Smart Crop Predictor using Machine Learning

[1]Gayatri Awate, [2]Shreneel Kala, [3]Jayshree Aher, [4]Aditi Dhumal, [5]Sanyukta Tamhankar

[1]School of Computer Engineering and Technology
[1]MIT World Peace University, Pune, India

Abstract: **Agricultural forecast and statistics is an important resource that has not been explored comparably to its impact. The aim of our project is to make this process computerized by implementing principles of data analytics and representing it on a dashboard. These relational trends will act as solutions for farmers, especially in drought-prone areas. Data Science is coming to light in the research field in agricultural crop yield analysis. In this project, our emphasis is on the enactment of several machine learning approaches in the very essential agricultural domain.**

Keywords**: agriculture, farmers, machine learning, yield**

I. INTRODUCTION

In this paper, we aim at identifying specific areas that are suitable for specific crops, so as to avoid a deficit of crops. The ultimate objective is to stand as a system that when fed with data regarding various parameters, successfully produces trends and correlations that can help the farmers choose the best type of crops that they can plan in the given conditions. The aim of our module is for the collection of the raw data in directive to standardize it, analyze it, and then feed it to a system that will help us provide the relational trends. These relational trends will act as solutions for farmers, especially in drought-affected areas. Using various parameters such as the amount of rainfall, production of crops, area-wise statistics, soil conditions, NPK (Nitrogen, Phosphorus, Potassium) ratio, agricultural inputs, etc. These agricultural tendencies will be formulated and a report-based system will be introduced which will provide various solutions to the farmers for eschewing droughts and sustaining the development of their crops. Machine Learning is a very blooming research field in agricultural crop yield analysis. In this project, our focus is on the various systematic employment of machine learning strategies and approaches in the agricultural field. There are different machine learning approaches in use, such as KMeans and Support Vector Machines (SVM). Our crop prediction analysis aims to find data models and achieve high accuracy.

II. LITERATURE SURVEY

Analyzing and estimating weather, soil, rainfall in the scope of agriculture is of enormous importance in a densely populated nation like India where the prevalence of the population is dependent on agriculture. Many of the researchers and programmers have created multiple systems to help the farmers of India in understanding the best crop to grow with the availability of resources that they have. This section focuses on the works done previously in this field.

Pratik Srichandan et al. [1] this paper presents an analysis of agricultural production systems to balance the reserve and market of agricultural products during the development of environmental detectors. Agricultural production systems based on IoT sensors consist of three distinguishable parts. Their system will help by comparing different factors and various parameters which will be affecting the agriculture process directly or indirectly, but it may or may not consider all the parameters. Data and its privacy points need to be properly addressed, but when these issues are considered and used too rigorously, they can delay.

Sjaak Wolfert et al. [2] in this paper, have created a structure to present a methodical classification of problems for the analysis of the various big data applications in a wide range of areas of smart farming, in terms of very important anthropometry. Big data is also a challenge in this task, as large amounts of disorganized data often require a lot of expertise. Farnaz

Babaie Sarijaloo et al. [3] in this article, the author uses various types of machine learning strategies to study crop yield. They mainly focus on combinations. The outcome in the testing stage indicated that the XGBoost algorithm is condescending and incomparable among all the executed methods in the model. Different Machine learning models can be used. More types of crops and geographical regions can be covered.

Jharna Majumdar et al.'s [4] paper has is based on the survey of the data on agriculture and puts a light on the preferable parameters to increase yield production using the various data mining approaches like multiple linear regression, DBSCAN, PAM, and CLARA. It consisted of a limited dataset and some algorithms mentioned are not very efficient for predictions.

Mishra Subhadra et al. [5] present reflections on the applicability of several machine learning strategies in the important areas of agriculture production. This article deals with different types of statistical analysis and methods. This presented a new approach to agricultural crop management production. This document mentions that precise and punctual predictions of farming production are essential for critical approach conclusions such as import and export, then prices etc. issued by the Department of Economics and Statistics.

Dahikar SS et al. [6] The article shows how to extract and synthesize algorithms for better crop forecasting. They selected the studies carefully, analyzed the different techniques and characteristics used, and made recommendations for further study. They

also mention that raw data and large amounts of unstructured data can be difficult to process. They also mention that filtering out the exact features can be difficult.

S.R.Rajeswari et al. [7] the focus is used to forecast the yield and the methods used to predict the value of the crop. It uses processes and algorithms such as preprocessing, feature extraction, random forest, Bayesian networks, and ANNs. This project has a very broad scope for the future. Their modules are implemented so that farmers can work when they are not at home and can be connected to various mobile applications. They also propose using drones to accumulate real-time data and monitor yields.

MT Shakoor et al. [8] in this research paper presents predictions based on validating static data sets utilizing SVM strategies. Algorithms such as Decision Tree, ID3 Learning, and KNNR were applied in this study. Since this study is limited to some fixed datasets, the number of predictors required for accurate prediction is limited to some extent. This made it impossible to track progress and led to the use of predictive models that implement the best algorithm for a limited number of predictor statistics.

D.S. Zingade et al [9] launched an app that collects data correlated to the soil, climate, and last year's crops and suggests the most profitable crops and yields as it helps farmers decide which crops to grow. The algorithms used in this study were various regression methods such as data collection, data preprocessing, linear, nonlinear, and polylinear regression. In the future, all agricultural equipment will be able to connect via the Internet using IoT. The sensors used on the farm collect information about the current condition of the farm, and given these statistics, the device can increase humidity, acidity, etc. accordingly.

Potnuru Sai Nishant et al. [10] present a system that will help by comparing different factors and various parameters which will be affecting the agriculture process directly or indirectly, but it may or may not consider all the parameters. Different parameters such as state, season, district, location have been evaluated.

After studying some research papers, it was observed that most of the research was limited to certain regions and only a few parameters were considered. Some proposed systems could not generate significant inferences in the crop predictions. They also were unable to understand the necessary parameters that are essential for the crops. Crop data considered was limited to very few regions or parameters in certain papers. Many Machine Learning models were not tested for analysis and prediction of the crops. The work was also limited to certain areas and parameters, hence the conclusions obtained cannot be generalized for the entire area or the different crops. Thus, a large dataset that will consider a wide-spread geographical location should be used and a variety of classifications algorithms should be trained and tested on the dataset to get reliable results for reliability and accuracy.

III. PROPOSED SYSTEM.

System Architecture

Figure 1 shows a flow-chart-based visualization of the proposed system and the procedure it will use. The dataset called 'Crop_recommendation' possesses the data which would allow the users to build a predictive model to approve the most sustainable yield to grow in distinct areas established on several additional parameters. This dataset was assembled by expanding the datasets of rain, weather, and fertilizers data accessible for India. The data is being retrieved from government websites. Following which the segmentation and standardization are done. Data is stored and pre-processed to correspond to null values and missing values. The dataset applied is trained and split into the following two parts - training and testing data. It is divided into the ratio of 80:20, where 20% is the testing data and 80% is training data. After this, the data will be tested using several machine learning algorithms. The algorithms used are Decision Tree, Naive Bayes, (SVM) Support Vector Machine, Random Forest, and XGBoost.
Eventually, the results will be visualized.

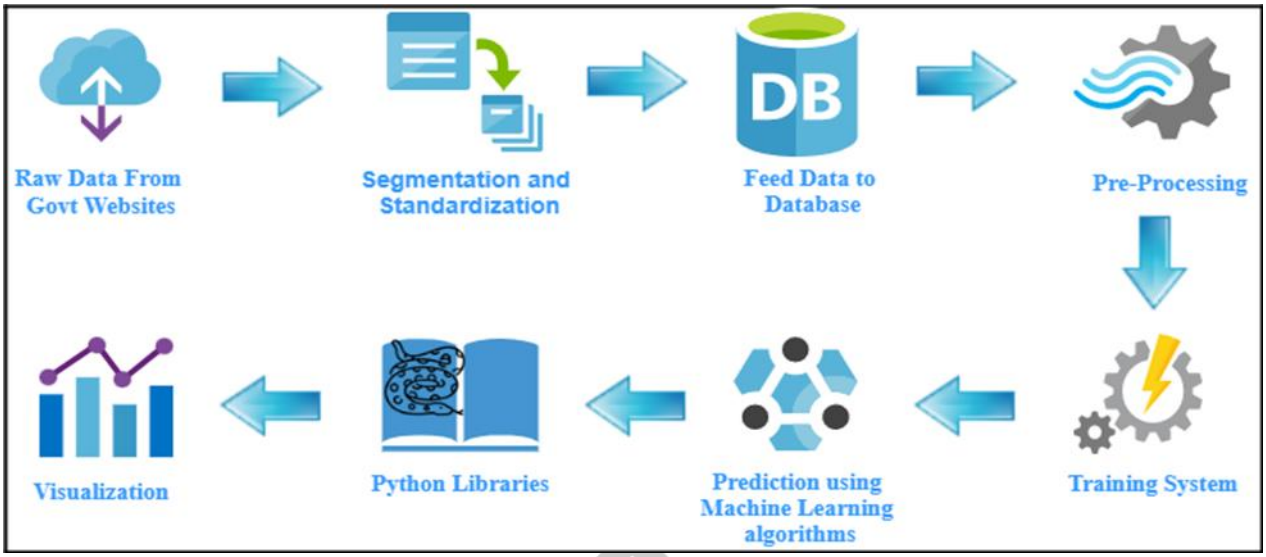

Fig. 1 System Architecture

Methodology

The yield production can be analyzed using the following steps:

1) Acquire the dataset: The dataset is retrieved from trusted and reliable sources so that the data can be efficient.

2) Import all the libraries: In order to perform pre-processing and machine learning on the data, some libraries need to be imported inside Python. This will enable smooth operations while working on the implementation.

3) Import the dataset: Dataset will be imported employing the read_csv() operation of the pandas library, which is employed to read a CSV file and execute functions on it. Employing this, we will read a CSV file locally.

4) Identifying and handling the missing values: If the dataset includes some missing data, then it might cause a massive issue for our machine learning model. Hence it is very essential to address missing values that are present in the dataset.

5) Splitting the dataset: In data pre-processing, we have divided our dataset into a training set and test set. This is one of the essential phases of data pre-processing as by doing this, we can enhance the implementation of our model.

6) Applying the machine learning models: The test data is then operated on the various selected machine learning models and their precisions are reviewed.

7) Displaying prediction using Visualization: The outcome of the prediction is then visualized using different visualizations

Dataset

To overcome the limitations that can be seen from the previously studied platforms, we have retrieved the data from government websites which will be a reliable source.

Dataset Schema

Features	Description
N	Representing Nitrogen percentage in the soil
P	Representing Phosphorus percentage in the soil
K	Representing Potassium percentage in the soil
Temperature	Indicates the temperature needed for the crop
Humidity	Indicates the humidity level
pH	Power of Hydrogen
Rainfall	Indicates the rainfall in mm.
Label	Name of crop

Table no. 1 Dataset Schema

Crop Prediction through Machine learning models:

Various Machine Learning algorithms are employed for predicting the crops established on different parameters such as NPK (Nitrogen, Phosphorus, Potassium) ratio range in the soil, the temperature in degree Celsius, relative humidity in %, ph level of the soil and rainfall in mm. Since multiple attributes are used for the prediction of crops, diverse multiclass prediction models are used. Five different types of prediction algorithms are used in this proposed module to predict the harvest. The various algorithms employed are Decision Tree, Naive Bayes, Support Vector Machine (SVM), Random Forest, XGBoost.

1) Decision Tree:

The Decision tree algorithm is a tool for decision support that utilizes a tree-like standard of conclusions and their feasible outcomes, along with chance event results, assets costs, utility, etc. The primary objective of this algorithm is to assemble a model that will indicate the value of a given target variable by understanding the different uncomplicated decision rules concluded from the data attributes. A tree can be noticed as a piecewise consistent required analysis. The algorithm carries out well even when its presumptions are relatively contradicted by the accurate model from which the data were developed.

2) Naive Bayes:

Naive Bayes classification approach is utilized for binary as well as multi-class classification problems. This classifier considers that the existence of a distinct attribute in a class is independent of the presence of some additional feature. This model is straightforward to assemble and execute and extremely beneficial for very huge data sets. Besides the clarity, this algorithm is also comprehended to exceed even the positively experienced classification methods and algorithms. This theorem delivers a method of calculating posterior probability P(c|x) from

$$P(c|x) = \frac{P(X|C).P(c)}{P(x)} \quad \text{were}$$

P(c): class prior probability
P(x): predictor prior probability
P(x|c): likelihood

3) Support Vector Machine (SVM):

The SVM is a Machine Learning model which is skilled to summarize between two distinct classes if the collection of labeled data is delivered in the training set to the algorithm. In SVM, we plot each data object as an attribute in n-dimensional with the value of each element being the value of an individual coordinate. In this technique, the data is classified by locating and encountering a hyperplane that maximizes the difference between data points illustrated in an N-dimensional space

4) Random Forest:

A random forest is a bagging approach that is used to solve regression and classification problems. It utilizes ensemble learning, which is a procedure that integrates many classifiers to deliver solutions to complicated problems. It has multiple decision trees and this algorithm assembles the outcome which is established on the predictions of the decision trees. Thus, it then predicts by considering the mean of the output from multiple trees, and extending the number of trees enhances the accuracy of the result. This algorithm removes the shortcomings of a decision tree algorithm. The major advantages of random forest algorithms are the handling of missing values and never over-fitting the model.

5) XGBoost

Extreme gradient boosting is an algorithm that has been incorporated to boost the performance and speed utilizing machine learning. It makes use of parallel tree boosting to go to a broad scope of data science problems speedy and precisely. This algorithm has its built-in capabilities which are developed to deal with missing data and after each iteration, the user can perform cross-validation.[17]

Result Analysis

The dataset shows the ratio of 80% training data to 20% test data. Evaluation of different machine learning models for multi-class classification trained in datasets. The accurateness, precision, recollection, and f1 score of the metrics are taken into account. The most accurate model is selected as a recommendation system for future harvests.

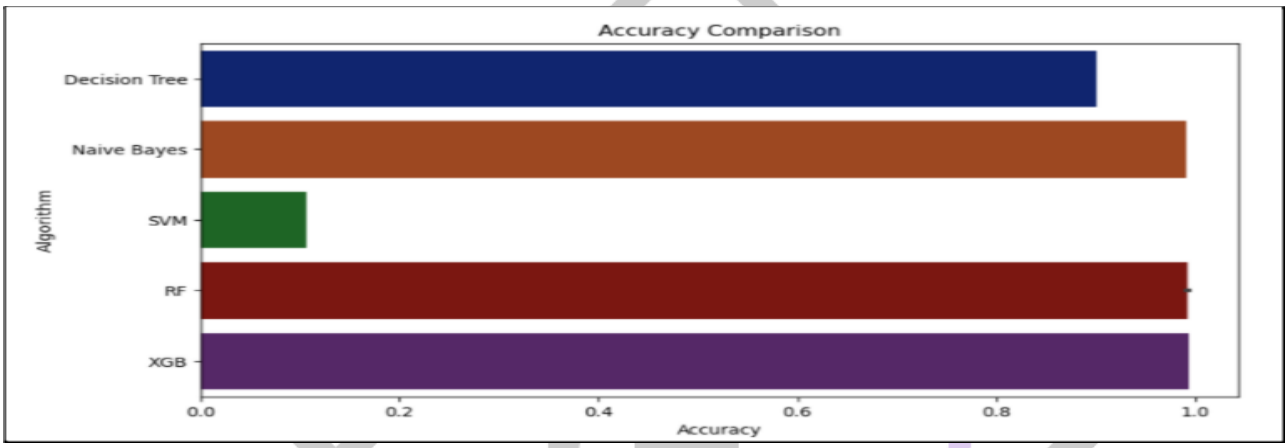

Fig. 2. Accuracy Comparison

After following the obtained accuracy from Fig. 2 obtained on all the classifier models, it is observable that the XG Boost Algorithm has the best potential accuracy among Naïve Bayes, KNN, Random Forest Algorithm, Decision Treen and SVM. The XG Boost has obtained the highest accuracy of 99.55% on the test data followed by Random Forest with 99.32% accuracy. Hence, it is concluded that ensemble models have performed better than other classification models considered in the system.

IV. CONCLUSION

This project highlights the implementation of machine learning and different algorithms in the domain of agriculture. Crop yield prediction with relevant inputs, fertilization prediction as well as rainfall prediction will be created in order to enable agriculturalists to expand the crop production. Decision-making based on analytical models will accompany conventional methods of farming. It is a learning-based system, so, the more attributes the user will add to it, the more meticulous the result will be predicted. This platform can help the farmers as only a single forum that can help them to access and analyze the crop they would like to produce and how they should time their procedure of initiating their yield for the season.

In the future, this application can be used in predicting various other elements like leaf disease detection, fruit disease detection, and other models that can be applied to this platform to increase the scope of the project.

REFERENCES

[1] Srichandan, Pratik & Mishra, Ashis & Singh, Harkishen. (2018). Data Science and Analytic Technology in Agriculture. International Journal of Computer Applications. 179. 21-28. 10.5120/ijca2018916850.

[2] Sjaak Wolfert, Lan Ge, Cor Verdouw, Marc-Jeroen Bogaardt, Big Data in Smart Farming – A review, Agricultural Systems, Volume 153, 2017, Pages 69-80, ISSN 0308-521X, https://doi.org/10.1016/j.agsy.2017.01.023.

[3] Farnaz Babaie Sarijaloo, Michele Porta, Bijan Taslimi, Panos M. Pardalos, Yield performance estimation of corn hybrids using machine learning algorithms, Artificial Intelligence in Agriculture, Volume 5,2021, Pages 82-89, ISSN 2589-7217, https://doi.org/10.1016/j.aiia.2021.05.001.

[4] Majumdar, J., Naraseeyappa, S. & Ankalaki, S. Analysis of agriculture data using data mining techniques: application of big data. J Big Data 4, 20 (2017). https://doi.org/10.1186/s40537-017-0077-4

[5] Mishra, Subhadra & Mishra, Debahuti & Santra, Gour. (2016). Applications of Machine Learning Techniques in Agricultural Crop Production: A Review Paper. Indian Journal of Science and Technology. 9. 10.17485/ijst/2016/v9i38/95032.

[6] S. S. Dahikar, and S. V. Rode, "Agricultural crop yield prediction using artificial neural network approach", International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering, vol. 2, no. 1, pp. 683-686, 2014.

[7] S.R.Rajeswari, Parth Khunteta, Subham Kumar, Amrit Raj Singh, Vaibhav Pandey, Smart Farming Prediction Using Machine Learning, International Journal of Innovative Technology and Exploring Engineering (IJITEE), Volume-8 Issue-7 May 2019, ISSN: 2278-3075

[8] Shakoor, M.T., Rahman, K., Rayta, S.N., & Chakrabarty, A. (2017). Agricultural production output prediction using Supervised Machine Learning techniques. 2017 1st International Conference on Next Generation Computing Applications (NextComp), 182-187.

[9] Prof. D.S.Zingade, Omkar Buchade, Nilesh Mehta, Shubham Ghodekar, Chandan Mehta "Crop Prediction System using Machine Learning", International Journal of Advance Engineering and Research, Volume 4, Dec.-2017, Special Issue 5, e-ISSN: 2348-4470

[10] P. S. Nishant, P. Sai Venkat, B. L. Avinash and B. Jabber, "Crop Yield Prediction based on Indian Agriculture using Machine Learning," 2020 International Conference for Emerging Technology (INCET), 2020, pp. 1-4, doi: 10.1109/INCET49848.2020.9154036.

[11] Ghorpade-Aher J., Sonkamble B. (2022) Effective Feature Selection Using Ensemble Techniques and Genetic Algorithm. In: Yang XS., Sherratt S., Dey N., Joshi A. (eds) Proceedings of Sixth International Congress on Information and Communication Technology. Lecture Notes in Networks and Systems, vol 236. Springer, Singapore. https://doi.org/10.1007/978-981-16-2380-6_32

[12] S. Jain and D. Ramesh, "Machine Learning convergence for weather based crop selection," 2020 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS), 2020, pp. 1-6, doi: 10.1109/SCEECS48394.2020.75.

[13] Doshi, Z., Nadkarni, S., Agrawal, R., & Shah, N. (2018). "AgroConsultant: Intelligent Crop Recommendation System Using Machine Learning Algorithms". 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA).doi:10.1109/iccubea.2018.8697349

[14] Kulkarni, N. H., Srinivasan, G. N., Sagar, B. M., & Cauvery, N. K. (2018). "Improving Crop Productivity Through A Crop Recommendation System Using Ensembling Technique". 2018 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS). doi:10.1109/csitss.2018.8768790

[15] R. Priya, D. Ramesh and E. Khosla, "Crop Prediction on the Region Belts of India: A Naïve Bayes MapReduce Precision Agricultural Model," 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2018, pp. 99-104, doi: 10.1109/ICACCI.2018.8554948.

[16] Panchamurthi, S. (2019). "Soil analysis and prediction of suitable crop for agriculture using machine learning". Int. J. Res. Appl. Sci. Eng. Technol, 7(3), 2328-2335.

[17] XGBoost,https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d