JOURNAL PUNJAB ACADEMY OF SCIENCES (Peer Reviewed Open Access Journal)
JPAS, 23: 208-216 (January-December, 2023)
Available online at www.jpas.in
ISSN 2229-7014 (P)

A SURVEY ON CROP YIELD PREDICTION USING MACHINE LEARNING

Rajni Devi[1], Dr Harpreet Kaur*
Department of Computer Science and Engineering,
Punjabi University Patiala[1,*]
*E-mail: rajnidigra100@gmail.com

ABSTRACT

Agriculture is the science and practice of raising plants and animals. India is the second-largest agricultural nation in the world, with 60.45% of its land used for farming. In India, agriculture is one of the most common and least-paid professions. Being an agricultural nation, India's economy is heavily dependent on rising agricultural yields and agro-industrial goods. Machine learning can bring a boom in the agriculture field by changing the income scenario for optimal crop. This paper focuses on predicting the yield of the crop by applying various machine learning algorithms. Machine learning algorithms' predictions will assist farmers in selecting the best crops for their farms based on factors including soil type, temperature, humidity, water level, spacing depth, soil PH, season, fertilizer, and months. This paper focuses on a concise comparative work of several papers that discuss many methods for assessing crop yield. It seeks to predict agricultural yield through consideration and investigation of the datasets of previous years of the crop. The study also describes various current methods for auditing crop yield. It also includes a comparison of several algorithms and their advantages and disadvantages.

Keywords: Agriculture, Crop Prediction, Machine Learning, KNN, Decision Tree, Random Forest, Naive Bayes, SVM, Logistic Regression.

INTRODUCTION

India has a long history of agriculture, going back to the Indus [1]. Valley Civilization period, India comes at second in this industry. India is generally an agricultural country. With the largest net cropped area, India tops the world rankings, followed by the US and China. In terms of economic contribution, India's economy is primarily based on agriculture. It is the pillar of the Indian economy and more than 50% of India's population are dependent on agriculture for their survival. Agriculture is considered as the main and the foremost culture in India. It is one of the major and the least paid occupations in India. The primary objective of agricultural planning is to maximize crop output rates while utilizing a certain amount of land resources. The geography of a location, the weather, the kind of soil, and the harvesting techniques all affect the rate at which crops are produced. Different prediction models employ different subsets of these influencing characteristics for various crops [2]. Variations in weather, climate, and other similar environmental factors have become a serious threat to agriculture's ability to thrive. In agricultural planning, choosing each crop is important. The need for agricultural products will skyrocket as the world population, which was forecast to be 1.8 billion in 2009, is expected to reach 4.9 billion by 2030.The need for agricultural products will increase among people in the future, necessitating effective

farmland development and an increase in crop output [3]

Nowadays, some farmers don't have awareness about the crop suits their soil as per the soil quality, soil nutrients and the soil composition. Depending upon these parameters, they may choose a crop which will not give them as expected production rate. There is a lot of study done on agriculture planning, and the aim is to develop a model that can estimate crop yields with efficiency and accuracy crop classification, soil classification, weather forecasting, crop disease forecasting, and crop classification based on growing phase. Recent days have seen several variations in the weather. Crop development is therefore challenging in this climate [4]. In addition, fertilizers are an important factor and one of the key considerations when growing crops, therefore we want to use some technology to understand the yield nuances and help farmers grow crops effectively.

The two main parameters that affect crop productivity are temperature and rainfall. Because it might be difficult to grasp these factors, machine learning has been developed to help. The farmer can predict the yield in the past based on his or her expertise and previous experiences, but the current climate circumstances may vary significantly, making it impossible for them to do so. Modern technological advancements like artificial intelligence and machine learning, which can forecast outcomes, have risen as a result of the new technological era. They were also trained and put to the test to forecast and provide the most precise and close assumption for events that would take place in the future. With the help of this machine learning algorithm, we can calculate and predict the most productive result of the yield [4].

Farmers may benefit from modern technology advances in yield prediction by being able to estimate costs based on production. In terms of yield prediction, there are mainly two categories: classification and prediction phase. The classification and prediction approaches are briefly described in this paper. Large collections of existing databases can be examined to obtain predictions in order to create new knowledge. It is regarded as one of the tested approaches to problem solving, resulting in a prediction that is automatic and roughly accurate. There are a number of steps that can be involved in crop yield prediction are data acquisition, data pre-processing, feature selection, classification and prediction.

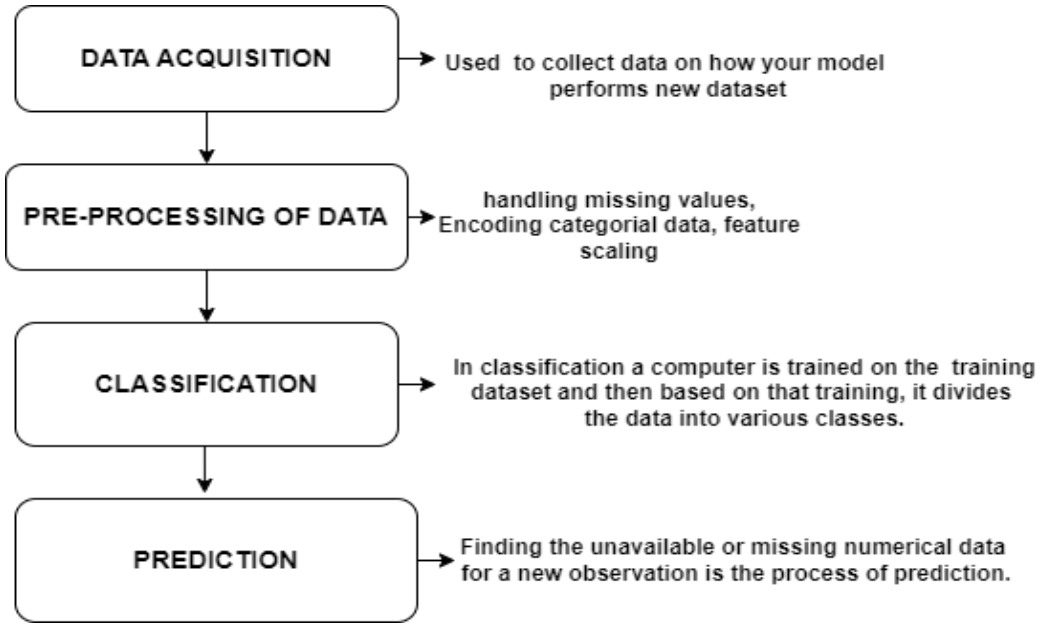

Fig. 1: Architecture of crop yield prediction

A. DATA ACQUISITION:

It is a technique for collecting data on yield production and supplying it as input. To operate, it needs two things: data and models. There should be sufficient features in the collected data (data point that can help with a prediction).

B. PRE-PROCESSING OF DATA:

Pre-processing data is a technique for converting impure data into a clean data set (i.e., a suitable syntax). Simply said, if information is acquired from various sources and grouped, it is done so in a raw form that is not likely to be useful for examination.

C. CLASSIFICATION:

The procedure involves determining how a new factor belongs into a group of classifications using a training set of data that includes observations with established class correlates.

D. PREDICTION:

it is a method of referencing an algorithm's output after it has been trained on a collection of factual data and applied to advanced data in order to predict the possibility of a particular result.

E. MACHINE LEARNING IN AGRICULTURE:

Machine learning is a subset of Artificial Intelligence. It is a practical and methodical methodology that can help us anticipate crop yields more accurately in accordance with various meteorological variables. The datasets are utilized to train the model that starts the outcome based on historical data [5].

The data set needs to be divided into training and test sets for prediction purposes. The training data set is used to build a model in a data set, and the testing data set is used to validate the model that was produced using the training data set [5]. Machine learning algorithm will comprehend the pattern of the crop and yield based on the various variables and forecast the yield of the area in which he

is going to crop [6]. Depending on the study challenge and research objectives, an ML model can be either descriptive or predictive. Predictive models are used to make forecasts about the future, whereas in order to understand what has happened and draw conclusions from the data obtained, descriptive models have been applied (Alpaydin, 2010). To create a high-performance predictive model, ML studies must overcome a variety of obstacles. The right algorithms must be selected in order to solve the current problem, and both the underlying platforms and algorithms must be capable of handling the volume of data [7].

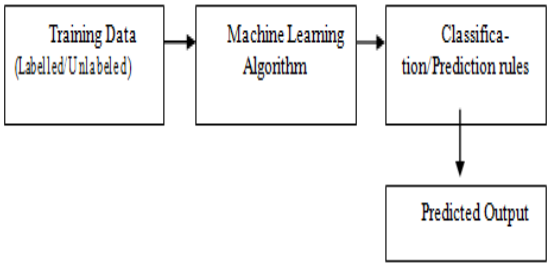

Fig 2. Machine Learning Process [8]

Machine learning techniques can be used to increase crop yields. Crop yield rates can be increased with the use of various machine learning algorithms. Machine learning algorithms will predict the most efficient output of the yield. Previously yield was predicted on the basis of the farmer's prior experience but now Since weather conditions might suddenly change, they cannot determine the yield.

In this paper, we apply various machine learning methods to predict the agricultural yield. Machine learning algorithms will help the farmers to decide which crop to grow to get the maximum yield by taking into account elements like temperature, rainfall, location, etc. Regression trees, random forests, convolution neural networks, and the K-nearest algorithm are some of the machine learning techniques that are frequently used in prediction techniques.

LITERATURE SURVEY

Various studies have been carried out in literature regarding crop yield prediction. In this section light has been shed on various researches pursued in this field:

Balamurugan [1] describes the prediction of agricultural yield using only a random forest classifier. The crop output was predicted using a variety of factors, including rainfall, temperature, and season. No further machine learning techniques were used on the datasets. Comparison and quantification were missing due to the lack of other algorithms, making it impossible to give the appropriate algorithm. Vishnu Vardhan [2] has provided information about the methods and applications used in agriculture. Agriculture planning methods include K-Nearest Neighbor (KNN), Artificial Neural Networks (ANN), and Support Vector Machines (SVM). The challenge of crop prediction, which is based on the facts at hand, is a crucial agricultural concern in this work. Different types of approaches are assessed for this goal using various data sets.Bondre and Mahagaonkar [3] describe ML techniques to predict the crop yield and suggested manure application. A significant problem in agriculture was yield prediction, which was solved by creating a machine learning system. The effectiveness of the created model for calculating crop yield in agriculture was assessed. An advantage of the developed model was the use of prior data for crop prediction and the recommendation of an appropriate fertilizer for each crop by means of ML algorithms like random forest and SVM. The smart irrigation system for farms, which would have increased yields, was not, however, put into practice. S. V. Bhosale [4] used three algorithms, clustering k-means, apriorism, and Bayes, then hybridized the algorithms for better efficiency of yield prediction. They took into account parameters like area, rainfall, and

soil type, and their system was also able to determine which crop is suitable for cultivation based on the mentioned features. Jude Immaculate [5] describes the Algorithms for machine learning in the agricultural sector. This calculates the efficacy of each machine learning system that can forecast agricultural yield. This also discusses how the algorithm is implemented and how it functions. Eeswari [6] describes the crop's output depending on the average, minimum, and maximum temperature. They have also gone one further step aside from that crop evapotranspiration is a characteristic. This crops the weather and the amount of water that is absorbed into the period of a plant's growth. In order to make an informed choice regarding the yield of the groupings, this attribute is taken into account.

They all gathered a data set with these attributes, sent it as input to a Bayesian network, and then classified it into two categories called true and false classes.

The observed classifications in the model were then compared with the predicted classifications in the model with a confusion matrix, bringing the accuracy.

Finally, they came to the conclusion that crop yield prediction using Naive Bayes and a Bayesian network provides higher accuracy when compared to SMO classifier, and that it will be advantageous to anticipate crop production prediction under various climatic and agricultural conditions. K.Ruth Ramya [7] describes crop yield projection using a classifier from the random forest. They took into account the minimum temperature, humidity, and rainfall as parameters; the outcomes acquired with a 99.7% accuracy rate. Limitation: Only Limitation The absence of an ensemble algorithm in this paper is a drawback. Which, in comparison, decreases precision Ponnuru Sai Nishant [9] describes a machine learning model to forecast crop yields. Farmers can immediately use the method because it is so straightforward. To forecast crop yield, they used regression models such as Kernel Ridge, E Net, and Lasso. The regression model was improved via stacking regression.

CROP YIELD PREDICTION USING MACHINE LEARNING LGORITHMS

Predictions of agricultural yield are influenced by a number of different variables. The most significant factors are temperature and rainfall. Temperature and rainfall data are sequential data; therefore, time series machine learning algorithms are used for them. The algorithms are listed below:

A. KNN (K- Nearest Neighbor):

KNN method makes the assumption that similar items exist within the area. KNN is one of the easiest Machine Learning methods based on Supervised Learning technique.

The k-Nearest Neighbor method can be used for both regression and classification predictive problems but mostly it is used for the Classification problems. KNN algorithm classifies new data points using existing data and similarity metrics [2].

The class with the closest neighbor receives the data. The similar objects are located close to one another. It computes the distance between each fresh input sample predictor and each training sample predictor using a distance function (such as the Euclidean, Manhattan, or Makowski distance functions), and then chooses the distances that are closest to the targets and are also the shortest. The difficulty of selecting k depends on how sensitive the data set is. More variance and lower bias are related to lower values of k, and inversely. It uses the locality principle, which is useful for nonlinear problems with great adaptability.

B. Decision Tree:

The decision tree system falls under the order of supervised learning technique. Both classification and regression problems can be solved using them.

Decision tree learning divides the full sample space recursively into smaller sub-sample spaces that are sufficient for a straightforward model to be developed. The entire sample space is held by the root node, which is the first node in the tree. dividing a sample space into a smaller sub-sample space requires splitting the root node into children nodes, each of which can then be split into a leaf node (a node on which further splitting is not feasible).

The nodes in the tree, with the exception of the leaf node, divide the sample space according to a set of conditions depending on the values of the input attributes, and the leaf node assigns an output value for those input attributes that are on the path from the root to the leaf in the tree. The main objective of utilizing a decision tree to sub sample data is to reduce the mixing of distinct output values and assign a single output value to the sub sample space [10].

C. Random Forest:

Random Forest is a ML algorithm. Another name for it is ensemble because it mixes the same or various algorithms for categorizing items.

Ensemble learning is a method of learning where you combine various algorithms or run the same algorithm repeatedly to create a more accurate prediction model. Several algorithms of the same kind are combined in the random forest algorithm. Both classification and regression issues can be solved using the Random Forest method.

D. Naive Bayes Algorithm:

A class of classification algorithms known as Naive Bayes classifiers are based on the Bayes theorem. Each classification is independent of the others, which is a mixture of various having a common premise.

A classifier is used in a machine learning algorithm to differentiate between various items based on specific features. For classification, a Naive Bayes classifier, a probability-based machine learning model, is used.

Bayes Theorem:

$P(A|B) = P(B|A) * P(A) / P(B)$…….(I)

The Bayes theorem can be used to calculate the likelihood that A will occur given the occurrence of B. The evidence in this case is B, but the hypothesis is A.

The predictors and traits are assumed to be independent in this scenario.

In other words, the existence of one attribute has no impact on the other. so that is referred to as naive [11].

E. Support Vector Machine:

Support Vector Machine, sometimes known as SVM, is one of the most widely used algorithms for supervised learning.

It is used to solve both classification and regression issues.

In machine learning, it is generally employed to solve classification issues. The SVM separates the data into decision surfaces, which further divide the data into two hyperplane groups.

Training points define the supporting vector for the hyperplane [12].

Presumably due to larger margins, a hyperplane that is farthest from the closest learning data point typically has superior margins, less errors, and a high classifier generalization. SVM is mostly applicable to tasks like face recognition, text classification, handwriting recognition, and picture classification [13].

F. Logistic Regression:

Logistic regression is the Machine Learning algorithm; it falls under the category of supervised learning. When classifying the classes, one statistical model that is employed

as a method of problem-solving is called logistic regression.

It employs a logistic function to represent a binary dependent variable, that is, in the form of 0s and 1s, when we need to distinguish one class from another, we use it. It can be used for hotel booking, text editing, medical purposes, and credit scoring. Logistic regression can be used only when there is certainty about the continuity of the flow of values [13]. With this method, we can determine whether an email is valid or spam. When one class is different from another class, it is used, like making a political election prediction.

Table 1 Comparison table of several crop yield prediction techniques.

Algorithm	**Advantages**	**Disadvantages**
K Nearest Neighbor (K-NN) Classifier	1) Easy to understand and implement 2) Training happens quickly 3) Zero cost in the learning process 4) Examines the output of several crops from earlier years' production.	1) Testing is slow 2) Memory restrictions 3) As it is supervised lazy learner, it runs slowly 4) Expensive for large data set
Decision Tree	1) Decision trees are quick and easy to use. 2) It generates an accurate result 3) It requires less memory. 4) It can handle noisy data.	1) Long training periods are required. 2) It has an issue of over fitting. 3) Unstable classifier 4) Generate categorical output
Naive Bayes	1) Short computational time is needed for training. 2) It performs well. 3) By eliminating the unnecessary elements, it increases classification performance.	1) In order to get decent results from the Naive Bayes classifier, a lot of records are needed. 2) less accurate on some datasets when compared to other classifiers. 3) Require large data for good results
Random Forest	1) The random forest algorithm is not biased, since there are many trees and every tree is trained on a subset of data. 2) If a new data point is added to the data-set, the Random Forest method is stable and remains unaffected. 3) It reduces the issue of over-fitting	1) Due to the fact that it creates several trees in order to integrate their outputs, it consumes a lot of resources and computational power. 2) It takes a lot of time to train because it combines many decision trees to decide the class.

Logistic Regression	1) It is simpler to set up and train 2) It doesn't assume anything regarding the distribution of classes in feature space. 3) It can quickly identify unknown items.	1) Boundaries are created linearly. 2) The presumption that the relationship between the dependent variable and the independent variables is linear.
Support vector machine (SVM)	1) Create a very accurate classifier. 2) less over fitting robust to noise 3) Easy to interpret result	1) Run very slowly 2) Computationally expensive 3) Sensitive to running kernel choice

CONCLUSIONS

Agriculture is the primary resource of both food and money for the farmers. Agricultural examinations would support agricultural bodies in assisting farmers in making wise and lucrative decisions. Today's diseases that affect plants are more common, which has resulted in a 0–100% reduction in plant output. In this paper we have studied agricultural yield prediction using several machine learning methods. Additionally, we contrasted various crop yield prediction algorithms and covered their advantages and disadvantages. We have studied that a random variable takes a lot of time to train because it combines many decision trees to decide the class. In the case of SVM it takes less time and it creates very accurate results.

REFERENCES

Nigam, A., Garg, S., Agrawal, A. and Agrawal, P., 2019, November. Crop yield prediction using machine learning algorithms. In *2019 Fifth International Conference on* Image Information Processing (ICIIP) (pp. 125-130). IEEE.

.

Jirage, P.S., Patil, P.R., Mali, S.S., Koshti, M.P., Kandekari, S.S. and Akulwari, P.K., A Survey On Crop Yield Prediction Using Machine Learning.

Reddy, D.J. and Kumar, M.R., 2021, May. Crop yield prediction using machine learning algorithm. In *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)* (pp. 1466-1470). IEEE.

Jeevaganesh, R., Harish, D. and Priya, B., 2022, April. A Machine Learning-based Approach for Crop Yield Prediction and Fertilizer Recommendation. In *2022 6th International Conference on Trends in Electronics and Informatics (ICOEI)* (pp. 1330-1334). IEEE.

Keerthana, M., Meghana, K.J.M., Pravallika, S. and Kavitha, M., 2021, February. An ensemble algorithm for crop yield prediction. In *2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)* (pp. 963-970). IEEE.

Bhanumathi, S., Vineeth, M. and Rohit, N., 2019, April. Crop yield prediction and efficient use of fertilizers. In *2019 International Conference on Communication and Signal Processing (ICCSP)* (pp.0769-0773). IEEE.

Van Klompenburg, T., Kassahun, A. and Catal, C., 2020. Crop yield prediction using machine learning: A systematic literature review. *Computers and Electronics in Agriculture*, *177*, p.105709.

Nathgosavi, V., 2021. A survey on crop yield prediction using machine learning. *Turkish Journal of Computer*

and Mathematics Education (TURCOMAT), *12*(13), pp.2343-2347

Islam, T., Chisty, T.A. and Chakrabarty, A., 2018, December. A deep neural network approach for crop selection and yield prediction in Bangladesh. In *2018 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)* (pp. 1-6). IEEE.

Rajeswari, V. and Arunesh, K., 2016. Analyzing soil data using data mining classification techniques. *Indian journal of science and Technology*, *9*(19), pp.1-4.

Singh, N., Pant, D., Singh, D.P. and Pant, B., Crop Prediction Method To Maximize Crop Yield Rate Using Machine Learning Technique: A Case Study For Uttrakhand Region.

Kumar, R., Singh, M.P., Kumar, P. and Singh, J.P., 2015, May. Crop Selection Method to maximize crop yield rate using machine learning technique. In *2015 international conference on smart technologies and management for computing, communication, controls, energy and materials (ICSTM)* (pp. 138-145). IEEE.

Suruliandi, A., Mariammal, G. and Raja, S.P., 2021. Crop prediction based on soil and environmental characteristics using feature selection techniques. *Mathematical and Computer Modelling of Dynamical Systems*, *27*(1), pp.117-140.