International
Journal
Of Advanced Research in Engineering& Management (IJAREM)
ISSN: 2456-2033 || PP. 13-19

A Comprehensive Review on Crop Yield prediction using Data mining & Machine learning Techniques

Shalu Kushwah
VITM, Gwalior, MP, India
Research Scholar

Sandeep Kumar Tiwari
Dept. of Computer Science & Engineering
VITM, Gwalior, MP, India

Anand Singh Bisen
Dept. of Computer Science & Engineering
VITM, Gwalior, MP, India

Abstract: Agriculture is the main concern and emerging field for research in every country. In India the population is increasing very rapidly with increase in population the need for the food is also increasing.In agriculture sector where farmers and agribusinesses have to make numerous decisions every day and intricate complexities involves the various factors influencing them. An essential issue for agricultural planning intention is the accurate yield estimation for the numerous crops involved in the planning. Machine learning techniques are necessary approach for accomplishing practical and effective solutions for this problem. Agriculture has been an obvious target for big data. Environmental conditions, variability in soil, input levels, combinations and commodity prices have made it all the more relevant for farmers to use information and get help to make critical farming decisions.

Keywords: ANN, SVM, Decision tree, Clustering, Association rule mining.

I. Introduction:

"Agriculture is the art and science of growing plants and other crops and for food, other human needs, or economic gain". India is an agricultural country. Two- third of the population is dependent on agriculture directly or indirectly. Agriculture provides the highest contribution to national income. There are many serious problems like- Erosion, Diseases, Pests, Weeds, Drought, Rainfall that people face trying to grow food today.

The components for cultivation are as follows:

- Irrigation and Rainfall
- Soil type
- pH (Potential of Hydrogen)
- Moisture content of soil
- Fertilizer
- Pesticides (Biopesticides and Chemical Pesticides)
- Insecticides
- Soil nutrients
- Temperature
- Humidity

Machine learning techniques can be very useful for better estimation of crop production rate on the basis of various parameters.Classification and prediction techniques are applied on metrological-related data and crop related data. Various predictions can be made on the basis of predicted result which can help in increasing crop production rate. There are various machine learning algorithms which can be used for prediction of crop production rate on the basis of all the available parameters.

In this research paper we found that various data mining and machine learning techniques can be used for estimating the crop yield production rate and also how to find the optimal parameters or features for better prediction.

The aim of this research paper is to study the state of the art in the agriculture field. This papers elaborates how we can used machine learning techniques for better crop yield prediction and what are the challenges in this field for farmers.

In this paper Section II describes the machine learning techniques and their types and in section III we have done the literature survey of research papers these papers are taken from IEEE, Science direct and springer journal websites.

Section IV descries the problem identification and research gap and finally section V describes the conclusion and future work.

II. Machine Learning Technique:

Machine learning is a subset of artificial intelligence. It focuses mainly on the designing of system thereby allowing them to learn and make predictions based on some experience which is data in case of machines. Machine learning enables computer to act and make data driven decisions rather than being explicitly programmed to carry out certain task these programs are designed to learn and improve over time when exposed to new data. The process starts with good quality data and then training machines by building machine learning models using the data and different algorithms. Machine learning models can be broadly classified into three types [1]:

(1).Supervised Learning- Supervised learning is that learning in which the model is trained on a labelled data set. Labelled dataset is one which have both input and output values. Supervised learning as the name indicates the presence of a supervisor as a teacher. Basically supervised learning is a learning in which train or test the machine using data.

Supervised learning algorithm consist of two types of technique- Classification and Regression.

Classification : It is a Supervised Learning task where output is having defined labels(discrete value). It can be either binary or multi class classification. In binary classification, model predicts either 0 or 1 ; yes or no but in case of multi class classification, model predicts more than one class.

Regression : It is a Supervised Learning task where output is having continuous value.

(2).Unsupervised Learning- It is a type of machine learning algorithm used to draw consequences from data set consisting of input data without output values. It is the training of machine using data that is neither classified nor labelled and algorithm to act on that data without any supervision. It means that teacher is not present in this type of learning.

Unsupervised learning consists of two types of techniques Clustering and association rule mining.

(3).Reinforcement Learning- It is a type of machine learning algorithm where an agent learns to behave in an environment by performing actions. In this learning agent decides what action is performed for a task, according to the action's agent got rewards by the environment. These rewards may be positive or negative.

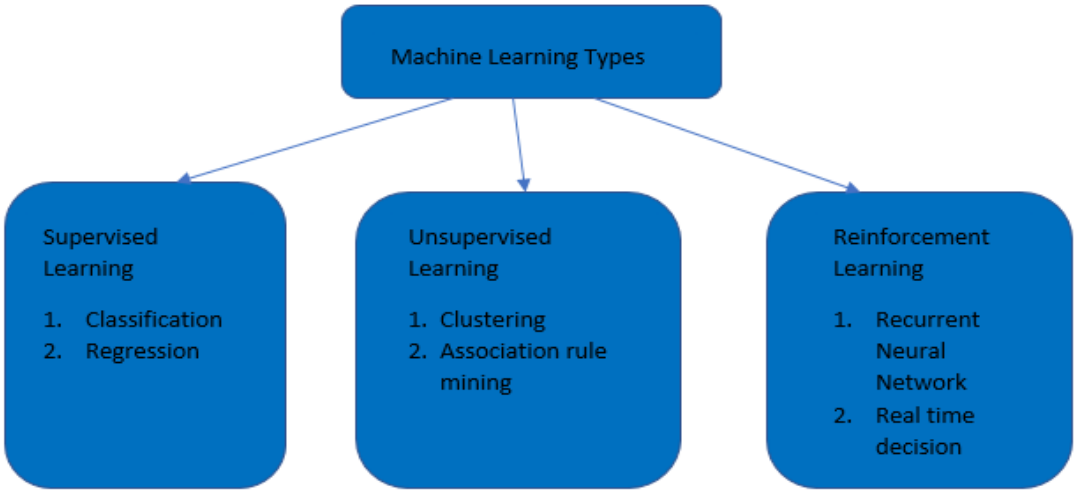

Figure 1: Machine Learning Types

Machine Learning Algorithms-

There are various machine learning algorithms in which some important machine learning algorithms are as follows:

(1). Logistic Regression: Logistic Regression is one of the classification algorithms, used to predict a binary values in a given set of independent variables (1 / 0, Yes / No, True / False). To represent binary / categorical values, dummy variables are used. For the purpose of special case in the logistic regression is a linear regression, when the resulting variable is categorical then the log of odds are used for dependent variable and also it predicts the probability of occurrence of an event by fitting data to a logistic function.

(2). Linear regression: Linear regression is a supervised learning algorithm; it performs a regression task. It describes a linear relationship between input variable and output variable. It is used for predictive analysis. Linear regression is a linear approach for modeling the relationship between the criterion or the scalar response and the multiple predictors or explanatory variables. Linear regression focuses on the conditional probability distribution of the response given the values of the predictors[19].

(3). Multiple Linear regression: Multiple linear regression is a supervised learning algorithm which is used for regression. It describes the relationship between two or more independent variable and single dependent variable by fitting an algorithm to data[19].

(4). Polynomial Regression: Polynomial regression is a type of linear regression in which the relationship between the independent variable and dependent variable is describes as an mth degree of polynomial. It fits the nonlinear relationship between the dependent and independent variable. It is used for curvilinear data. Polynomial regression is fit with the method of least squares. The goal of regression analysis to model the expected value of a dependent variable y in regards to the independent variable x[19].

(5). Decision tree : Decision tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label. Decision Tree is a tree structured frame work. Decision Tree(DT) is a white box type of ML algorithm. Primarily, it is used for classification. However DTcan also be used for regression. It works on the principle called "Decision-making logic". The decision tree is a distribution-free or non-parametric method, which does not depend upon probability distribution assumptions. Decision trees can handle high dimensional data with good accuracy [15].

(6). Random Forest: It belongs to the family of supervised learning approaches, suitable for the classification and regression problems as well. Basic working ideas behind this approach are multiple collections of tree-structured classifiers. Random forest is an ensemble learning method. It is used when size of dataset is large and the very large number of input variables approximately hundreds or thousands[20].

(7). Support Vector Machine: In machine learning, support vector machines (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyze data used for

classification and regression analysis.A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyper plane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyper plane which categorizes new examples.An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible.

In addition to performing linear classification, SVMs can efficiently perform a non-linear classification, implicitly mapping their inputs into high-dimensional feature spaces [8].

III. Literature Review:

Vinita CN et.al. proposed a model for predicting the rice production rate. In this research paper researcher described the various data mining techniques and algorithms like support vector machine algorithm, Bayesian classifier, selective attribute network algorithm, K-Nearest neighbors' algorithm and clustering techniques K-means clustering algorithm. In this paper author also proposed a SVM system flow diagram for providing the alert to farmers. And researcher applied the SVM algorithm for making the decisions for farmers and also applied the selective attribute network algorithm and K-means clustering algorithm for clustering the data. The researcher predicted the crop yield on the basis of various climatic conditions. Researcher described that for improving the accuracy of results and for better crop yield prediction artificial neural network and neural network can be used in future.It can understand of the high dimensional relation between yearly and seasonal climatic patterns which determine crop yield helpsboth farmers and otherdecision makers to be able to predict the effects of drought and other climatic conditions[1].

JharnaMajumdaret.al. focuses on the analysis of agriculture data and finding optimal parameters to maximize the crop production using data mining techniques. Authors elaborate that data mining techniques and described that data mining techniques paly a significant role in big data analysis of agriculture. Data mining is the process of extracting optimal patterns from large amount of dataset. In this proposed work the PAM, CLARA and DBSCAN clustering algorithms are used. The researcher used the dataset of Karnataka state. In this proposed work PAM, CLARA, and DBSCAN clustering algorithms are applied on dataset for categorizing the dataset of Karnataka which have similar crop yield production rate. The clusters are made according to the temperature i.e. two clusters are made and on the basis of rainfall range i.e. six clusters are formed and according to the soil type i.e. five clusters formed. For more better results the researcher proposed the modified DBSCAN algorithm. It is method is used to cluster the data based on districts which are having similar temperature, rainfall and soil type and for find out the optimal parameters like temperature, soil type, rainfall for better production of crop yield[2].

Kirtanjhaet.al. reviewed the problems that are occurring in the agriculture field and described that machine learning and data mining techniques can be used for increasing the production rate. The researcher reviewed the previous workin the agriculture field and emphasis on the problems inthe agriculture field. The researcher proposed a model in which artificial neural network is used for predicting the crop yield after that automation and wireless network system is used in agriculture field. The researcher concluded that machine, learning, deep learning, fuzzy logic and IOT, artificial intelligence can be used for automation in this field[3].

A.T.M Shakil Ahamed et. al tested few data mining techniques for prediction of the annual yield of major crops in Bangladesh. In this paper, the clustering technique is used to predict the results. There are parameters such as temperature, humidity, minimum temperature, maximum temperature, average sunshine, Soil PH and salinity are used to predict the Annual crop production. K- means clustering is used by the author for recommending plant crops in the districts of Bangladesh[6].

Be labed Image et. al., proposed an approach for extracting information by using data mining approaches in three domains, bioinformatics, medicine, and agriculture industry. Initially, the variables are clustered to increase the functionality, and the association rules are used between the target variables and the previously identified set of variables[7].

Shreya et. al. reviewed and identified the problems and challenges that are faced by farmers in India. They have also collected and analyzed the dataset available online by using data visualization techniques of data mining for better understanding the data then they applied unsupervised learning approach i.e. K-Means clustering algorithm for finding out appropriate and valid clusters of dataset. They have also used Apriori algorithm of association rule mining for counting the frequently occurring features and then naïve bayes algorithm is implemented for the estimation of crop yield prediction[4].

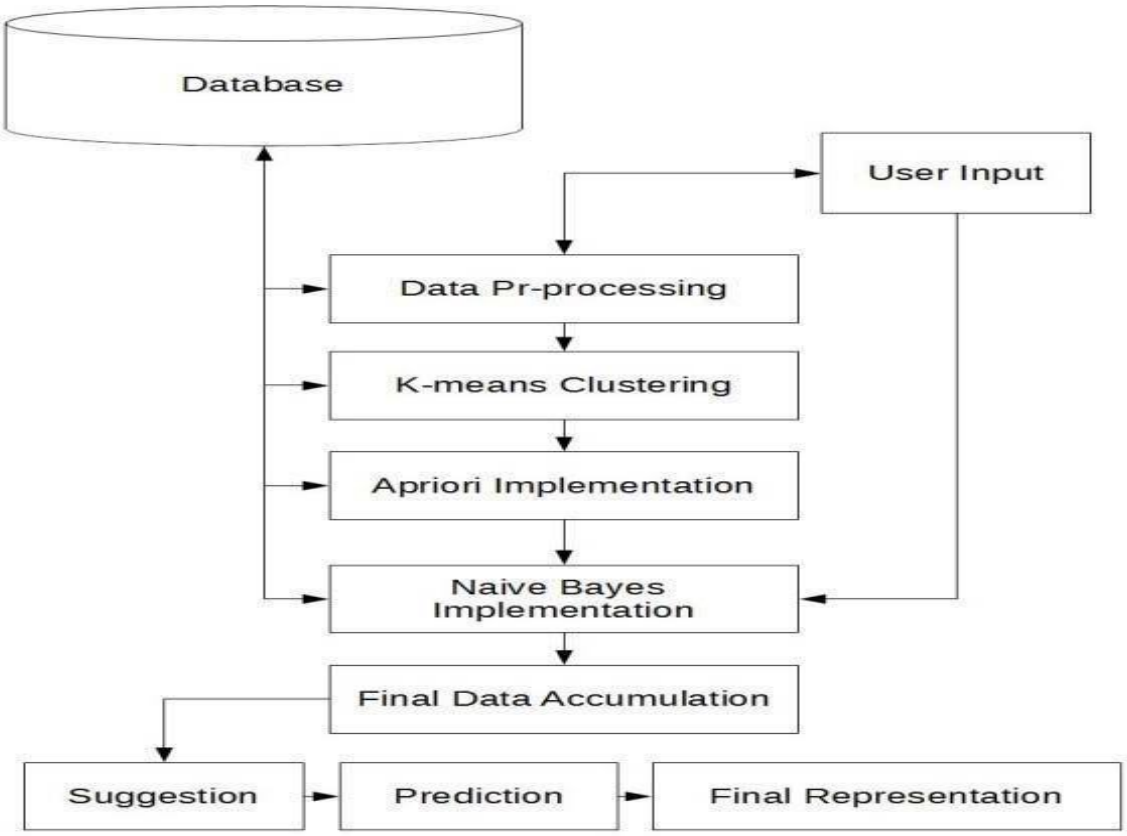

Figure2: Proposed methodology

Mercelin Franciset.al. proposed a model for disease diagnosis along with classification in agriculture field .He made classification based on the spots found on the leaves so that agricultural productivity can be improved. He applied methods of computer vision and deep learning techniques in order to find disease identification and classificationof public available apple leaf images. In their framework, authors first applied their methodology on the multi-space image reconstruction inputs and generated a new set of images containing gradient images. Next in their work they extracted the high level semantic features from the original and reconstructed images using the convolutional and depth-wise separable convolutional layers. Classification was achieved using the SoftMax classifier. The hyper parameters and computational cost are computed mathematically which provides an insight of creativeness to the researchers [5].

T.R.Lekhaaet.al. studied various machine learning and data mining techniques. He suggested thatWEKA(Waikato setting for information Analysis) may be a usedfor implementing machine learning algorithms.He explained about a strategy display for breaking down learning. Weka is an absolute apartment of Java programs affiliated beneath a acceptedinterface to admittance analysis and analysis on datasets application advanced techniques. Recentlyaggregation has aswell formed on MOA, an ambiance for mining abstracts streams. Weka provides the absolute ambiance for advancing analysis in abstracts mining.

He described that WEKA tool can be used for data preprocessing, data visualization, and for predicting the crop production rate [18].

Michael L. Mann et.al. reviewed and estimate that timely and accurately agricultural impact assessments for droughts are critical for designing appropriateinterventions. The assessments may be ad hoc, late, or spatially imprecise, with reporting at the zonal or regional level. He said that this is problematic as when we will find variability in losses at the village-level, which is missing when reporting at the zonal level. In the proposed methodology, he proposed a data fusion method in which they combined remotely sensed data with agricultural survey datathat may overcome these challenges. Theyused the method to Ethiopia, which is hit by droughts and is a substantial recipient of ad hoc imported food aid. then used remotely sensed data found near mid-season to

predict substantial crop losses of greater than or equal to 25% due to drought at the village level for five primary cereal crops. They trained machine learning models to predict the likelihood estimation of losses and find out the most influential features.The researcher suggested that these models might be used to help monitor and predict yields for disaster response teams and policy makers. In this work they examined the benefits of using remotely sensed data and machine learning models to predict agricultural losses due to drought in Ethiopia. They predicted, by the midpoint of the growing season, which sub-kebeles will suffer substantial crop losses at the time of harvest. They also developed a custom set of algorithms to summarize changes in plant phenology, as measured by NDVI, precipitation, and potential and actual evapotranspiration up until the date of maximal greenness, which corresponds approximately to the middle of the growing season [17].

Abraham SudharsonPonrajet.al. studied that asrapidly growth of Internet of Things (IoT) devices in cities, homes, buildings, industries, health care sectorand also in agricultural field have sure the way for deployment of wide range of sensors in them. They explained that machine learning technique has created new opportunities for big data analysis. Machine learning will help the farm management system to achieve its goal by exploiting the data that is continuously made available with the help of Agricultural IoT(AIoT) platform and helps the farmer with insights,decisive action and support. They analyzed various existing supervised and unsupervised machine learning techniques applied in agricultural field and also compared performance of one technique with another by using accuracy and a confusion matrix parameter. While AIoT is the next big thing in the modern agricultural farm management system, applying machine learning algorithm to the data generated from the different inputs of a farm set up with the help of AIoT makes the system more intelligent, provides decisive information and predicts the upcoming outcome. In this methodology various machine learning algorithm were analyzed, each have their own pros and cons from the process to the outcome[10].

IV. Problem identification and research gap

A large number of previous work has been done by researchers for prediction of crop production of various types of crops but still there is a boom forimprovement in accuracy of prediction and also build an efficient model for agriculture department.

In the past decades researchers have used various data mining and machine learning algorithms likes DBSCAN, CLARA, Decision tree, artificial neural networks and naïve Bayes classifier for prediction of crop production rate but they all predicted on the basis few parameters like rainfall, area harvested, soil type. In agriculture sector where farmers and agribusinesses have to make innumerable decisions every day and intricate complexities involves the various factors influencing them. An essential issue for agricultural planning intention is the accurate yield estimation for the numerous crops involved in the planning. As the impact of climate change on agriculture could result in problems with food security and may threaten the livelihood activities upon which much of the population depends. Climate change can affect crop yields (both positively and negatively), as well as the types of crops that can be grown in certain areas. In India found that increase in temperature reduced potential yield grains, and environmental changes are also vary day to day. That's why there is need for an efficient model that can predict the crop production rate on the basis of all the environmental conditions.

We also find that we can design a system for agricultural department for providing the suggestions to farmers regarding that what are the optimal parameters like rainfall, soiltype for a particular type of crop and which fertilizer will be best for crop and also for disease detection of crops by using the machine learning and deep learning techniques.

V. Conclusion

In this research paper we have studied about the various machine learning techniques and also reviewed how to apply machine learning and data mining techniques for better estimation of crop yield.Machine learning techniques are necessary approach for accomplishing practical and effective solutions for this problem. Agriculture has been an obvious target for big data. Environmental conditions, variability in soil, input levels, combinations and commodity prices have made it all the more relevant for farmers to use information and get help to make critical farming decisions. That's why there is a need for an efficient model in agriculture field which can predict the crop production rate on the basis of all the parameters which also provides some recommendations to farmers like which fertilizer is best for this crop and how to maximize crop production rate and also for disease detection in crops. That's why various machine learning and deep learning algorithms can be used for better prediction results.

References:

[1]. C. N. Vanitha, N. Archana and R. Sowmiya, "Agriculture Analysis Using Data Mining And Machine Learning Techniques," *2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS)*, Coimbatore, India, 2019.
[2]. Jharna Majumdar, Sneha Naraseeyappa and Shilpa Ankalaki, "Analysis of agriculture data using data mining techniques: application of big data," Springer, 2017.
[3]. KirtanJha, Aalap Doshi, Poojan Patel, "A Comprehensive review on automation in agriculture using artificial intelligence," Artificial intelligence in agriculture, Science direct, 2019.
[4]. P. Surya, Dr. I. Laurence Aroquiaraj, "Crop Yield Prediction in Agriculture Using Data Mining Predictive Analytic Techniques", IJRAR December 2018.
[5]. Mercelin Frances, C. Deisy, "Mathematical and Visual Understanding of the Deep Learning Model Towards m-Agriculture for Disease Diagnosis, " Springer 2020.
[6]. A. T. M. S. Ahamed et al., "Applying data mining techniques to predict annual yield of major crops and recommend planting different crops in different districts in Bangladesh," 2015 IEEE/ACIS 16th International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), Takamatsu, 2015, pp. 1-6.
[7]. B. Imane, B. Abdelmajid, T. A. Mohammed, T. A. Mohammed and T. A. Youssef, "Data mining approach based on clustering and association rules applicable to different fields," 2018 International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS), Kenitra, 2018, pp. 1-5.
[8]. SML Venkata Narasimhamurthy, AVS Pavan Kumar, "Rice Crop Yield Forecasting Using Random Forest Algorithm", IJRASET, October 2017.
[9]. Michael L. Mann, James M. Warner, Predicting high-magnitude, low-frequency crop losses using machine learning: an application to cereal crops in Ethiopia, springer 2019.
[10]. Abraham Sudharson Ponraj, Vigneswaran T, "Machine Learning Approach for Agricultural IoT", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-7, Issue-6, March 2019.
[11]. S. Mishra, P. Paygude, S. Chaudhary and S. Idate, "Use of data mining in crop yield prediction," *2018 2nd International Conference on Inventive Systems and Control (ICISC)*, Coimbatore, 2018, pp. 796-802.
[12]. D.Rajesh, "Application of Spatial Data Mining for Agriculture", International Journal of Computer Applications (0975 – 8887) Volume 15– No.2, February 2011.
[13]. M. Paul, S. K. Vishwakarma and A. Verma, "Analysis of Soil Behaviour and Prediction of Crop Yield Using Data Mining Approach," *2015 International Conference on Computational Intelligence and Communication Networks (CICN)*, Jabalpur, 2015, pp. 766-771.
[14]. A. Savla, N. Israni, P. Dhawan, A. Mandholia, H. Bhadada and S. Bhardwaj, "Survey of classification algorithms for formulating yield prediction accuracy in precision agriculture," *2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, Coimbatore, 2015, pp. 1-7.
[15]. Yogesh Gandge, Sandhya, "A Study on Various Data Mining Techniques for Crop Yield Prediction", 2017 International Conference on Electronics, Communication, Computer and Organization Techniques.
[16]. K. Pavya, Dr. B. Srinivasan, "Feature Selection Techniques in Data Mining: A Study", IJSDR, June 2017.
[17]. Michael L. Mann & James M. Warner & Arun S. Malik, "Predicting high-magnitude, low-frequency crop losses using machine learning: an application to cereal crops in Ethiopia" springer 2020.
[18]. T. R. Lekhaa, A. Aruna, M. Malarmathi, "Budding Novel applications in agriculture Victimization data processing", 2018 International Conference on Soft-computing and Network Security (ICSNS), 2018.
[19]. https://www.geeksforgeeks.org/Machine _learning.
[20]. Akanksha Gahoi, "Data Analytics for Predicting Wheat Crop Production", Artificial & Computational Intelligence, 2019.