Proceedings of the 4th International Conference on Computing and Data Science (CONF-CDS 2022)
DOI: 10.54254/2755-2721/2/20220639

Predict the Growth of Crop Yield Using Machine Learning

Anupama C G[1*[0000-0002-8048-5102]],**Karthick.S**[1[0000-0002-5346-2546]] **and S.Selakumara Samy**[1[0000-0001-5052-9357]]

[1]Department of Computational Intelligence
SRM Institute of Science and Technology,Chengalpattu,India

anupamag@srmist.edu.in

Abstract. India is an Agricultural based country. Our 70% of the primary and secondary sector depend upon farming which is done both in urban and rural areas. So basically due to such a large dependency on agriculture, many times the farmers are not able to judge which crop to grow or which crop can give them the best yield with respect to different conditions. So first we took the data from the reliable sources for the different agricultural products. Then we clean and clear the data from the empty values. As well as we did a primary analysis of the raw data. The data consist of different variables that we had to consider, some were removed as they were not helping in prediction. Variables which we used for the predictions are Nitrogen Phosphorus, Potassium, Humidity, Rainfall. So for solving the problem of farmers we came up with Machine Learning using all these variables we used different models in Machine Learning. We train our models by different types of combinations and get to know which variables are good for predicting. Then we took our results and checked that it matches the final results of our test data or not.

Keywords: Crop Yield Prediction, Crop Recommendation, Random Forest Classification, Decision Tree, KNNclassifier.

1. Introduction

India is an agriculture-dependent country. Agronomic yield is influenced by organic, financial, and seasonal variables. In a country with 1.38 billion mouths to feed, estimating agricultural production may be a troublesome task for our country. Predicting crop production will facilitate farmers build the required plans for purposes such as storing and promoting. Crop yield forecasting entails a large quantity of information, creating it an ideal contender for data processing ways. DataMining is the technique of collecting antecedently hidden estimated info since large records. dataprocessing supports within the investigation of upcoming outlines and appeal, allowing firms towards create knowledgeable selections. The method of analyzing, cleaning, and modeling information to get helpful data and conclusions is thought as information study.

Approaches are accustomed to alter the client's information into valued info. This analysis is often lengthy to agronomy similarly. Furthermost planters trusted on their long-run field expertise with exact yields to forecast a larger crop within the returning period. Still, they do not collect a good worth for his or her yields. It usually happens attributable to lean irrigation or poor crop choice, however it's going to additionally occur once crop yields area unit under expected because of a spread of things, the farmers

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2/20220639

United Nations agency form up the bulk don't deliver the goods the anticipated Crop. That information set of harvest yield that contains of several parts. By learning the mud and air for the particular space, through that growth yield manufacture, optimum crops are often calculable. Advantage of our analysis primarily is Farmers can take pleasure in this forecast to see that crops area unit best for his or her farm supported soil kind, ph., and plant food.

2. Literature Survey

Shailesh Shetty S [1] The initiative aids farmers in their evaluation that crop to grow in an exceedingly exact space at a particular period and predicting whether or not it'll be gainful or not. It offers the specifics by stipulating whether or not the crop is lucrative. As a result, this device aids farmers in their decision-making method, permitting them to save lots of period. Suvidha Jambekar [2] regression investigation is useful as a prognostic modeling device to guess yield manufacture for harvest production. The regression algorithm functional were, variable reconciling Regression Projections, so numerous statistical regression, Random Forest Regression in keeping with the outcomes. Random Forest Regression could also be wont to exactly estimation wheat berry, and rice, and corn making. R. Vidhya [3] They discovered accurateness rate recovers once a dataset with a lot of options is employed. As critical alternative approaches, like call trees, linear regression, random forest algorithms are shown to be higher to alternative prediction algorithms. The enclosed dataset incorporates plenty of variables leading to a lot of accurate prediction.

Hetal Patel, Dharmendra Patel [4] They evaluated how well the classification algorithms performed. Naive Thomas Bayes, J48, and easy Cart. This comparative examination of crop prediction makes use of an outsized dataset and 10- cross-fold authentication to administer a sign of the predictive capabilities of the datamining ways used. N.Rohit M.Vineeth and S.Bhanumathi, [5] The projected system predicts the harvest crop using random forest and deeplearning models. And proposes the number of fertilizers ought to be cast-off for top yield. a pair of totally dissimilar dataset for crop information and nourishment data area unit used. 20% of the dataset is utilised for training, and 80% for instruction is employed for testing the data. The district, area, season, and production factors area unit used to build a machine learning model that forecasts crop. By seeing phosphoric, potassium, element quantity in mud, the amount of nourishment needed is set.

3. System Architecture and Design

Initially we searched for databases related to crop Production. So we found out 2-3 databases from government websites as well as from kaggle. From that database we decided to make two kinds of modules i.e. Yield Prediction Module and Crop Pre-diction Module[6]. So Input Variables that datasets have were nitrogen , phosphorous , potassium , temperature , humidity , rainfall , crops , state name , district name , seasons , crop area , production [7] .To get the most important variables we did some pre analysis , then we chose some of them for our final analysis. But before our final training, we clean the data using label encoders, removal of outliers etc. Then after all the cleaning, we had a perfect data to do the predictions. So we decided to use four models and choose the best fit for our final output which will be shown [8]. We used Decision Tree, K-Nearest Neighbor, Linear Regression, and Random Forest. After all the predictions we got to know that Random forest technique is giving the maximum accuracy [9]. So we gave out final result as of Random Forest Prediction. Final output of our prediction contains Yield prediction as well as Crop Prediction [10]. Yield Prediction was analyzed using all four models,

DOI: 10.54254/2755-2721/2/20220639

and Crop Prediction was analyzed using Linear Regression and Random Forest [11][12]. System Architecture Design is given in Fig. 1.

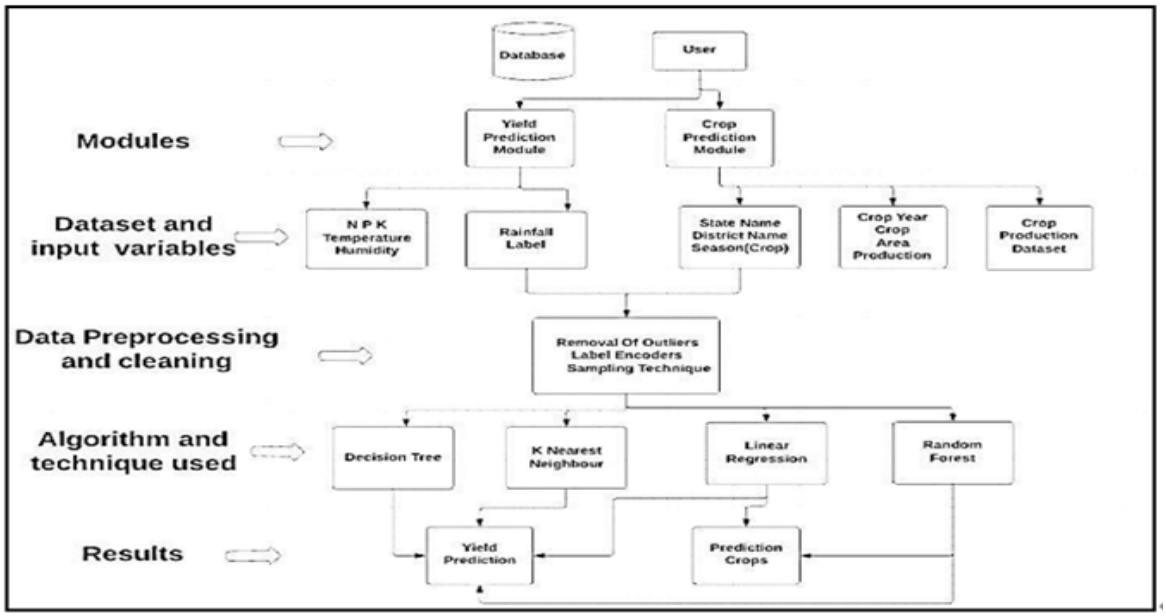

Fig. 1. System Architecture Diagram.

4. Methodology

Machine learning has opened ways to different decision making solutions for harvest yield forecast, together with supportive selections on what plants to cultivate and what to do throughout the crop season [13]. And data mining assists in analysis of large datasets of previous records easily. Thus, enabling us to predict future behavior of crop production. Many machine learning To assist agricultural production fore-cast analysis, algorithms are used [14]. We performed a scientific Literature Review (SLR) to gather and combine the alternatives and algorithms that are used in experiments to forecast agricultural production [15].

One of the most important factors to determine the effectiveness of an algorithm is its accuracy. To make an accurate prognosis and take into account the erratic temperature changes, rain numerous machine learning classifiers like linear Regression, K-NN Classifier, Decision Tree, Random Forest etc. are applied to urge a outline. By studying the results of the machine learning classifiers, we came to a assumption that the Random Forest algorithmic program delivers the most correct results [16].

4.1. Data Collection

Data mining depends on effective data collection and warehousing as well as computer dispensation. Using a technique called data- preprocessing, the raw data into a hygienic dataset. A huge dataset is required for the datamining application. The data collected from multiple sources is usually in raw type which isn't possible for the analysis. It might embrace data that's incomplete, obsolete, or inconsistent, i.e. it might contain redundant data or have blank fields [17]. The data ought to be normalized. By smearing various techniques like replacement misplaced values and null values, we can remodel knowledge into an obvious arrangement. Through data min-ing, we filter out the relevant data from the stack and drop the redundant data and clear out the rows with vacant fields for discovering actionable information from large sets of data [18].

DOI: 10.54254/2755-2721/2/20220639

4.2. Data Preprocessing

Data mining depends on effective data collection and warehousing as well as computer dispensation. Using a technique called data- preprocessing, the raw data into a hygienic dataset. A huge dataset is required for the datamining application. The data collected from multiple sources is usually in raw type which isn't possible for the analysis. It might embrace data that's incomplete, obsolete, or inconsistent, i.e. it might contain redundant data or have blank fields [17]. The data ought to be normalized. By smearing various techniques like replacement misplaced values and null values, we can remodel knowledge into an obvious arrangement. Through data min-ing, we filter out the relevant data from the stack and drop the redundant data and clear out the rows with vacant fields for discovering actionable information from large sets of data [18].

4.3. Train and Test Model

The training and testing of data is the last stage in the preparation of data. The info typically tends to be split unevenly after developing the model typically needs the maximum amount of data points as attainable. In supervised learning, an algorithm uses a large number of instances to develop a model in an effort to discover the model with the lowest possible loss [19].

Table 1. Agricultural dataset.

s.no	temperature	humidity	ph	rainfall	labels
0	20.879744	82.002744	6.502985	202.935536	rice
1	21.770462	80.319644	7.038096	226.655537	rice
2	23.004459	82.320763	7.840207	263.964248	rice
3	26.491096	80.158363	6.980401	242.864034	rice
4	20.130175	81.604873	7.628473	262.717340	rice
…	…	…	…	…	…
3095	25.287846	89.636679	6.765095	58.286977	watermelon
3096	26,638386	84.695469	6.28924	48.324286	watermelon
3097	25.331045	84.305338	6.904242	41.532187	watermelon
3098	26.897502	83.892415	6.463271	43.971937	watermelon
3099	26.986037	89.413849	6.260839	58.548767	watermelon

The machine needs to learn the behavior of the data before producing results. After the results are obtained, a set of data is used to check for its correctness. The training dataset is that the initial dataset used to train the machine learning algorithms to be told and turn out right predictions.

The dataset is split into two parts: the training set consists of 75-80% of the total dataset. The training set serves as the basis for the model's learning; in other words, it is utilised to determine the model's weights and biases. 20-25% is set aside for the testing model. The model has no knowledge of this dataset and is free of any of its biases.

4.4. Machine Learning Algorithm

Once data ripping is finished the next method is making a training a scikit-learn model. training as a practise a machine learning model needs machine learning algorithmic long stretch of time. The government approved dataset is collected that con-sists of the parameters like characteristics like the

DOI: 10.54254/2755-2721/2/20220639

name of the state, the district, the yield, the temperature, etc., and taken as input in csv format. rule at the side of data for pattern training. Here, we often use the Random Forest algorithm, a well-known supervised learning approach that relies on bagging techniques [20].

Random Forest enables us to analyze crop growth associated with these climatic conditions and biophysical modification. The Random Forest method combines many different decision trees. This method is a classifier-supported classification algorithm. The dataset will be split into training and testing data. More training dataset is employed to make the decision tree [21]. By taking into account training data, the model may construct a decision tree and remove the weaker node from the training data to get a much better model. By taking into account training data, the model may construct a decision tree and remove the weaker node from the training data to get a much better model [22]. The ultimate goal of the bagging process is to boost overall results via a combination of mastering output. During training, the random forest algorithm creates several decision trees. The ultimate output is the one with the most votes after predictions from various decision trees are gathered [23].

Tree predictors are gathered in random forests in such a manner that every tree depends on the values of a random set sampled severally and with a similar distribution for all trees within the forest. Random Forest used the bagging method to train the info that will increase the accuracy of the result. The platform utilised to create the trained model using the Random Forest technique is called a Jupyter notebook.

4.5. Prediction

To forecast the output based on fresh input, the trained model is used. In this case, the trained model was preserved in a separate file so that it could be predicted based on fresh input. In this approach, the trained Machine Learning model, which saves the object in a binary stream, is stored using the pickle format created in Jupyter, and the model is tested using a dataset. This prediction model comprises random forest algorithms, which use information it will need to make predictions to learn attributes from training data [24].

4.6. Accuracy Determination

Accuracy is the one of the metrics used for evaluating classification models. Accuracy is calculated by dividing the number of correct predictions by the total number of predictions.

The formula is given below:

Accuracy=Number of correct predictions/Total number of predictions

Accuracies of different models are compared and the best model is determined.

5. Results and Discussions

5.1. Data Analysis and Visualization

Datasets with multiple parameters and variables are often difficult to interpret for general crowd. Even for data scientists, there is a lot of numerical representation which when put together becomes overwhelming to be understood. So, to analyze the data with ease, we prefer data visualization. Transforming data into a visual context, such as a map or graph, makes it simpler for people to comprehend it and derive meaning from it. Data visualization aims to make it simpler to see trends, patterns, and outliers in large datasets. We have used different graphs to represent the trends in the production with respect to each of the factors in the dataset [25]. We also studied the change in growth of crops over the years and different seasons in a year. Sample result is given in Fig. 2.

DOI: 10.54254/2755-2721/2/20220639

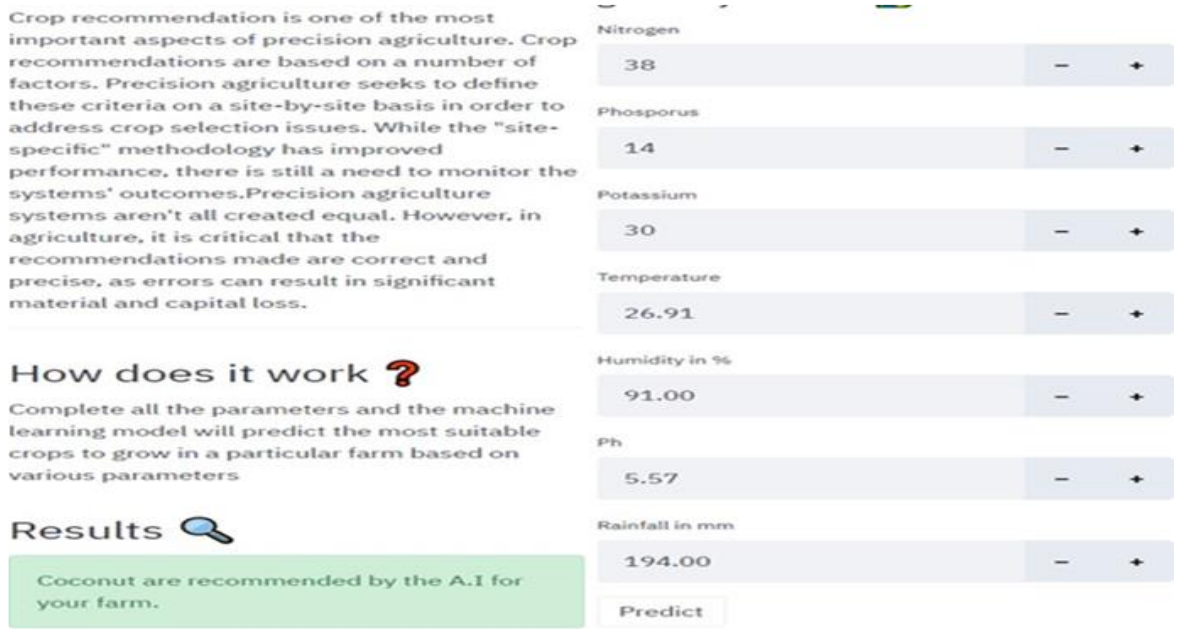

Fig. 2. Crop Prediction using Randomforest.

5.2. Crop Name Prediction

The crop that will be fully grown in a particular district at the appropriate time was predicted using the Random Forest Classifier, which has the greatest accuracy.

Abundantly growing crops in a very state are often selected, and based on the concepts of area, production, temperature, humidity, rainfall, and wind speed, yield was determined. The Random Forest classifier was trained on the preprocessed dataset. Instant weather data from the chosen state is used for forecasting. The selected state's crops were correctly predicted by the trained model.

5.3. Crop Yield Prediction

The assembly of the anticipated crop was mapped to the crop that the Random Forest Classifier predicted. The user-entered zone was then split off from the production to encourage crop output. The crop yield of an area can be calculated as: Yield= Production/Area. Farmers may choose the best time to produce the right crop to get the best results by using the expected yield of each crop.

6. Conclusion

For estimating agricultural production, the study discusses machine learning techniques. supported temperature, season, and placement. By integrating Precipitation, Temperature, and other factors like season and location, a Yield projection will be generated for a certain district. Once When all the variables are taken into account, Random Forest is the best classifier. The dataset that is in use with a lot of options will increase the accuracy rate. When compared to competing technologies such as multiple simple regressions and call trees, random forest is the better prediction rule. Numerous factors in our dataset allowed for a large number of accurate predictions. By launching this initiative, farmers will be able to reduce losses and boost crop yields, therefore extending their agricultural resources. This will not only assist farmers in choosing the best crop to plant in the next season, but it will also assist in bridging the gap between technology and agriculture.

DOI: 10.54254/2755-2721/2/20220639

References

[1] Shruthi G Sangeeta Design and implementation of crop yield prediction model in agri-culture Int. J. Sci. Technol. Res., 8 (01) (JANUARY 2020).

[2] B. Devika, B. Ananthi Analysis of crop yield prediction using data mining technique to predict annual yield of major crops

[3] Hetal Patel, Dharmendra Patel A comparative study on various data mining algorithms with special reference to crop yield prediction

[4] R. Vidhya, Pragya Mathur, Shivani Sai Valluri Crop yield predictzion using random forest Int. J. Adv. Sci. Technol., 29 (9s) (2020), pp. 3084-3086

[5] Priya, P., Muthaiah, U., Balamurugan, M .Predicting Yield of the Crop Using Machine Learning Algorithm.

[6] Guo, T., Yu, K., Aloqaily, M. and Wan, S., 2022. Constructing a prior-dependent graph for data clustering and dimension reduction in the edge of AIoT. Future Generation Computer Systems, 128, pp.381-394.

[7] Yang, L., Yu, K., Yang, S.X., Chakraborty, C., Lu, Y. and Guo, T., 2021. An intelli-gent trust cloud management method for secure clustering in 5G enabled internet of medical things. IEEE Transactions on Industrial Informatics

[8] Sathishkumar V E, Changsun Shin, Youngyun Cho, “Efficient energy consumption prediction model for a data analytic-enabled industry building in a smart city”, Building Research & Information, Vol. 49. no. 1, pp. 127-143, 2021.

[9] Sathishkumar V E, Youngyun Cho, “A rule-based model for Seoul Bike sharing de-mand prediction using Weather data”, European Journal of Remote Sensing, Vol. 52, no. 1, pp. 166-183, 2020.

[10] Sathishkumar V E, Jangwoo Park, Youngyun Cho, “Seoul Bike Trip duration predic-tion using data mining techniques”, IET Intelligent Transport Systems, Vol. 14, no. 11, pp. 1465-1474, 2020.

[11] Sathishkumar V E, Jangwoo Park, Youngyun Cho, “Using data mining techniques for bike sharing demand prediction in Metropolitan city”, Computer Communications, Vol. 153, pp. 353-366, 2020.

[12] Sathishkumar V E, Yongyun Cho, “Season wise bike sharing demand analysis using random forest algorithm”, Computational Intelligence, pp. 1-26, 2020.

[13] Sathishkumar, V. E., Wesam Atef Hatamleh, Abeer Ali Alnuaim, Mohamed Abdel-hady, B. Venkatesh, and S. Santhoshkumar. "Secure Dynamic Group Data Sharing in Semi-trusted Third Party Cloud Environment." Arabian Journal for Science and Engi-neering (2021): 1-9.