IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

PREDICTION OF CROP YIELD USING MACHINE LEARNING BASED ON INDIAN AGRICULTURE

Akash B, Hanish Kumar R
Department of Information Technology
St. Joseph's College of Engineering
Chennai, India

Uthrakumari P
Department of Information Technology
St. Joseph's College of Engineering
Chennai, India

ABSTRACT: *Agribusiness has made a significant global commitment to enhancing the nation's financial responsibility. Nevertheless, due to the lack of organic framework control innovations, the majority of farming areas are still in their early stages of development. These problems prevent enhanced yield creation, which has an effect on the farming economy. From this point on, a redesign of the cultivating benefit is dependent on the assumption of plant yield. Agricultural zones must use ML methodology to predict the production from a given dataset in order to combat this problem. the coordinated use of ML algorithms to examine the dataset. Here, we'll forecast the yield in the event that a particular crop is chosen; otherwise, we'll use the parameters to forecast the yield of every crop. District name, season and year.*

Keywords: *Dataset, MachineLearning-Regression methods, Mean absolute error, R2-score*.

I INTRODUCTION

Our study revealed that the majority of earlier publications made use of meteorological variables like sunlight and rainfall. When concentrating on soil, they employed variables such as crop sensitivity, soil PH, and type. They used factors including soil, sunshine, fertilizer, temperature, rainfall in the paddy, and pests to anticipate the rice production. The farmers may find it challenging to understand these terms, so in order to simplify the process, we are developing an application that predicts crop yield using parameters such as district name, season, and year. This makes it simple for farmers to use the application because the terms are straightforward.

Data scientists typically apply different kinds of machine learning computations to the massive informational sets. We hav e more than hundreds of crops cultivated throughout the entire nation of India. The website data world provided the data used in this study. Certain parameters are included in the data set: state name, district name, crop, season, year, area, and production.

All economies are built on agriculture. Agricultural scientists in elective nations discovered that experiments designed to increase crop productivity by using pesticide-friendly state approaches led to dangerously high drug use. A link between drug usage and crop productivity has been revealed by these investigations.

The field of agribusiness is one that is greatly benefiting from the opportunities presented by information science, artificial intelligence (AI), and machine learning techniques in recent years. These developments are a response to the environmental and population concerns that the public is aware of.

However, reports indicate that strong global agribusiness yield growth is required to produce food for an expanding population on a planet experiencing increased temperatures. The great bulk of research in the field of yield prediction using cubic centimeters makes use of remote sensing data from the homestead. The goal of agribusiness is to increase and diversify crop yields, and consequently, the types of yields that sustain human existence. Nevertheless, in the present era, people will generally need a lot of shot-appreciated places. The number of people concerned about agricultural development is declining. Furthermore, as the population of humans continues to grow, it becomes even more important to develop yields at the right time and place because the environment is dynamic, and deviations from traditional climate design are happening more often than in the past. If the PC code could be created to demonstrate the intelligent influence of environmental components, particularly the effect of major events (such as warmth, rainfall, and excess water) occurring during extremely unique crop development stages, the data gap between antiquated methods of development and new agrarian innovations might be survived. Changes in temperature undoubtedly affect both local and worldwide food production. Consequently, new methods of temperature change analysis, the creation of conducive environments for temperature change transformation, and lawmakers who will curb the severe consequences of climate change on the food supply are needed to build up computer code to present crop predictions. Climate and vermin will eventually cause the dirt to change, thus CEOs will need to deal with an abundance of data that is either directly or indirectly tied to one another. Accordingly, it will consider a practical example to provide a quick assessment of the impact of temperature change in agriculture.

Farming should adapt to these changes in the environment, and it will do so by developing models that will, in theory, improve executive practices, increase the turns of the new produce to cope with soil progressions, and implement innovative rearing projects. An extremely easy technique to find and record the occasional changes in the environment might be achieved by increasing the value of anticipating. Afterwards, one may easily assess the impact of temperature changes and verify scenarios that include observed variations in weather and water distribution by utilizing PC code supported AI. Information {processing} is the process of breaking down test data collected throughout a number of different locations from completely different angles, separating patterns or examples {of information of information on info}, and transform them into client-supporting information. The references, associations, and linkages between this data will also be reactivated into information that is sent to the customer in the form of historical instances and potential trends. The information provided by AI will help ranchers grow their crops by predicting the likelihood of agricultural disasters or completely preventing them.

Ranchers may lessen their losses and obtain the greatest prices for their harvests by using yield forecasts. The most important profession in our nation is agriculture. It is the largest financial sector and has a major role in the overall development of the nation. Harvest determination is a crucial factor in horticultural arrangement. The terrain, climate, kind of soil, and methods of harvesting all affect the rate at which yield is created. Various expectation models use different subsets of these affecting limits for different yields. AI approach is used to predict crop yield using a variety of computations. The AI techniques like relapse tree, irregular forest, convolution neural organization, and K-closest computation are commonly applied in expectation strategy. Making the decision of which particular yield should grow during a certain season on a given piece of land is always fraught with risk for ranchers. It depends on a number of parameters, such as cost, rate of generation, and types of government initiatives. Regardless of the amount of money invested on the land, water, and seed type used to create the yield rate, the harvest may explode, bringing the rancher and his family horrific calamities.

The pre-owned AI technique has been developed to anticipate plant yield or harvest, taking advantage of the different information available in horticulture, including crop, soil, and climatic data. It is suggested to use plant development forecasting to verify the viability of plant output using AI techniques.

II RELATED WORK

Virendra Panpatil et al[1] had made huge work for Indian farms by building profitable yield recommendation schema.

The intended schema is employed to determine the ideal planting season as well as plant fulfillment and advancement. They produced higher occurrence accuracy by using a discrete classifier. The most appealing section of the framework that may be applied to check on different yields in almost any situation.

Mayank et al[2] developed an improvised schema for crop yield using executed AI computations and with target to give effortless to use User Interface, rise the exactness of crop yield estimate, examining discrete climatic parameters.

Zhihao et al[3] two relapse administered AI strategies SVM, RVM to show viability in soil quality forecast. a shrewd remote gadget for detecting soil dampness and meteorological information.

Sabri Arik et al[4] includes a analysis regarding Soil Fertility , Plant Nutrient by utilizing back spread computation . The outcomes are exact and empowers advancement in soil properties.

It works best when contrasted with conventional techniques. Be that as it may, framework is moderate wasteful and not steady.

Shivnath et al[5] proposed about BPN to assess the test informational collection. BPN utilizes a concealed layer which supports in better execution in foreseeing soil properties.
BPN present, is utilized to build up a self-prepared capacity to foresee soil properties with boundaries.

This results in more exactness and executes better compared to the customarily utilized strategies, in any case, at times the framework turns be moderate and irregularity is found in the yield.

Raval et al[6] examine regarding the Knowledge Discovery Process and the rudiments of different Data Mining approaches, for example, Association rules,
Classification etc.,

Agrawal et al examine regarding different Data Mining devices, for example,
Dashboards, Text-Mining instruments. They give an outline about these apparatuses and the different situations where they can be conveyed.

Grajales et al[7] recommended a web app that uses open dataset like verifiable creation, land cover, nearby environment surroundings and coordinates them to give simple admittance to the ranchers. The suggested engineering basically centers around open-source devices for the advancement of the application. The client can choose area for which the subtleties are accessible at a single tick.

Bendre et al[8] gathers information about GIS, GPS, VRT and RS are controlled utilizing Map Reduce calculation and direct relapse calculation to figure the climate information that can be utilized in accuracy agribusiness.

Verma, A. et al[9] used Naïve Bayes and KNN algorithms as classification strategies for crop prediction on soil datasets containing nutrients such as zinc, phosphorus, copper, pH, iron, sulfur, manganese, organic carbon, nitrogen, and potassium.

Chakrabarty, A. et al[10] produced a crop forecast for Bangladesh. They took into account elements such as soil composition, fertilizer type, soil type and structure, consistency, texture, and reactivity.

III PROPOSED SYSTEM

The information must be loaded into the proposed work, checked for null and duplicate values, and then the dataset must be cleaned and trimmed in preparation for analysis. It is important that we properly document our resources and validate your cleaning choices. Applying Machine Learning to any Dataset reaches its most stimulating phase at this point. Another name for it is "algorithm choice" for optimum result prediction. Data scientists typically use different machine learning algorithms on large datasets. However, each of those different computations may, at a deep level, be classified as either of two groups: controlled or individual learning. Managed learning: Supervised learning is a schema in which the required data is provided together with the yield data. Information and yield data are utilized for processing data in the future. Learning problems fall into two categories. Classification and Regression Problems.

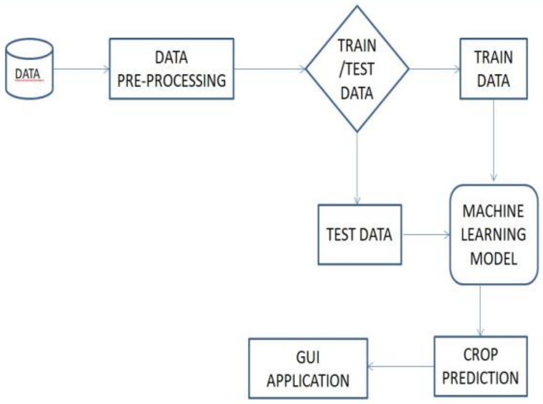


Fig. Architecture of proposed Model

3.1 TRAIN THE DATASET

Initially, we import every module that is needed. Next, we clean our data using the proper cleaning techniques. Next, we divided our data into groups based on the desired outcome. We used the train_test_split approach to divide our dataset into train and test data. Target esteems are indicated by the y prefix, while element esteems are indicated by the X prefix. This divides the dataset into the previously indicated proportions of train and test data. Here, we're employing an 80:20 ratio. We have now embodied every computation. In order to help our system prepare itself, we have included our preparation knowledge into this computation. The training portion is now complete.

3.2 TEST THE DATASET

We utilize part of the dataset's data as test data to gain confidence in our trained model. In this instance, we hav split the dataset in an 80:20 ratio across the train and test sets. By transmitting the feature values from the testing data, we use the trained model to forecast the yield values. We currently use measures like R2 score, mean squared error, and so forth to compare the anticipated and actual values. These measures show us how closely the real data can be predicted. The lower the mean squared, the better. The greater the R2 score, the better.The correlation between expected and real values is determined by the r2 score.

IV MODULES DESCRIPTION

4.1 DATA VALIDATION AND CLEANING

4.1.1 DATA SET DESCRIPTION

We used the dataworld website to obtain our dataset. There are more than 250000 data points in the dataset. This dataset's vast data is really helpful since machine learning models perform better when we have more data. The state, district, region, production, season, and year are among the parameters of the dataset.

4.1.2 DATA CLEANING

We remove the state property from the list of possible characteristics as we will use the district attribute instead. We don't need it for training because the state depends on the district. We are removing rows with negative production and area numbers as they should be positive and are thus inconsistent. Moreover, as these rows may interfere with training, we remove any na, nan, and empty values prior to instruction. Because we are interested in yield per area, we employ a derived attribute production Per Area for the aim of our training model. Since production and area are already contained in our derived goal property Production Per Area, we remove them.

We utilize part of the dataset's data as test data to gain confidence in our trained model. Here, we have split the dataset in a ratio of 67:33 between the test and train data. After training the model, we submit the feature values from the testing data to forecast the yield values. Using measures like the R2 score, mean squared error, and others, we now compare the anticipated and actual values. Our ability to forecast the real data is indicated by these measures. Higher the lower the mean squared. An increased R2 score is preferable. The correlation between expected and actual values is computed using the r2 score.

4.2 DATA NORMALIZATION

4.2.1 PREPROCESSING AND SCALING

We utilize a label encoder to transform string datatype training data into numeric data. The process of label encoding involves allocating a number, commencing from, to every data point. The district name, crop, and season are the attributes that we are label encoding. Next, we employ normalization and the MinMax transform as two different scaling strategies to adjust each data characteristic. Production per region is scaled by normalization, and the district name, crop, and season are scaled using the MinMax transform. To perform the MinMax transform, subtract each datapoint from the minimum value among the datapoints and divide the result by the maximum-minimum.Datapoints are normalized by dividing them by the standard deviation. Since the min max transform is producing decent results and many of the data points are low, we have decided to normalize to production per region.

4.3 RANDOM FOREST

This well-known machine learning approach works well for problems involving both regression and classification. The idea behind this approach is ensemble learning, which increases the accuracy and efficacy of the applied model by combining many classifiers to provide a solution for complex issues. A classifier or regressor called Random Forest uses numerous decision trees on various subsets of a given dataset and averages them to increase the dataset's high accuracy in classification problems and reduce error values in regression issues. Put another way, this algorithm takes into account predictions from each decision tree rather than relying just on one, and it predicts the outcome based on the predictions. On observation we can say that: .

$of\ the\ modelNo\ of\ Trees\ in\ forest \propto Accuracy$

4.4 DECISION TREE

This well-known machine learning approach is applied to problems involving both regression and classification. Three nodes make up its composition. The root node is the first node, or the starting node. While the branches stand in for the decision rules, the interior nodes explain the characteristics of the data collection. The leaf node displays the outcome in the end.

4.5 K-NEAREST NEIGHBOR (KNN)

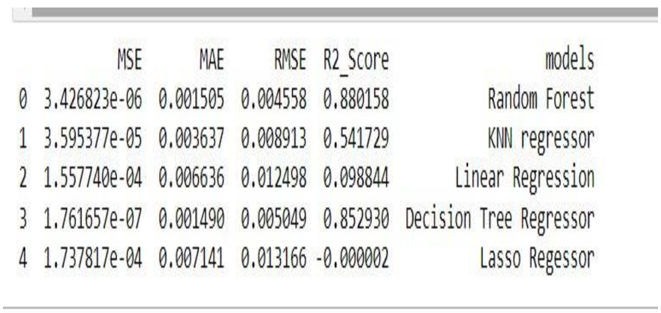
It is a controlled artificial intelligence figure that maintains all occurrences associated with n-dimensional planning data centers. In order to choose the most widely accepted class, it divides the closest k saved cases into groups and returns the mean of the k nearest neighbours for authenticated data. When a neighbour request is made with a distance weight, all k neighbours' responsibilities are stacked according to their distance, with the closest neighbours bearing a greater share of uses the whole preparation set to make assumptions about the endorsement set. KNN finds the k "closest" occurrences by scanning the full set in order to estimate the value of another instance. "Closeness" is determined by evaluating closeness across all features.

4.6. LASSO REGRESSION

When determining the mean, data points are decreased towards a midpoint in a process known as lasso regression, a type of linear regression that makes use of depreciation. This specific model is quite helpful and fits well with models that exhibit significant degrees of multicollinearity, as this dataset does. L1 regularization is performed by it. In Lasso solutions, the quadratic programming problems $\lambda\Sigma|\beta j|pj=1\ \ jni=1\ \ yi\ \ xij\beta j$ are solved using the following formula: Σ(~− ɰ)2+

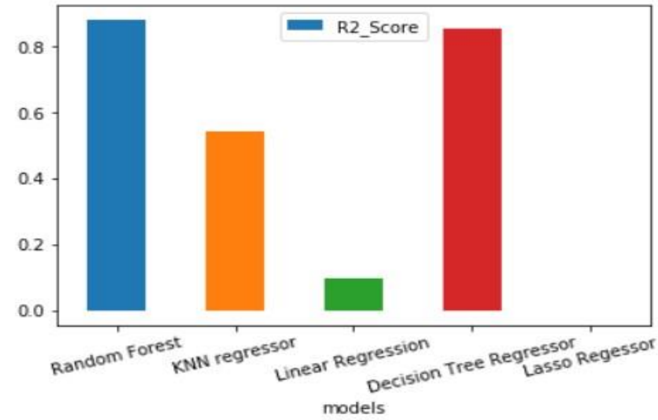
A tuning parameter regulates the L1 penalty's strength It basically describes how much the object is contracting. It's analogous to linear regression, where a predictive model is formed using just the residual sum of squares. If it is 0, we know that all characteristics are considered. It indicates that no characteristics are considered if the value is infinite. As λ grows, the bias rises while the variance falls, and vice versa.

V Result and Discussion

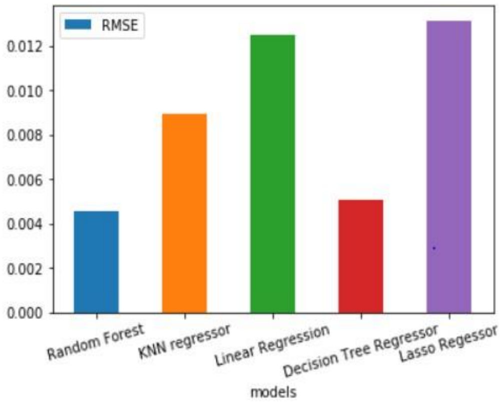


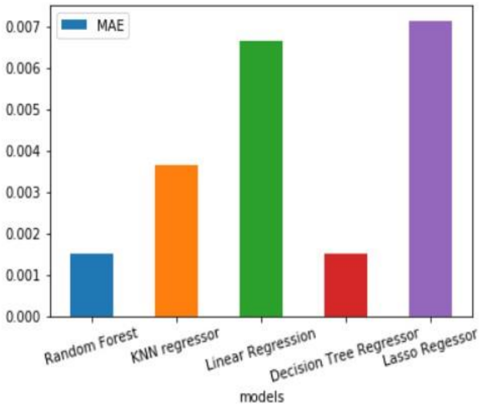
```
            MSE       MAE      RMSE  R2_Score                   models
0  3.426823e-06  0.001505  0.004558  0.880158            Random Forest
1  3.595377e-05  0.003637  0.008913  0.541729            KNN regressor
2  1.557740e-04  0.006636  0.012498  0.098844        Linear Regression
3  1.761657e-07  0.001490  0.005049  0.852930  Decision Tree Regressor
4  1.737817e-04  0.007141  0.013166 -0.000002           Lasso Regessor
```

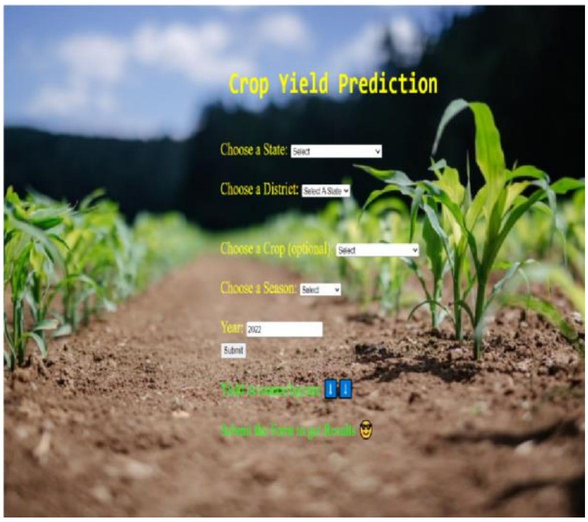
R2_Score plot



RMSE PLOT







Crop Yield Prediction

Choose a State: Select

Choose a District: Select A State

Choose a Crop (optional): Select

Choose a Season: Select

Year: 2022

Submit the Form to get Results 😎

VI CONCLUSION

The natural progression began with data cleansing, data visualization, preprocessing, training and testing the dataset, and lastly, model creation and assessment. Lastly, we use machine learning methods with varying outcomes to predict the yield. This brings up some of the ongoing discussions over yield gauge. The structure will cover the majority of notable yield types, so farmers may become more knowledgeable about the collect that may never have been produced and that permeates every potential yield. This assists farmers in a unique way.

If a particular crop is provided, it displays the yield; if not, it displays the yield of all crops.

REFERENCES

[1] Crop Data (1997 – 2010) dataworld.com

[2] ML algorithms: Analyticsvidhya.com

[3] javascript codewithhugo.com

[4] Deployment heroku

[5] PavanPalle,Praful Nikam, Abhijeet Pandhe, Vijay Pagare,
Prof. DilipDalgade Crop Yield Prediction based Climatic boundaries 2019.

[6] Vishal Vats, Naveen Kumar, Arun Kumar, crop yield prediction using machine learning algorithms,2018.

[7] Alaslani, Maram Elrefaei, Lamiaa. (2019). Learning With CNN for IRIS Recognition. International Journal of Artificial Intelligence Applications. 10. 49-66.

8]Yuan, Jun Ni, Bingbing Kassim, Ashraf. (2014). Half-CNN: A General schema for complete-Image Regression

[9] Training Recurrent Neural Networks, Ilya Sutskever, PhD Thesis, 2012.

[10] Thomas van Klompenburg,Ayalew Kassahun and Cagatay Catal, “Crop yield prediction using machine learning: A systematic literature review”, Computers and Electronics in Agriculture, Volume 177.

[11] Saeed Khaki,Lizhi Wang and Sotirios V. Archontoulis, CNN-RNN schema for Crop Yield Prediction ,
Frontiers in Plant Science, v. 10, 2019