*As per UGC guidelines an electronic bar code is provided to seure your paper*

*International Journal for Modern Trends in Science and Technology*, 8(05):499-505, 2022

ISSN: 2455-3778 online
DOI: https://doi.org/10.46501/IJMTST0805075
Available online at: http://www.ijmtst.com/vol8issue05.html

# Crop Analysis, Clustering and Predection using Machine Learning

**Greeshma M Shajan[1] | Neenu Kuriakose[2] | Sangeetha J[3]**

[1], PG Scholar, Dept. of Computer Science, St. Albert'sCollege (Autonomous)Ernakulam
[2]Asst. Professor, Dept. of Computer Science, St. Albert'sCollege (Autonomous)Ernakulam,
[3]Asst. Professor, Dept. of Computer Science, St. Albert's College (Autonomous)Ernakulam
Corresponding author Email ID: greeshmamshajan@gmail.com

**To Cite this Article**
Greeshma M Shajan, Neenu Kuriakose and Sangeetha J. Crop Analysis, Clustering and Predection using Machine Learning. International Journal for Modern Trends in Science and Technology 2022, 8(05), pp. 499-505. https://doi.org/10.46501/IJMTST0805075

**Article Info**
Received: 22 April 2022; Accepted: 16 May 2022; Published: 23 May 2022.

## ABSTRACT

***Agriculture plays a vital part in the socio-economic fabric of India. Failure of farmers to decide on the best-suited crop for the land using traditional and non-scientific styles is a serious issue for a country where roughly 58 percent of the population is involved in husbandry. Occasionally, growers failed to choose the right crops grounded on the soil conditions, sowing season, and geographical position. This results in self-murder, quitting the husbandry field, and moving towards civic areas for livelihood. To overcome this issue, this exploration work has proposed a system to help the growers in crop selection by considering all the factors like sowing season, soil, and geographical position. Likewise, perfection husbandry is being enforced with ultramodern agrarian technology and it's evolving in developing countries that concentrate on point-specific crop operation. The need of the hour is to design a system that could give predictive insights to the Indian growers, thereby helping them make an informed decision about which crop to grow. Predictive analytics is a branch of advanced analytics that makes predictions about future outcomes using historical data combined with statistical modelling, data mining techniques, and machine learning With this in mind, I propose a system, an intelligent system that would consider environmental parameters (temperature, downfall, geographical position in terms of state) and soil characteristics (pH value, soil type and nutrients attention) before recommending the most suitable crop to the stoner.***

*KEYWORDS: Machine Learning, Predictive Analytics, Prediction of Favourable Crops, Crop Production, Climate, and Temperature*

## 1. INTRODUCTION

Most Agriculture is a major area for Indian frugality and mortal survival. It likewise contributes a large part to our day-to-day life. In utmost cases, growers kill because of product loss because they are ineffective to pay the bank loans taken for tilling purposes. We have noticed that the climate is changing persistently, which is dangerous to the crops and leads growers toward debts and self-murder. These pitfalls are also minimized when statistical styles are applied to data and by using these styles, we're suitable to recommend the sole crop to the planter for his agrarian land so that it helps him to induce maximum profit. Currently, husbandry has developed a lot in India. The advice of crops relies on certain parameters. Numerous exploration workshops are being conducted, to grasp an accurate and more effective

model for crop growing. Machine Learning focuses on algorithms like supervised, unsupervised, and reinforcement learning, and every one of them has its advantages and downsides This paper aims to recommend the foremost suitable crop-supported input parameters like Nitrogen (N), Phosphorous (P), Potassium (K), PH value of soil, Moisture, Temperature, and Rainfall. This paper predicts the precision of the future production of 11 different crops like rice, jute, cotton, maize, coconut, papaya, orange, apple, muskmelon, watermelon, grapes, mango, banana, pomegranate, lentil, black gram, mung bean, moth beans, order sap, chickpea, pigeon peas, kidney beans and coffee crops using different supervised machine learning approaches and recommends the foremost suitable crop for India. This proposed system applied different forms of Machine Learning algorithms a touch just like the k-nearest neighbors (KNN) algorithm, a simple, supervised machine learning algorithm that will break both bracket and retrogression problems. For clustering, here I used K-Means Clustering and which is an unsupervised learning algorithm that is habituated to break the clustering problems in machine learning or data wisdom, and also Hierarchical clustering, A Hierarchical clustering system works via grouping data into a tree of clusters. Hierarchical clustering begins by treating every detail as a separate cluster. In Hierarchical Clustering, the end is to supply a hierarchical series of nested clusters.

**STRUCTURE OF PAPER**

The paper is organized as follows: In Section 1, the introduction of the paper is provided along with the structure, important terms, objectives and overall description. In Section 2 we discuss related work. In Section 3 we have the complete information about design and methodology. Section 4 shares information about the experiment result and performance analysis. Section 5 tells about performance aanalysis. Section 6 tells us about the future scope and concludes the paper with references.

**OBJECTIVES**

Extensive work has been done, and plenty of strategies are applied within the agriculture sector. The biggest challenge in agriculture is to extend farm production and supply it to the end-user with the simplest attainable worth and quality. It is conjointly determined that a minimum of five hundredths of the farm product gets wasted by overestimating a soil's potential to support a specific crop. Field surveys, crop growth models, remote sensing, applied math models, and their combos are usually accustomed predict crop yield.

PROPOSED METHOD

Machine Learning is well equipped once it involves analyzing information relating to soil conditions, together with temperature, moisture level, and chemical makeup, all of that affect crop growth and livestock wellbeing. Today in agriculture, this could permit crops to be fully grown at abundant higher preciseness, sanctionative farmers to treat plants and animals virtually one by one, that successively considerably will increase the effectiveness of farmers' choices. Using this could develop means that to even predict harvest yields and valuate crop quality for individual plant species to notice crop sickness and weed infestations that were antecedent, not possible.This project aims to address some of the problems in current systems by greatly minimizing the human intervention in the process and thus reducing costs and errors. The aim is to ease the task of both the buyer and the seller.

## 2. RELATED WORK

In the past, various studies have been conducted on the development and design of crop-based forecasting systems. Each previously completed and reported task has its strengths and weaknesses in developing a system based on the specific problem of crop production improvement.

According to Aditya Shastry, H. A. Sanjay, and E. Bhanusree (2017), regression approaches can be used to forecast yield in the region with satisfactory results. The experiment demonstrates that the regression method may be used to anticipate crop yields for a specific geographical region with satisfactory results. In Asia, it stands out amongst the majority of countries in terms of production and the use of various technologies. Crops are widely noticed in various parts of our country. The regression model, often known as the forecast model, is a method for predicting the outcome of Wheat, maize, and cotton are all grown for specific years. The result shows that the proposed regression model is a technique for predicting yield.

In the agricultural field, Arun Kumar, Vishal Vates, and Naveen Kumar(2018) apply descriptive analytics concepts. The data from the analytical work is used to inform whatever data analytics are used on sugarcane crop datasets. Three important supervised algorithms are discussed in this study. To train and develop the model, they used machine learning techniques such as k-NN, SVM, and LS-SVM. The focus of this paper is on when we apply this strategy to datasets and consider the relative learning of multiple procedures. It shows the validity of each dataset training method as well as the mean squared error (MSE)during the test data cross-validation phase This experimentation endeavor could be stepped up once more. In the following stage, they can build a crop recommender system using this information.

S. Mamatha Jajur, Soumya N. G., and G. T. Raju's (2019) work will help agronomists increase agricultural yields, reduce soil degradation in farmed grasslands, and reduce fertilizer applied in crop yield by prescribing the right crop based on numerous factors. They provided labor assistance, to agronomists, in selecting the wrong crop for cultivation and achieving unceasingly. The suggested set-up can be enhanced in the future to include retail requirements and retail infrastructure availability, anticipated surplus, and opportunities, and upright ingathering mechanization of storage and processing This would allow for far-reaching forecasts based on location, meteorological conditions, and profit-making potential.

When wheat and barley are the key crops in this location, Z.H. Khalil and S.M. Abdullaev (2020) revealed the crash of meteorological patterns on winter output in Al-Diwanyah Iraq. It's been noted that the changing environmental conditions in Diwaniyah, which are characterized by high temperatures and low precipitation, have an impact on agriculture. This could result in rising temperatures, low precipitation, and weak winds, which cause an increase in evaporation rate and, as a result, the appearance of soil concerns. Fully utilized variability to improve and maximize crop production.

## 3. DESIGN AND METHODOLOGY

The main ideology of the study revolves around the concept of using machine learning models to identify the best crop to grow on the Indian land. Therefore, the results prove to be very beneficial to the agriculture of farmers. I used a Kaggle dataset containing 2201values from 22 unique plants and I applied the machine learning algorithm to the dataset. Using the dataset, the model is trained according to the actual values and the accuracy of the model is tested.

**Process Description**

The following diagram makes it easier to understand how we proceed.

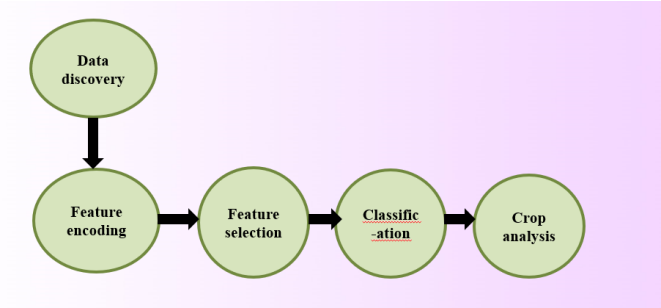


Figure 1 Architecture

1. Data discovery

The dataset consists of parameters such as nitrogen (N), phosphorus (P), potassium (K), soil pH, moisture, temperature, and precipitation. The record was taken from the Kaggle website. Kaggle is a subsidiary of Google that provides users a platform to urge and publish knowledge sets. Excluding this, it conjointly permits the users to create models in an environment that's usually web-based and data-science-oriented. I've taken the dataset from Kaggle that wanted to train the model. The dataset contains 2201instances or dates from past historical dates. This dataset includes rice, corn, chickpeas, bean, pigeon pea, moth bean, mung bean, broad bean, lentil, pomegranate, banana, grape, watermelon, apple, orange, coconut, cotton, jute, mango, coffee, papaya.

```
[ ] # lets read the dataset
    data = pd.read_csv("/content/Crop_recommendation (1).csv")

    # lets check teh shape of the dataset
    print("Shape of the Dataset :", data.shape)

    Shape of the Dataset : (2200, 8)

    # lets check the head of the dataset
    data.head()
```

| | N | P | K | temperature | humidity | ph | rainfall | label |
|---|---|---|---|---|---|---|---|---|
| 0 | 90 | 42 | 43 | 20.879744 | 82.002744 | 6.502985 | 202.935536 | rice |
| 1 | 85 | 58 | 41 | 21.770462 | 80.319644 | 7.038096 | 226.655537 | rice |
| 2 | 60 | 55 | 44 | 23.004459 | 82.320763 | 7.840207 | 263.964248 | rice |
| 3 | 74 | 35 | 40 | 26.491096 | 80.158363 | 6.980401 | 242.864034 | rice |
| 4 | 78 | 42 | 42 | 20.130175 | 81.604873 | 7.628473 | 262.717340 | rice |

Figure 2 Data set

2. Feature encoding

For a successful application, feature encoding is required. Data obtained from different resources are sometimes in

raw form. It may contain incomplete, redundant, inconsistent data. Therefore, in this step, these redundant data must be filtered out. Data must be normalized.

3. Feature selection

Feature extraction is an activity that manually selects the predictors or attributes that contributes most to the results. Intangible attributes of data reduce the accuracy of prototypes, and you can learn to prototype and rely on intangible properties. Feature extraction from five datasets from five different crops simplifies the amount of data needed to represent a large dataset. RF Classifier technique was applied to select the attribute. In this method, the attribute with increased entropy value is selected as the primary function for accurately predicting crop yields. This step focuses on identifying and using the most relevant attributes from the dataset. This process removes irrelevant and redundant information about the use of the classifier.

4. Classification

Classification algorithms in machine learning use input data knowledge to predict the probability that the resulting knowledge can represent one among the planned classes. Clustering is an unsupervised problem of finding natural groups in the feature space of input data. Prediction studies of the system are used to analyze the best yield production according to climatic conditions and soil conditions according to clustered datasets. Innovation improvements provide a vast amount of information. This information is selectively removed as it helps to examine and distinguish patterns in the information. Machine learning builds a very valuable large information structure to store the vast amount of information associated with it for prediction. The last forecast record from groups the crop states.

5. Crop analysis

By analyzing the parameters, this project can give information that is correct and accurate which leads to the prediction of crop sustainability. Combined principles of crop modeling with machine learning for crop yield forecasting. This framework prescribed to connected yields suggestion helps farmers connect on harvests and climate estimates. Be that because it could, imperative in Agriculture is, all yields generation depends on the soils since soils are basic horticulture improvement and harvest creation. On the off chance that soil is not affordable for a specific harvest, ranchers cannot get profit creation. Thus recommending the yields by estimating climate and connected to soil can facilitate agriculturists effectively distinguish affordable crops.

## 4. EXPERIMENTALRESULT AND PERFORMANCE ANALYSIS

1.System Setup and Configurations

Here I utilized Scikit-learn, Python's most popular machine learning toolkit, to calculate accuracy, and split the training and testing sets. Scikit-learn is a Python machine learning library. It uses algorithms such as logistic regression, decision tree classifier, random forest classifiers, and support vector machines for classification and regression. It can be used in conjunction with other Python packages such as NumPy and pandas. Scikit-learn is a great library for supervised learning, which entails training the model by loading a sample dataset that it can observe and structure its learning around. It also allows us to create training and testing datasets using train test split.

2. Exploration of Different Crops

We evaluate each crop's variable and conduct philosophical experiments concerning their meaning and value for our vision during the data exploration stage. We concentrate on and comprehend the dependent and independent variables. Then, for each crop, clean the data sets and deal with missing data, outliers, and categorical variables, before determining whether or not our data sets match the assumptions required by multivariate approaches. Outliers are values in a dataset that are outliers compared to the rest of the data. Outliers might be the result of perusal delusion or error, system liability, manual delusion, or misleading. In this paper, we apply the Z-score method in Python to remove outliers.

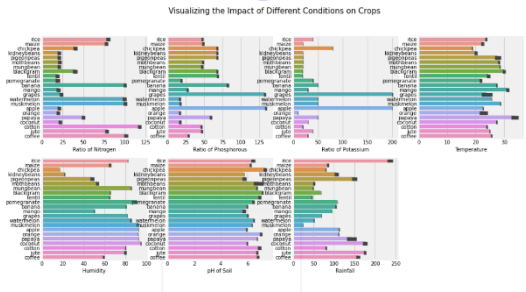


Figure 3 crops visulization

3.ALGORITHMS USED

oTrain_test_split Library :

I've imported train_test_split from sci-kit-discover ways to break up our dataset into one-of-a-kind units consistent with our needs- one for education of the version and the alternative for trying out the operating and accuracy of the skilled version so that we will select the satisfactory viable set of rules. Then the fine-appearing set of rules is chosen for the model.

oK-Means Clustering: It is an unsupervised learning algorithm. It permits us to cluster the facts into specific agencies and a convenient manner to find out the types of groups

withinside the unlabelled dataset on its very own without the need for any training. It is a centroid-primarily based set of rules, wherein every cluster is related to a centroid. The foremost intention of this set of rules is to reduce the sum of distances among the facts factor and their corresponding clusters.

- Hierarchical clustering: It is an unsupervised system getting to know the algorithm, that's used to group the unlabelled datasets right into a cluster and is additionally referred to as hierarchical cluster analysis or HCA. In this algorithm, we increase the hierarchy of clusters withinside the shape of a tree, and this tree-formed shape is referred to as the dendrogram. Sometimes the outcomes of K-manner clustering and hierarchical clustering can also additionally appear similar, however, they each vary relying on how they work. As there may be no requirement to predetermine the range of clusters as we did withinside the K-Means algorithm.

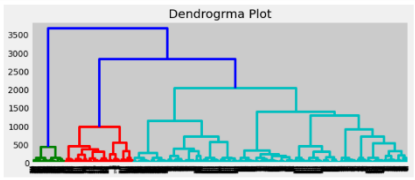


Figure 4Dendogram

- K-Nearest Neighbour: It is one of the simplest Machine Learning algorithms based on the Supervised Learning technique. K-NN algorithm assumes the similarity between the new case/data and available cases and puts the new case into the category that is most similar to the available categories. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. It can be used for Regression as well as for Classification but mostly it is used for Classification problems. It is a non-parametric algorithm, which means it does not make any assumptions on underlying data.

For Predictive Analysis :

- Random Forest: RF is a widespread machine learning rule that belongs to the supervised learning technique. It may be used for each Classification and Regression issue in ML. It's supported the idea of ensemble learning, which is a method of mixing multiple classifiers to unravel a posh downside and boost the performance of the model. As the name suggests, "Random Forest is a classifier that contains a variety of trees on numerous subsets of the given dataset and takes the common to improve the prognostic accuracy of that dataset." The larger number of trees within the forest leads to higher accuracy and prevents the matter of overfitting.
- Decision Tree: DT is a Supervised learning approach that can be used for each class and Regression problem, however in the main it's far favored for fixing Classification problems. It is a tree-established classifier, in which inner nodes constitute the functions of a dataset, branches constitute the selection policies and every leaf node represents the outcome. Decision Trees are a kind of Supervised Machine Learning in which the statistics are constantly broken up in keeping with a positive parameter. The tree may be defined through entities, specifically selection nodes, and leaves. The leaves are the selections or the very last outcomes. And the selection nodes are in which the statistics are broken up.
- Gradient boosting: GB formula is one of the foremost powerful algorithms within the field of machine learning. As we all know that the errors in machine learning algorithms are loosely classified into 2 classes i.e. Variance Error and bias error. As gradient boosting is one of the boosting algorithms it's wont to minimize the error of the model.
- XGBoost: The beauty of this important algorithm lies in its scalability, which drives fast literacy through resemblant and distributed computing and offers effective memory usage. It's no wonder also that CERN honored it because of the stylish approach to classifying

signals from the massive Hadron Collider. XGBoost surfaced because of the most useful, straightforward, and robust result.

## 5. PERFORMANCE ANALYSIS

We have implemented different Machine Learning models like decision trees, random forest regression, etc for the accurate prediction of different crop yields. I started with data exploration to know the relation between the variables and to develop an appropriate model. The proposed model contains two phases: The training phase and the test phase. We have taken 70% of the data for training the Dataset and the remaining 30% for testing the dataset. The Splitting of the dataset is done by Train_ Test_ Split function which splits the data arrays into two subsets that is x_train and y_train. We do not need to divide the dataset manually. Train_ Test_ Split function will make the random partition for two subsets. This research will provide a deep insight into the varied conditions affecting the crop yield which may help the farmers to reduce losses and generate greater revenue. This may also positively contribute to reducing environmental depletion and maintaining a balance between continued agricultural growth and therefore the ecological health of the land upon which humans depend.

| | Model | Accuracy | Train_acc |
|---|---|---|---|
| 2 | RFC | 0.997727 | 1.000000 |
| 3 | GBC | 0.995455 | 1.000000 |
| 4 | XGB | 0.995455 | 1.000000 |
| 1 | DT | 0.990909 | 1.000000 |
| 0 | KNN | 0.977273 | 0.988636 |

Figure 5 Model accuracy

This work gives accuracy as shown above. The KNN fits best and all others are overfitting to this data. Overfitting occurs when the machine learning model attempts to cover all data points or data points that exceed the required data points in a particular dataset. As a result, the model begins to cache noise and inaccurate values in the dataset, all of which reduce the efficiency and accuracy of the model. The overfitting model has a low bias and a large variance. After knowing the prediction accuracy of different crops using different ML models, I've selected the best-predicted accuracy for each crop using different models and recommended that the crops get better production as compared to other crops.

## 6. FUTURE SCOPE AND CONCLUSION

The proposed work will benefit farmers to maximize productivity in agriculture, reduce soil degradation in cultivated fields, and reduce fertilizer use in crop production by recommending the proper crop considering various attributes. The system assists the agriculturists in picking an applicable crop for their husbandry land grounded on the essential variables. The system is to plan 24 and grow a recommendation model to make the recommendations for crops reckoned on geological and climatic attributes using machine learning procedures. The proposed work aids framers in accurately selecting the crop for cultivation and attaining sustainability. In the future, the proposed system is often extended to think about market demand and availability of market infrastructure, expected profit and risk, and post-harvest storage and processing technologies. this can provide a comprehensive prediction of the idea of geographical, environmental, and therefore economic aspects. The main future work's aim is to improved dataset with larger number of attributes and to build website and mobile app for easy to use.

### Conflict of interest statement

Authors declare that they do not have any conflict of interest.

## REFERENCES


[1] Agricultural Recommendation System for Crops Using Different Machine Learning Regression Methods January 2021, International Journal of Agricultural and Environmental Information Systems 12(1):1-20 DOI:10.4018/IJAEIS.20210101.oa1, https://www.researchgate.net/publication/349715434_Agricultural_Recommendation_System_for

[2] A Survey on Crop Recommendation Using Machine Learning, M.V.R. Vivek, D.V.V.S.S. Sri Harsha, P. Sardar Maran, International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-7, Issue-5C, February 2019

[3] Crop Recommendation System, October 2020, ThewhatteigeHarinduthaRuchirwarya, Sri Lanka Institute Of Technology, https://www.researchgate.net/publication/346627389_Crop_Recommendation_System

[4] Machine Learning-Based Crop Recommendation System Dhruv Piyush Parikh1, Jugal Jain, Tanishq Gupta, and Rishit Hemant Dabhade School of Electronics Engineering, School of Computer Science Engineering, Vellore Institute of Technology, Chennai, India, Thakur College of Engineering & Technology, Mumbai,

India, International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)Volume 6, Issue 1, June 2021.

[5] Crop Recommendation using Machine Learning Techniques S. Mamatha Jajur, Soumya N. G., G. T. Raju, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-9 Issue-2S, December 2019

[6] https://www.javatpoint.com/k-means-clustering-algorithm-in-machine-learning

[7] https://www.javatpoint.com/k-means-clustering-algorithm-in-machine-learning

[8] https://www.javatpoint.com/k-means-clustering-algorithm-in-machine-learning

[9] https://www.javatpoint.com/k-means-clustering-algorithm-in-machine-learning

[10] [9]https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm

[11] Doshi, Z. (2018). Agro Consultant: Intelligent Crop Recommendation System Using Machine Learning Algorithms. Institute of Electrical and Electronics Engineers, 3(2), 123–135.

[12] Everingham, Y., Sexton, J., Skocaj, D., & Bamber, G. I. (2016). Accurate prediction of sugarcane yield using a random forest algorithm. Agronomy for Sustainable Development, 36(2), 27-35.

[13] Fathima, M., Sowmya, K., Barker, S., & Kulkarni, S. (2020). Analysis of crop yield prediction using data mining technique. International Research Journal of Engineering and Technology, 7(5), 7708-7713. Gandhi, N., Armstrong, L. J., Petkar, O., & Tripathy, A. K. (2016). Rice crop yield prediction in India using support vector machines.

[14] International Joint Conference on Computer Science and Software Engineering, 3(2), 1-5. doi:10.1109/JCSSE.2016.7748856 Garanayak, M., Mohanty, S. N., Jagadev, A. K., & Sahoo, S. (2019). A recommender system using item-based collaborative filtering (CF) and K-Means. International Journal of Knowledge-Based and Intelligent Engineering Systems, 23(2), 93–101. doi:10.3233/KES-190402.

[15] 2017 International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT), "A Study on Various Data Mining Techniques for Crop Yield Prediction", Yogesh Gange, Sandhya

[16] 2017 IEEE Region 10 Humanitarian Technology Conference, "RSF: A Recommendation System for Farmers", Miftahul Jannat Mokarrama; Mohammad Shamsul Arefin. 2017 International Conference on I2C2, "Agriculture decision support system using data mining", Prof. Rakesh Shirsath; Neha Khadke; Divya More.