculture is a complicated process [1] and multiple models have been proposed and tested to this end. The problem calls for the use of assorted datasets, given that crop cultivation depends on biotic and a biotic factors [2]. Biotic factors include those elements of the environment that occur as a result of the impact of living organisms (microorganisms, plants, animals, parasites, predators, pests), directly or indirectly, on other living organisms. This group also includes anthropogenic factors (fertilization, plant protection, irrigation, air pollution, water pollution and soils, etc.). These factors may contribute to the occurrence of many changes in the yield of crops, cause internal defects, shape defects and changes in the

chemical composition of the plant yield. The shaping of the environment as well as the growth and quality of plants is influenced by a biotic and biotic factors .A biotic factors can be divided into physical, chemical, and other. The recognized physical factors include: mechanical vibrations (vibration, noise), radiation (e.g., ionizing, electromagnetic, ultraviolet, infrared); climatic conditions (atmospheric pressure, temperature, humidity, air movements, sunlight); soil type, topography, soil rockiness, atmosphere, and water chemistry, especially salinity. The chemical factors include: priority environmental poisons, such as sulfur dioxide and derivatives, PAHs; nitrogen oxides and derivatives, fluorine, and its compounds, lead and its compounds, cadmium and its compounds, nitrogen fertilizers, pesticides, carbon monoxide. The others are: mercury, arsenic, dioxins and furans, asbestos, and a atoxins [3]. A biotic factors also include bedrock, relief, climate, and water conditions - all of which affect its properties. Soil-forming factors have a diversified effect on the formation of soils and their agricultural value [4].

Predicting crops yields is neither simple nor easy. The methodology for predicting the area under cultivation is, according to Myers *et al.* [5] and Muriithi [6], a set of statistical and mathematical techniques useful in an evolving and improving optimization process. It also has important uses in design, development, and formulation new as well as improving existing products. Presentation or performance of statistical analysis requires the possession of numerical data. Based on them, conclusions are drawn as to various phenomena and further, on this basis, binding economic decisions can be made. According to Muriithi [6], the better you describe certain phenomena in terms of numbers, the more you can say about them, and with increasing data accuracy you can also obtain more accurate information and make more accurate decisions.

The biggest problem in the temperate climate zone is assessment of agro climatic factors in terms of shaping the yield of winter plant species, mainly cereals. The key factor influencing wintering yield, which provides access to days with a temperature over of 5_ C, their number and frequency, and the number of days in the wintering period with temperatures above 0_C and 5_C. A number of these can be estimated on the basis of public statistics and yield regression statistics in years. Developed models for checking the situation that assess whether they want to be a probation of state policy in the field of intervention in the cereal market. Efficient forecasting of productivity requires forecasting of agro meteorological factors. Aspects related to the variability of these factors may pose a particular problem [7]. Many researchers have dealt with this issue with varying degrees of success [8]_[10].

Grabowska *et al.* [9] predicted narrow-leaf lupine yields for 2050-2060 using weather models and three climate change scenarios for Central Europe: E-GISS model, HadCM3 and GFDL. The _t of the models was assessed by means of the determination coefficient R2, corrected coefficient of determination R2adj, standard error of estimation and the

coefficient of determination R2pred calculated using the Cross Validation procedure. The selected equation was used to forecast lupine yield under the conditions of doubling the CO2 content in the atmosphere. These authors stated that the influence of meteorological factors on the yield of narrow-leaved lupine varied depending on the location of the station. The temperature (maximum, average, minimum) at the beginning of the growing season, as well as rainfall during the _owering - technical maturity period, most often had a significant influence on the yield. It has been shown that the predicted climate changes will have a positive effect on the lupine yield. The simulated protability was higher than observed in 1990-2008, and HadCM3 was the most favorable scenario.

D¡browska-Zieli«ska *et al.* [8] assessed the usefulness of plant biophysical parameters, calculated from the ranges of rejected electromagnetic radiation recorded by the new generation satellites Sentinel-2 and Proba-V, for forecasting crop yields in Poland. In 2016-2018, ground measurements were carried out in arable fields in the area included in the global crop monitoring network GEO Joint Experiment of Crop Assessment and Monitoring JECAM. Classification of crops was performed using optical and radar images Sentinel-1 and RadarSat-2. The prototypical model of Biomass and Evapo transpiration PRO was used to simulate the growth of winter wheat cultivation, to forecast its biomass size. Got high accuracy of 94% of the size of biomass modeled with real biomass.

2. LITERATURE SURVEY

Rice Yield Prediction Using Support Vector Machines, 2019. Support vector machines (SVMs) are now widely used in the world of computer software. Because they can reproduce, they are often employed inside the provision. The development of SVM-primarily based category fashions for Indian rice yield prediction is the focus of this work. The kernel polynomial function of SVM education, okay-cross validation, and the multiple classification technique have all been used in experiments. The Department of Economics and Statistics of the Indian Department of Agriculture provided data on rice output in India for this study. The four-triple cross-validation method produced a great predicted accuracy of 75.06% for the four-year relative average growth. MATLAB software is utilized in this work's experiments.

The protection of American agriculture's food supply and economic growth are significantly aided by agricultural coverage. One of the main challenges in agricultural planning is selecting the crop or crops to plant. It is dependent on a number of factors, such as government coverage, market pricing, and production pace. Numerous scholars have examined crop forecasting, climate forecasting, soil classification, and crop class in order to inform agricultural planning through the application of statistical techniques or system learning approaches. Crop selection becomes a conundrum when there are multiple options for when to plant

and use available land. In order to solve the crop selection problem, optimize crop internet at a certain time in the yield duration, and ultimately obtain the greatest financial increase of the location, this study employs a technique known as the Crop Selection Method (CSM).

Machine Learning Crop Data System, 2020. As far as we know, India is the second most famous country in the United States. The majority of people in India and around the world work in agriculture. Farmers repeatedly plant the same plants without experimenting with other crop varieties, and they apply fertilizers in unknown amounts without knowing enough about the type and quantity of the fertilizer. As a result, it simultaneously affects yield, causes soil acidity, and damages the top layer. As a result, we developed a system that uses device learning algorithms to raise farmers' productivity levels. Based solely on soil composition and meteorological factors, our method will suggest the greatest crop that is appropriate for a certain soil. The gadget also provides information on the type and quantity of fertilizers needed for seed cultivation. Therefore, farmers may cultivate novel plant varieties, increase yields, and prevent soil pollution by using our equipment.

Give an overview of the usage of machine gaining knowledge of strategies, 2019. In India, one of the most important yet lowest paid jobs are agriculture. By cultivating the most desirable crops, machine learning could change the profitability situation and spur an increase in agriculture. The goal of this paper is to provide a yield prediction through the application of several system study techniques. The mean absolute errors are used to compare the overall performance of those approaches. Farmers will be able to choose which crop to grow in order to maximize yield with the help of machine learning algorithms that consider variables such as area, temperature, rainfall, and more.

3. EXISTING SYSTEM

You *et al.* [15] posited an adaptable and precise technique to anticipate yields by employing openly accessible remote sensing data.

The methodology enhances existing procedures in three different ways. To begin with, a remote detecting network is applied to propose a working methodology. Next, a novel dimensionality reduction procedure is presented that uses a convolutional neural network (CNN) alongside long-term memory. Finally, a Gaussian process is used to investigate and examine the spatio-transient structure of the data and enhance its accuracy. Anantha *et al.* [16] implemented a recommendation system using an associate ensemble model with majority voting. The random tree, Chi-square Automatic Interaction Detection (CHAID), kNN, and Naive Bayes (NB) are used as learners to help determine the most appropriate crop, taking into consideration soil parameters, with the results showing high accuracy and potency. The classified image generated by these techniques consists of ground truth-applied

mathematics information Further, it incorporates such data as the parameters of the square measure in terms of the weather and crop yield, as well as state and district-wise crop produce.

All of the above are employed to predict specific crop yields in a given set of circumstances. Rale *et al.* [17] developed a forecasting model which uses the default settings along with RF regression for crop yield production.

Fernando *et al.* [19] studied data on annual coconut production from 1971 to 2001 in a particular region and assessed its economic impact. The research revealed that the loss sustained by the economy in crop shortage terms was around US $50 million. Ji *et al.* [20] advanced an estimation technique to predict rice yields. The study attempted to determine the effectiveness of artificial neural networks (ANN) in predicting rice yield in mountainous regions. It assessed the efficacy of the ANN, relative to biological parametric variations, and compared the efficiency of multiple bilinear regression models with the ANN model. Boryan *et al.* [21] proposed a decision tree-based technique to depict openly accessible state-level crop cover groups, in accordance with guidelines laid down by the Cropland Data Layer (CDL)
and National Agricultural Statistics Service (NASS), and utilizing ground truth collected during the June Agricultural Survey. The proposed work outlines theNASS CDL program.

It presents information dealing with handling strategies, order and approval, precision evaluation, and CDL item particulars, and product cost estimation procedure. Hansen and Loveland [22] proposed the use of Landsat to acquire satellite imagery that facilitates remote sensing of the environment.

Disadvantages

- ❖ The system is not implemented RECURSIVE FEATURE ELIMINATION (RFE).

- ❖ The system is not implemented Sampling techniques which are applied during preprocessing to balance the dataset and maximize the prediction performance.

4. PROPOSED SYSTEM

• Boruta is a random forest-based classification algorithm [38] that involves the voting of versatile unbiased indistinct classifiers in decision trees. The importance of a characteristic is estimated by calculating the loss of classification exactness caused by the random permutation of attributes within objects. The average and standard deviation of the loss of accuracy are calculated, and the average loss is divided by the
standard deviation to obtain the Z score to measure average fluctuations in mean accuracy loss among crops.

A `shadow' attribute is made for each tree by randomly rearranging the values of the initial attributes across objects. The importance of every attribute is determined by analyzing all the attributes in the system. Given the random nature of the fluctuations, the shadow attributes are used as a reference

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

to point to the most important ones. As is to be expected, the degree of accuracy depends greatly on the shadow attributes. Consequently, the values will be re-shuffled constantly to obtain optimal results.

The Boruta algorithm comprises the following steps: 1. The data system, which is extended by affixing copies of all the shadow attributes, is always prolonged by 5 shadow attributes.

2. The added attributes are shuffled with the original attribute to remove any correlation with the response.

3. The Z score is computed by running a random forest algorithm on the widespread information system.

4. The Maximum Z Score Attributes (MZSA) are calculated and any attribute with a value higher than the MZSA is assigned a ``hit".

5. For attributes with undetermined importance, a two-sided test of equality with the MZSA is carried out.

6. Attributes with importance significantly lower than the MZSA are identified as `unimportant' and permanently eliminated from the information system.

7. Attributes with importance significantly higher than the MZSA are marked `important'.

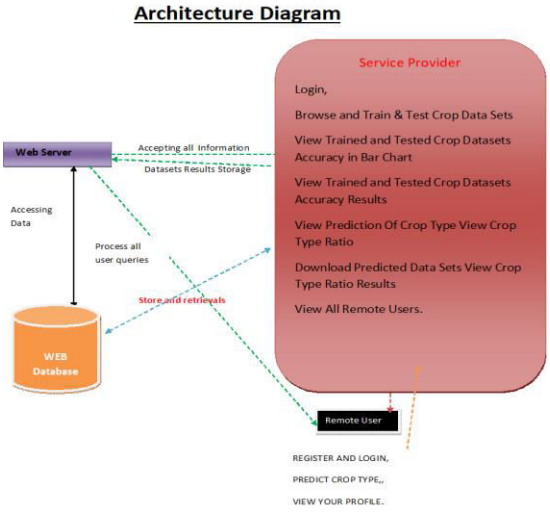
8. Shadow attributes are thus eliminated from the information system.

9. The process is repeated until all attributes are marked with a level of importance.

Advantages

- The RFE technique is a wrapper feature selection technique that starts with the entire dataset. The ranking method crucial to the RFE technique orders the dataset from the best to the worst, based on which salient features are selected.
- The main advantage the RFE has over other methods is that it categorically verifies every feature's role in processing the output of the model and eliminates features only based on their performance.

5. SYSTEM ARCHITECTURE



6. MODULES

Service Provider

In this module, the Service Provider has to login by using valid user name and password. After login successful he can do some operations such as Login, Browse and Train & Test Crop Data Sets, View Trained and Tested Crop Datasets Accuracy in Bar Chart, View Trained and Tested Crop Datasets Accuracy Results, View Prediction Of Crop Type ,View Crop Type Ratio, Download Predicted Data Sets, View Crop Type Ratio Results, View All Remote Users.

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

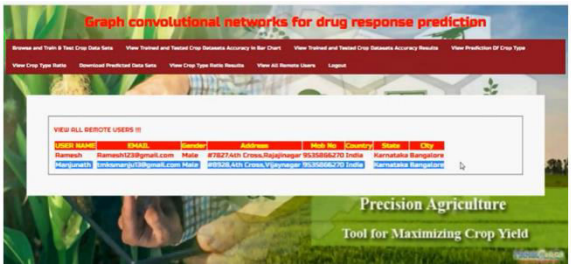
View and Authorize Users

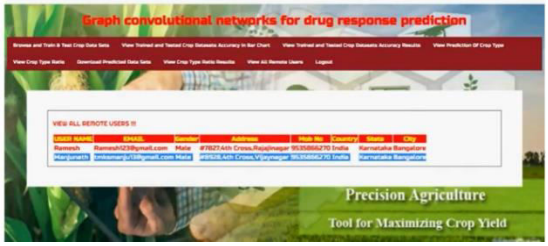
In this module, the admin can view the list of users who all registered. In this, the admin can view the user's details such as, user name, email, address and admin authorizes the users.

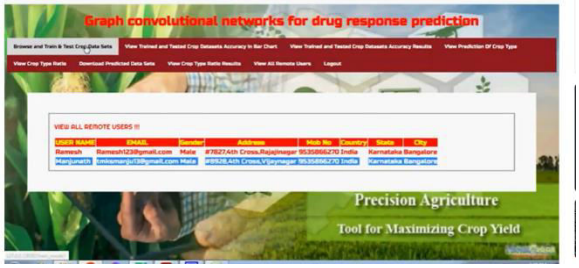
Remote User

In this module, there are n numbers of users are present. User should register before doing any operations. Once user registers, their details will be stored to the database. After registration successful, he has to login by using authorized user name and password. Once Login is successful user will do some operations like REGISTER AND LOGIN, PREDICT CROP TYPE, VIEW YOUR PROFILE.

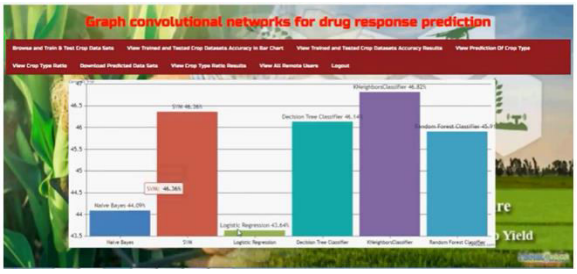
7. RESULTS

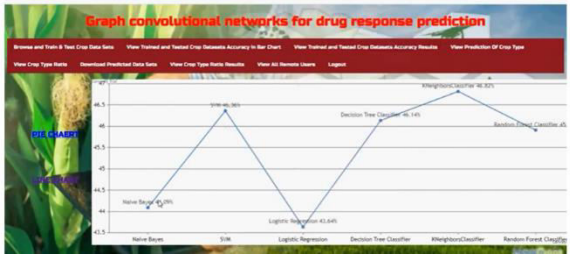


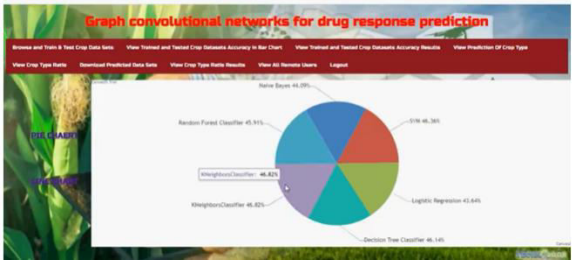


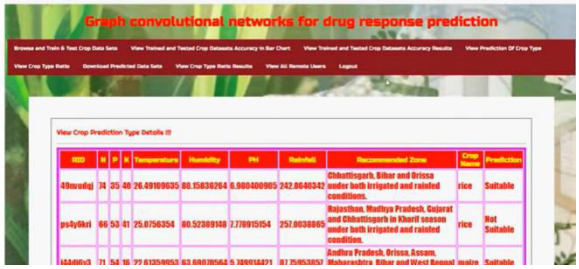


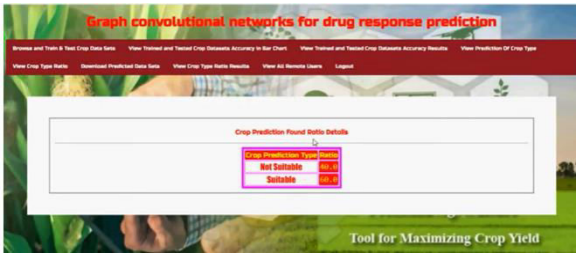
International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
IJARST
ISSN: 2457-0362

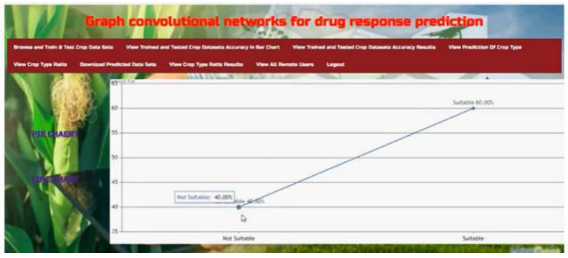












8. CONCLUSION

It's challenging to forecast which crops will be grown in agriculture. The yield size of plant cultivations has been predicted in this research using a variety of feature selection and classification strategies. The findings show that an ensemble approach outperforms the current classification method in terms of prediction accuracy. On a farm and national level, the structure of the planting of grains, potatoes, and other energy crops may be planned by forecasting their area. Utilising contemporary forecasting methods can result in quantifiable financial gains.

REFERENCES

[1] R. Jahan, ``Applying naive Bayes classi_cation technique for classi_cation of improved agricultural land soils," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 6, no. 5, pp. 189_193, May 2018.

[2] B. B. Sawicka and B. Krochmal-Marczak, ``Biotic components in_uencing the yield and quality of potato tubers," *Herbalism*, vol. 1, no. 3, pp. 125_136, 2017.

[3] B. Sawicka, A. H. Noaema, and A. Gáowacka, ``The predicting the size of the potato acreage as a raw material for bioethanol production," in *Alternative Energy Sources*, B. Zdunek, M. Olszáwka, Eds. Lublin, Poland: Wydawnictwo Naukowe TYGIEL, 2016, pp. 158_172.

[4] B. Sawicka, A. H. Noaema, T. S. Hameed, and B. Krochmal-Marczak, ``Biotic and abiotic factors in_uencing on the environment and growth of plants," (in Polish), in *Proc. Bioró»norodno±¢ .rodowiska Znaczenie, Problemy, Wyzwania. Materia ªy Konferencyjne*, Pu ªawy, May 2017.
[Online]. Available: https://bookcrossing.pl/ksiazka/321192
[5] R. H. Myers, D. C. Montgomery, G. G. Vining, C. M. Borror, and S. M. Kowalski, ``Response surface methodology: A retrospective and literature survey," *J. Qual. Technol.*, vol. 36, no. 1, pp. 53_77, Jan. 2004.
[6] D. K. Muriithi, ``Application of response surface methodology for optimization of potato tuber yield," *Amer. J. Theor. Appl. Statist.*, vol. 4, no. 4, pp. 300_304, 2015, doi: 10.11648/j.ajtas.20150404.20.
[7] M. Marenych, O. Verevska, A. Kalinichenko, and M. Dacko, ``Assessment of the impact of weather conditions on the yield of winter wheat in Ukraine in terms of regional," *Assoc. Agricult. Agribusiness Econ. Ann. Sci.*, vol. 16, no. 2, pp. 183_188, 2014.
[8] J. R. Ol¦dzki, ``The report on the state of remotesensing in Poland in 2011_2014," (in Polish), *Remote Sens. Environ.*, vol. 53, no. 2, pp. 113_174, 2015.
[9] K. Grabowska, A. Dymerska, K. Poáarska, and J. Grabowski, ``Predicting of blue lupine yields based on the selected climate change scenarios," *Acta Agroph.*, vol. 23, no. 3, pp. 363_380, 2016.
[10] D. Li, Y. Miao, S. K. Gupta, C. J. Rosen, F. Yuan, C. Wang, L. Wang, and Y. Huang, ``Improving potato yield prediction by combining cultivar information and UAV remote sensing data using machine learning," *Remote Sens.*, vol. 13, no. 16, p. 3322, Aug. 2021, doi: 10.3390/rs13163322.
[11] N. Chanamarn, K. Tamee, and P. Sittidech, ``Stacking technique for academic achievement prediction," in *Proc. Int. Workshop Smart Info-Media Syst.*, 2016, pp. 14_17.
[12] W. Paja, K. Pancerz, and P. Grochowalski, ``Generational feature elimination and some other ranking feature selection methods," in *Advances in Feature Selection for Data and Pattern Recognition*, vol. 138. Cham, Switzerland: Springer, 2018, pp. 97_112.
[13] D. C. Duro, S. E. Franklin, and M. G. Dubé, ``A comparison of pixelbased and object-based image analysis with selected machine learning algorithms for the classi_cation of agricultural landscapes using SPOT-5 HRG imagery," *Remote Sens. Environ.*, vol. 118, pp. 259_272, Mar. 2012.
[14] S. K. Honawad, S. S. Chinchali, K. Pawar, and P. Deshpande, ``Soil classi_cation and suitable crop prediction," in *Proc. Nat. Conf. Comput. Biol., Commun., Data Anal.* 2017, pp. 25_29.
[15] J. You, X. Li, M. Low, D. Lobell, and S. Ermon, ``Deep Gaussian process for crop yield prediction based on remote sensing data," in *Proc. AAAI Conf. Artif. Intell.*, 2017, vol. 31, no. 1, pp. 4559_4565.